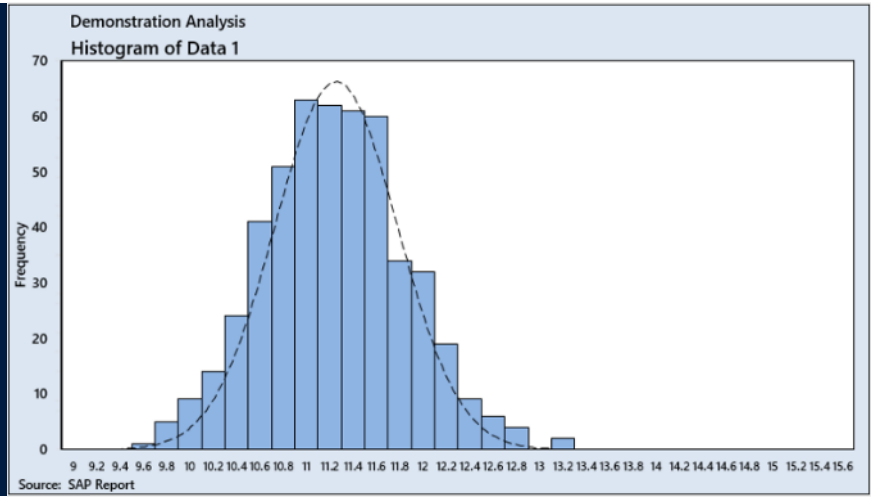
# Advanced Analytics Solutions
# Data Analysis Toolkit User Guide




2021 Edition

V2 April 2021

Advanced Analytics Solutions

# Contents

**This User Guide provides comprehensive instructions on all the worksheets in the Data Analysis Toolkit:**



Problems opening your Toolkit for the first time? The solution is here.

# About this User Guide

**This User Guide will show you every feature of the Advanced Analytics Solutions Data Analysis Toolkit.**

This is not a text book on graphical analysis, statistics or Lean Six Sigma

- Explaining the meaning of all the analysis here would require a much larger document, and there are many books available that perform this service well
- If you would like help with the underlying theory, please get in touch: toolkit@advancedanalyticssolutions.co.uk

The User Guide is intended to be used for reference, not to be read from start to finish

- After reading about the first few tools, you will soon realise that the Toolkit is simple and intuitive to use
- Having grasped the way it works, you will probably work out what to do for yourself in most cases
- We hope you find the answers to your questions in this document, but if you need help or clarification, please email us: toolkit@advancedanalyticssolutions.co.uk

# How does the 2021 Edition Compare to the 2020 Edition?

**The 2021 Edition includes a number of new advanced features:**

- Descriptive Statistics has its own hyperlink and can be selected independently from the Time Series Plot
- Control charts for attribute data (NP and C charts for number of defective units and counts of incidents respectively)
- Process Capability now offers the option to conduct square root or natural logarithm transformations of your data
  - Unlike other data analysis software which will stop with an error message if you try to transform data that has negative values, the Toolkit will automatically recognise these and adjust the transformation so that it works with your data
- Fitted Line Plot has been renamed Scatter Plot and Regression, and now gives you the ability to create a stratified scatter plot in addition to performing simple linear regression
- A new tool, Multiple Regression, has been added to perform regression several continuous Xs.
  - You can also get the toolkit to model any of the Xs with a curved response with a single click.
  - Multiple Regression also includes a Matrix Plot for up to 8 Xs, which enables you to visually check for relationships between them (ie, multicollinearity)
  - Variance Inflation Factors are provided, to quantify multicollinearity
- The Pareto Chart and Chi Squared Analysis will both handle larger numbers of factors, increasing their flexibility and power
- It's easier to choose between data entry methods for GR&R and DOE
- Portuguese and Spanish language versions have been added (the Toolkit now supports 10 languages: Bulgarian, English, French, German, Hungarian, Italian, Malay, Portuguese, Russian and Spanish)

# Instructions and License

# Purchasing and Activating Your Toolkit

## The purchase and activation process involves:*

- Purchase the toolkit from https://advancedanalyticssolutions.co.uk/data-analysis-toolkit/
- After you have paid for the toolkit, you will be asked to download the User Code Generator
- Use this file to generate a 5-digit code that is unique to your computer
- Submit this 5-digit code in the same web page that provided the User Code Generator
- You will receive the Toolkit by email a few hours later
- Save the toolkit to your hard drive and when you first open it, click ‘Enable Editing’ if prompted to do so in order to activate it.

**Please see the Troubleshooting section here if you have any problem with this process**

*none of these steps applies to the free 30-day trial version, only to the full purchased version.

# Instructions and License

Instructions and License

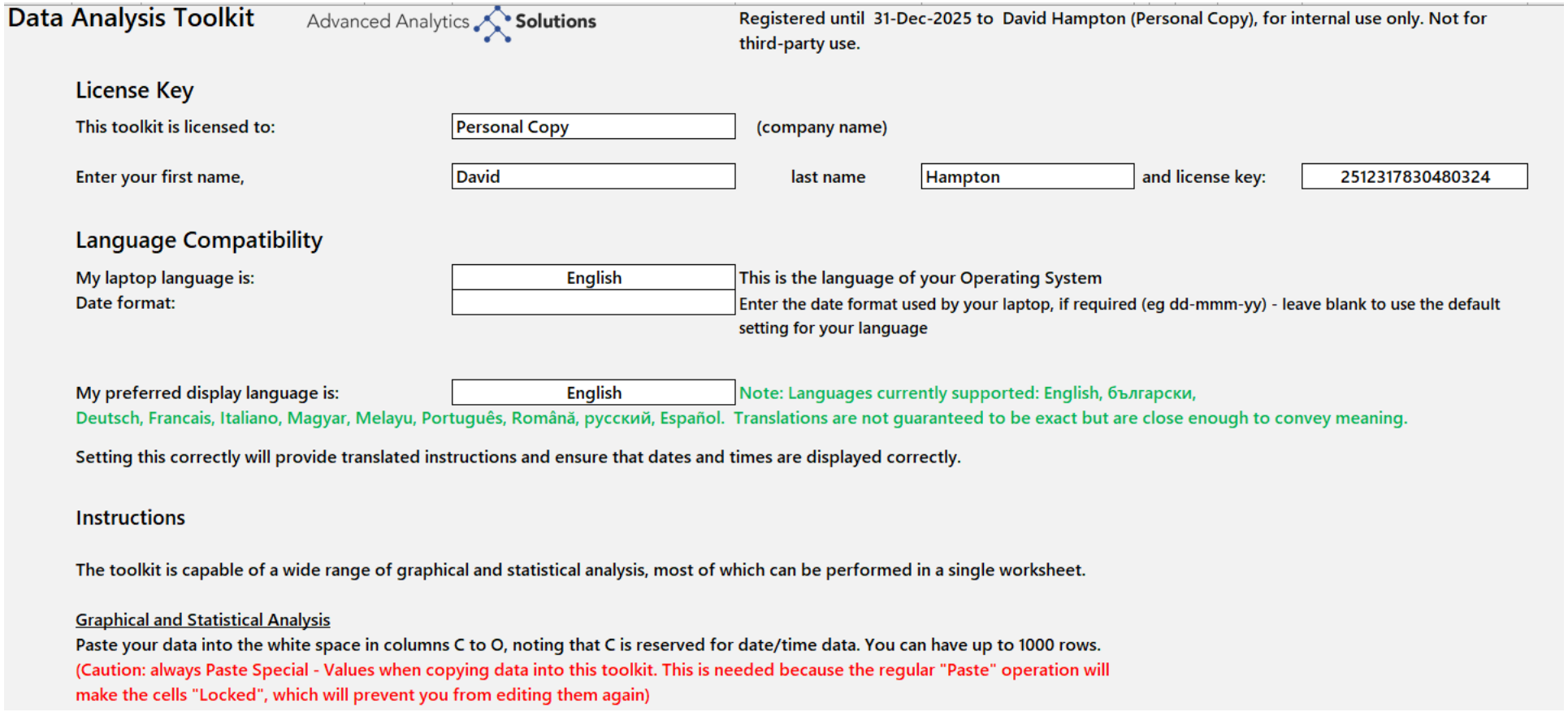
## This worksheet is used for:

- User and license information
- Language choices
- Basic instructions

## User and License Information

When you first purchase the Toolkit, the company name (corporate clients only), user name and license key are already configured.

If you make *any* changes to this data, the toolkit will stop working. So please don't.

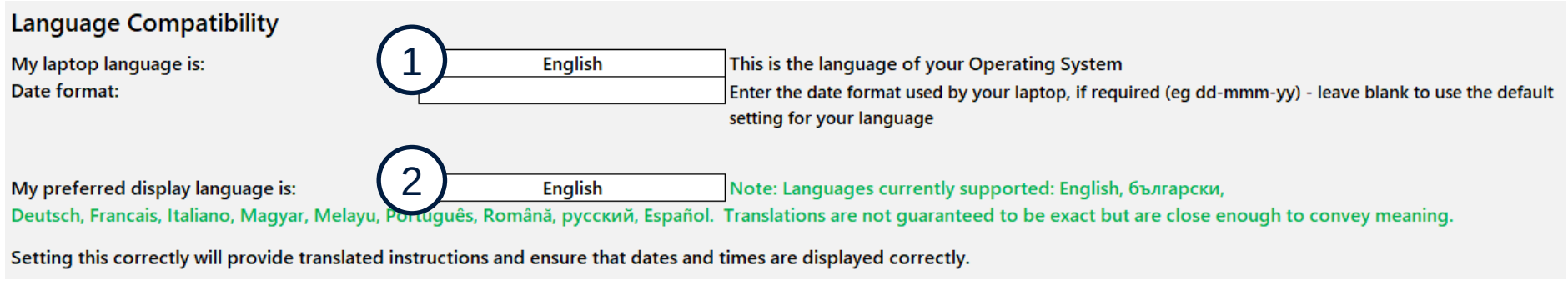


# Instructions and License



## Language Choices

The Toolkit has the capability to work in any of the following languages: Bulgarian, French, German, Hungarian, Italian, Malay, Portuguese, Romanian, Russian and Spanish.  More languages are planned, based on customer demand.




## How to set the correct language options

There are two things you need to provide:

- What language does your laptop operate in? (1)
  - This will be the language that Windows uses for dialogue boxes, error messages etc
  - Linked to this, what is your date format? – we shall cover this in the next slide
- What language would you like to use to display messages in (2)?
  - This is the one that will make the most obvious change – it sets the language of all the instructions and messages in the toolkit.

# Instructions and License



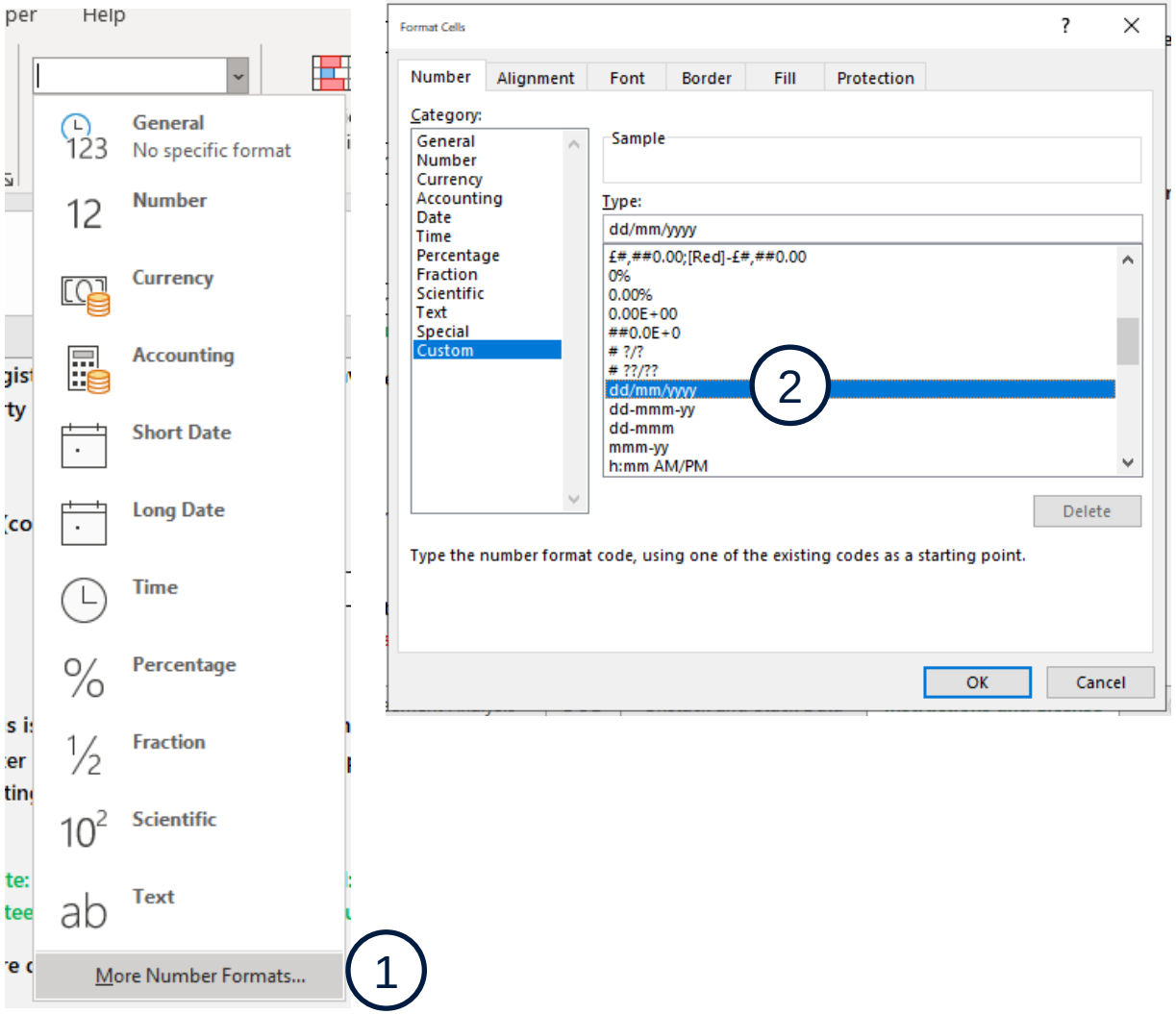
## Language Choices – setting your date format

If your laptop language is English, you do not need to take any action here.

If your laptop language is different, you will need to enter your date format into cell E15 so that the time axis on Time Series Plots and Control Charts display correctly.

To see the text that you should enter into the date format cell:

- In the number section of the Excel menu bar, select More Number Formats as shown (1)
- Within the sub-menu, select Customer as shown (2)
- Scroll down until you see, in text form, the options for date formats
- This will give you the text to be entered in the date format box
  - In this example on an English language laptop, it's dd/mmm/yyyy but you could equally choose dd-mmm-yy
  - On a French laptop (for example) you would see jj/mmm/aaaa - short for jour (day), mois (month) and annee (year)
- If you have difficulties with this process, please contact toolkit@advancedanalyticssolutions..co.uk



# Getting Started

**How to enter data and how Graphical and Statistical Analysis is organised**

# How To Enter Data

## We'll start with the worksheet "Graphs and Statistical Analysis"

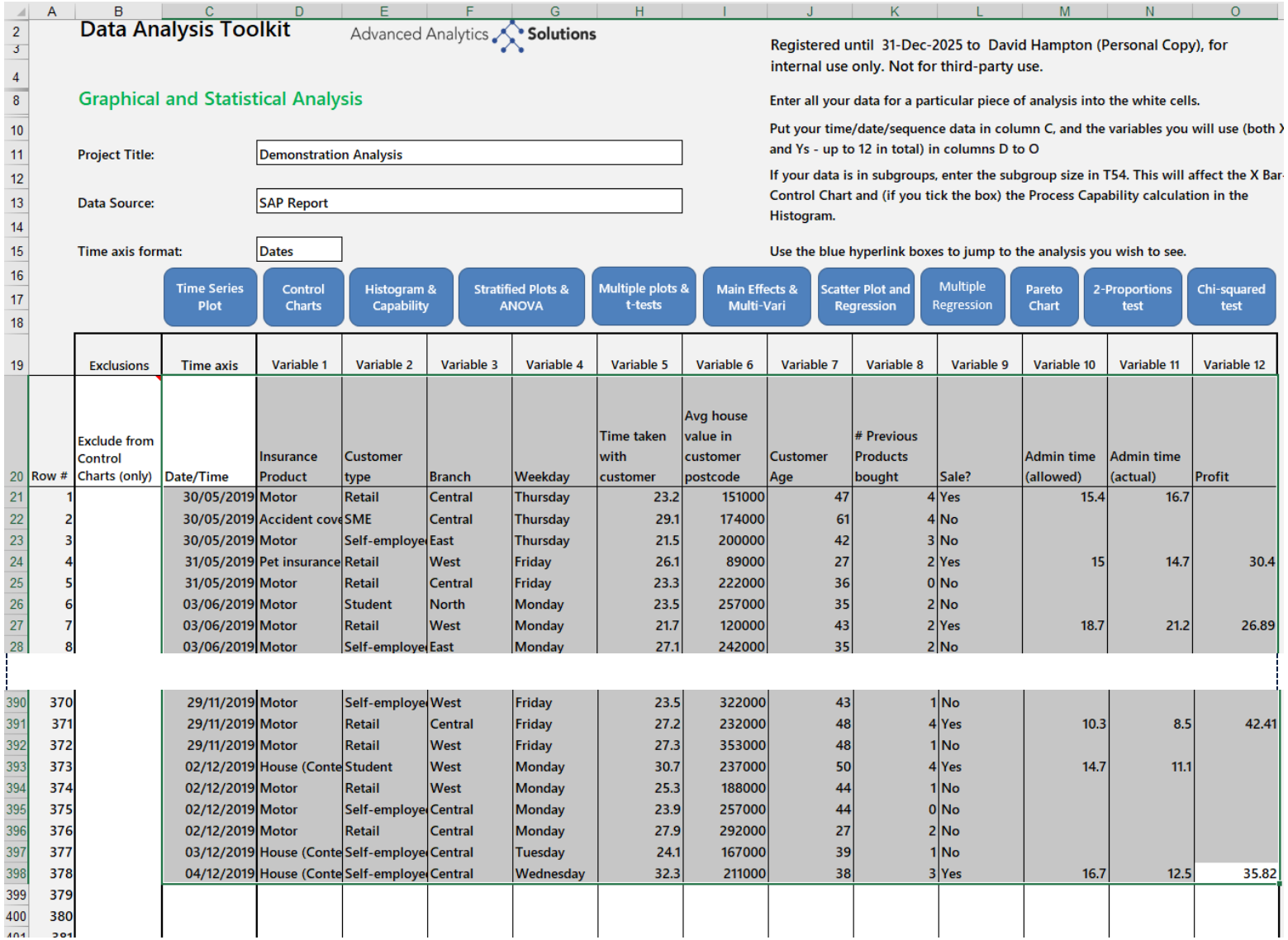
- All the commonly-used graphical and statistical tools are here

The worksheet has some demonstration data already, which makes it easier to see how it works intuitively

- For the purpose of this User Guide, we shall use fresh data
- Highlight all the data in this worksheet and delete it (the existing data has 378 rows)

Hints & Tips

- Cells that have a white background can be edited. Cells that have a grey background cannot.
- To quickly highlight all the data, select the cell in the top-left corner (here, it is the one called "Date/Time"), then hold down the Control and Shift keys, then press the right arrow and the down arrow



Data Analysis Toolkit — Advanced Analytics Solutions

Registered until 31-Dec-2025 to David Hampton (Personal Copy), for internal use only. Not for third-party use.

Graphical and Statistical Analysis

Enter all your data for a particular piece of analysis into the white cells.

Put your time/date/sequence data in column C, and the variables you will use (both X and Ys - up to 12 in total) in columns D to O

If your data is in subgroups, enter the subgroup size in T54. This will affect the X Bar Control Chart and (if you tick the box) the Process Capability calculation in the Histogram.

Project Title: Demonstration Analysis

Data Source: SAP Report

Time axis format: Dates

Use the blue hyperlink boxes to jump to the analysis you wish to see.

Time Series Plot | Control Charts | Histogram & Capability | Stratified Plots & ANOVA | Multiple plots & t-tests | Main Effects & Multi-Vari | Scatter Plot and Regression | Multiple Regression | Pareto Chart | 2-Proportions test | Chi-squared test

| | Exclusions | Time axis | Variable 1 | Variable 2 | Variable 3 | Variable 4 | Variable 5 | Variable 6 | Variable 7 | Variable 8 | Variable 9 | Variable 10 | Variable 11 | Variable 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Row # | Exclude from Control Charts (only) | Date/Time | Insurance Product | Customer type | Branch | Weekday | Time taken with customer | Avg house value in customer postcode | Customer Age | # Previous Products bought | Sale? | Admin time (allowed) | Admin time (actual) | Profit |
| 1 | | 30/05/2019 | Motor | Retail | Central | Thursday | 23.2 | 151000 | 47 | 4 | Yes | 15.4 | 16.7 | |
| 2 | | 30/05/2019 | Accident cove | SME | Central | Thursday | 29.1 | 174000 | 61 | 4 | No | | | |
| 3 | | 30/05/2019 | Motor | Self-employe | East | Thursday | 21.5 | 200000 | 42 | 3 | No | | | |
| 4 | | 31/05/2019 | Pet insurance | Retail | West | Friday | 26.1 | 89000 | 27 | 2 | Yes | 15 | 14.7 | 30.4 |
| 5 | | 31/05/2019 | Motor | Retail | Central | Friday | 23.3 | 222000 | 36 | 0 | No | | | |
| 6 | | 03/06/2019 | Motor | Student | North | Monday | 23.5 | 257000 | 35 | 2 | No | | | |
| 7 | | 03/06/2019 | Motor | Retail | West | Monday | 21.7 | 120000 | 43 | 2 | Yes | 18.7 | 21.2 | 26.89 |
| 8 | | 03/06/2019 | Motor | Self-employe | East | Monday | 27.1 | 242000 | 35 | 2 | No | | | |
| 370 | | 29/11/2019 | Motor | Self-employe | West | Friday | 23.5 | 322000 | 43 | 1 | No | | | |
| 371 | | 29/11/2019 | Motor | Retail | Central | Friday | 27.2 | 232000 | 48 | 4 | Yes | 10.3 | 8.5 | 42.41 |
| 372 | | 29/11/2019 | Motor | Retail | West | Friday | 27.3 | 353000 | 48 | 1 | No | | | |
| 373 | | 02/12/2019 | House (Conte | Student | West | Monday | 30.7 | 237000 | 50 | 4 | Yes | 14.7 | 11.1 | |
| 374 | | 02/12/2019 | Motor | Retail | West | Monday | 25.3 | 188000 | 44 | 1 | No | | | |
| 375 | | 02/12/2019 | Motor | Self-employe | Central | Monday | 23.9 | 257000 | 44 | 0 | No | | | |
| 376 | | 02/12/2019 | Motor | Retail | Central | Monday | 27.9 | 292000 | 27 | 2 | No | | | |
| 377 | | 03/12/2019 | House (Conte | Self-employe | Central | Tuesday | 24.1 | 167000 | 39 | 1 | No | | | |
| 378 | | 04/12/2019 | House (Conte | Self-employe | Central | Wednesday | 32.3 | 211000 | 38 | 3 | Yes | 16.7 | 12.5 | 35.82 |
| 379 | | | | | | | | | | | | | | |
| 380 | | | | | | | | | | | | | | |

# How To Enter Data (continued)

## Open this embedded Excel file “Practice Data”

- This contains fictional data about the number of calls handled in a Call Centre.  Additional information includes weekday and IT system version.
- Highlight all the data in the first worksheet and Copy
- In the toolkit, select cell C20 then select the drop-down arow under Paste and Paste Values as shown

### Why is this needed?

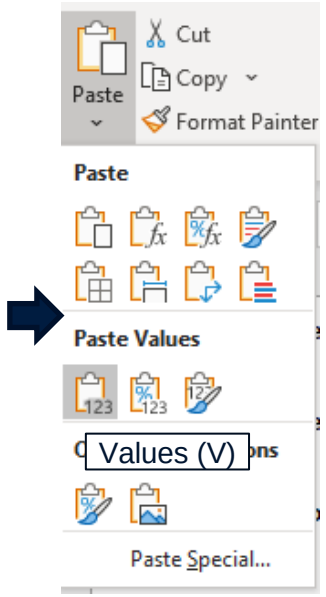
- If you simply ‘Paste’, you will lose the formatting and spoil the appearance of the toolkit … and in some cases the cells will become locked, so that you cannot change them
- It’s fiddly at first but you will soon get used to it

### Hints & Tips

- All Excel workbooks can be spoiled if you make a mistake. Make a backup copy of your toolkit NOW

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Call Date | Weekday | Calls Handled | Calls not resolved within 24 hours | Complaints Received | IT System |
| 2 | 30/03/2020 | Monday | 1897 | 190 | 3 | Original |
| 3 | 31/03/2020 | Tuesday | 1532 | 152 | 4 | Original |
| 4 | 01/04/2020 | Wednesday | 1437 | 146 | 3 | Original |
| 5 | 02/04/2020 | Thursday | 1570 | 156 | 1 | Original |
| 6 | 03/04/2020 | Friday | 1798 | 183 | 4 | Original |
| 7 | 06/04/2020 | Monday | 1841 | 187 | 1 | Original |
| 8 | 07/04/2020 | Tuesday | 1745 | 175 | 3 | Original |
| 9 | 08/04/2020 | Wednesday | 1768 | 173 | 5 | Original |
| 10 | 09/04/2020 | Thursday | 1651 | 163 | 4 | Original |
| 11 | 10/04/2020 | Friday | 1674 | 165 | 2 | Original |
| 12 | 13/04/2020 | Monday | 1811 | 179 | 6 | Original |
| 13 | 14/04/2020 | Tuesday | 1492 | 151 | 4 | Original |
| 14 | 15/04/2020 | Wednesday | 1589 | 158 | 2 | Original |
| 15 | 16/04/2020 | Thursday | 1497 | 149 | 2 | Original |
| 16 | 17/04/2020 | Friday | 1695 | 170 | 5 | Original |
| 17 | 20/04/2020 | Monday | 1757 | 176 | 4 | Original |
| 18 | 21/04/2020 | Tuesday | 1594 | 163 | 4 | Original |
| 19 | 22/04/2020 | Wednesday | 1561 | 157 | 4 | Original |
| 20 | 23/04/2020 | Thursday | 1832 | 181 | 3 | Original |
| 21 | 24/04/2020 | Friday | 1699 | 169 | 5 | Original |
| 22 | 27/04/2020 | Monday | 1705 | 170 | 5 | Original |
| 23 | 28/04/2020 | Tuesday | 1623 | 160 | 3 | Original |
| 24 | 29/04/2020 | Wednesday | 1745 | 170 | 4 | Original |
| 25 | 30/04/2020 | Thursday | 1519 | 148 | 6 | Original |
| 26 | 01/05/2020 | Friday | 1540 | 154 | 5 | Original |
| 27 | 04/05/2020 | Monday | 1500 | 160 | 5 | New |
| 28 | 05/05/2020 | Tuesday | 1379 | 177 | 3 | New |
| 29 | 06/05/2020 | Wednesday | 1498 | 192 | 2 | New |
| 30 | 07/05/2020 | Thursday | 1581 | 198 | 2 | New |

(Not all the data is shown)

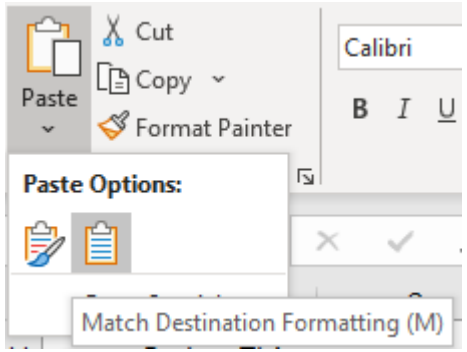



| | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|
| 19 | Exclusions | Time axis | Variable 1 | Variable 2 | Variable 3 | Variable 4 | Variable 5 |
| 20 | Exclude from Control Charts (only) | Call Date | Weekday | Calls Handled | Calls not resolved within 24 hours | Complaints Received | IT System |
| 21 | | 30/03/2020 | Monday | 1897 | 190 | 3 | Original |
| 22 | | 31/03/2020 | Tuesday | 1532 | 152 | 4 | Original |
| 23 | | 01/04/2020 | Wednesday | 1437 | 146 | 3 | Original |
| 24 | | 02/04/2020 | Thursday | 1570 | 156 | 1 | Original |
| 25 | | 03/04/2020 | Friday | 1798 | 183 | 4 | Original |
| 26 | | 06/04/2020 | Monday | 1841 | 187 | 1 | Original |
| 27 | | 07/04/2020 | Tuesday | 1745 | 175 | 3 | Original |
| 28 | | 08/04/2020 | Wednesday | 1768 | 173 | 5 | Original |
| 29 | | 09/04/2020 | Thursday | 1651 | 163 | 4 | Original |
| 30 | | 10/04/2020 | Friday | 1674 | 165 | 2 | Original |
| 31 | | 13/04/2020 | Monday | 1811 | 179 | 6 | Original |
| 32 | | 14/04/2020 | Tuesday | 1492 | 151 | 4 | Original |
| 33 | | 15/04/2020 | Wednesday | 1589 | 158 | 2 | Original |
| 34 | | 16/04/2020 | Thursday | 1497 | 149 | 2 | Original |
| 35 | | 17/04/2020 | Friday | 1695 | 170 | 5 | Original |
| 36 | | 20/04/2020 | Monday | 1757 | 176 | 4 | Original |
| 37 | | 21/04/2020 | Tuesday | 1594 | 163 | 4 | Original |
| 38 | | 22/04/2020 | Wednesday | 1561 | 157 | 4 | Original |
| 39 | | 23/04/2020 | Thursday | 1832 | 181 | 3 | Original |
| 40 | | 24/04/2020 | Friday | 1699 | 169 | 5 | Original |
| 41 | | 27/04/2020 | Monday | 1705 | 170 | 5 | Original |
| 42 | | 28/04/2020 | Tuesday | 1623 | 160 | 3 | Original |
| 43 | | 29/04/2020 | Wednesday | 1745 | 170 | 4 | Original |
| 44 | | 30/04/2020 | Thursday | 1519 | 148 | 6 | Original |
| 45 | | 01/05/2020 | Friday | 1540 | 154 | 5 | Original |

(Not all the data is shown)

# How To Enter Data (continued)

**If you are pasting from another application, use Paste > Match Destination Formatting (M)**



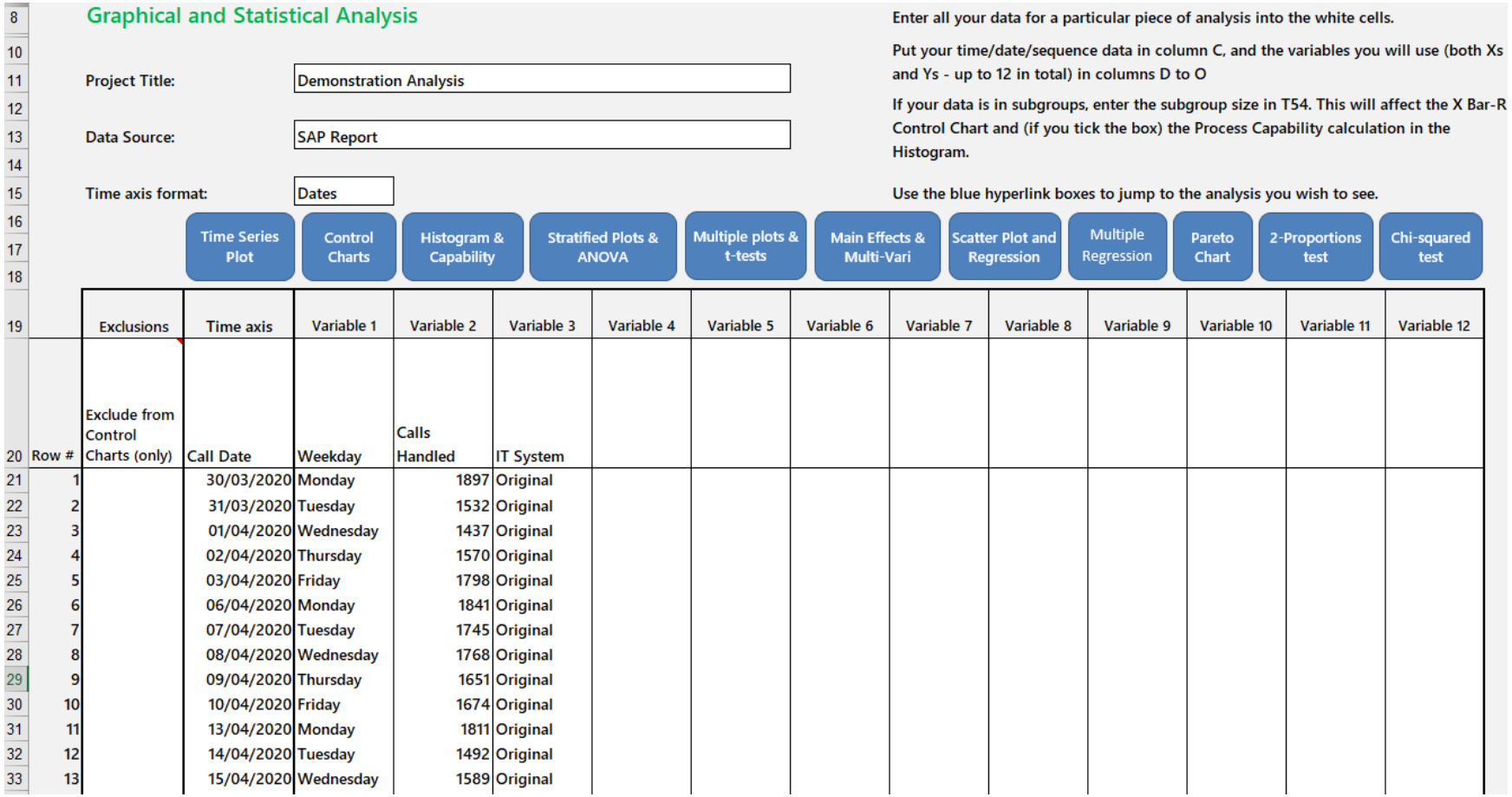

***This is essential* because if you use simple copy-paste from another application, the cells can become locked and cannot be changed again**

- This is a known problem with Excel but is unlikely to be fixed soon

# The design of the toolkit

## Note that the toolkit is designed for use with one dataset at a time

- Everything will be on the same time/date scale so there is only one time axis column



Graphical and Statistical Analysis

Enter all your data for a particular piece of analysis into the white cells.

Put your time/date/sequence data in column C, and the variables you will use (both Xs and Ys - up to 12 in total) in columns D to O

If your data is in subgroups, enter the subgroup size in T54. This will affect the X Bar-R Control Chart and (if you tick the box) the Process Capability calculation in the Histogram.

Use the blue hyperlink boxes to jump to the analysis you wish to see.

Project Title: Demonstration Analysis

Data Source: SAP Report

Time axis format: Dates

Time Series Plot | Control Charts | Histogram & Capability | Stratified Plots & ANOVA | Multiple plots & t-tests | Main Effects & Multi-Vari | Scatter Plot and Regression | Multiple Regression | Pareto Chart | 2-Proportions test | Chi-squared test

| | Exclusions | Time axis | Variable 1 | Variable 2 | Variable 3 | Variable 4 | Variable 5 | Variable 6 | Variable 7 | Variable 8 | Variable 9 | Variable 10 | Variable 11 | Variable 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Row # | Exclude from Control Charts (only) | Call Date | Weekday | Calls Handled | IT System | | | | | | | | | |
| 1 | | 30/03/2020 | Monday | 1897 | Original | | | | | | | | | |
| 2 | | 31/03/2020 | Tuesday | 1532 | Original | | | | | | | | | |
| 3 | | 01/04/2020 | Wednesday | 1437 | Original | | | | | | | | | |
| 4 | | 02/04/2020 | Thursday | 1570 | Original | | | | | | | | | |
| 5 | | 03/04/2020 | Friday | 1798 | Original | | | | | | | | | |
| 6 | | 06/04/2020 | Monday | 1841 | Original | | | | | | | | | |
| 7 | | 07/04/2020 | Tuesday | 1745 | Original | | | | | | | | | |
| 8 | | 08/04/2020 | Wednesday | 1768 | Original | | | | | | | | | |
| 9 | | 09/04/2020 | Thursday | 1651 | Original | | | | | | | | | |
| 10 | | 10/04/2020 | Friday | 1674 | Original | | | | | | | | | |
| 11 | | 13/04/2020 | Monday | 1811 | Original | | | | | | | | | |
| 12 | | 14/04/2020 | Tuesday | 1492 | Original | | | | | | | | | |
| 13 | | 15/04/2020 | Wednesday | 1589 | Original | | | | | | | | | |

Note the cells where you can enter the title of your project and the data source

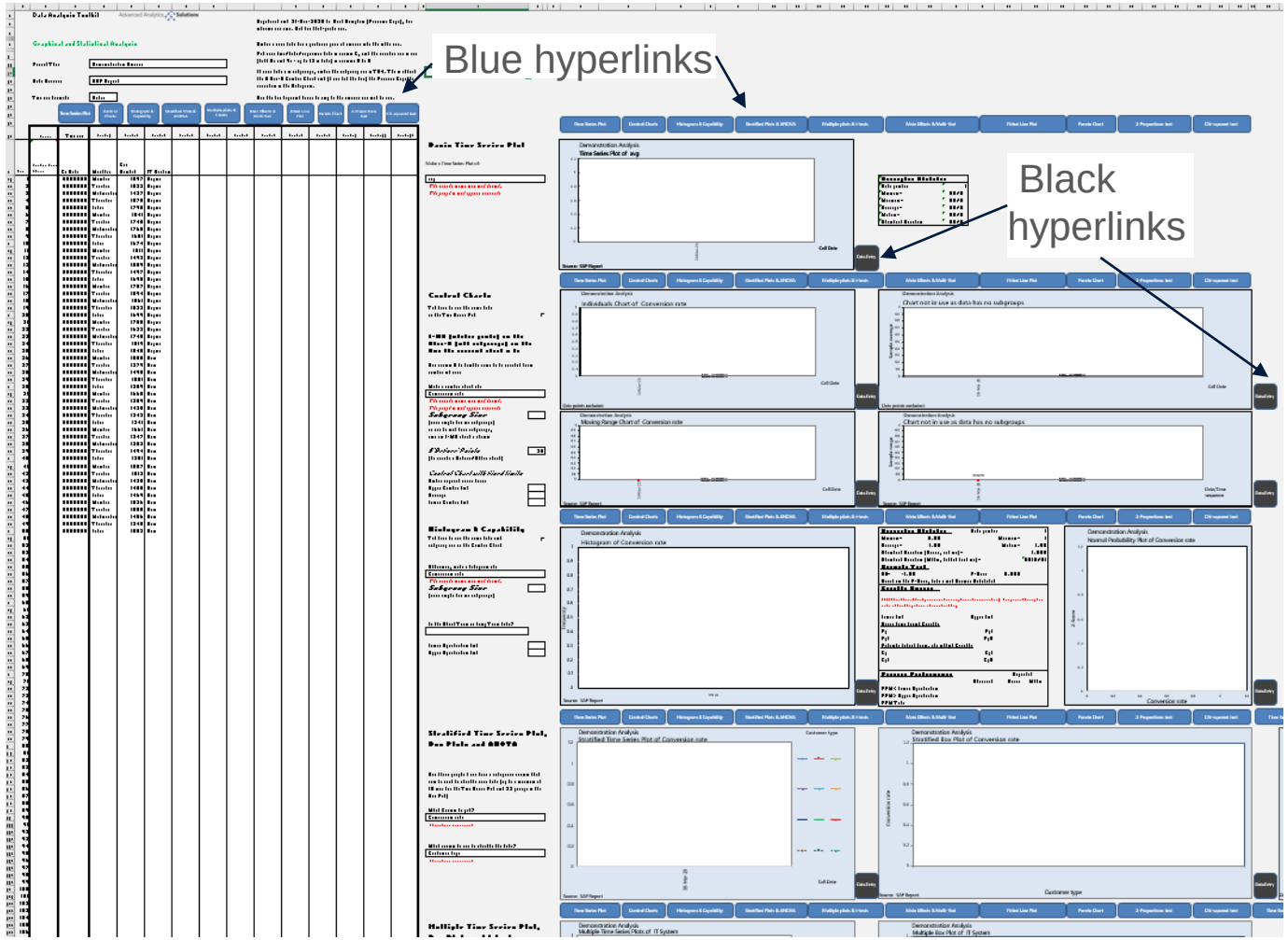
- We'll cover 'Time axis format' later

The blue buttons are hyperlinks that jump to the various types of analysis available

# The bigger picture

## There is a huge range of graphical and statistical tools available

- There are blue hyperlinks at regular intervals – these enable you to jump straight to the analysis you require
- There are also several black hyperlinks – these take you back to the data entry area



# Graphical and Statistical Analysis

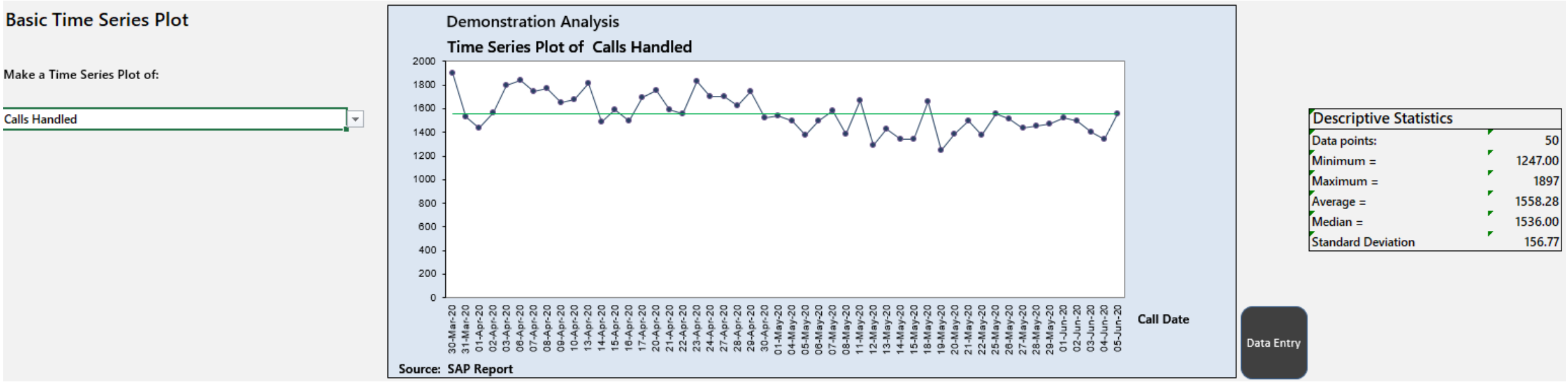
# Time Series Plots

**Used to visualise variation over time**

# Time Series Plot

**Let's make our first graph.**


- Click the hyperlink to Time Series Plot
- Select the column 'Calls Handled' from the drop-down list and the graph will appear
- The main descriptive statistics for your data appear to the right of the Time Series Plot
- Note the black hyperlink that will take you back to Data Entry

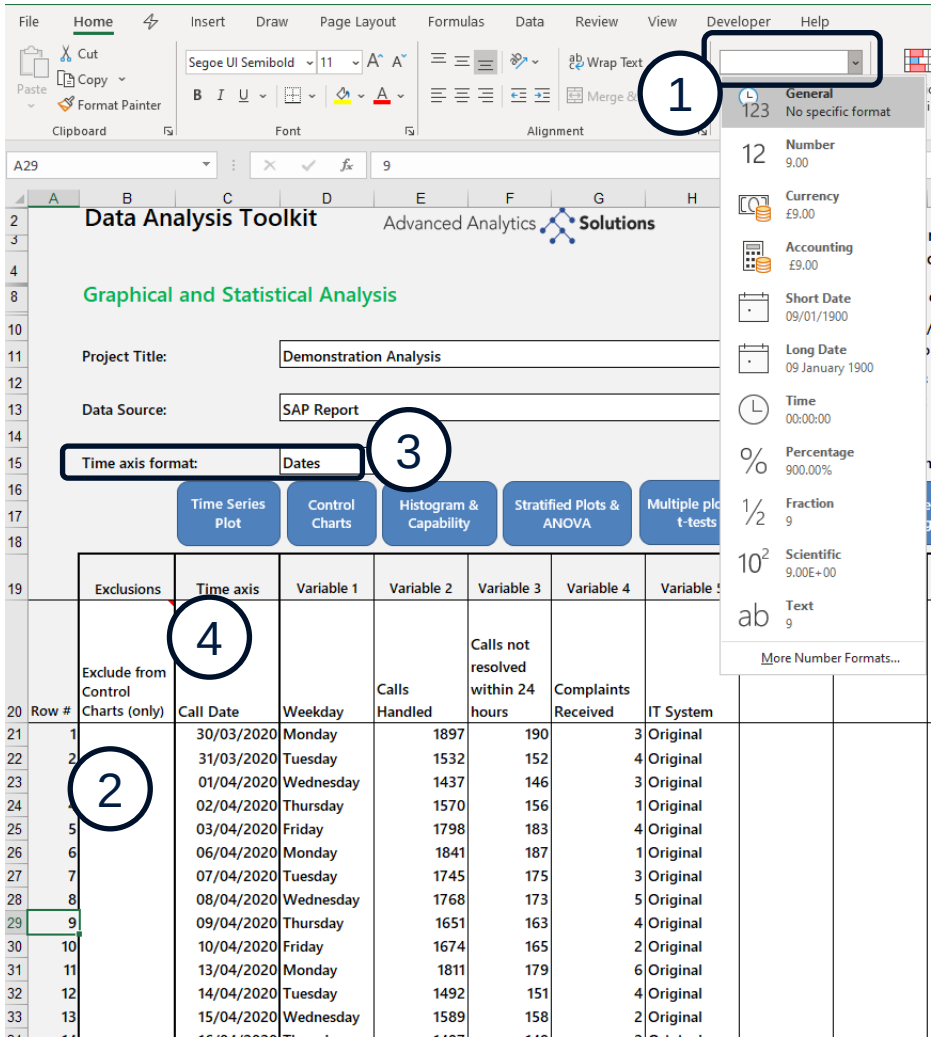


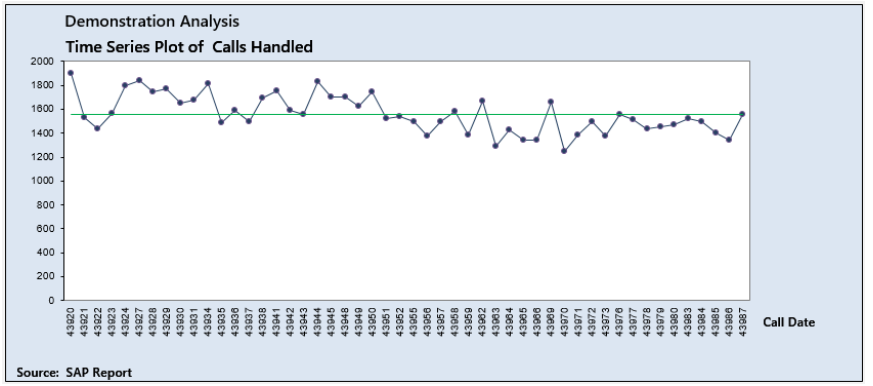
# Time Series Plot



## Changing the date format

- By default, the X-axis is in date format.
- If your data has a different format, you can change its appearance in Excel in the normal way, by highlighting the dates and then using the Number Format box as shown (1).
  - Note: this will only affect the formatting of the data (2), not the display in the graph
- Once you are comfortable with how to change the format and display, please change back to Date format (these are, after all, dates)




- To change the format of the X-axis in the Time Series plot, change the 'Time axis format' box (3), like this:


- This will cause the X-axis of the Time Series Plot to change, too
- Note that the X-axis label can be directly set in cell C20 (4)

# Copying graphs

Time Series Plot

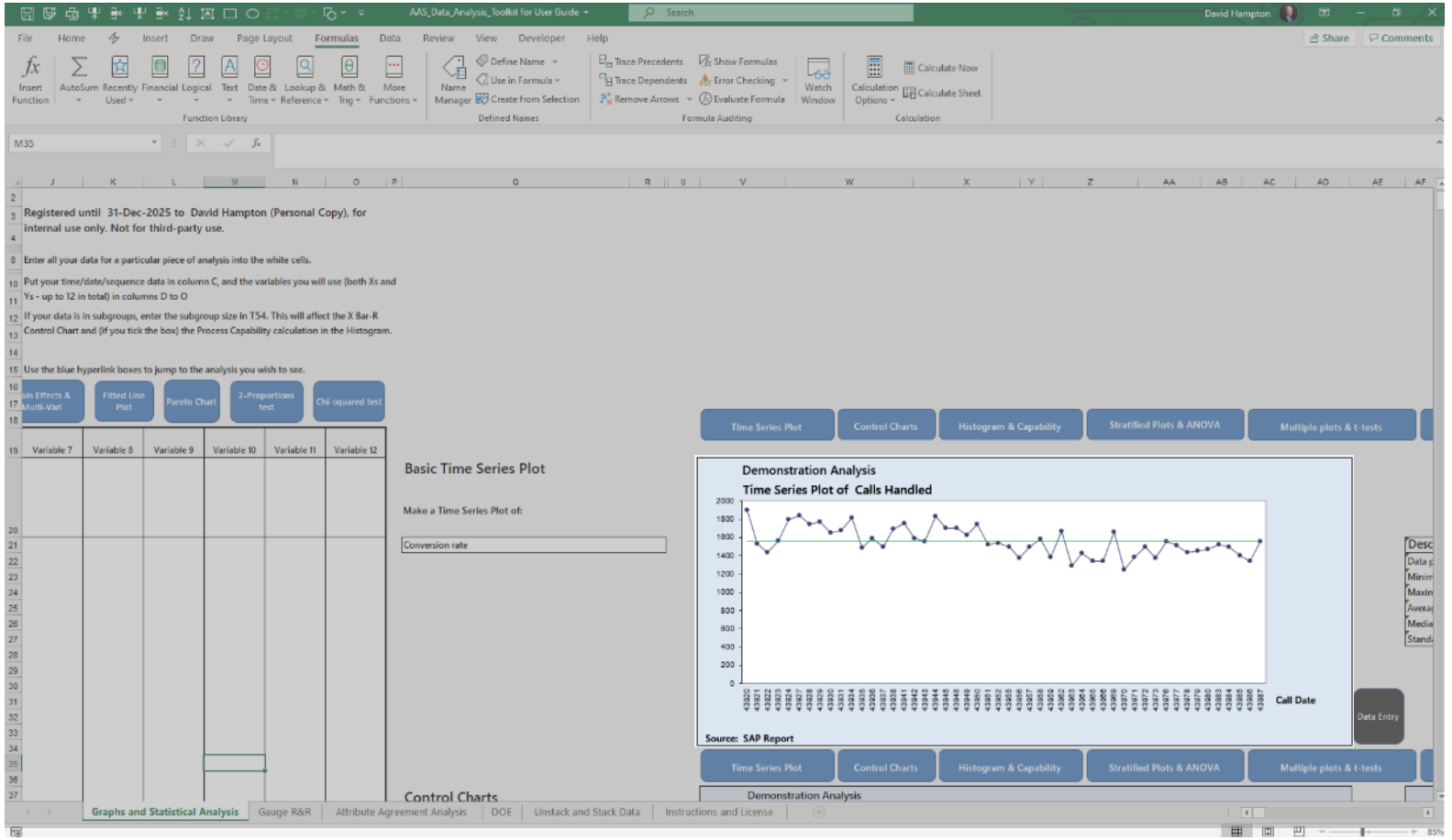
**To copy a graph, use Windows Snip & Sketch**

Shortcut: Shift -  - S

This tool snips a section of your screen and puts it onto your clipboard.

DON'T right-click on the graph and copy

- This copies the entire 12MB of the Data Analysis Toolkit – that will soon turn your presentation into a huge file
- Some graphs are comprised of several layers and you will only get the top layer



# Descriptive Statistics

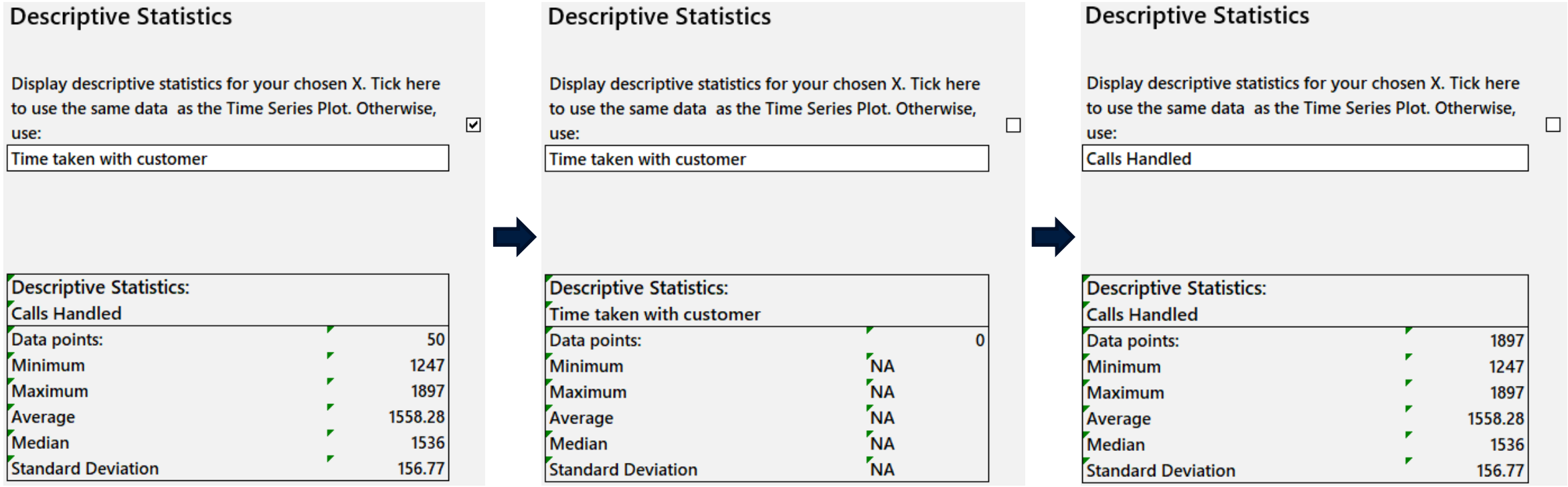
**Show me the numbers!**

# Descriptive Statistics

Descriptive Statistics

## Get the key numbers for your data

- The Descriptive Statistics provided are: the number of data points, Minimum, Maximum, Mean (Average), Median and Standard Deviation
- Click the hyperlink to Descriptive Statistics – this section is to the right of the Time Series Plot
- There is a tick box labelled “tick here to use the same data as the Time Series Plot”, and because this is already ticked, you’re looking at the descriptive statistics for Calls Handled. You can untick the box – now the Toolkit is trying to find Time Taken with Customer, which you deleted – so nothing is shown. Select Calls Handled from the drop-down list and the data reappears.



# Control Charts

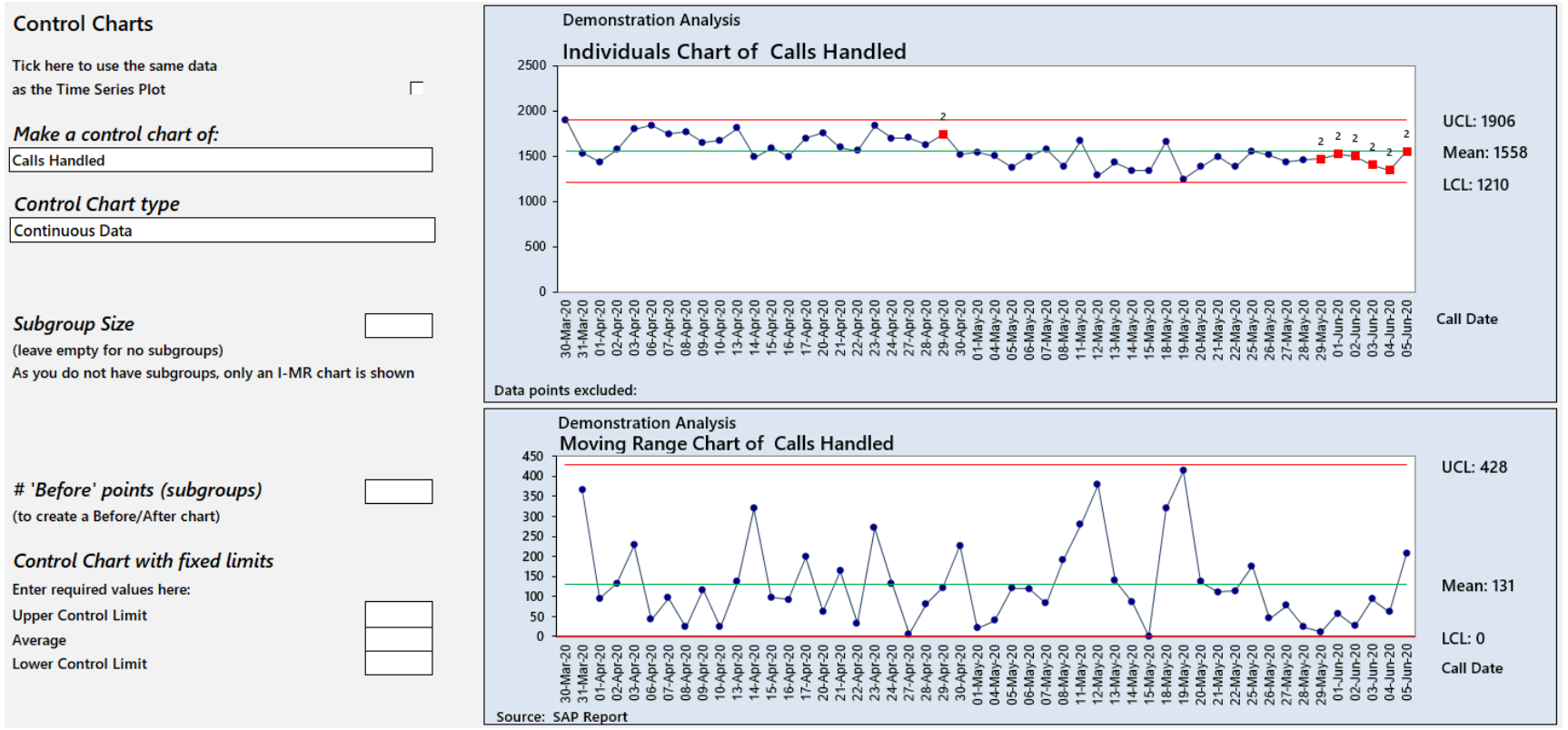
## Used to find Special Causes in the data

# Control Charts for Continuous Data

Control Charts

## Individual and Moving Range Chart

- Click the Control Chart hyperlink and use the drop-down box to select ‘Calls Handled’ as before
  - Alternatively, you can tick the box labelled ‘tick here to use the same data as the Time Series Plot’
- This is an Individual and Moving Range chart
  - The top chart shows the individual data points
  - The lower chart shows the moving range
- Special causes are highlighted with red squares and the cause type




Special Cause types:

1. Outlier (one point more than 3s from the mean)
2. Shift (nine consecutive points on the same side of the mean)
3. Trend (6 consecutive jumps all up or down)
4. Alternating (14 consecutive points alternating up and down)

# Control Charts for Continuous Data

Control Charts

## X Bar-R Control Charts

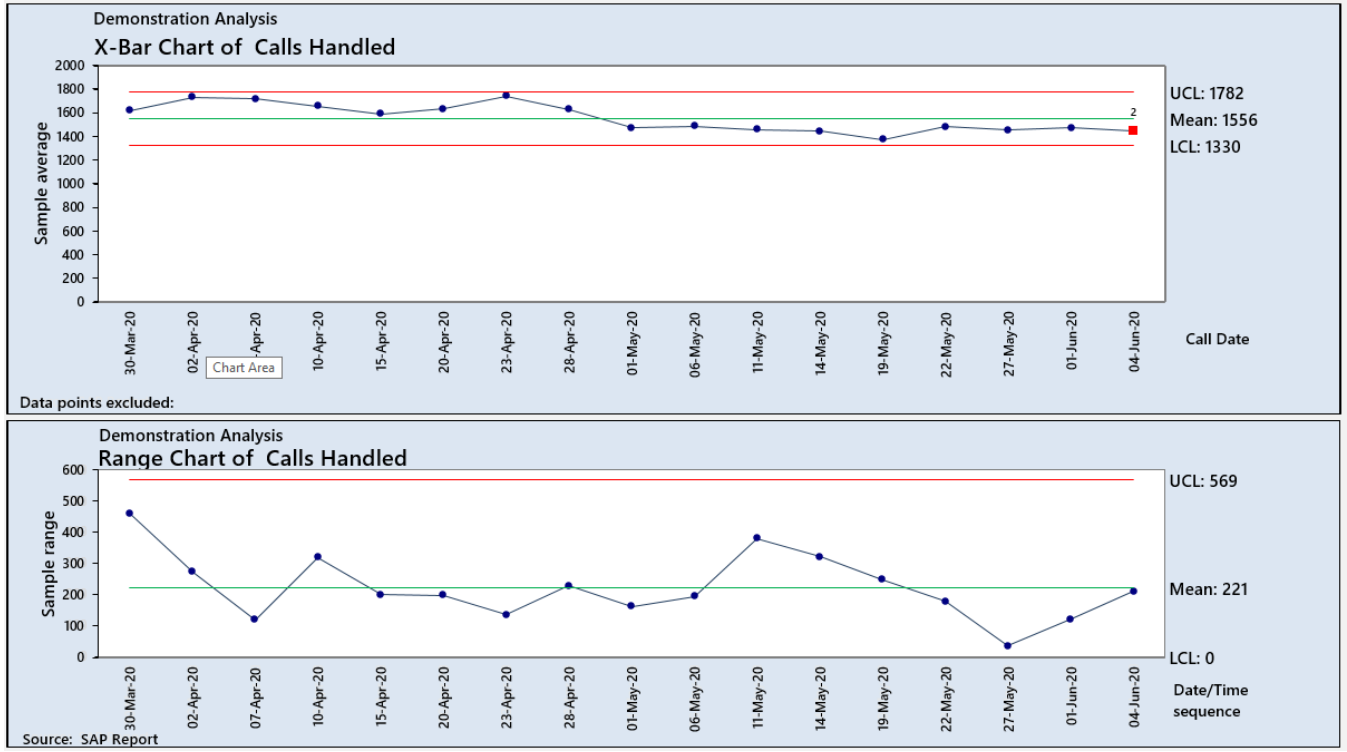
- If your data is in subgroups, use the Subgroup Size box to specify the subgroup size.
- Specify a subgroup size of 3 as shown, and watch what happens
  - The individual and Moving Range chart is no longer active because the data is in subgroups
  - Scroll right to see the X-Bar R chart
  - The top chart shows the subgroup averages, the bottom chart shows their ranges
- Please now delete the 3 in ‘Subgroup Size’
  - This isn’t really subgrouped data; we have just temporarily created subgroups to show how they are treated
  - All the remaining options are the same whether you have individual values or subgroups

***Subgroup Size*** 3

(leave empty for no subgroups)

as your data is in subgroups,

only an X bar-R chart is shown



# Control Charts for Continuous Data

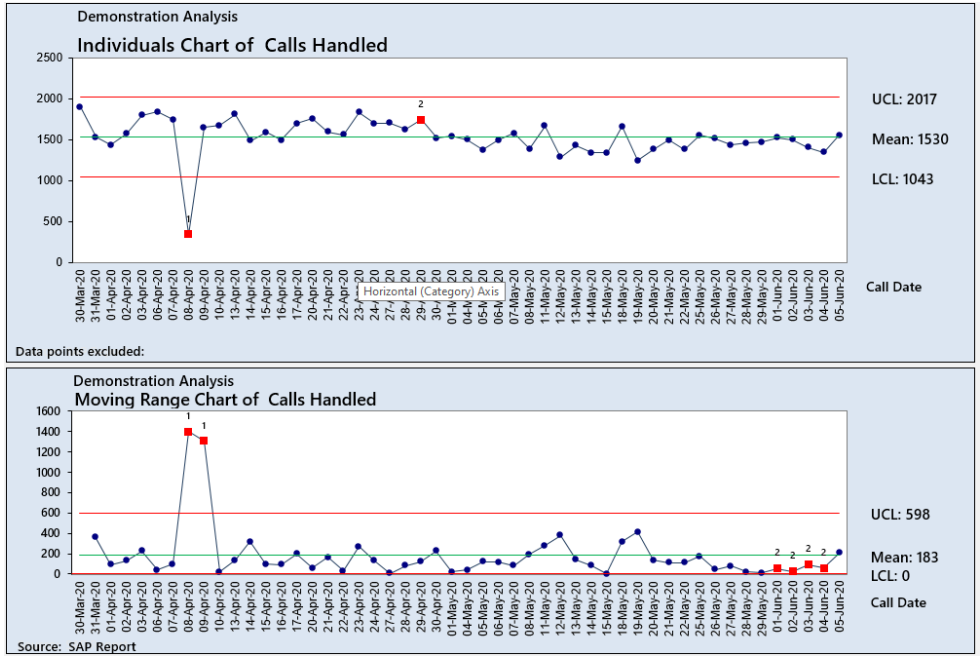


## Excluding data points

Suppose there was a major incident (a long power outage?) and this created a huge Special Cause on one day – here, we have changed the number of calls handled on 8 April to 345

- Notice what happens to the Control Limits
  - The LCL has dropped from 1211 (two slides back) to 1043, and the UCL has increased (because of the increase in variation) from 1905 to 2016.
  - These limits are now wrong because they no longer truly reflect the true natural variation in the process

| | Exclusions | Time axis | Variable 1 | Variable 2 | Variable 3 |
|---|---|---|---|---|---|
| Row # | Exclude from Control Charts (only) | Call Date | Weekday | Calls Handled | IT System |
| 1 | | 30/03/2020 | Monday | 1897 | Original |
| 2 | | 31/03/2020 | Tuesday | 1532 | Original |
| 3 | | 01/04/2020 | Wednesday | 1437 | Original |
| 4 | | 02/04/2020 | Thursday | 1570 | Original |
| 5 | | 03/04/2020 | Friday | 1798 | Original |
| 6 | | 06/04/2020 | Monday | 1841 | Original |
| 7 | | 07/04/2020 | Tuesday | 1745 | Original |
| 8 | | 08/04/2020 | Wednesday | 345 | Original |
| 9 | | 09/04/2020 | Thursday | 1651 | Original |
| 10 | | 10/04/2020 | Friday | 1674 | Original |
| 11 | | 13/04/2020 | Monday | 1811 | Original |
| 12 | | 14/04/2020 | Tuesday | 1492 | Original |
| 13 | | 15/04/2020 | Wednesday | 1589 | Original |
| 14 | | 16/04/2020 | Thursday | 1497 | Original |




- We need to do something about this!

# Control Charts for Continuous Data

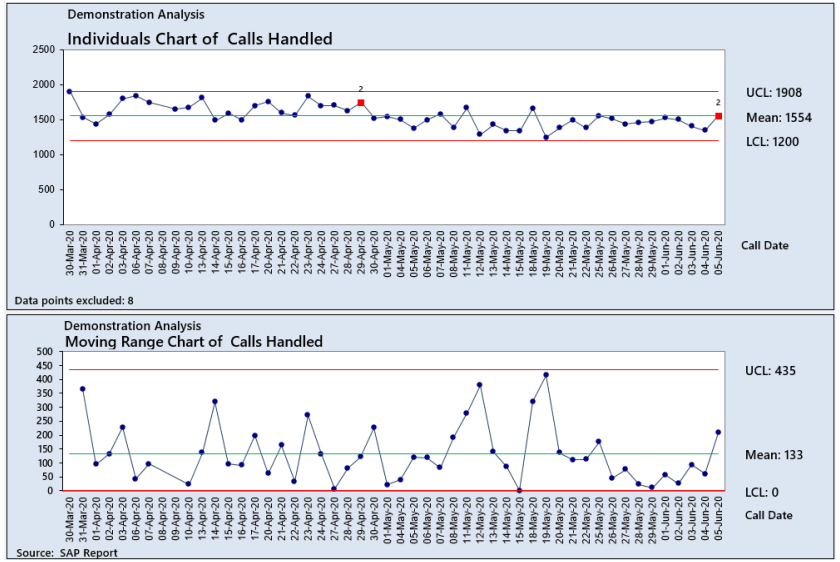


## Excluding data points (continued)

The natural way to deal with this situation is:

- Exclude the ‘rogue’ data point from our calculations (we know it was a one-off, not representative of the ongoing process)
- Still plot the data point (we are not trying to fiddle anything, just keep the chart accurate)
- To do this, put an ‘x’ in the Exclusions column as shown (note, this ONLY affects the Control Charts, not other graphs)

| | Exclusions | Time axis | Variable 1 | Variable 2 | Variable 3 |
|---|---|---|---|---|---|
| Row # | Exclude from Control Charts (only) | Call Date | Weekday | Calls Handled | IT System |
| 1 | | 30/03/2020 | Monday | 1897 | Original |
| 2 | | 31/03/2020 | Tuesday | 1532 | Original |
| 3 | | 01/04/2020 | Wednesday | 1437 | Original |
| 4 | | 02/04/2020 | Thursday | 1570 | Original |
| 5 | | 03/04/2020 | Friday | 1798 | Original |
| 6 | | 06/04/2020 | Monday | 1841 | Original |
| 7 | | 07/04/2020 | Tuesday | 1745 | Original |
| 8 | x | 08/04/2020 | Wednesday | 345 | Original |
| 9 | | 09/04/2020 | Thursday | 1651 | Original |
| 10 | | 10/04/2020 | Friday | 1674 | Original |
| 11 | | 13/04/2020 | Monday | 1811 | Original |
| 12 | | 14/04/2020 | Tuesday | 1492 | Original |
| 13 | | 15/04/2020 | Wednesday | 1589 | Original |
| 14 | | 16/04/2020 | Thursday | 1497 | Original |




The data point is still plotted but the Average and Control Limits no longer take it into account

- Note, they have still changed very slightly because the original data point has gone
- Please now remove the x and change the data point back to its original value, 1768

# Control Charts for Continuous Data

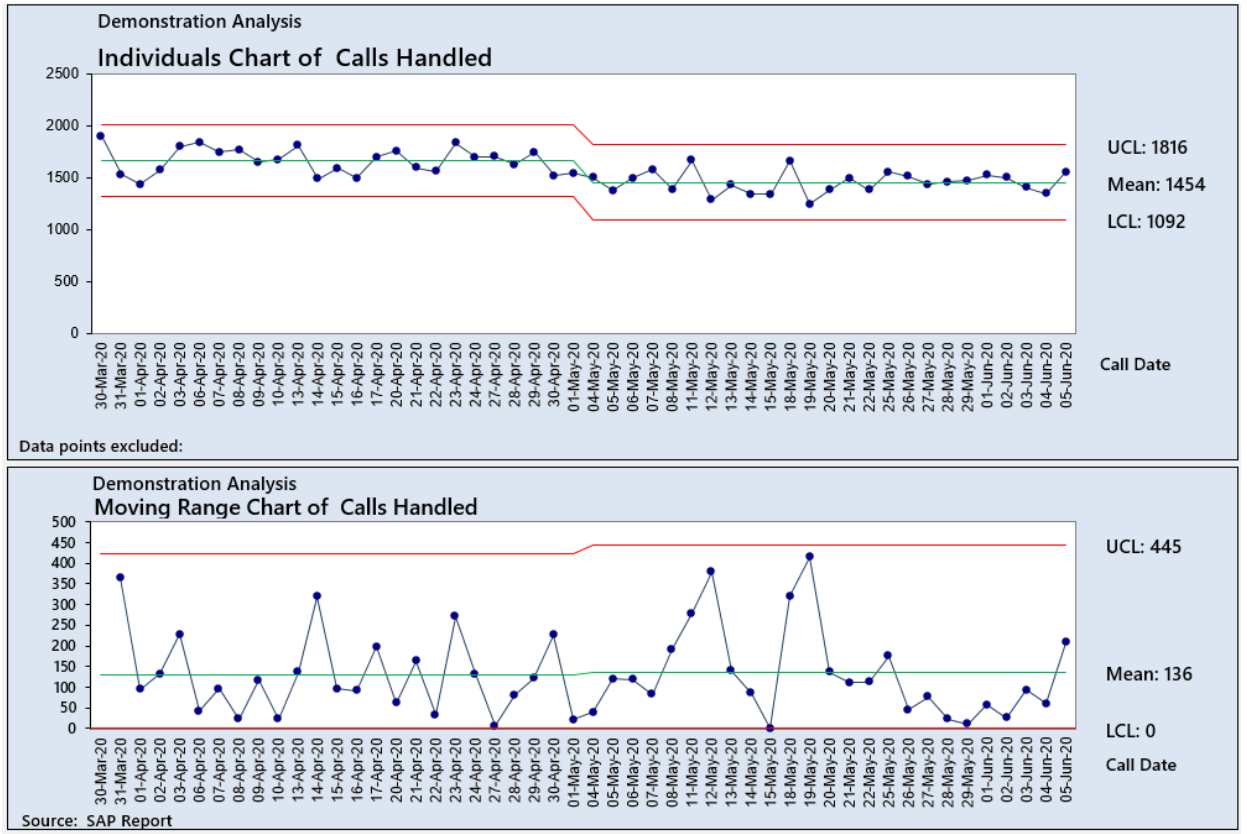


## Before and After Charts

You can break the data into two sections, to show 'Before' and 'After'

| # 'Before' Points | 25 |
|---|---|
| (to create a Before/After chart) | |

- Notice that the data has a column 'IT System' that shows whether we were using the original or the new system.
- Enter the number of points in the 'Before' stage (here, it's 25) and the toolkit will create the two stages as shown
- In this example you can see that the new system is associated with a reduction in the number of calls handled
- Efforts will be needed to find out why!



# Control Charts for Continuous Data



## Fixing Limits on the Control Chart

When you have established a stable process, it is often a good idea to fix the limits on the Control Chart

This means that the limits will not be automatically adjusted as you enter new data…

… which prevents them from widening to accommodate a change in the process…

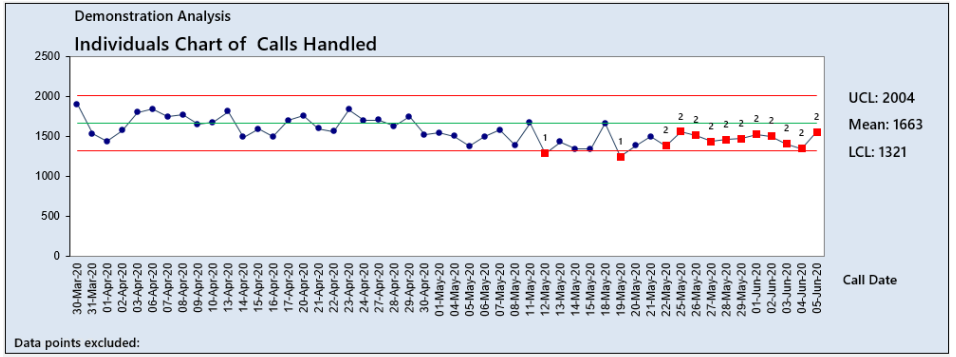
… which can make your Control Chart more sensitive to gradual changes

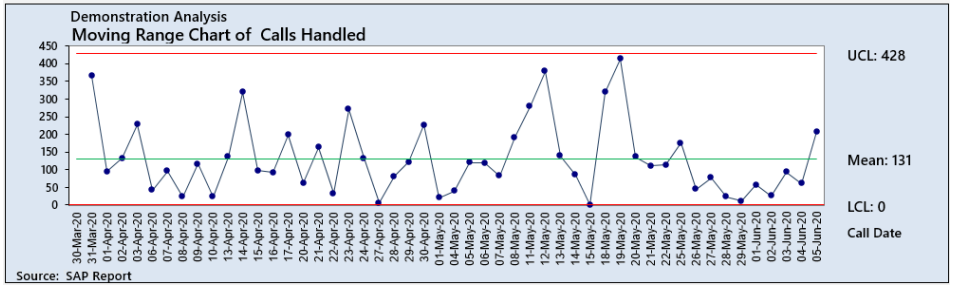
The call handling process was stable before the IT system was changed. We shall enter the previous UCL, Average and LCL, as shown in the data entry box

***Control Chart with fixed limits***

Enter required values here:

| | |
|---|---|
| Upper Control Limit | 2004 |
| Average | 1663 |
| Lower Control Limit | 1321 |




- You can see that the Control Limits and Average are now fixed at the values we entered
- The drop-off in the number of calls handled is even more obvious than it was with the original chart
  - We could have caught the problem more quickly in this way

# Control Charts for Attribute Data

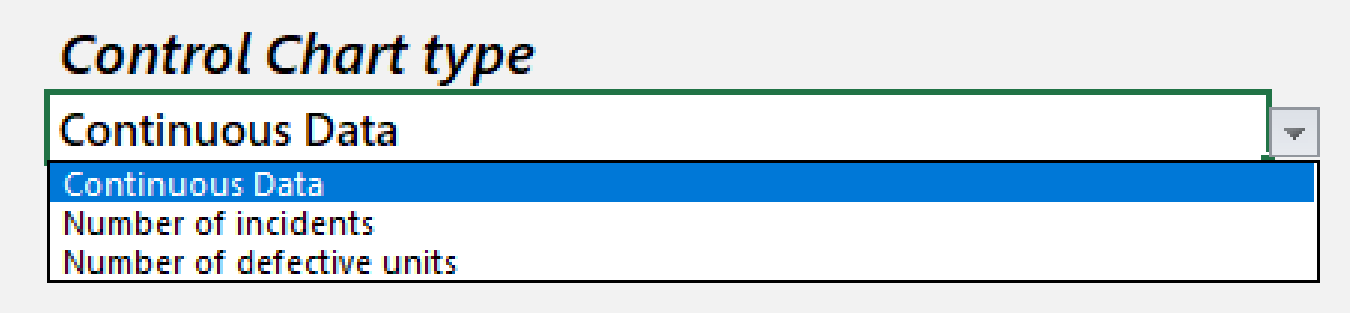
Control Charts

## C and NP Charts

So far, we have looked only at Control Charts for continuous data. We can also create charts for two types of attribute data:

- The number of incidents (used where you count the number of incidents within a specific time frame) – this is a C Chart
- The number of defective units (for when you sample a certain number of items and record how many of them are defective) – this is an NP Chart

To create these charts, use the drop-down menu under ‘Control Chart type’ to select the desired chart



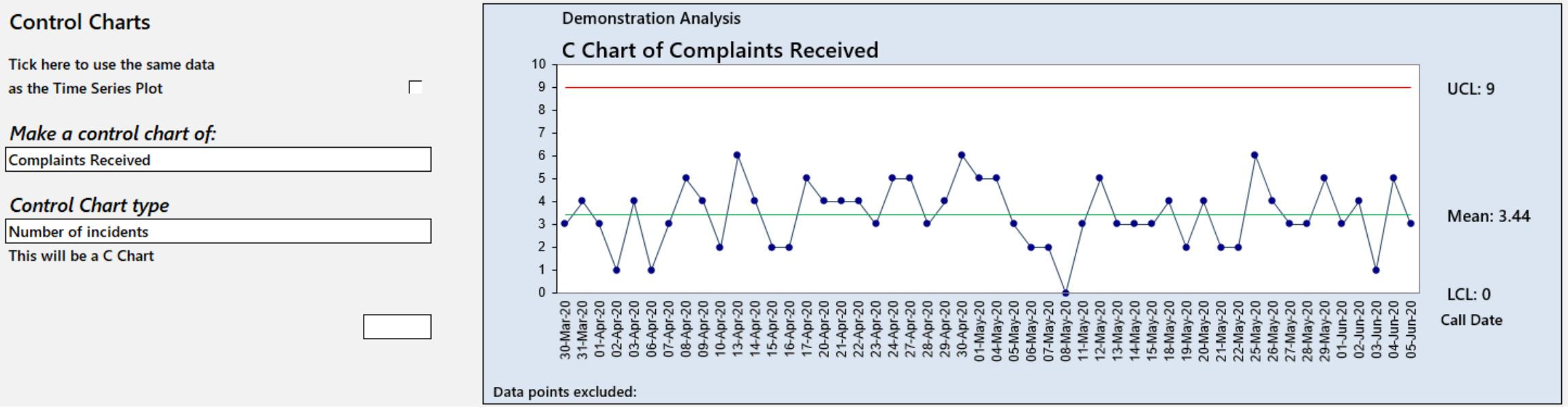

Note: Before we look at these charts, please delete the fixed control limits settings that you used in the previous slide, we will not need them again.

# Control Charts for Attribute Data



## C Charts

- C Charts are used to count the number of incidents in a specified period. They are based on a Poisson distribution.
  - We refer to the number of incidents here, but it could equally be the number of cars counted in a traffic survey, the number of safety incidents or the number of calls to a helpline. C Charts are used in cases where you can count the number of occurrences of something, but you cannot count the number of non-occurrences (you can't count the number of safety incidents that did not happen)
  - In any case, you will need to have a fixed, consistent time period for the count data. Note, C Charts do not use subgroups.
  - In our Calls Handled case, we'll use the number of complaints received
- Select 'Complaints Received' and change the Control Chart type to Number of incidents, as shown below:



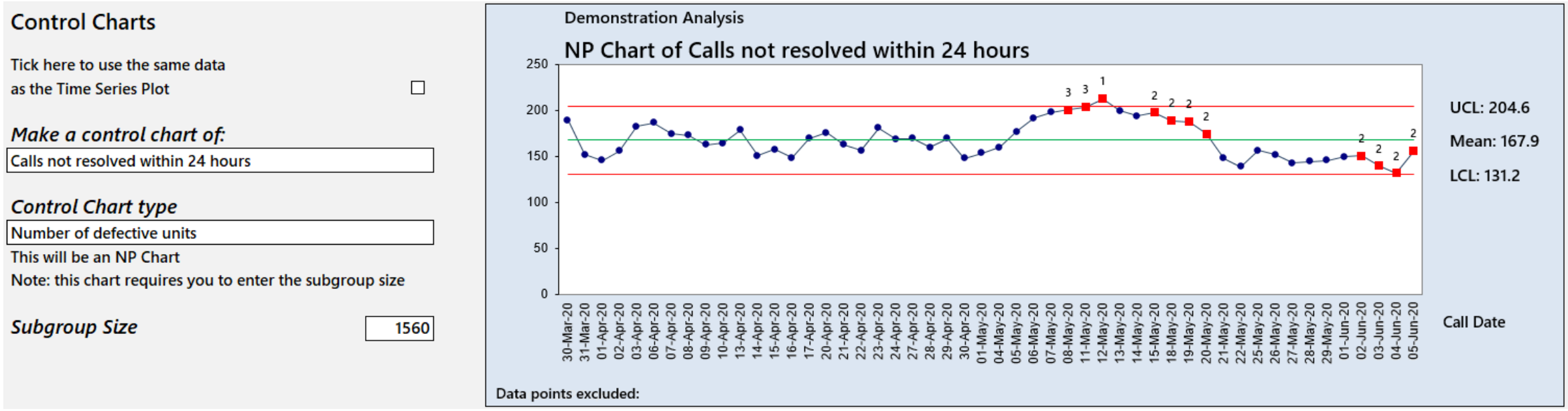

- Note, you do not need subgroups for this type of chart, so the 'Subgroup Size' heading has disappeared.

# Control Charts for Attribute Data



## NP Charts

- NP Charts are used to count the number of defective units in a sample. They are based on a binomial distribution.
  - These charts are used when you can count both OK and Not OK units – so every unit is either defective or not defective, for example
  - The number of units checked is entered in the Subgroup Size box (this needs to be roughly constant)
  - In our Calls Handled case, we'll use the number of calls not resolved within 24 hours
- Select 'Calls not resolved within 24 hours', change the Control Chart type to Number of defective units and enter the subgroup size as 1560 (which is roughly the average number of calls per day):
  - Note, unlike C Charts, NP Charts require a Subgroup Size



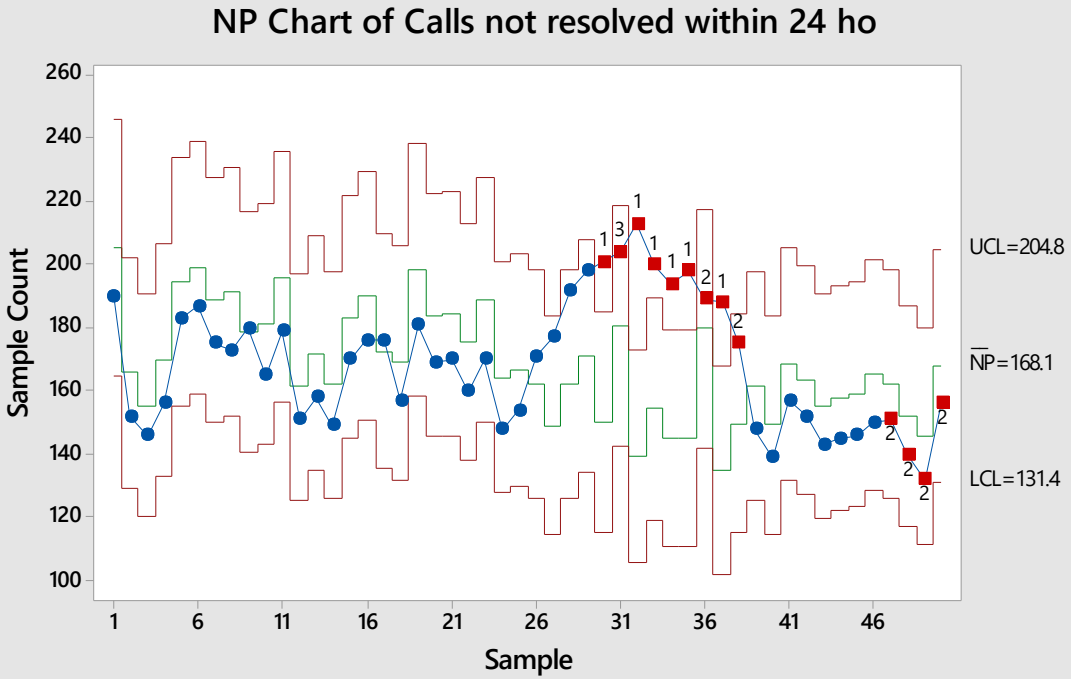

- There were evidently delays in resolving calls in the middle two weeks of May

# Wrinkles with Control Charts for Attribute Data



## Wrinkle number 1: the NP Subgroup size wasn't really constant

- We'll admit to taking a slight liberty with the data in the previous slide. We said you should enter the subgroup size of 1560 "which is roughly the average number of calls per day" – but actually, the number of calls received per day varies a lot – from about 1250 to about 1900.
- Strictly speaking, the mean and control limits should be adjusted every time the subgroup size changes, but the Toolkit will not do this. There are two reasons for this simplification:
  - In most cases, you will have a constant subgroup size (for example, a fixed number of units is checked every day and the number of defective units is recorded), and the subgroup size is simply the number of units checked – so there is no need
  - Adjusting the control limits every time the subgroup size change makes the chart very difficult to interpret (the illustration on the right was created using Minitab, with the same data), and the essence of Control Charts is to engage the people doing the work in process monitoring – these complex charts will get in the way of that. You can see that more or less the same special causes are identified by the Toolkit and Minitab, even in this very extreme example with widely varying subgroup sizes.



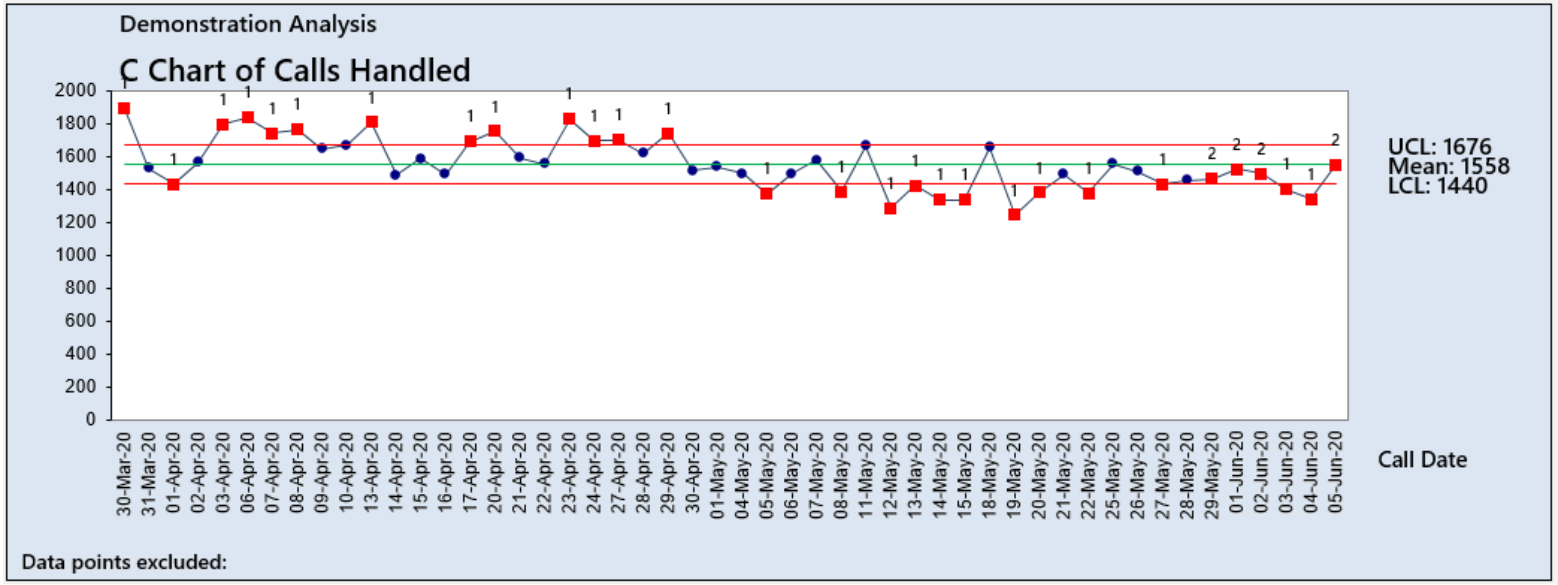

*Tests are performed with unequal sample sizes.*

# Wrinkles with Control Charts for Attribute Data



## Wrinkle number 2: Isn't the number of calls handled actually Count Data?

- Of course, strictly speaking, the number of calls handled – which we used to create an I-MR chart – is really count data, not continuous data.
- In our 'Number of complaints' data, this should clearly be treated as attribute data, as there aren't many of them – between 0 and 6 – so a C Chart is appropriate. When the counts are high, as with the number of calls handled, the data looks like continuous data (there are hundreds of possible values) and an I-MR Chart is better.
- Why not use a C Chart for Calls Handled, if it is technically correct? The answer is that, when you have a large number of 'incidents' to count, a C Chart becomes over-sensitive.
- You don't have to take our word for it – try making a C Chart for Calls Handled and you will see Special Causes firing off everywhere.
- This may be technically accurate, but it is of no practical use – the point of special causes is to trigger an investigation, and if every data point is a special cause, you are unable to prioritise your investigations.



# Histograms and Process Capability

**Used to check the shape of the distribution and determine performance**

# Histograms and Process Capability

Histogram & Capability

## Histograms

The Histogram and Capability hyperlink takes you to a number of pieces of analysis based around the histogram.

We'll use a different dataset for this, so delete the existing data and copy and paste values from the worksheet 'Impurities' with the 'Date/Time' heading going into C20 as before

- The data comes from a chemical manufacturing process in which a small percentage of impurities is generated.
- In addition to the production date and the impurity content, we have the batch of raw material that was used and the machine used to make the chemical (there are three machines - numbered 110, 180 and 220)
- If you wish, you can practice using Time Series Plots, Control Charts and Stratified Plots using this data

| 19 | | Exclusions | Time axis | Variable 1 | Variable 2 | Variable 3 | Variable 4 |
|---|---|---|---|---|---|---|---|
| 20 | Row # | Exclude from Control Charts (only) | Date/Time | Impurity Content | Batch | Shift | Machine |
| 21 | 1 | | 26/11/2019 | 3.21 | a | Day | 110 |
| 22 | 2 | | 26/11/2019 | 3.01 | a | Night | 180 |
| 23 | 3 | | 26/11/2019 | 2.78 | a | Day | 220 |
| 24 | 4 | | 27/11/2019 | 2.94 | a | Night | 110 |
| 25 | 5 | | 28/11/2019 | 2.97 | a | Day | 180 |
| 26 | 6 | | 29/11/2019 | 2.95 | a | Night | 220 |
| 27 | 7 | | 29/11/2019 | 2.95 | a | Day | 110 |
| 28 | 8 | | 30/11/2019 | 3.3 | a | Night | 180 |
| 29 | 9 | | 30/11/2019 | 3.38 | a | Day | 220 |
| 30 | 10 | | 01/12/2019 | 3.05 | a | Night | 110 |
| 31 | 11 | | 02/12/2019 | 3.05 | a | Day | 180 |
| 32 | 12 | | 02/12/2019 | 3.24 | a | Night | 220 |
| 33 | 13 | | 03/12/2019 | 2.99 | a | Day | 110 |
| 34 | 14 | | 03/12/2019 | 3.37 | a | Night | 180 |
| 35 | 15 | | 03/12/2019 | 2.98 | a | Day | 220 |
| 36 | 16 | | 03/12/2019 | 3.09 | a | Night | 110 |
| 37 | 17 | | 04/12/2019 | 3.6 | a | Day | 180 |
| 38 | 18 | | 04/12/2019 | 3.23 | a | Night | 220 |
| 39 | 19 | | 05/12/2019 | 3.18 | a | Day | 110 |
| 40 | 20 | | 05/12/2019 | 3.29 | a | Night | 180 |
| 41 | 21 | | 06/12/2019 | 3.67 | a | Day | 220 |
| 42 | 22 | | 07/12/2019 | 3.19 | a | Night | 110 |
| 43 | 23 | | 07/12/2019 | 3.21 | a | Day | 180 |

# Histograms

Histogram & Capability

- To make a histogram with this data, click the ‘Histogram and Capability’ hyperlink
- Select the column ‘Impurity content’ and, as usual, the histogram appears immediately
- The overlaid curves show Normal distributions with the same mean and standard deviation as the histogram
- The red line uses the ‘overall’ variation (see the explanation of standard deviation on the next slide)
- The dotted black line uses the ‘within’ variation (see the explanation of standard deviation on the next slide)

**Histogram & Capability**

Tick here to use the same data and subgroup size as the Control Chart ☐

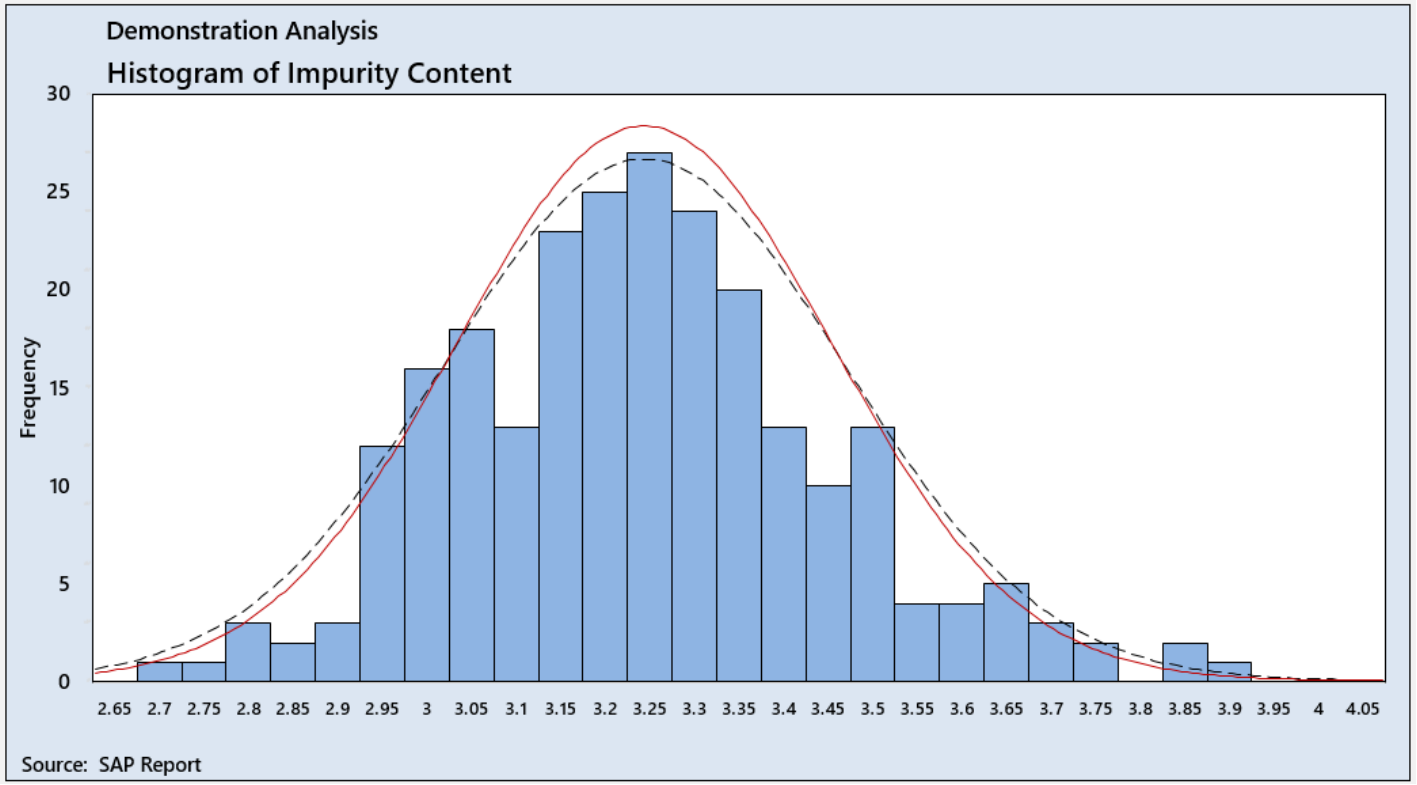
Otherwise, make a histogram of:
Impurity Content

***Subgroup Size***
(leave empty for no subgroups)

Is this Short Term or Long Term data?

Lower Specification Limit
Upper Specification Limit

***Data Transformation (advanced technique)***
No transformation ◉
Square root ○
Natural logarithm ○



# Histograms

Histogram & Capability

## Descriptive Statistics

Scroll to the right and you will see a box with some descriptive statistics for this data

- Number of data points
- Mean
- Minimum, Maximum
- Median
- Standard deviation
  - The Overall calculation is based on the true standard deviation of the data
  - The Within version is based on an estimate of the short-term variation using the average difference between consecutive data points. This uses the same estimating method as the I-MR chart

| Descriptive Statistics | | Data points: | 245 |
|---|---|---|---|
| Minimum = | 2.72 | Maximum = | 3.89 |
| Average = | 3.24 | Median = | 3.23 |
| Standard Deviation (Overall, red line) = | | | 0.211 |
| Standard Deviation (Within, dotted black line) = | | | 0.224 |
| **Normality Test** | | | |
| AD= | 0.69 | P-Value | 0.070 |

# Histograms

Histogram & Capability

## Normality Test

Scroll further to the right and you will see a Normal Probability Plot

- If the blue dots form a roughly straight line, that indicates a Normal distribution
- This data looks pretty Normal
- The bottom section of the Descriptive Statistics box provides the Anderson-Darling test for Normality
- If the P-value is above 0.05 then it is reasonable to say that the data is Normally distributed
  - As the p-value here is 0.07, you can assume the data is Normally distributed.

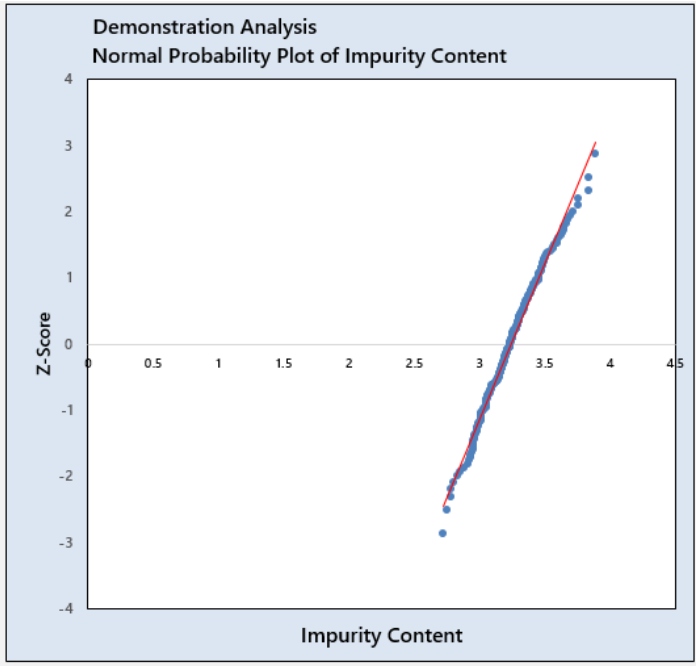
| Descriptive Statistics | | Data points: | 245 |
|---|---|---|---|
| Minimum = | 2.72 | Maximum = | 3.89 |
| Average = | 3.24 | Median = | 3.23 |
| Standard Deviation (Overall, red line) = | | | 0.211 |
| Standard Deviation (Within, dotted black line) = | | | 0.224 |
| **Normality Test** | | | |
| AD= | 0.69 | P-Value | 0.070 |

Capability Analysis

NOTE: Check Control Charts for Special causes before using these results.

| | Lower Limit | Upper Limit |
|---|---|---|
| Overall (long term) Capability | | |
| Pp | | Ppk |
| PpL | | PpU |
| Potential (short term, aka within) Capability | | |
| Cp | | Cpk |
| CpL | | CpU |

| Process Performance | Observed | Expected Overall | Expected Within |
|---|---|---|---|
| PPM < Lower Specification | | | |
| PPM > Upper Specification | | | |
| PPM Total | | | |



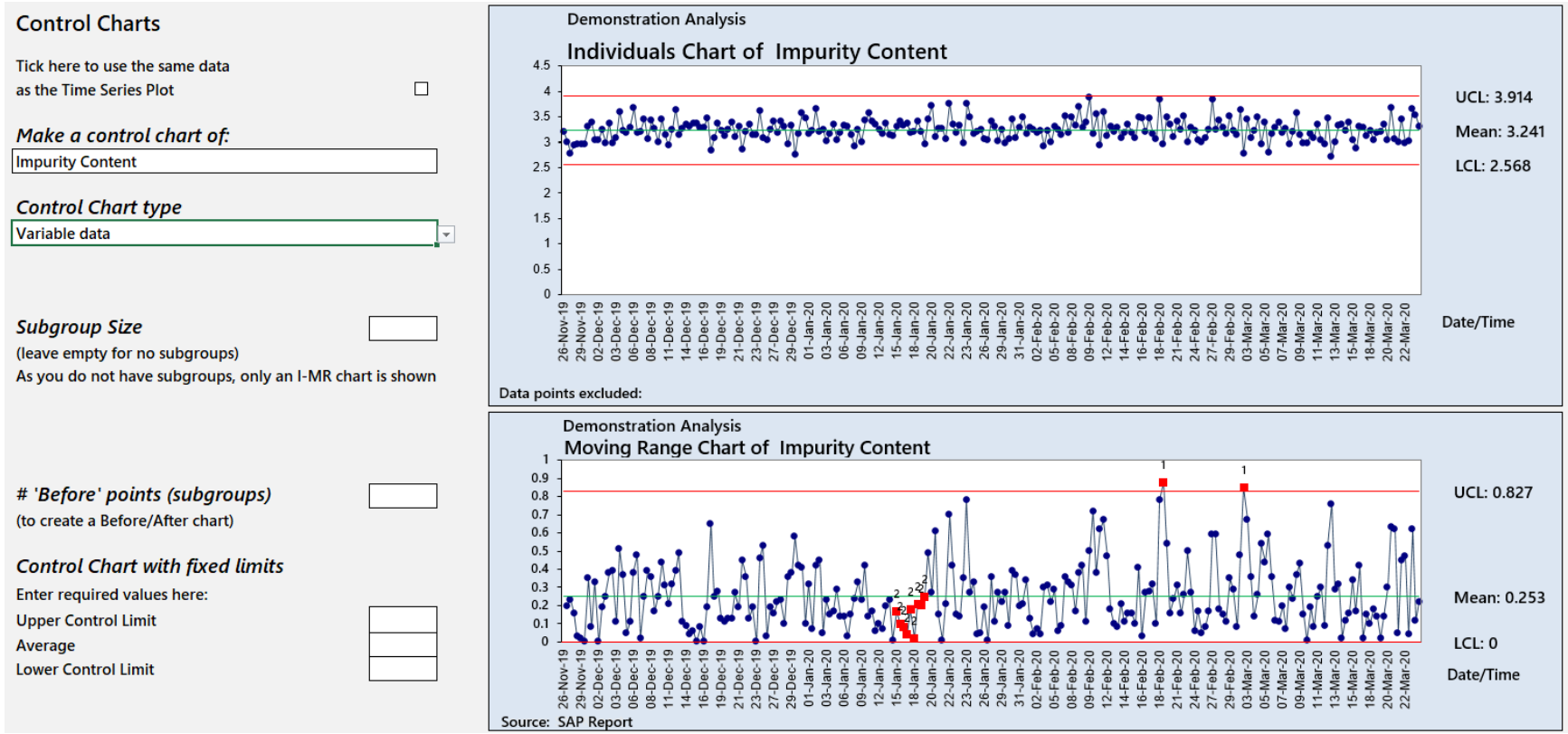
# Process Capability

Histogram & Capability

## Checks before calculating Process Capability

Return to the Control Charts section via the hyperlink and change the selection to ‘Impurity Content’ (also change the Control Chart type to Continuous data) to check for special causes

- The individuals chart looks fine, but the Moving Range chart has a number of special causes
- These should be investigated, but for the purposes of demonstrating the toolkit we shall ignore them in this case.



# Process Capability

Histogram & Capability

**We shall calculate Process Capability using an Upper Specification Limit of 4.**

- Enter the Upper Specification Limit as shown
- Calculations for Cpk and Ppk will now appear, together with Process Performance (Parts per Million data)

**Histogram & Capability**

Tick here to use the same data and subgroup size as the Control Chart ☐

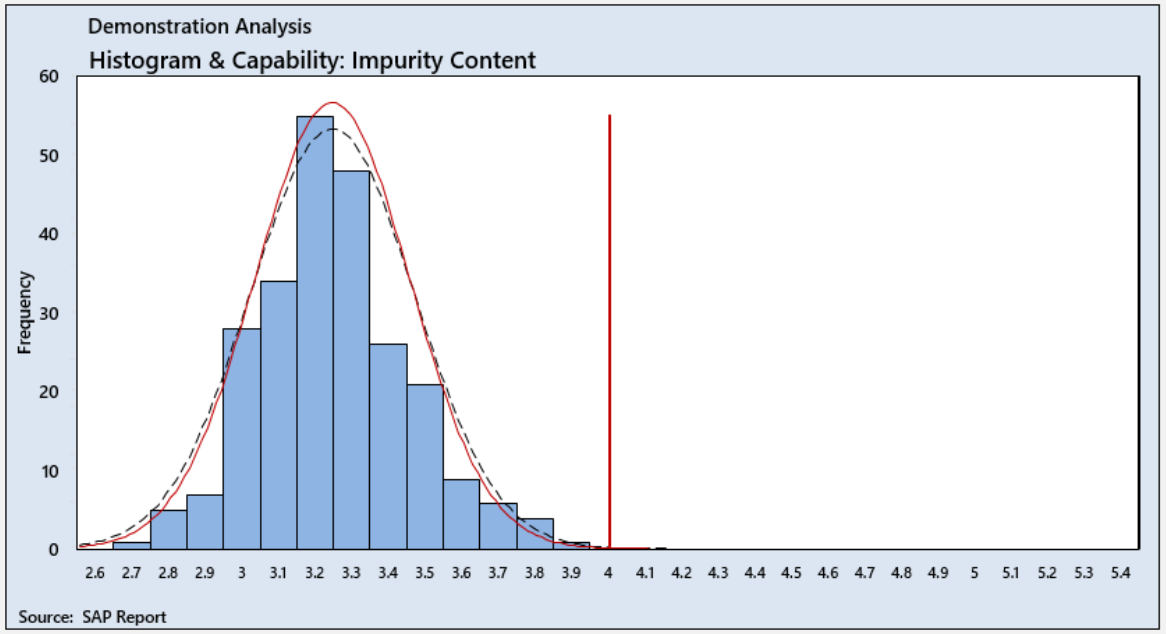
Otherwise, make a histogram of:
Impurity Content

***Subgroup Size***
(leave empty for no subgroups)

Is this Short Term or Long Term data?

Lower Specification Limit
Upper Specification Limit 4

***Data Transformation (advanced technique)***
No transformation ◉
Square root ○
Natural logarithm ○




Data Entry

**Descriptive Statistics** Data points: 245
Minimum 2.72 Maximum 3.89
Average 3.2408 Median 3.23
Standard Deviation (Overall, red line) 0.2111
Standard Deviation (Within, dotted black line) 0.2244
**Normality Test**
AD 0.69 P-Value 0.070

**Capability Analysis**

NOTE: Check Control Charts for Special causes before using these results.

Lower Limit Upper Limit 4
Overall (long term) Capability
Pp Ppk 1.20
PpL N/A PpU 1.20
Potential (short term, aka within) Capability
Cp Cpk 1.13
CpL N/A CpU 1.13

| Process Performance | Observed | Expected Overall | Expected Within |
|---|---|---|---|
| PPM < Lower Specification | N/A | N/A | N/A |
| PPM > Upper Specification | 0 | 161 | 358 |
| PPM Total | 0 | 161 | 358 |

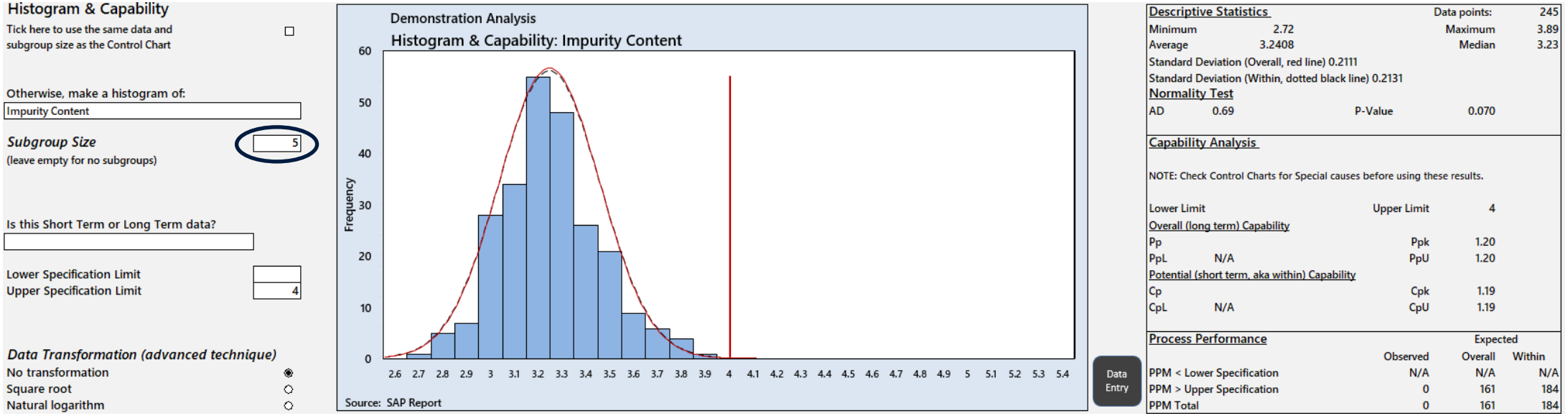
- Typically, acceptable levels of Process Capability in manufacturing processes are Cpk = 1.67 and Ppk = 1.33 – so this process is a ‘fail’.
  - The defect rate may only be a few hundred per million but even a small deterioration would result in a large defect rate

# Process Capability

Histogram & Capability

## Subgrouped data

- If the data has been collected in subgroups (eg 5 consecutive measurements every hour), you can calculate within-subgroup and overall capability.
- Enter a 5 in the Subgroup Size section as shown



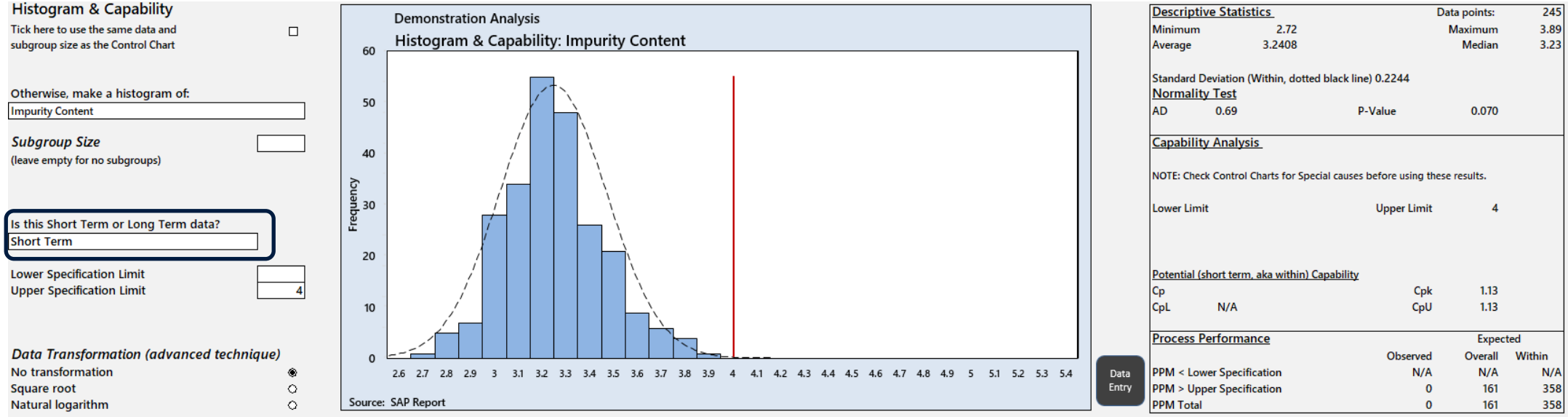

- Cpk now refers to variation within the subgroups, and Ppk refers to overall variation
- This is not really subgrouped data – we are just illustrating the functionality – so please delete the 5 before continuing

# Process Capability

Histogram & Capability

## Short Term/Long Term Data

- If the data has been collected over a short period of time, Cpk should be reported rather than Ppk
  - Capability is expected to be higher over a short time period than a long one, so the acceptance standards for Cpk and Ppk are normally different
  - Use the drop-down box to specify that this is Short Term data



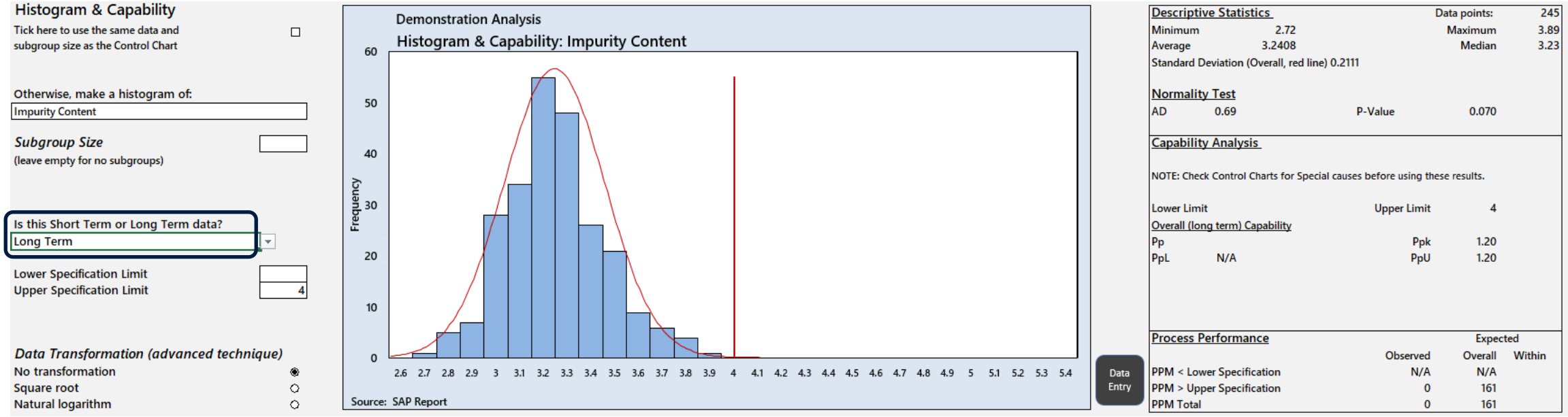

  - You can see that only Cpk is now reported, Ppk is hidden

# Process Capability

Histogram & Capability

## Short Term/Long Term Data

- If the data has been collected over a long period of time, Ppk should be reported rather than Cpk
  - Capability is expected to be lower over a long time period than a short one, so the acceptance standards for Cpk and Ppk are normally different)
  - Change the entry in the drop-down box from Short Term to Long Term




  - You can see that only Ppk is now reported, Cpk is hidden

# Process Capability

Histogram & Capability

## Data Transformation

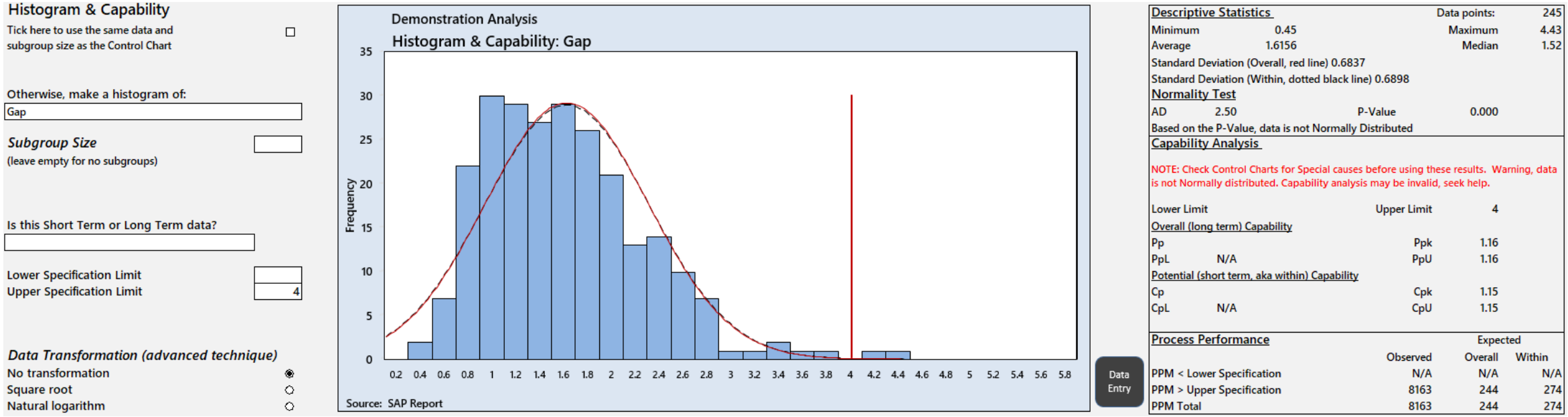
- Please note, this is an advanced topic.  If you are not fully familiar with it, please skip the remainder of this section. Transforming data inappropriately can lead to incorrect decisions.
- Data transformations enable us to take data which is naturally skewed and *try* to remove the skew, so that we can calculate Process Capability based on the shape of the Normal Distribution
  - This technique can accidentally hide Special Causes, which is expressly not what we are trying to do – hence, it needs to be used with caution
- The toolkit provides the two most commonly-used transformations – Square Root and Natural logarithm.
- To illustrate this, we shall use the data in the Practice Data worksheet ‘Skewed Data’ – please copy the two columns from this sheet (Gap and Time to complete maintenance task) into two spare columns of the toolkit (there’s no need to remove the existing data for now)

# Process Capability

Histogram & Capability

## Square Root Transformation

- We shall start by looking at Gap. This is the gap between two assembled parts. It is physically impossible to have a negative gap – they won't fit – and the Upper Specification Limit is 4mm.
- Enter this information, as shown below



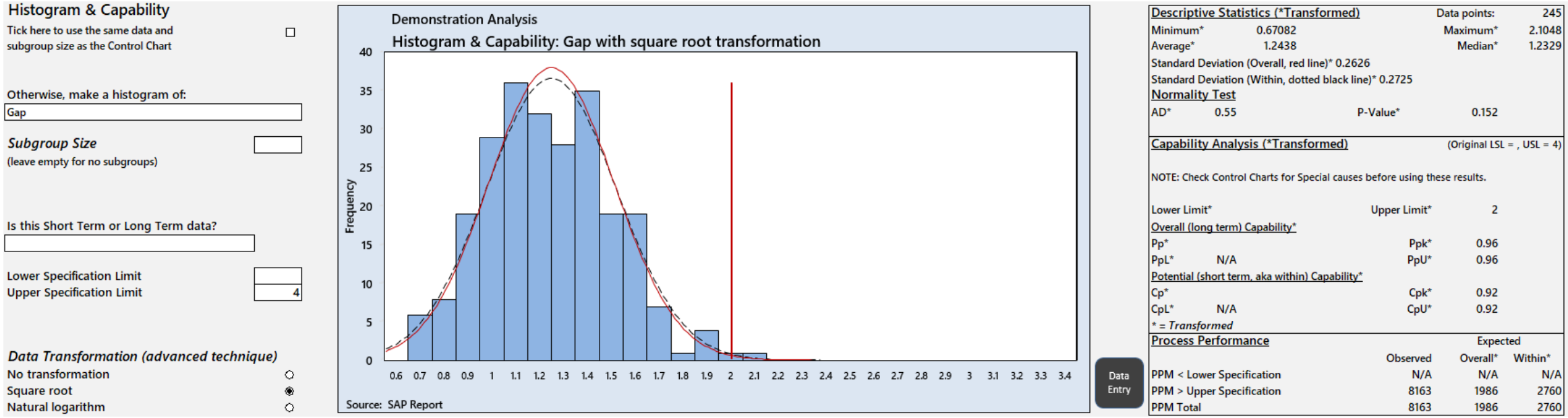

- Note that the capability calculation does not look right – our data contains 8163 PPM out of specification, but the expected overall PPM is only 244. The Normal distribution does not fit the data. The Anderson-Darling test gives a p-value of 0.000.
- Also note that the Toolkit is giving you a helpful warning about the data Normality (see the red writing in the illustration)

# Process Capability

Histogram & Capability

## Square Root Transformation

- Select the radio button for Square Root transformation, as shown.




- Note that the p-value is now 0.152, so the square root transformation was effective in this case.
- The upper spec limit has been transformed to 2 (ie the square root of 4) and the capability calculations are now realistic – with 1986 PPM expected overall and a Ppk of 0.96.
- Note: The Toolkit will offer you a transformation, even if your data contains negative values. In such cases, the toolkit will automatically increase all values (and the spec limits) so that the lowest value is zero, before taking square roots. You will see a message to let you know what was done.

# Process Capability

Histogram & Capability

## Natural Logarithm Transformation

- Next, we'll look at 'Time to complete maintenance task'. This is a "quick and easy" task that is performed during break times in a manufacturing facility; if the job takes more than 10 minutes, it will prevent the line from restarting at the end of the break.
- Please select the appropriate column, change the Upper Specification Limit and revert to no data transformation as shown.

**Histogram & Capability**

Tick here to use the same data and subgroup size as the Control Chart ☐

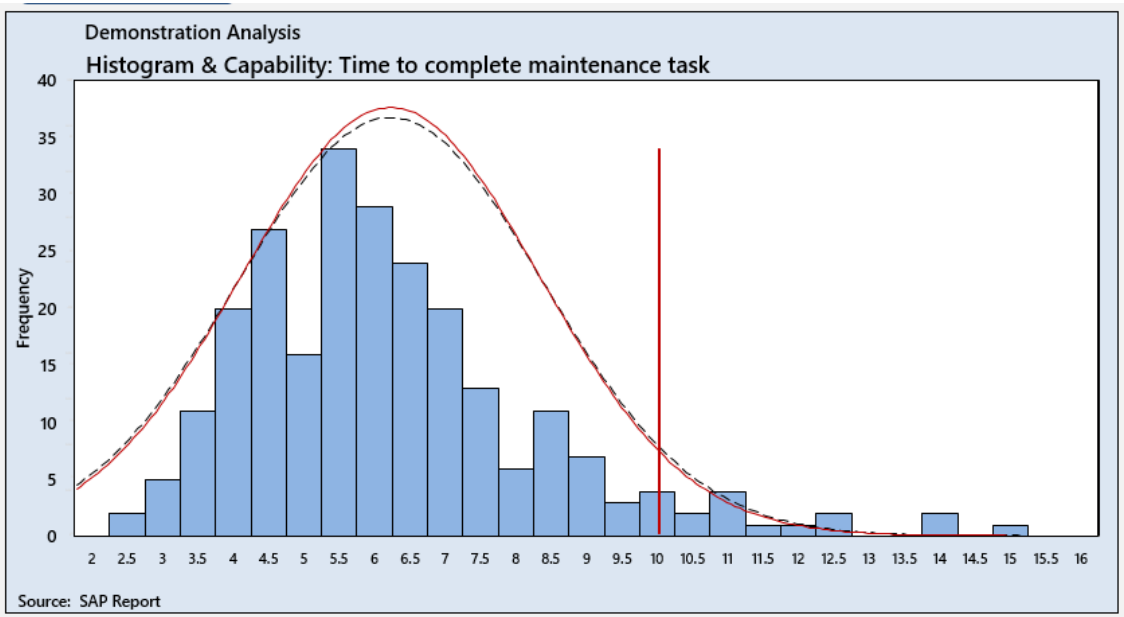
Otherwise, make a histogram of:
Time to complete maintenance task

*Subgroup Size*
(leave empty for no subgroups)

Is this Short Term or Long Term data?

Lower Specification Limit
Upper Specification Limit 10

*Data Transformation (advanced technique)*
No transformation ◉
Square root ○
Natural logarithm ○




Data Entry

**Descriptive Statistics**

| | | | |
|---|---|---|---|
| Data points: | 245 | | |
| Minimum | 2.5 | Maximum | 14.8 |
| Average | 6.2151 | Median | 5.9 |

Standard Deviation (Overall, red line) 2.12
Standard Deviation (Within, dotted black line) 2.171

**Normality Test**

| | | | |
|---|---|---|---|
| AD | 4.35 | P-Value | 0.000 |

Based on the P-Value, data is not Normally Distributed

**Capability Analysis**

NOTE: Check Control Charts for Special causes before using these results. Warning, data is not Normally distributed. Capability analysis may be invalid, seek help.

| | | | |
|---|---|---|---|
| Lower Limit | | Upper Limit | 10 |
| Overall (long term) Capability | | | |
| Pp | | Ppk | 0.60 |
| PpL | N/A | PpU | 0.60 |
| Potential (short term, aka within) Capability | | | |
| Cp | | Cpk | 0.58 |
| CpL | N/A | CpU | 0.58 |

**Process Performance**

| | Observed | Expected Overall | Expected Within |
|---|---|---|---|
| PPM < Lower Specification | N/A | N/A | N/A |
| PPM > Upper Specification | 65306 | 37117 | 40626 |
| PPM Total | 65306 | 37117 | 40626 |

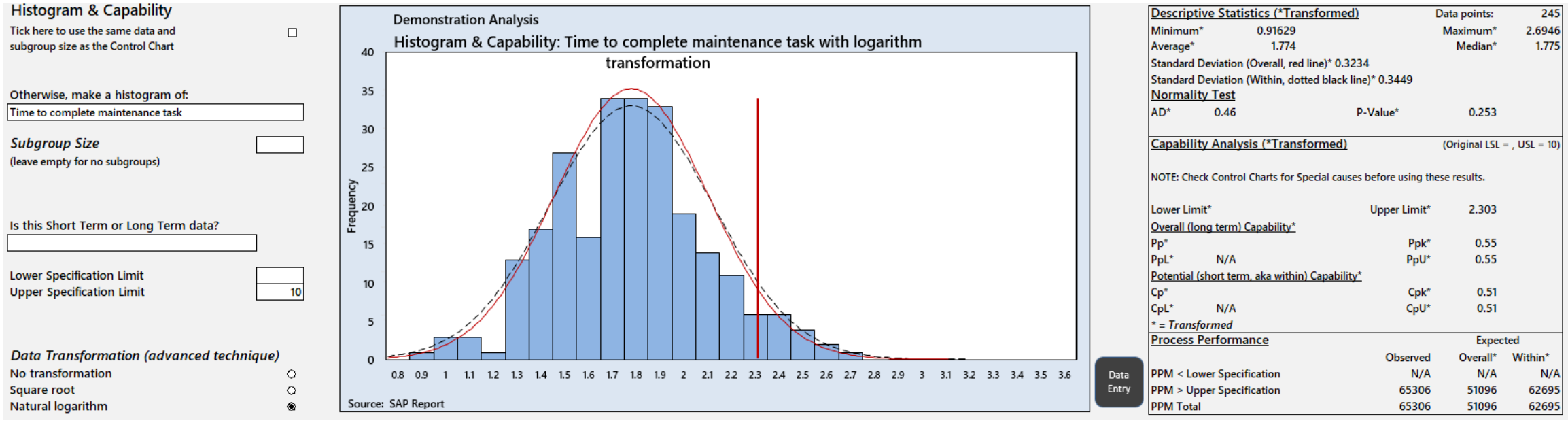
- As with the Gap example, the capability calculation does not look right and there is a red warning message – the Normal distribution does not fit this highly-skewed data at all.

# Process Capability

Histogram & Capability

## Natural Logarithm Transformation

- In this case, a Square Root transformation is not powerful enough to make the data Normal. So select the radio button for Natural Logarithm transformation, as shown




- Note that the p-value is now 0.253, so the natural logarithm transformation was successful.
- The upper spec limit has been transformed to 2.303 and the capability calculations are now realistic
- Note: The Toolkit will offer you a transformation, even if your data contains a 0 or negative values. In such cases, the toolkit will automatically increase all values (and the spec limits) so that the lowest value is 1, before taking square roots. You will see a message to let you know what was done.

# Process Capability

Histogram & Capability

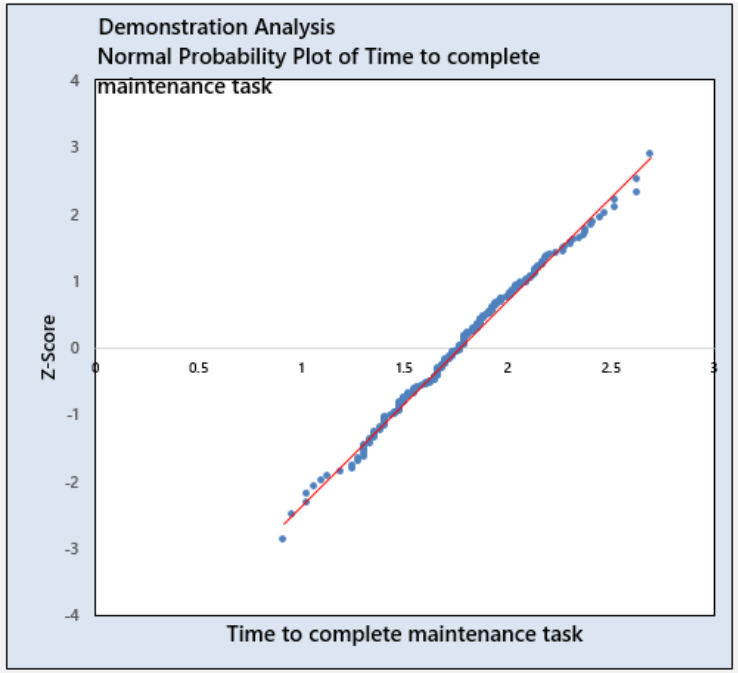
## Reviewing your data

- Note that all the transformed data is highlighted with an asterisk in the main reporting area.
- If you scroll to the right, you can see (on the other side of the Normal Probability Plot) the original descriptive statistics and specification limits, before the data transformation.

| Descriptive Statistics (*Transformed) | | | Data points: | 245 |
|---|---|---|---|---|
| Minimum* | 0.91629 | | Maximum* | 2.6946 |
| Average* | 1.774 | | Median* | 1.775 |
| Standard Deviation (Overall, red line)* 0.3234 | | | | |
| Standard Deviation (Within, dotted black line)* 0.3449 | | | | |
| Normality Test | | | | |
| AD* | 0.46 | P-Value* | 0.253 | |

| Capability Analysis (*Transformed) | | (Original LSL = , USL = 10) |
|---|---|---|
| NOTE: Check Control Charts for Special causes before using these results. | | |
| Lower Limit* | Upper Limit* | 2.303 |
| Overall (long term) Capability* | | |
| Pp* | Ppk* | 0.55 |
| PpL* N/A | PpU* | 0.55 |
| Potential (short term, aka within) Capability* | | |
| Cp* | Cpk* | 0.51 |
| CpL* N/A | CpU* | 0.51 |
| * = *Transformed* | | |

| Process Performance | Observed | Expected Overall* | Expected Within* |
|---|---|---|---|
| PPM < Lower Specification | N/A | N/A | N/A |
| PPM > Upper Specification | 65306 | 51096 | 62695 |
| PPM Total | 65306 | 51096 | 62695 |




Data Entry

Original data before transformation

| Descriptive Statistics | | | |
|---|---|---|---|
| Minimum | 2.5 | Data point | 245 |
| Average | 6.2151 | Maximum | 14.8 |
| | | Median | 5.9 |

Lower Specification Limit (LSL) =
Upper Specification Limit (USL) = 10

Standard Deviation (Overall, red line) 2.1202
Standard Deviation (Within, dotted black line)2.1709

# Stratified Plots and ANOVA

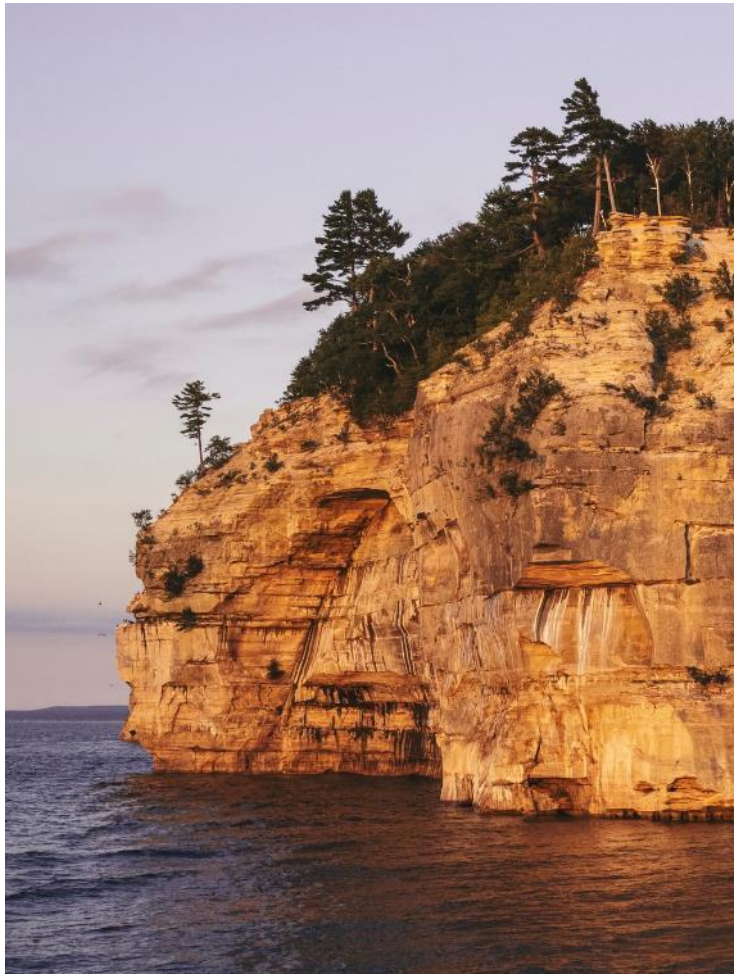
**Used to identify factors that affect the outcome**

# Stratified Plots

When there is too much variation, a good way to diagnose it is to collect data on factors that might affect the outcome.

- In our Impurities example, we have:
  - Data on the outcome – the impurity content
  - Data on two factors that might have an influence on the outcome: The batch (of raw materials), the shift and the machine number
- We can use these factors to ‘stratify’ the data (like the strata in rocks)

| 19 | | Exclusions | Time axis | Variable 1 | Variable 2 | Variable 3 | Variable 4 |
|---|---|---|---|---|---|---|---|
| 20 | Row # | Exclude from Control Charts (only) | Date/Time | Impurity Content | Batch | Shift | Machine |
| 21 | 1 | | 26/11/2019 | 3.21 | a | Day | 110 |
| 22 | 2 | | 26/11/2019 | 3.01 | a | Night | 180 |
| 23 | 3 | | 26/11/2019 | 2.78 | a | Day | 220 |
| 24 | 4 | | 27/11/2019 | 2.94 | a | Night | 110 |
| 25 | 5 | | 28/11/2019 | 2.97 | a | Day | 180 |
| 26 | 6 | | 29/11/2019 | 2.95 | a | Night | 220 |
| 27 | 7 | | 29/11/2019 | 2.95 | a | Day | 110 |
| 28 | 8 | | 30/11/2019 | 3.3 | a | Night | 180 |
| 29 | 9 | | 30/11/2019 | 3.38 | a | Day | 220 |
| 30 | 10 | | 01/12/2019 | 3.05 | a | Night | 110 |
| 31 | 11 | | 02/12/2019 | 3.05 | a | Day | 180 |

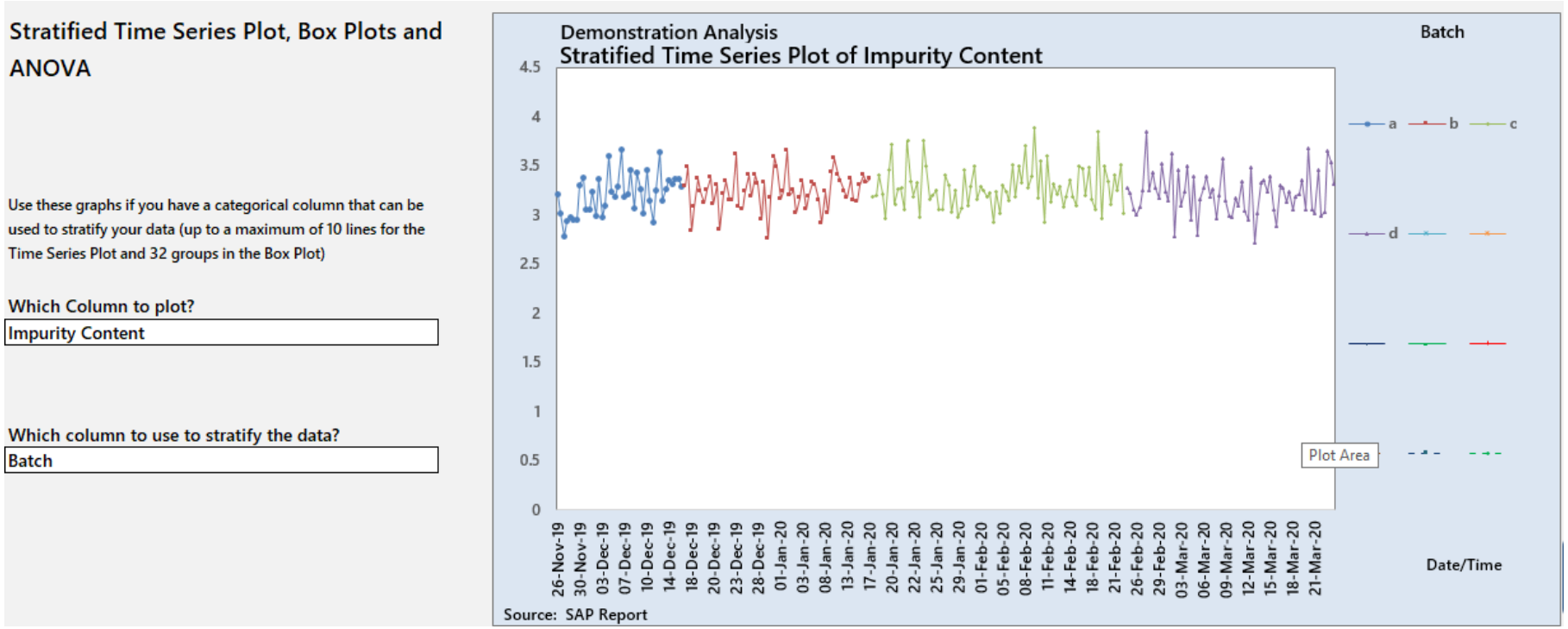


# Stratified Plots

## Stratified Time Series Plots

These graphs enable us to see both variation over time and the differences between the groups.

- Click the Hyperlink for Stratified Plots (we'll cover ANOVA shortly)
- For the column to plot, select 'Impurity Content' from the dropdown list
- For the column used to stratify the data, select 'batch'




There isn't much to see here, it seems

# Stratified Plots

## Stratified Time Series Plots

- Change the column used to stratify the data to ‘shift’

Stratified Time Series Plot, Box Plots and ANOVA

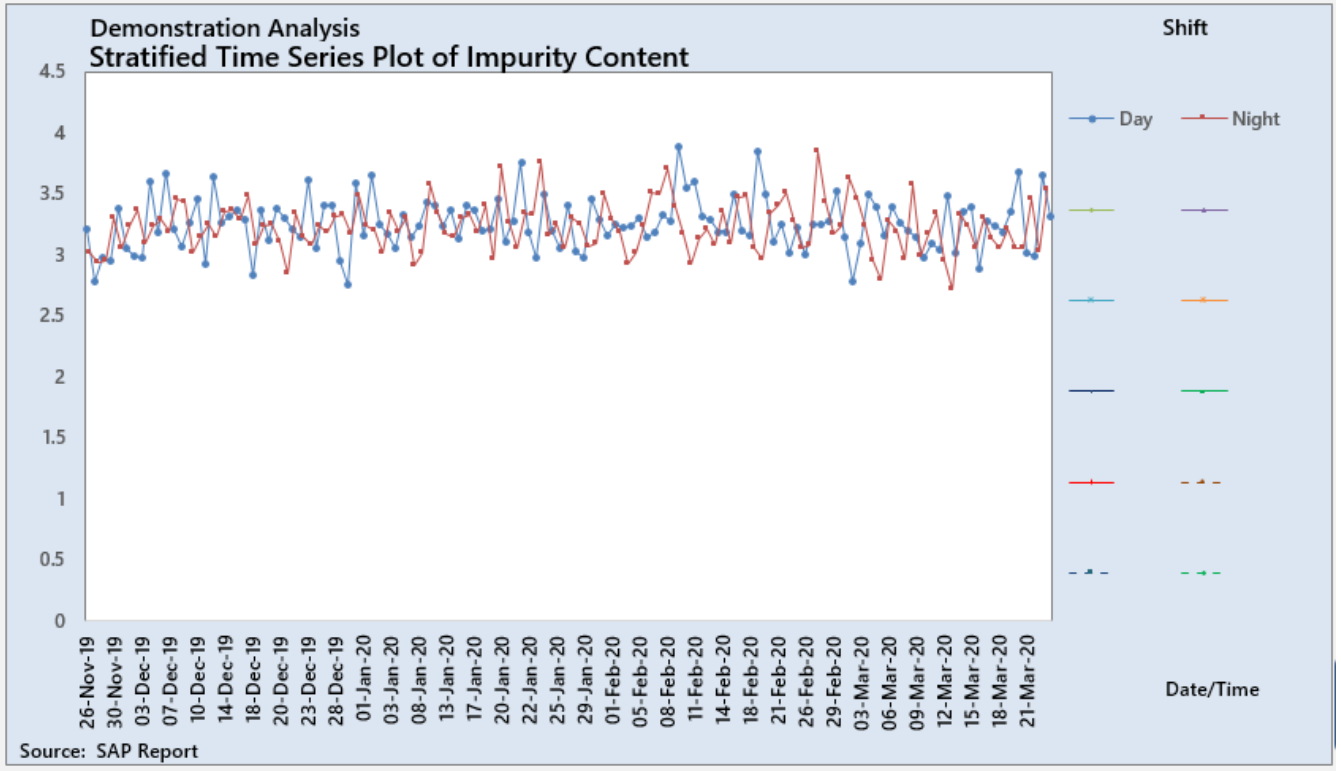
Use these graphs if you have a categorical column that can be used to stratify your data (up to a maximum of 10 lines for the Time Series Plot and 32 groups in the Box Plot)

Which Column to plot?

Impurity Content

Which column to use to stratify the data?

Shift




There doesn’t seem to be a difference here, either

# Stratified Plots

## Stratified Time Series Plots

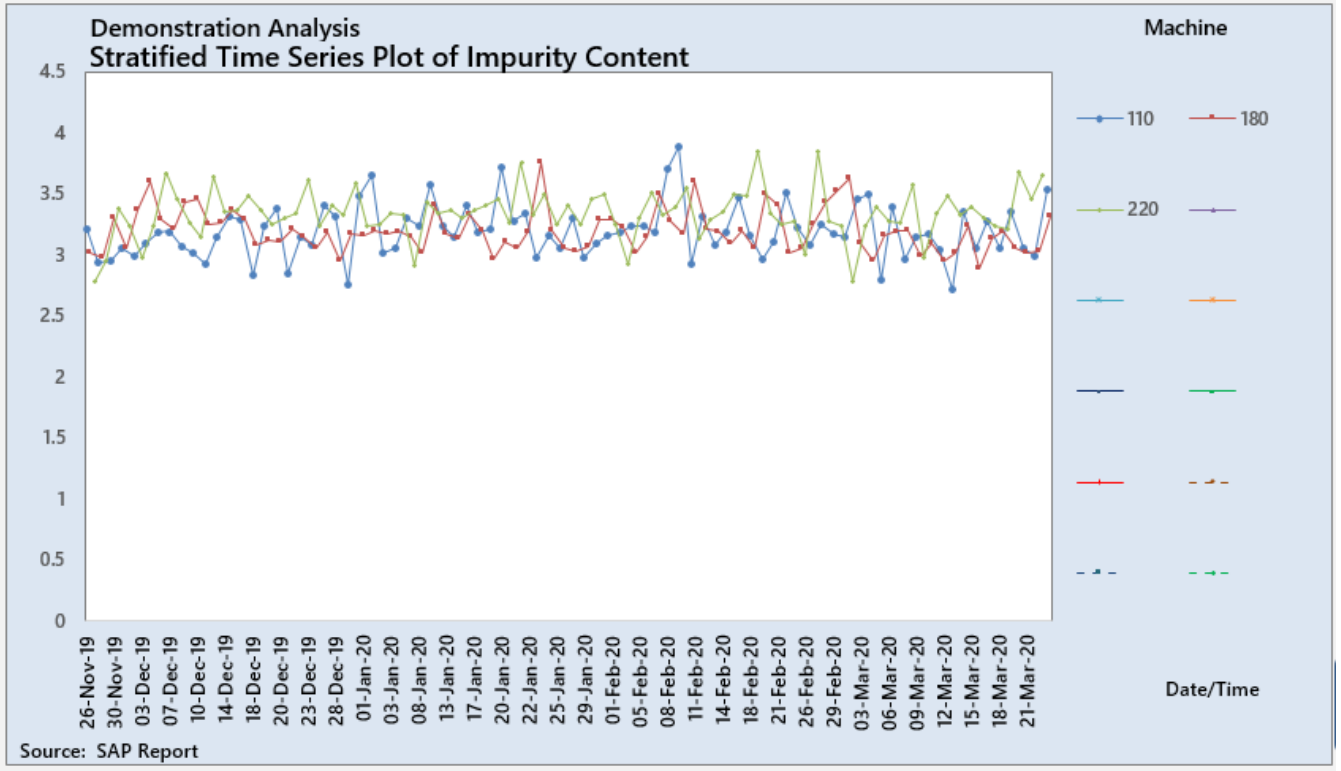
- Change the column used to stratify the data to ‘machine’

Stratified Time Series Plot, Box Plots and ANOVA

Use these graphs if you have a categorical column that can be used to stratify your data (up to a maximum of 10 lines for the Time Series Plot and 32 groups in the Box Plot)

Which Column to plot?
Impurity Content

Which column to use to stratify the data?
Machine




Once again, nothing jumps out.

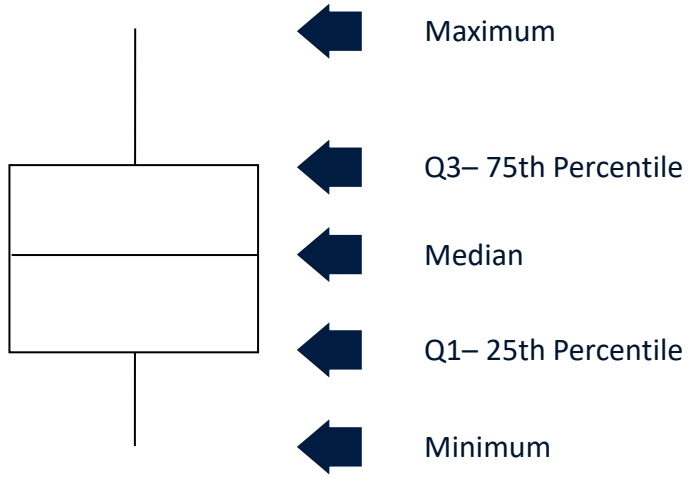
Stratified Time Series Plots help us to see variation over time, but we can get different insights if we take away the time dimension – and a great tool for that is Box Plots.

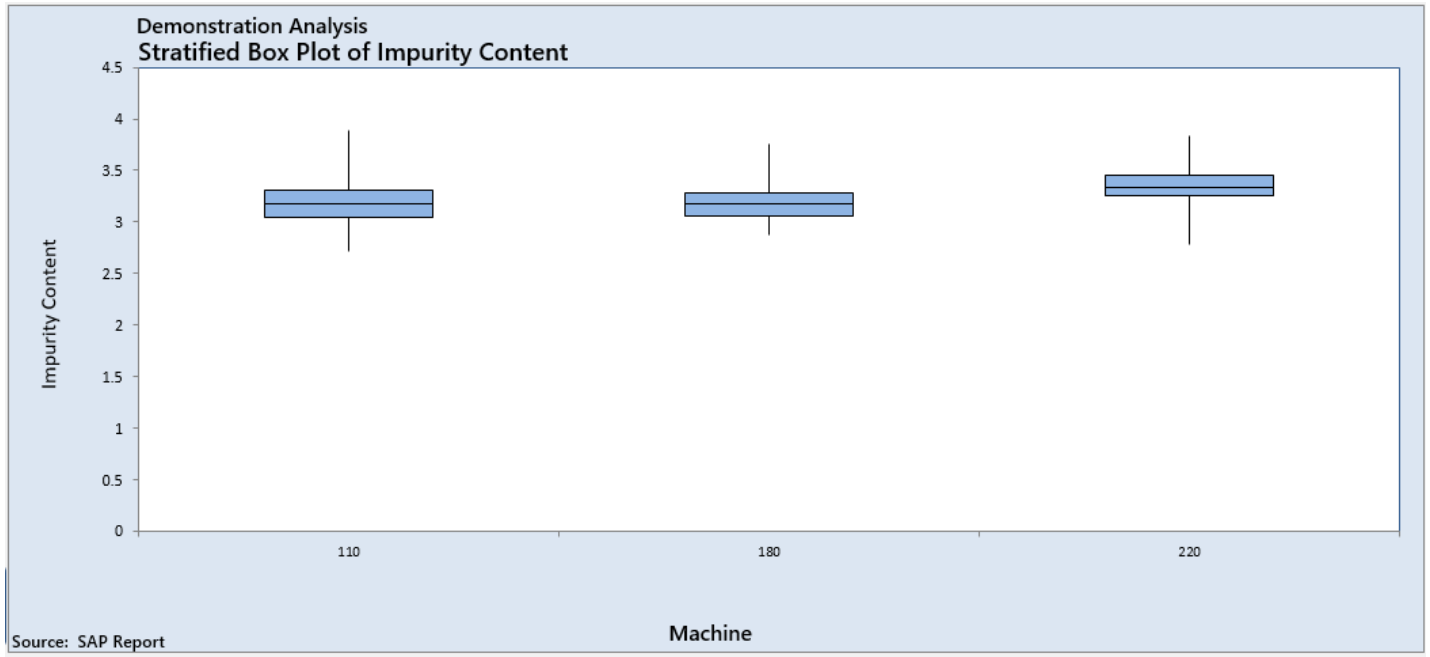
# Stratified Plots



## Box Plots

- Box Plots are a good way of visualising different groups within the data set side-by-side
  - They don't show variation over time – but that means less distractions than a Time Series Plot
- From the Stratified Time Series Plots we looked at just now, scroll to the right to see the Box Plots for Impurity Content, stratified by Machine




The difference is subtle but we can see that machine seems to make a difference

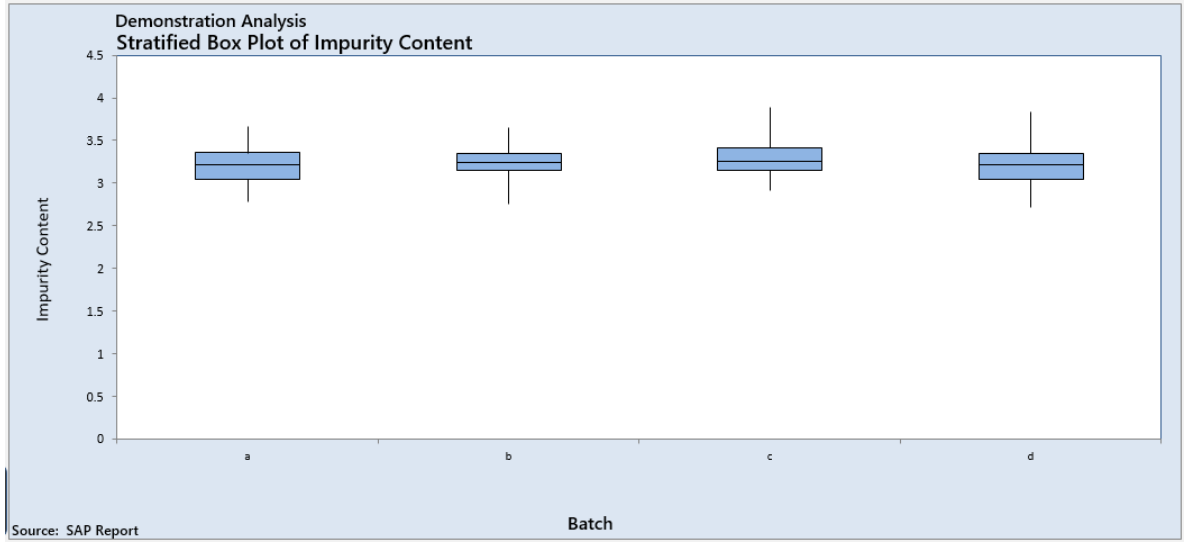
- Machine 220 might have a higher average than the other two

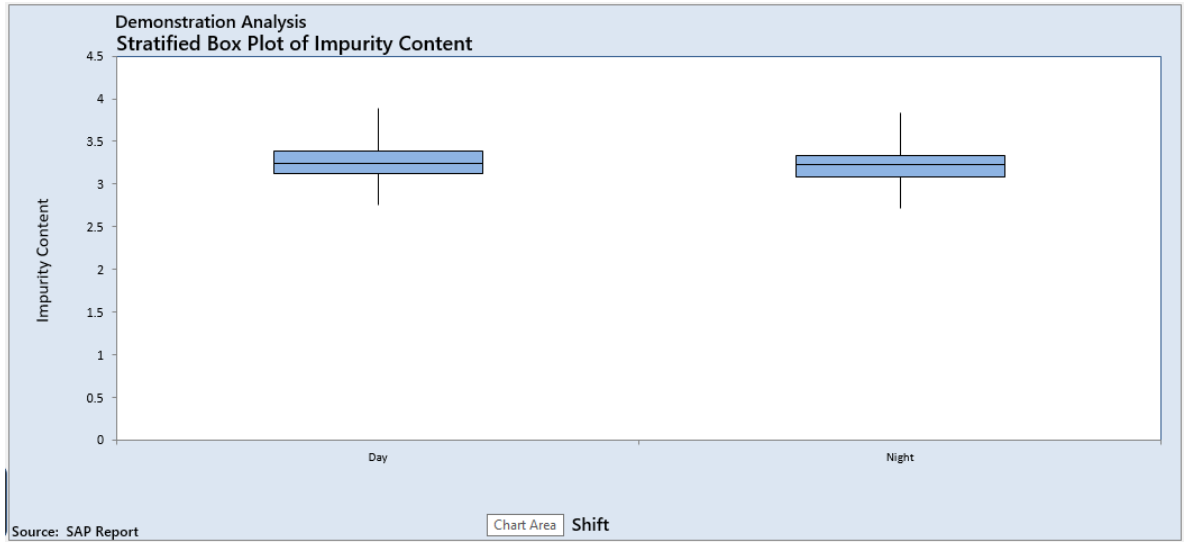
# Stratified Plots



## Box Plots

Here are the Box Plots for batch and shift – there doesn't seem to be a difference here.



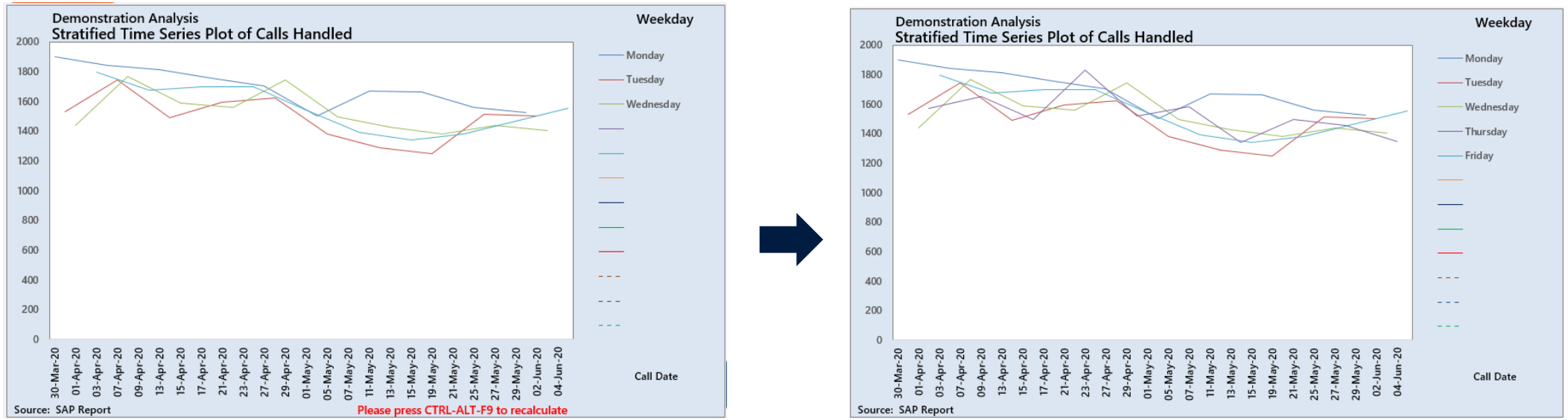
# Stratified Plots



## An occasional ‘wrinkle’

Excel has to perform a vast amount of calculations to create the plots that you see appearing almost instantaneously.

- Sometimes, these calculations are not fully completed, even with automatic recalculation selected
- If this happens, the lines and legend will not appear fully and you will see a red message: Please press CTRL-ALT-F9 to recalculate
- Simply hold down Control and Alt and press F9 and the graph will recalculate fully.



# Analysis of Variance (ANOVA)

**Statistical analysis of stratified data**

# Analysis of Variance (ANOVA)

Stratified Plots & ANOVA

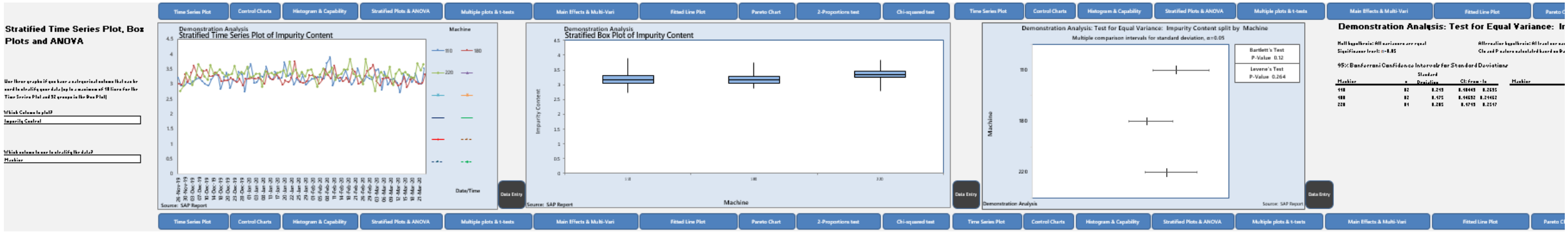
**ANOVA is used to statistically prove differences between two or more groups**

- You can see if you can prove a difference in variances, or in means
- The Null Hypothesis is that the variances (or means) are all the same and the Alternate is that they are not the same

As ANOVA involves an X with two or more settings, we have put it next to the Stratified Plots

- NOTE – if you are looking for 2-sample t-tests, these are located next to the Multiple plots (see this slide)
- NOTE – this is a relatively advanced tool, and most training courses at Green Belt level omit it.
  - You may choose to skip this section if you have not come across ANOVA before
  - The instructions in this User Guide are intended for people who are already familiar with the underlying theory

The ANOVA calculations can be found directly to the right of the stratified plots




Stratified Time Series Plots and Box Plots

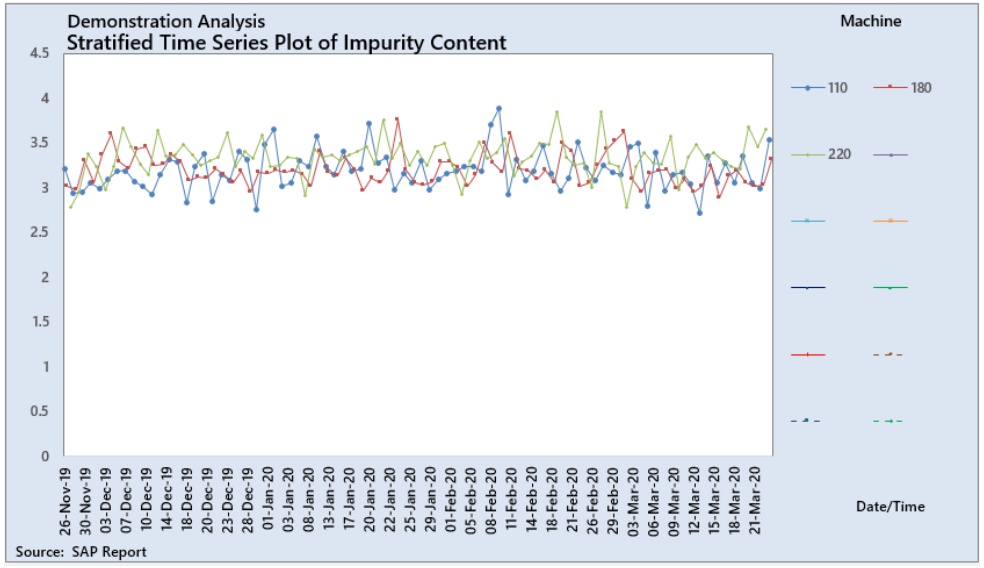
Analysis of Variance (continues to the right)

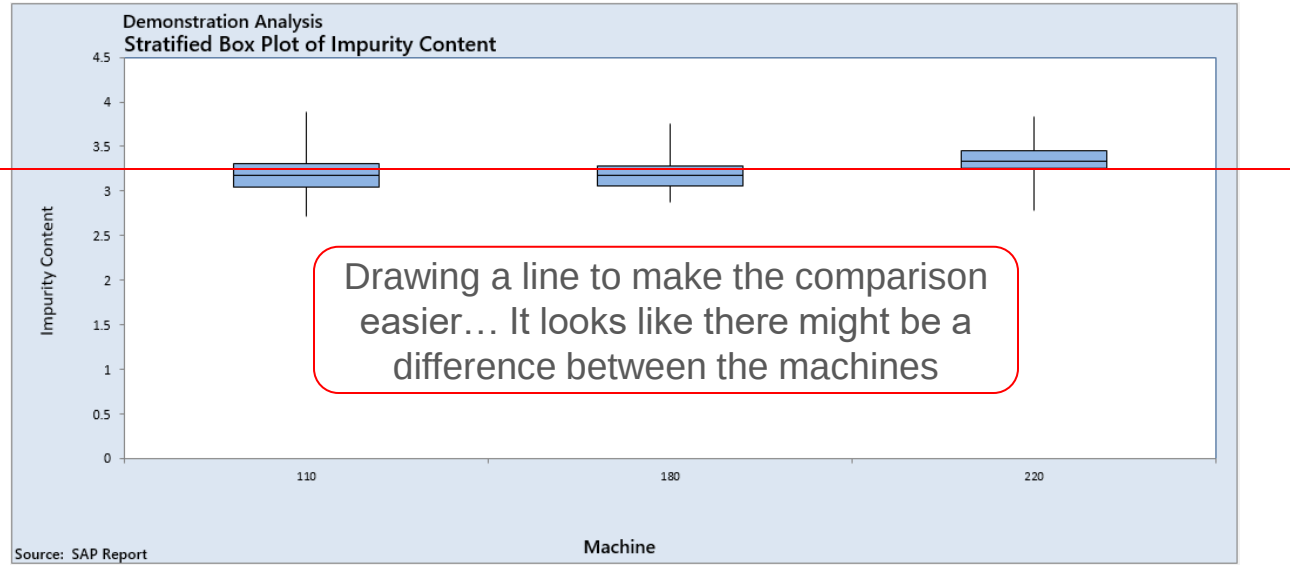
# Analysis of Variance (ANOVA)

Stratified Plots & ANOVA

**Is there a difference between the three machines?**

- In breaking down the data on impurity content based on the raw material batch, we saw no evidence of a difference
- Similarly, we could not see signs of a difference between day shift and night shift
- In comparing the three machines though, there was a suggestion that machine 220 had higher impurities than the other two
  - Was this just chance variation, or can we prove that there really is a difference between the machines?
- Please select ‘machine’ as the column to use to stratify the data, and we will investigate this one further



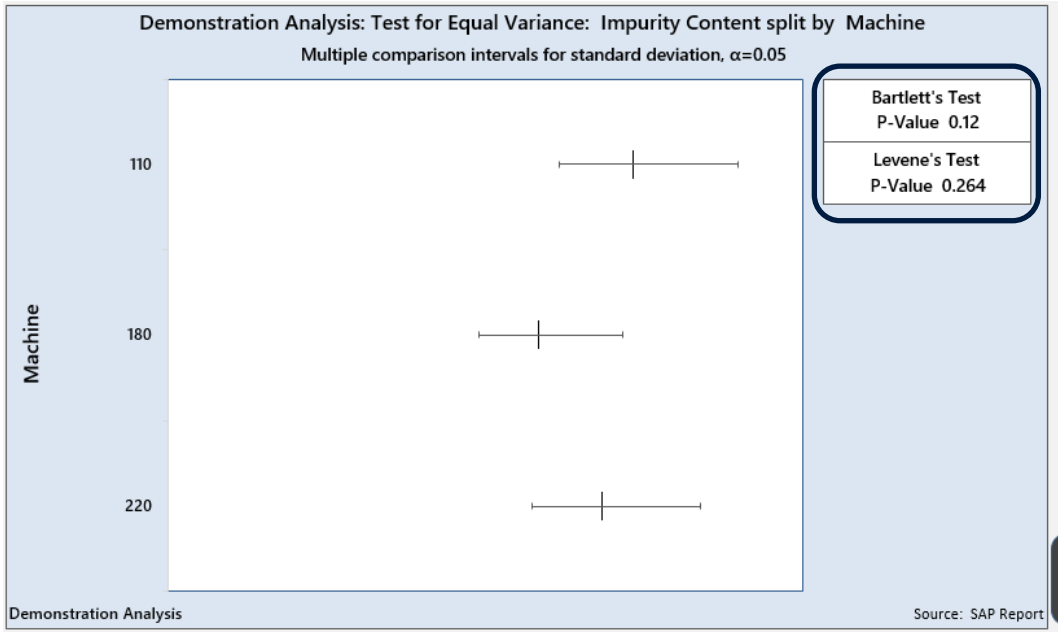



# Analysis of Variance (ANOVA)

Stratified Plots & ANOVA

## Test for Equal Variances

- Scroll to the right from the Box Plots and you will see the first part of the analysis: Test for Equal Variances
- This test has the Null Hypothesis that the variances in the groups are equal, and will determine if we can disprove the Null.
- Two p-values are given
  - Bartlett's test applies only to Normally distributed data
  - Levene's test applies to any continuous distribution
- We confirmed the data is Normally distributed when we looked at the Histogram so we'll use Bartlett's Test
- The P-Value is 0.12 – above 0.05 - so we have not proved any difference in the variances of these three groups.




Data Entry

**Demonstration Analysis: Test for Equal Variance: Impurity Content split by Machine**

Null hypothesis: All variances are equal
Significance level: α=0.05

Alternative hypothesis: At least one variance is different
CIs and P values calculated based on Bartlett's method for Normal data

**95% Bonferroni Confidence Intervals for Standard Deviations**

| Machine | n | Standard Deviation | CI: from | to |
|---|---|---|---|---|
| 110 | 82 | 0.219 | 0.1845 | 0.2695 |
| 180 | 82 | 0.175 | 0.1469 | 0.2146 |
| 220 | 81 | 0.205 | 0.1719 | 0.2517 |

| Machine | n | Standard Deviation | CI: from | to |
|---|---|---|---|---|

# Analysis of Variance (ANOVA)



## Test for Equal Variances (continued)

The data in the next section shows, for each machine:

- The number of samples from that machine (n)
- The standard deviation of the impurity content for that group
- The 95% confidence interval for the standard deviations of these machines
  - These are calculated using Bartlett's method for Normal data
  - These are Bonferroni Confidence Intervals, which means that the 5% risk of a false positive is shared between the three groups. This reduces the risk of a false positive if you have lots of groups

**Demonstration Analysis: Test for Equal Variance: Impurity Content split by Machine**

Null hypothesis: All variances are equal
Significance level: α=0.05

Alternative hypothesis: At least one variance is different
CIs and P values calculated based on Bartlett's method for Normal data

**95% Bonferroni Confidence Intervals for Standard Deviations**

| Machine | n | Standard Deviation | CI: from | to |
|---|---|---|---|---|
| 110 | 82 | 0.219 | 0.1845 | 0.2695 |
| 180 | 82 | 0.175 | 0.1469 | 0.2146 |
| 220 | 81 | 0.205 | 0.1719 | 0.2517 |

| Machine | n | Standard Deviation | CI: from - to |
|---|---|---|---|

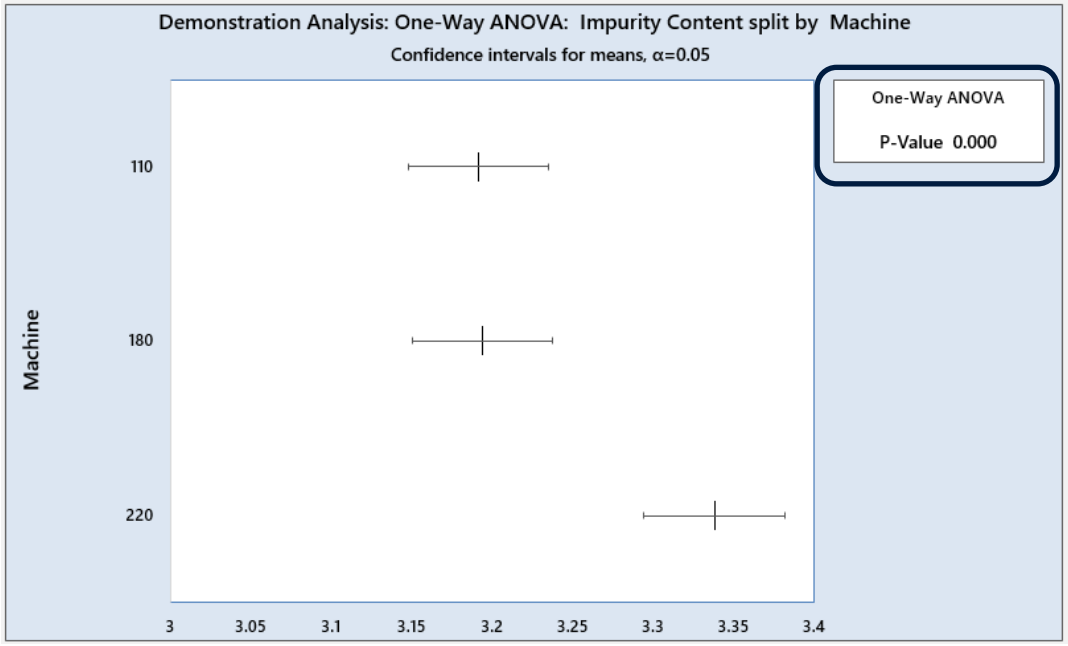
# Analysis of Variance (ANOVA)



## One Way ANOVA

The analysis in the next section (to the right of Test for Equal Variances) shows the results of a One Way ANOVA test.

- This tests to see if a difference between the group means can be proven.
- The graph on the left shows the 95% confidence intervals for the individual group means
- The panel in the graph shows the P-Value – here, it is 0.000 – clearly below 0.05, so this is proof that there is a difference in the group means.




Data Entry

**Demonstration Analysis: One-Way ANOVA: Impurity Content split by Machine**

Null hypothesis: All averages are equal  Alternative hypothesis: Averages are not equal  Significance level: α=0.05

Equal variances have been assumed for the analysis

**Analysis of Variance**

| Source | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| Machine | 2 | 1.147 | 0.573 | 14.267 | 0.000 |
| Error | 242 | 9.725 | 0.040 | | |
| Total | 244 | 10.871 | | | |

**Model Summary**

| s | R-sq | R-sq (adj) |
|---|---|---|
| 0.200 | 10.5% | 9.8% |

**Averages**

| Machine | N | Average | CI: from | - to |
|---|---|---|---|---|
| 110 | 82 | 3.191 | 3.148 | 3.235 |
| 180 | 82 | 3.194 | 3.151 | 3.237 |
| 220 | 81 | 3.34 | 3.29 | 3.38 |

# Analysis of Variance (ANOVA)



## One Way ANOVA

The data on the right gives further analytical details

**Demonstration Analysis: One-Way ANOVA: Impurity Content split by Machine**

Null hypothesis: All averages are equal  Alternative hypothesis: Averages are not equal  Significance level: α=0.05

Equal variances have been assumed for the analysis

**Analysis of Variance** (1)

| Source | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| Machine | 2 | 1.147 | 0.573 | 14.267 | 0.000 |
| Error | 242 | 9.725 | 0.040 | | |
| Total | 244 | 10.871 | | | |

**Model Summary** (2)

| s | R-sq | R-sq (adj) |
|---|---|---|
| 0.200 | 10.5% | 9.8% |

**Averages** (3)

| Machine | N | Average | CI: from | to |
|---|---|---|---|---|
| 110 | 82 | 3.191 | 3.148 | 3.235 |
| 180 | 82 | 3.194 | 3.151 | 3.237 |
| 220 | 81 | 3.34 | 3.29 | 3.38 |

The Analysis of Variance table (1):

- A breakdown of the variation between that accounted for by the machines, and random variation (Error)
- Degrees of freedom, Sum of Squares, Mean Squares and F ratio are given, together with the P value (as on the last slide)

The Model summary table (2):

- S (standard deviation of residual errors), R-squared (how much of the variation is explained my machine) and R-sq adjusted

The Averages table (3):

- For each machine, the number of samples, average (mean) of each group and the 95% Confidence Interval for the mean

# Multiple Plots and 2-sample t-tests

**Used to identify factors that affect the outcome**

# Multiple Plots



Multiple Plots do much the same thing as Stratified Plots, but with a different arrangement of the data

- Stratified plots use stacked data – all the data is 'stacked' in a single column, and additional columns can be used to break it into groups
- Multiple plots use unstacked data – each group has its own column

Here's how our Impurities data might look if converted from stacked to unstacked

- Data used for Stratified Plots

| Impurity Content | Shift |
|---|---|
| 3.21 | Day |
| 3.01 | Night |
| 2.78 | Day |
| 2.94 | Night |
| 2.97 | Day |
| 2.95 | Night |
| 2.95 | Day |
| 3.3 | Night |
| 3.38 | Day |
| 3.05 | Night |
| 3.05 | Day |
| 3.24 | Night |
| 2.99 | Day |

- Data used for Multiple Plots

| Impurity Content_Day | Impurity Content_Night |
|---|---|
| 3.21 | 3.01 |
| 2.78 | 2.94 |
| 2.97 | 2.95 |
| 2.95 | 3.3 |
| 3.38 | 3.05 |
| 3.05 | 3.24 |
| 2.99 | 3.37 |
| 2.98 | 3.09 |
| 3.6 | 3.23 |

NOTE:

- You can convert stack to unstacked data, and vice versa, using the worksheet "Unstack and stack data"
- This worksheet is explained here

# Multiple Plots



## Unstacked data is generally less useful than stacked data

Look again at the data below

- Stacked data enables us to stratify data in as many ways as we like (here, batch, shift and machine… but there could be more – just add extra columns… day of the week? Production supervisor? Time since the vessel was last cleaned?)
- Stacked data enables us to use date and time – one date/time value applies to the whole of each row
  - With unstacked data, the rows do not necessarily correspond in this way – the second row of this example features 2.78% on day shift of 26/11/2019 and 2.94% on night shift of 27/11/2019

- Stacked data used for Stratified Plots

| Date/Time | Impurity Content | Batch | Shift | Machine |
|---|---|---|---|---|
| 26/11/2019 | 3.21 | a | Day | 110 |
| 26/11/2019 | 3.01 | a | Night | 180 |
| 26/11/2019 | 2.78 | a | Day | 220 |
| 27/11/2019 | 2.94 | a | Night | 110 |
| 28/11/2019 | 2.97 | a | Day | 180 |
| 29/11/2019 | 2.95 | a | Night | 220 |
| 29/11/2019 | 2.95 | a | Day | 110 |
| 30/11/2019 | 3.3 | a | Night | 180 |

- Unstacked Data used for Multiple Plots

| Impurity Content_Day | Impurity Content_Night |
|---|---|
| 3.21 | 3.01 |
| 2.78 | 2.94 |
| 2.97 | 2.95 |
| 2.95 | 3.3 |
| 3.38 | 3.05 |
| 3.05 | 3.24 |

NOTE:

- The main purpose of the ‘Multiple Plots’ section is to provide flexibility for you if your data is already unstacked.
- If you have the choice, stacked data is usually best
- One tool – Paired t-tests – works better with unstacked data, so it is located in this Multiple Plots section

# Multiple Plots

Multiple plots & t-tests

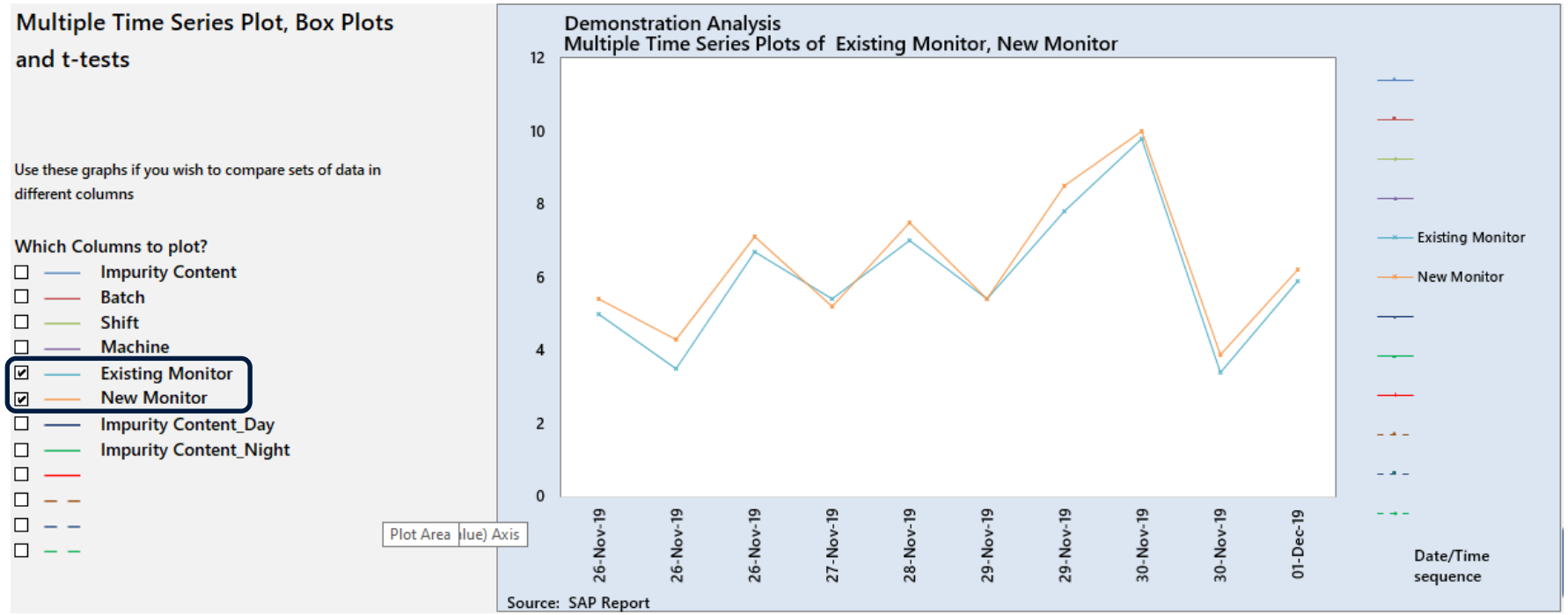
## Multiple Time Series Plots

- Copy the data from the Practice Data worksheet “Monitors” and paste the values into any free columns in the Toolkit
  - A hospital laboratory is investigating a new device for blood glucose monitoring. 10 blood specimens from different patients have been analysed by both the existing device and the new device.
  - The hospital is interested to see if there is a difference in the mean measurements between the two devices – if there is, that will call into question the validity of the results.

| Variable 5 | Variable 6 |
|---|---|
| Existing Monitor | New Monitor |
| 5 | 5.4 |
| 3.5 | 4.3 |
| 6.7 | 7.1 |
| 5.4 | 5.2 |
| 7 | 7.5 |
| 5.4 | 5.4 |
| 7.8 | 8.5 |
| 9.8 | 10 |
| 3.4 | 3.9 |
| 5.9 | 6.2 |

To make the plots, simply use the tick-boxes to select the columns you need – here, it’s “Existing Monitor” and “New Monitor”

From the time series plot, the readings from the new monitor seem to be higher than those from the existing one.



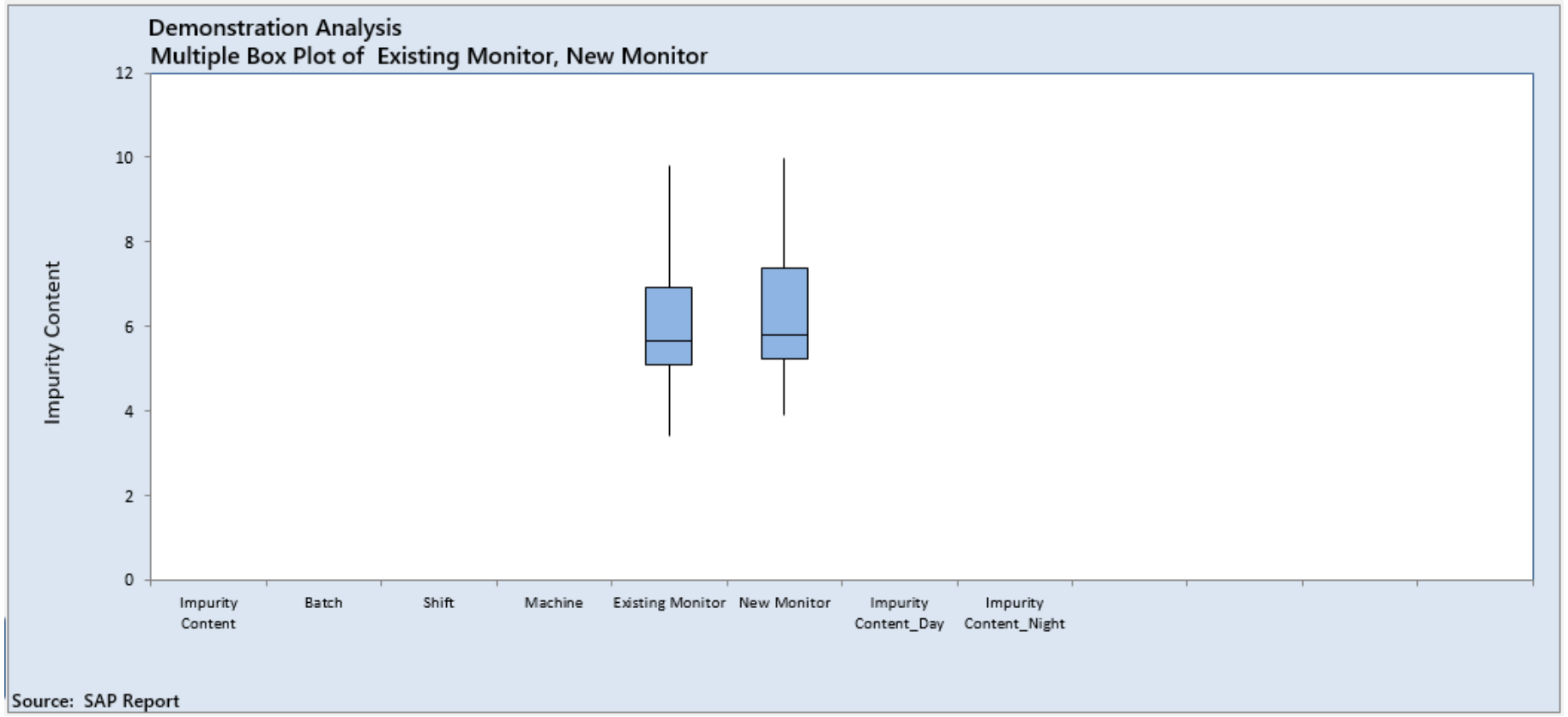
# Multiple Plots



## Box Plots

From the Time Series Plots, scroll to the right to see the Box Plots

- The layout and width of the boxes is slightly different to the ones created with stratified data, but the basic idea is exactly the same as for the box plots with stratified data.
- In this case, it is harder to see a difference with the box plots than it was with the time series plot.



# 2-sample t and Paired t-tests

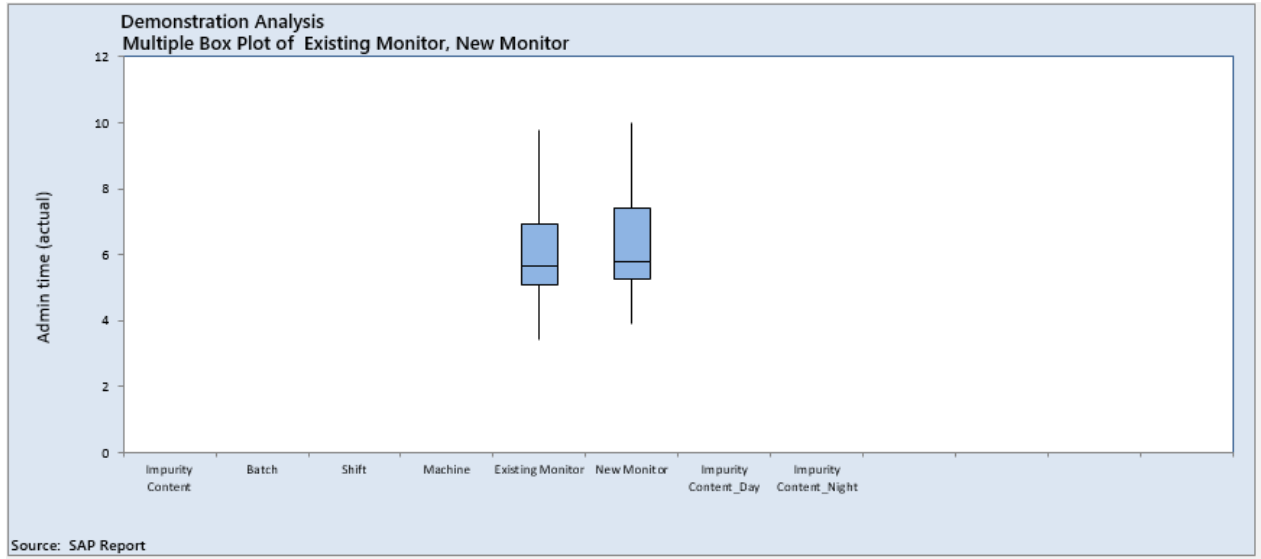
**Statistical comparison of the means of two samples of data**

# 2-sample t-tests

## A 2-sample t-test compares the means of two samples

- The sample sizes do not need to be the same
- The Null Hypothesis is that the means are the same and the Alternate Hypothesis is that they are different
- If you are not familiar with Hypothesis Tests, you can skip this section.

The analysis can be found to the right of the Multiple Box Plots.




Data Entry

**Two-Sample T test and CI for Existing Monitor and New Monitor**

(if you wish to compare more than two samples statistically, please used the stratified data section which is immediately above this one)

Null hypothesis: Averages are equal
Significance level: α=0.05

Alternative hypothesis: Averages are not equal
Equal variances are not assumed for this analysis

**Descriptive Statistics**

| Sample | n | Average | Standard Deviation | SE Mean |
|---|---|---|---|---|
| Existing Monitor | 10 | 5.990 | 1.941 | 0.614 |
| New Monitor | 10 | 6.350 | 1.920 | 0.607 |

Is this paired data? (data points in the two columns correspond to each other in some way)

**Estimate for Difference**

| Difference | 95% CI for Difference | |
|---|---|---|
| -0.360 | -2.182 | 1.462 |

**T test for Difference (2-tailed)**

| T-value | DF | P-Value |
|---|---|---|
| -0.42 | 17 | 0.682 |

As the P-value is above 0.05, you have not proved a difference

# 2-sample t-tests

**The output from the 2-sample t-test gives:**

- Descriptive statistics (1) gives the sample size, average, standard deviation and Standard Error of the Mean for each group
- The Estimate for Difference (2) gives the difference between the sample means and the 95% Confidence Interval for this
- The results of the t-test (3) (T-value, Degrees of Freedom and P-Value and a comment on the meaning of the P-value)
  - A 2-tailed test is used, this detects differences regardless of which sample has the larger mean

**Two-Sample T test and CI for Existing Monitor and New Monitor**

(if you wish to compare more than two samples statistically, please used the stratified data section which is immediately above this one)

Null hypothesis: Averages are equal
Significance level: α=0.05

Alternative hypothesis: Averages are not equal
Equal variances are not assumed for this analysis

**Descriptive Statistics** (1)

| Sample | n | Average | Standard Deviation | SE Mean |
|---|---|---|---|---|
| Existing Monitor | 10 | 5.990 | 1.941 | 0.614 |
| New Monitor | 10 | 6.350 | 1.920 | 0.607 |

Is this paired data? ☐
(data points in the two columns correspond to each other in some way)

**Estimate for Difference** (2)

| Difference | 95% CI for Difference | |
|---|---|---|
| -0.360 | -2.182 | 1.462 |

**T test for Difference (2-tailed)** (3)

| T-value | DF | P-Value |
|---|---|---|
| -0.42 | 17 | 0.682 |

As the P-value is above 0.05, you have not proved a difference

In this case, the P-value is 0.682, so we cannot say there is a difference between the two monitors.

NOTE:

- The two-sample t-test only works with exactly two samples. If you select only one sample, or more than two, this whole section will be blanked-out because it does not apply.
- If you wish to compare more than two samples, use ANOVA from the stratified data section here

# Paired t-tests

## Paired t-tests apply when (and only when) the data comes in pairs

- In this case, the data *is* paired (each row represents the same blood sample, analysed with two different monitors)
- If the sample sizes are different, the data cannot be paired and this option will not be available
- To indicate that the data is paired, select “Yes” from the paired data drop-down box, as shown

### Two-Sample T test and CI for Existing Monitor and New Monitor

(if you wish to compare more than two samples statistically, please used the stratified data section which is immediately above this one)

Null hypothesis: Averages are equal
Alternative hypothesis: Averages are not equal
Significance level: α=0.05
Equal variances are not assumed for this analysis

#### Descriptive Statistics

| Sample | n | Average | Standard Deviation | SE Mean |
|---|---|---|---|---|
| Existing Monitor | 10 | 5.990 | 1.941 | 0.614 |
| New Monitor | 10 | 6.350 | 1.920 | 0.607 |

Is this paired data? [No / Yes]
(data points in the two column correspond to each other in some way)

#### Estimate for Difference

| Difference | 95% CI for Difference | |
|---|---|---|
| -0.360 | -2.182 | 1.462 |

#### T test for Difference (2-tailed)

| T-value | DF | P-Value |
|---|---|---|
| -0.42 | 17 | 0.682 |

As the P-value is above 0.05, you have not proved a difference

# Paired t-tests

## Selecting ‘Yes’ for ‘Is this paired data?’ activates the Paired-t-test

- In addition to the 2-sample t-test data you can now see an additional row
  - Estimate for paired differences(1) gives the mean and confidence interval for the paired differences in the rows of data (one for each blood sample, in this case)
  - Paired t-Test (2-tailed) (2) gives the P-value for the paired test
- A paired t-test removes the variation between the samples, and so reduces the amount of random variation. This makes it more powerful than 2-sample t-tests, and in this case we are able to prove that there is a difference between the two monitors.

**Two-Sample T test and CI for Existing Monitor and New Monitor**

(if you wish to compare more than two samples statistically, please used the stratified data section which is immediately above this one)

Null hypothesis: Averages are equal
Significance level: α=0.05

Alternative hypothesis: Averages are not equal
Equal variances are not assumed for this analysis

**Descriptive Statistics**

| Sample | n | Average | Standard Deviation | SE Mean |
|---|---|---|---|---|
| Existing Monitor | 10 | 5.990 | 1.941 | 0.614 |
| New Monitor | 10 | 6.350 | 1.920 | 0.607 |

| Is this paired data? | Yes |
|---|---|

(data points in the two columns correspond to each other in some way)

**Estimate for Difference**

| Difference | 95% CI for Difference | |
|---|---|---|
| -0.360 | -2.182 | 1.462 |

**T test for Difference (2-tailed)**

| T-value | DF | P-Value |
|---|---|---|
| -0.42 | 17 | 0.682 |

As the P-value is above 0.05, you have not proved a difference

**Estimate for Paired Difference** (1)

| Difference | 95% CI for Difference | |
|---|---|---|
| -0.360 | -0.576 | -0.144 |

**Paired T-Test (2-tailed)** (2)

| T-value | DF | P-Value |
|---|---|---|
| -3.762 | 9 | 0.004 |

Paired-T comment:
As the P-value is below 0.05, the difference is statistically significant

# Multi-Vari Charts

**Parallel comparison of the effects of up to 3 factors**

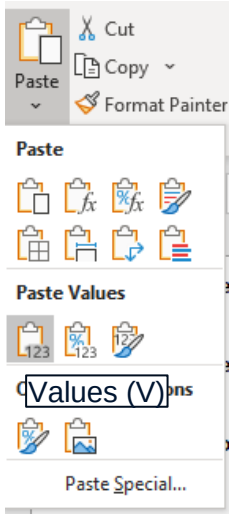
# Main Effects and Multi-Vari Charts

**Multi-Vari Charts enable you to graphically analyse the affect of up to three factors at once**

- This can help uncover hard-to-see relationships
  - Example: Suppose a new software system has been introduced. It improves productivity but only if adequate training has been given. For people who have not had training, it makes productivity worse. In this case, to understand what affects an individual's productivity you would need to be able to look at both the software version used *and* whether the person has received adequate training or not.

We shall use a new dataset to showcase Multi-Vari Charts, this is based on the price paid for motor insurance.

- Delete all the data currently in the Toolkit
- Using the Practice Data file again, copy all the data from the worksheet "Insurance" and paste the values into the Toolkit
  - Remember to use Paste Values as described here




| Time axis | Variable 1 | Variable 2 | Variable 3 | Variable 4 | Variable 5 | Variable 6 | Variable 7 |
|---|---|---|---|---|---|---|---|
| Date | Age | Age Group | Gender | Homeowner | Vehicle Age (years) | Vehicle value class | Insurance cost |
| 13/06/2020 | 28 | 25-49 | Female | Yes | 1 | Medium | 473 |
| 13/06/2020 | 27 | 25-49 | Female | Yes | 1 | Medium | 409 |
| 13/06/2020 | 49 | 25-49 | Male | No | 0 | high | 467 |
| 13/06/2020 | 47 | 25-49 | Male | No | 0 | Medium | 401 |
| 13/06/2020 | 64 | 50-74 | Male | Yes | 5 | Low | 399 |
| 13/06/2020 | 19 | Under 25 | Female | No | 4 | Low | 462 |
| 13/06/2020 | 72 | 50-74 | Male | Yes | 1 | high | 399 |
| 13/06/2020 | 22 | Under 25 | Female | No | 0 | Medium | 496 |
| 13/06/2020 | 32 | 25-49 | Female | Yes | 1 | Medium | 396 |
| 13/06/2020 | 37 | 25-49 | Male | Yes | 0 | Medium | 375 |
| 13/06/2020 | 29 | 25-49 | Female | No | 0 | Medium | 521 |

# Main Effects and Multi-Vari Charts

Main Effects & Multi-Vari

## Click the hyperlink “Main Effects & Multi-vari charts”

We shall look at the effect of three different factors on the price paid for insurance

- Age group
- Gender
- Homeowner

Tick the appropriate boxes as shown in the illustration

- Note, a maximum of 3 Xs can be selected at one time – if more are selected, only the first three will be used
- Note, the charts will handle a maximum of 5 possible values for each X. If there are more than 5 different values, only the five most commonly-occurring values will be used.


These are the Main Effects Plots. They show the simple averages of each setting for each X.

- We can see that the age group “Under 25” has a higher insurance cost than the others
- Gender does not seem to make a difference
- Homeowners seem to pay less than non-homeowners

# Multi-Vari Charts



**Scroll to the right to see the multi-vari chart for our chosen Xs.**

This is a complex graph to interpret, as it explores the interplay between three different Xs

The first X on the list is Age Group.

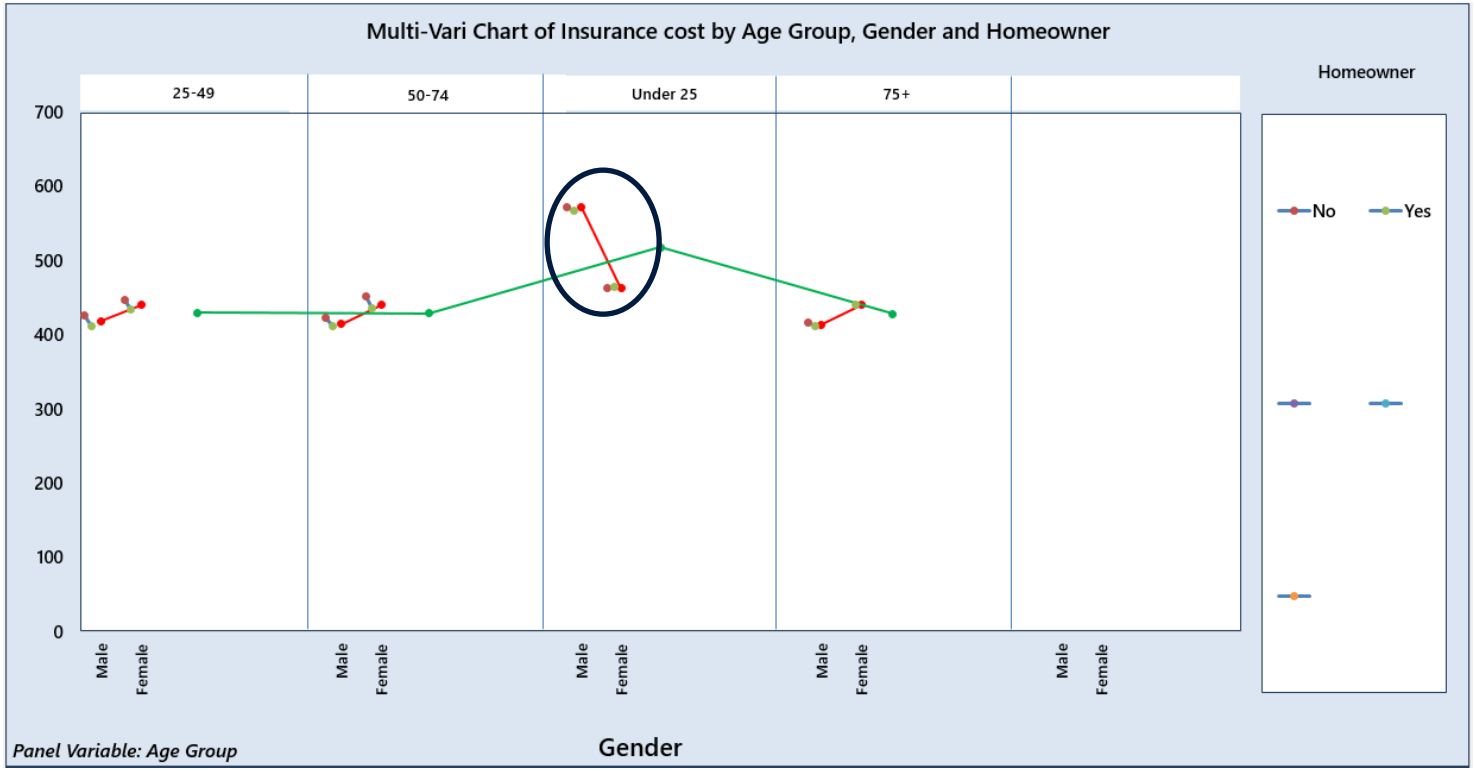
- The different values are shown in the five panels (25-49, 50-74, Under 25, 75+ and the fifth panel is blank)
- The averages for each panel are connected with a green line

The second X on the list is Gender

- The values of Gender – Female and Male – are shown as subgroups within each panel
- Within each panel, the average of the two genders are shown with a red line

The third X on the list is Homeowner

- Within each subgroup you will see coloured dots, representing the different values for the third X
- The colour-code for these is given in the legend on the right-hand side (in this case, brown for “Yes” and pale green for “No”)




The most notable feature is the sharp difference in cost between Male and Female in the Under 25 category – you cannot see these if you only review one factor at a time.

# Multi-Vari Charts

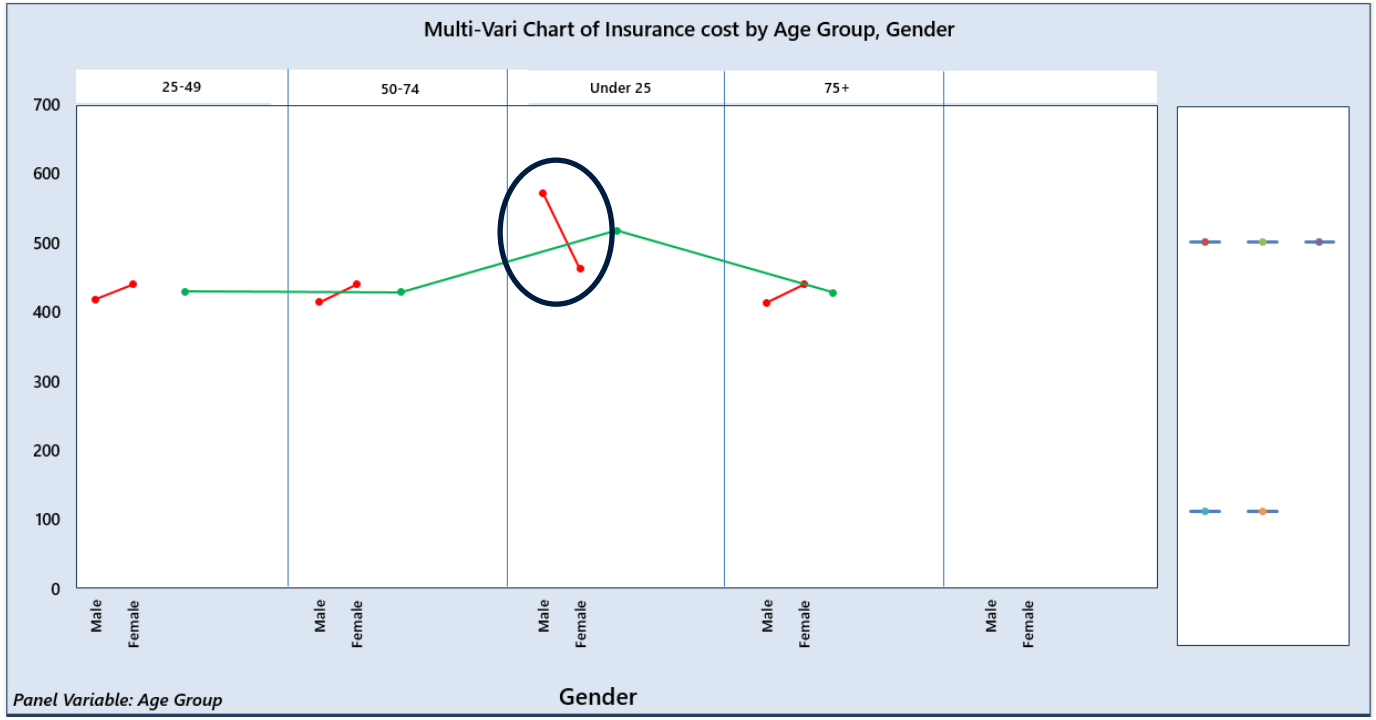


## Let's simplify the chart to just the two key factors – Age Group and Gender

- Un-tick the box for 'Homeowner' the make the graph clearer

Now the effect of being male and under 25 is even clearer.

Note that the legend box will be empty if you do not select a third X.  Generally speaking, two Xs are enough to discover the hidden relationships



# Scatter Plot and Regression

## Including Stratification, Correlation and Residuals Analysis

# Scatter Plot and Regression

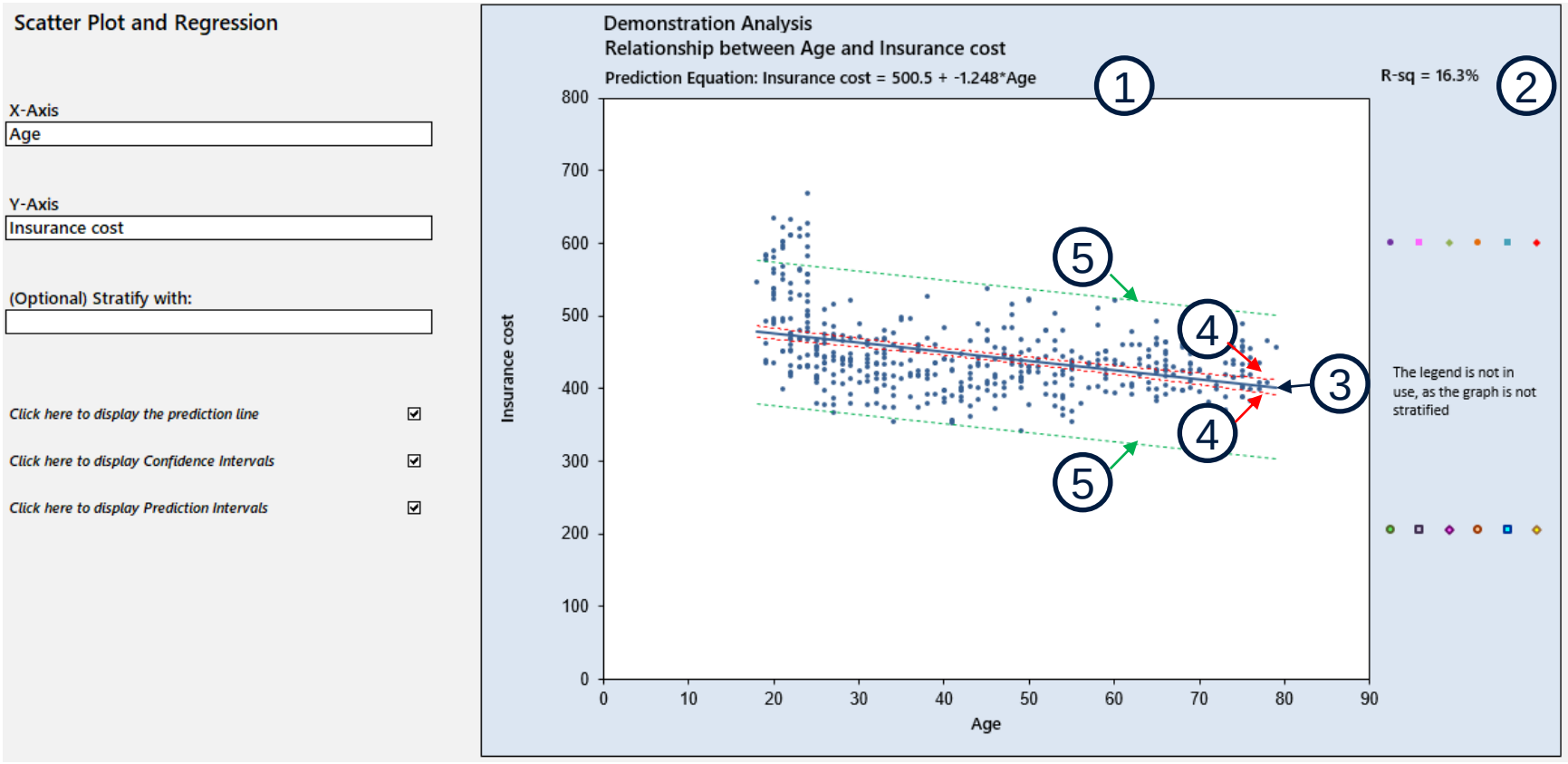
Scatter Plot and Regression

## A Scatter Plot shows the relationship between two continuous variables

Here, we'll use age as the X and Insurance cost as the Y. Click the hyperlink and select these from the drop-down list as shown. Delete the contents of the optional 'Stratify with' box for now, and tick the remaining boxes, as shown

The graph shows:

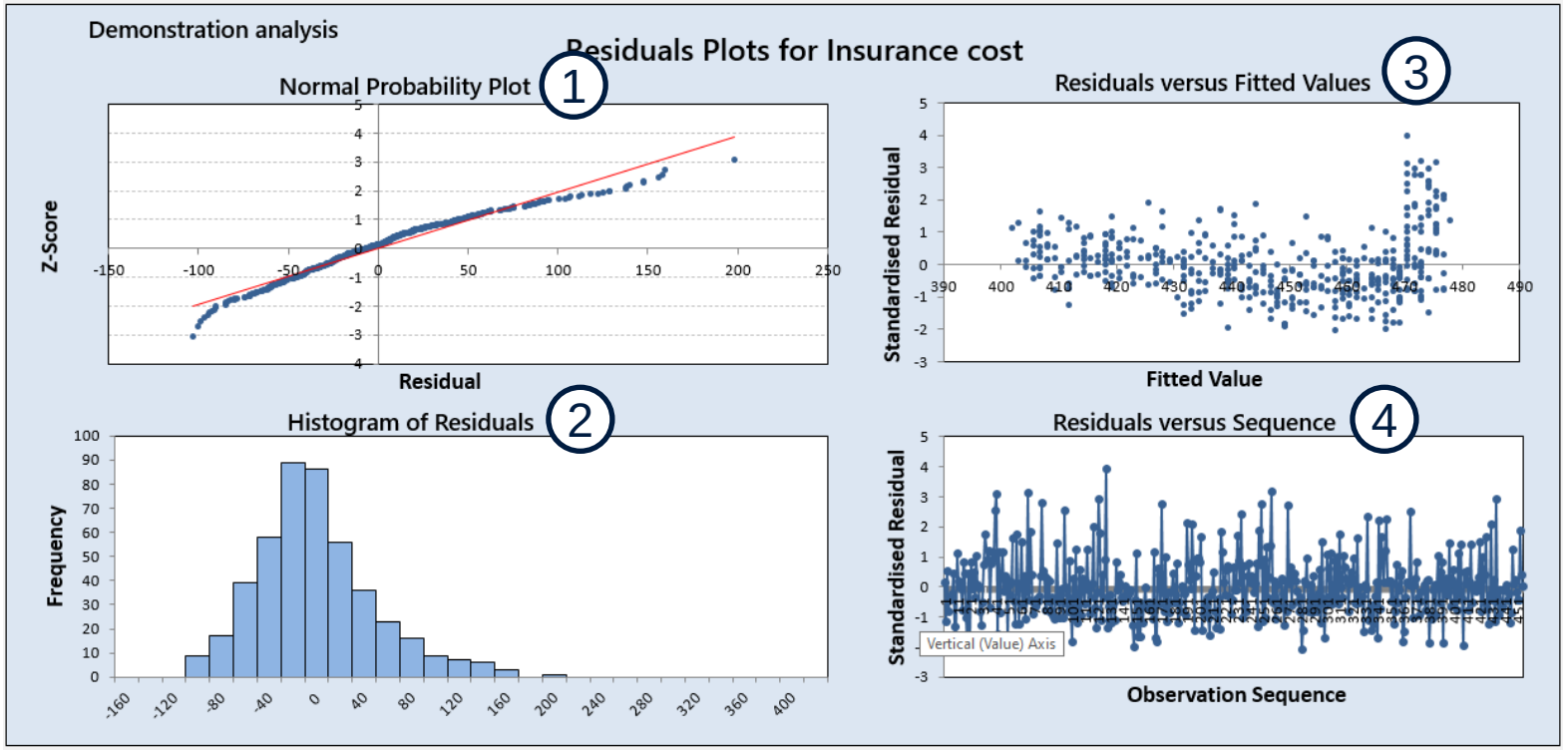
- The regression equation for the relationship between the X and the Y (1)
- The strength of the relationship, R-squared (2)
- The line of best fit for the X and the Y (3)
- The 95% Confidence Interval, showing the range within which the line would probably lie if you had additional data (4)
- The 95% prediction interval, showing the range within which future data points will probably lie (5)
- The regression line, confidence intervals and prediction intervals can be turned on and off with the three tick boxes



# Scatter Plot and Regression

## Residuals Analysis

The graphs to the right provide analysis of the residual errors. These are expected to be Normally distributed – as we can see, in this example, a number of concerns arise.




### Normal Probability Plot (1)

- This should be a straight line if the residuals are Normal – it has a slight curve in this case

### Histogram of Residuals (2)

- This will have a bell shape if the residuals are Normal, but it clearly has a skew to the right

### Residuals versus Fitted Values (3)

- You would like the residual errors to be similar, regardless of the fitted values. But we can see that the errors are much larger when the fitted values are higher

### Residuals versus sequence (4)

- This one, at least, seems fine

Comment:

- There are two causes for these problems with the residuals:
  - The data itself is skewed
  - Insurance costs have a specific discontinuity at age 25. Further analysis might best be done with under 25 and over 25 as separate groups.

# Fitted Line Plot

## Regression Analysis

Regression Analysis: Insurance cost versus Age

| Predictor (1) | | Coefficient (2) | Std Error | T | P |
|---|---|---|---|---|---|
| Constant | | 500.5 | 6.247 | 80.12 | 0.000 |
| Age | | -1.248 | 0.133 | -9.383 | 0.000 |
| s = 50.16 | R-sq = | 16.27% (3) | | | |

Analysis of Variance (4)

| Source | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| Regression | 1 | 221500 | 221500 | 88.04 | 0.000 |
| Residual Error | 453 | 1140000 | 2516 | | |
| Total | 454 | 1362000 | | | |

- The predictors (1) are the constant (intercept) and the X (here, Age).
- The coefficients of the prediction equation are given (2) together with additional statistical analysis giving standard error, T value and P value
  - Most notably, a P value below 0.05 means that this term is statistically significant
- The strength of the relationship (3) is quantified with s (the standard deviation of the residuals) and the R-squared value
- The Analysis of Variance table (4) gives the degrees of freedom (DF), Sum of Squares, Mean Square and P values.

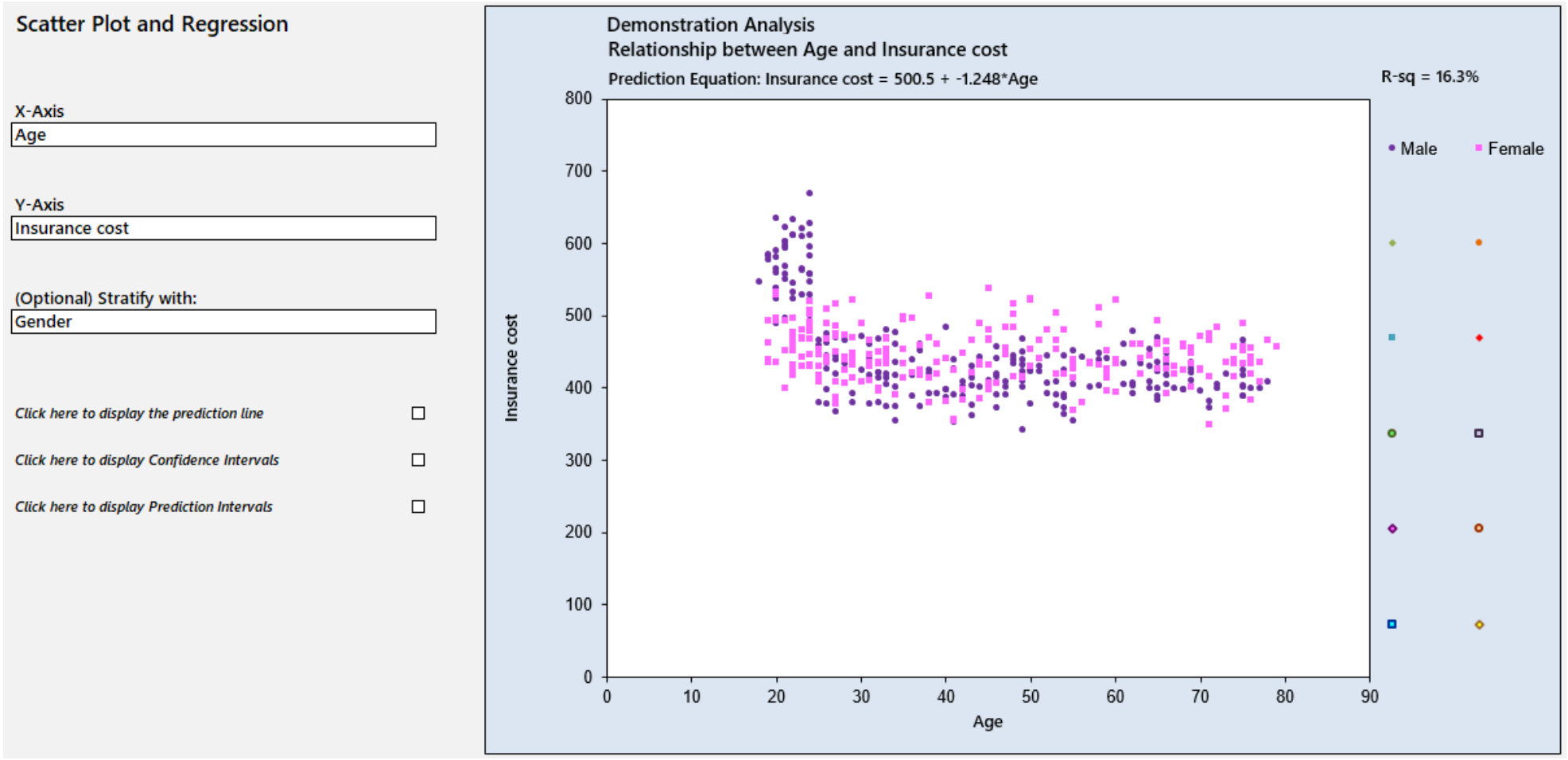
# Stratified Scatter Plot

Scatter Plot and Regression

**We shall now return to the scatter plot and look at the stratification feature**

Scroll left to return to the scatter plot

- In the optional ‘Stratify with’ box, select the column Gender
- To make the graph clearer, untick the boxes for prediction line, Confidence Intervals and Prediction Intervals
- The colour-coding of the stratified plot enables us to see that all those under-25s paying high premiums are male; for older drivers, females seem to pay slightly more than males



# Multiple Regression

## Including Matrix Plot, Correlation, Curvilinear Regression and Residuals Analysis

# Multiple Regression

Multiple Regression

**Multiple Regression enables us to model the Y with several Xs, to create a potentially more powerful model**

We shall use a new dataset for this section. Please delete all the data in the Toolkit and replace it with the data in the worksheet 'Life Expectancy' from the Practice Data. Remember to paste *values.*

The data comes from an entirely fictional study of the factors that influence life expectancy. Please do not read anything serious into the results!

We have:

- Variable 1 – Age at death – will be our Y
- Variable 2 – Female – will be a 0 if the person is male and a 1 if female (this enables us to create variable data out of something that has two categories)
- Variable 3 and 4 give the age at which the person's parents died
- Variables 5, 6 and 7 give the person's consumption of cigarettes, alcohol and fruit/vegetables per day
- Variable 8 gives the number of children they had
- Variables 9 and 10 give measures of the air quality where they lived – Nitrous oxides (NOX) and particulates in the air. Both of these are on a scale whereby 100 is the national average.

| Time axis | Variable 1 | Variable 2 | Variable 3 | Variable 4 | Variable 5 | Variable 6 | Variable 7 | Variable 8 | Variable 9 | Variable 10 | Variable 11 | Variable 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Age at death | Female | Father age at death | Mother age at death | Cigs per day | Units Alcohol per week | Portions fruit or veg per day | Number of children | NOX where living (normal=100) | Particulates where living (normal=100) | | |
| | 76 | 1 | 74 | 79 | 0 | 1 | 4 | 3 | 133 | 106 | | |
| | 74 | 0 | 71 | 97 | 4 | 4 | 2 | 3 | 131 | 155 | | |
| | 81 | 1 | 73 | 81 | 0 | 15 | 3 | 1 | 128 | 138 | | |
| | 73 | 0 | 72 | 79 | 3 | 21 | 3 | 1 | 25 | 27 | | |
| | 59 | 1 | 69 | 82 | 32 | 27 | 5 | 4 | 150 | 169 | | |
| | 78 | 0 | 73 | 85 | 10 | 5 | 1 | 3 | 137 | 123 | | |
| | 76 | 1 | 68 | 82 | 0 | 2 | 3 | 1 | 17 | 18 | | |
| | 72 | 0 | 71 | 87 | 0 | 11 | 3 | 4 | 45 | 50 | | |
| | 75 | 1 | 73 | 90 | 0 | 25 | 3 | 3 | 97 | 110 | | |
| | 76 | 0 | 68 | 95 | 1 | 3 | 1 | 3 | 169 | 190 | | |
| | 73 | 1 | 68 | 91 | 0 | 9 | 5 | 4 | 29 | 34 | | |
| | 55 | 0 | 68 | 82 | 0 | 22 | 1 | 0 | 23 | 27 | | |
| | 84 | 1 | 78 | 96 | 0 | 9 | 3 | 3 | 41 | 47 | | |
| | 63 | 0 | 94 | 84 | 55 | 30 | 2 | 4 | 135 | 128 | | |
| | 69 | 1 | 73 | 82 | 0 | 19 | 3 | 0 | 136 | 131 | | |
| | 76 | 0 | 68 | 91 | 46 | 33 | 1 | 1 | 137 | 150 | | |
| | 84 | 1 | 74 | 86 | 12 | 4 | 3 | 3 | 25 | 30 | | |
| | 69 | 0 | 74 | 78 | 16 | 9 | 1 | 1 | 157 | 173 | | |
| | 65 | 1 | 70 | 85 | 38 | 14 | 4 | 4 | 118 | 138 | | |
| | 68 | 0 | 69 | 83 | 25 | 11 | 1 | 3 | 22 | 23 | | |

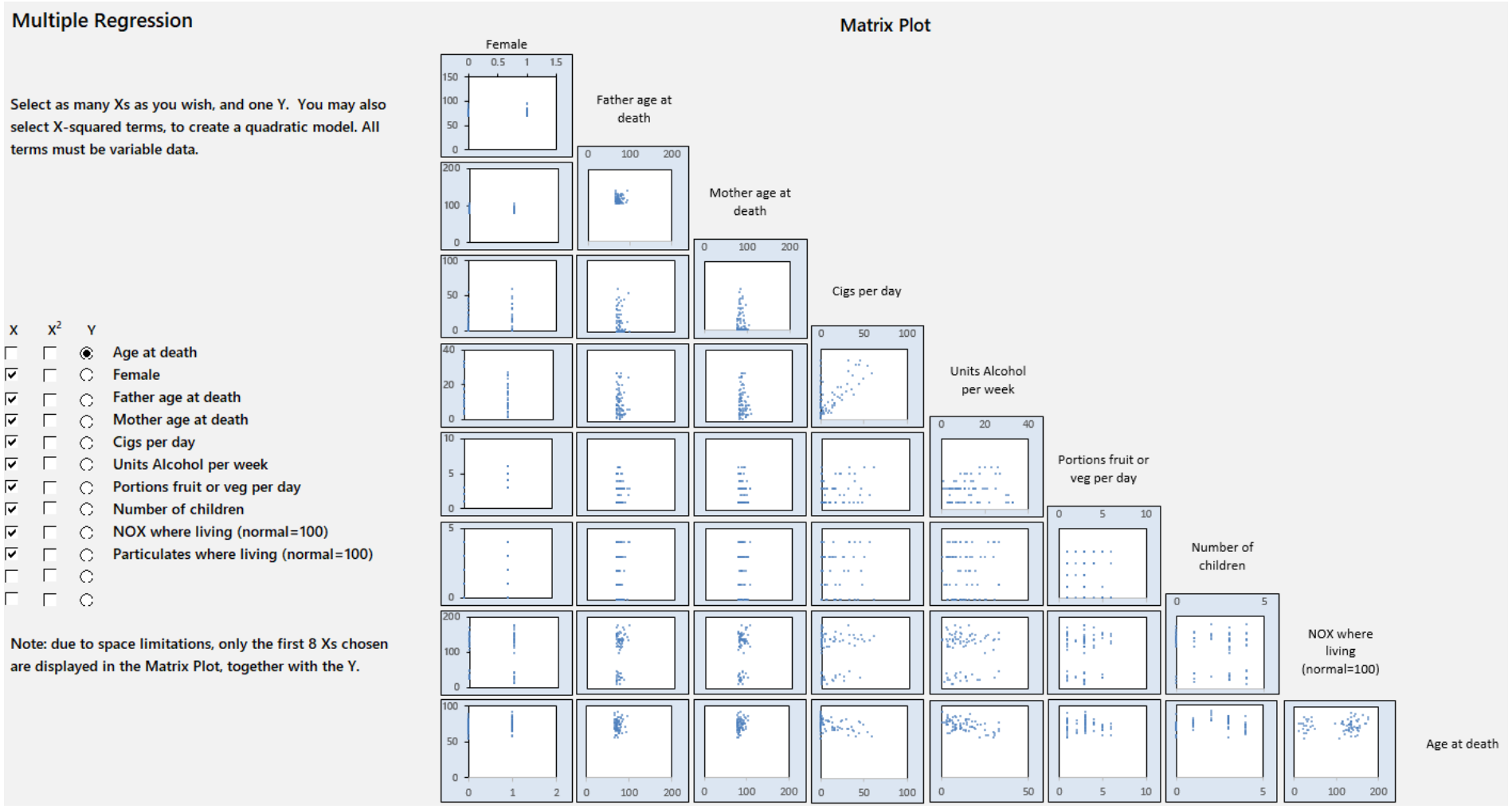
# Multiple Regression

Multiple Regression

## Start with a Matrix Plot

This enables you to visualise many variables' relationships at the same time.

- Click the blue Multiple Regression hyperlink and select the Y radio button as Age at death, and tick all the other variables as your potential Xs, as shown here
- The matrix plot will appear automatically. Note that it is limited to the first 8 Xs for space reasons, so in this case we do not see the X 'Particulates where living'.
- Notice that, although it is the first variable on the list, the Y is automatically placed in the last row of the matrix plot. This enables you to see the relationship between the Y and each of the Xs in one row, with the Y as the vertical axis



# Multiple Regression



## Multiple Regression calculations

Scroll to the right to see the regression analysis, including all the Xs (whether or not there was room for them in the matrix plot)

- You can see the standard deviation of the residuals, R-squared and R-squared adjusted (which is a better measure of multiple regression performance, as it takes account of the number of Xs being used) (1)
- The regression coefficients section tells you the coefficient of each factor in the prediction equation, the p-value to show statistical significance and the Variance Inflation Factor (VIF) which highlights multicollinearity. (2). See next slide for VIFs.
- The Analysis of Variance section shows how the variation in the Y can be broken down into the variation explained by the Xs and residual error (3)
- Finally, the Prediction Equation is given in full (4)

**Regression Performance Summary for Age at death** ①

| s | R-sq | R-sq adj |
|---|---|---|
| 6.7672 | 36.82% | 28.07% |

**Regression Coefficients** ②

| Term | Coefficient | Std Error | T | P | VIF |
|---|---|---|---|---|---|
| Constant | 57.6 | 14 | 4.12 | 0.000 | |
| Female | 6.46 | 2.6 | 2.48 | 0.016 | 2.77 |
| Father age at death | -0.0478 | 0.13 | -0.37 | 0.713 | 1.11 |
| Mother age at death | 0.231 | 0.123 | 1.88 | 0.065 | 1.14 |
| Cigs per day | -0.0642 | 0.0633 | -1.01 | 0.316 | 1.68 |
| Units Alcohol per week | -0.289 | 0.117 | -2.47 | 0.016 | 1.69 |
| Portions fruit or veg pe | -1.53 | 0.868 | -1.76 | 0.083 | 2.72 |
| Number of children | 1.08 | 0.525 | 2.05 | 0.044 | 1.07 |
| NOX where living (norm | 0.039 | 0.0462 | 0.84 | 0.404 | 9.15 |
| Particulates where living | -0.0195 | 0.0421 | -0.46 | 0.647 | 9.04 |

**Analysis of Variance** ③

| Source | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| Regression | 9 | 1735 | 192.8 | 4.2105 | 0.000 |
| Residual Error | 65 | 2977 | 45.79 | | |
| Total | 74 | 4712 | | | |

**Prediction Equation** ④

Age at death = 57.65 +6.456*Female -0.04785*Father age at death +0.2312*Mother age at death -0.06424*Cigs per day -0.2894*Units Alcohol per week -1.529*Portions fruit or veg per day +1.079*Number of children +0.03897*NOX where living (normal=100) - 0.01953*Particulates where living (normal=100)

# What is a Variance Inflation Factor (VIF)?



## A word about Multicollinearity

- Suppose you are a scout for a basketball team. You visit schools trying to identify future talent. You discover that you can predict point-scoring potential based on their height. And there's more - the length of a child's femur is also a good predictor. What's more, the length of their tibia (shin bone) is a predictor. And the length of their forearm…
- Should you put all of these into a regression model? Probably not, because they all linked – if you know the child is tall, you can pretty much guess that the longer bones in their body will be long, too
  - You will simply be explaining the same variation with several Xs
- Having several Xs closely associated with each other is called multicollinearity. And it is undesirable in regression – it can cause errors in the R-squared value and give misleading coefficients.
- The Variance Inflation Factor (VIF) is a measure of the multicollinearity between each X and all the other Xs.
- As a general rule, your VIFs should ideally not be greater than 5 and definitely not greater than 10. If your model breaks this rule, take out one of the collinear Xs
- We can make an exception to this rule if we are introducing X-squared terms – the multicollinearity will only apply over a limited range of values.

# Multiple Regression



## Removing terms with a high Variance Inflation Factor

Variance Inflation Factor (VIF) is a measure of the strength of association between each X and all the other Xs in the model. You don't want to see strong associations, because these are a sign that two Xs are giving us similar information, and high VIFs can cause the model to be distorted. VIFS should all be below 10, and ideally less than 5.

- We can see that the VIFs for NOX and particulates are rather high – both over 9. This is not surprising – if you live in an area with high air pollution, you are likely to have poor scores on multiple areas of air quality, so these Xs are strongly correlated.

- Scroll to the left to return to the area where you select the Xs, and untick the box for Particulates.
- You will see that, with Particulates removed, we have a model where all the VIFs are below 5.

**Regression Performance Summary for Age at death**

| s | R-sq | R-sq adj |
|---|---|---|
| 6.7672 | 36.82% | 28.07% |

**Regression Coefficients**

| Term | Coefficient | Std Error | T | P | VIF |
|---|---|---|---|---|---|
| Constant | 57.6 | 14 | 4.12 | 0.000 | |
| Female | 6.46 | 2.6 | 2.48 | 0.016 | 2.77 |
| Father age at death | -0.0478 | 0.13 | -0.37 | 0.713 | 1.11 |
| Mother age at death | 0.231 | 0.123 | 1.88 | 0.065 | 1.14 |
| Cigs per day | -0.0642 | 0.0633 | -1.01 | 0.316 | 1.68 |
| Units Alcohol per week | -0.289 | 0.117 | -2.47 | 0.016 | 1.69 |
| Portions fruit or veg pe | -1.53 | 0.868 | -1.76 | 0.083 | 2.72 |
| Number of children | 1.08 | 0.525 | 2.05 | 0.044 | 1.07 |
| NOX where living (norm | 0.039 | 0.0462 | 0.84 | 0.404 | 9.15 |
| Particulates where living | -0.0195 | 0.0421 | -0.46 | 0.647 | 9.04 |

**Analysis of Variance**

| Source | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| Regression | 9 | 1735 | 192.8 | 4.2105 | 0.000 |
| Residual Error | 65 | 2977 | 45.79 | | |
| Total | 74 | 4712 | | | |

**Prediction Equation**

Age at death = 57.65 +6.456*Female -0.04785*Father age at death +0.2312*Mother age at death -0.06424*Cigs per day -0.2894*Units Alcohol per week -1.529*Portions fruit or veg per day +1.079*Number of children +0.03897*NOX where living (normal=100) - 0.01953*Particulates where living (normal=100)

# Multiple Regression

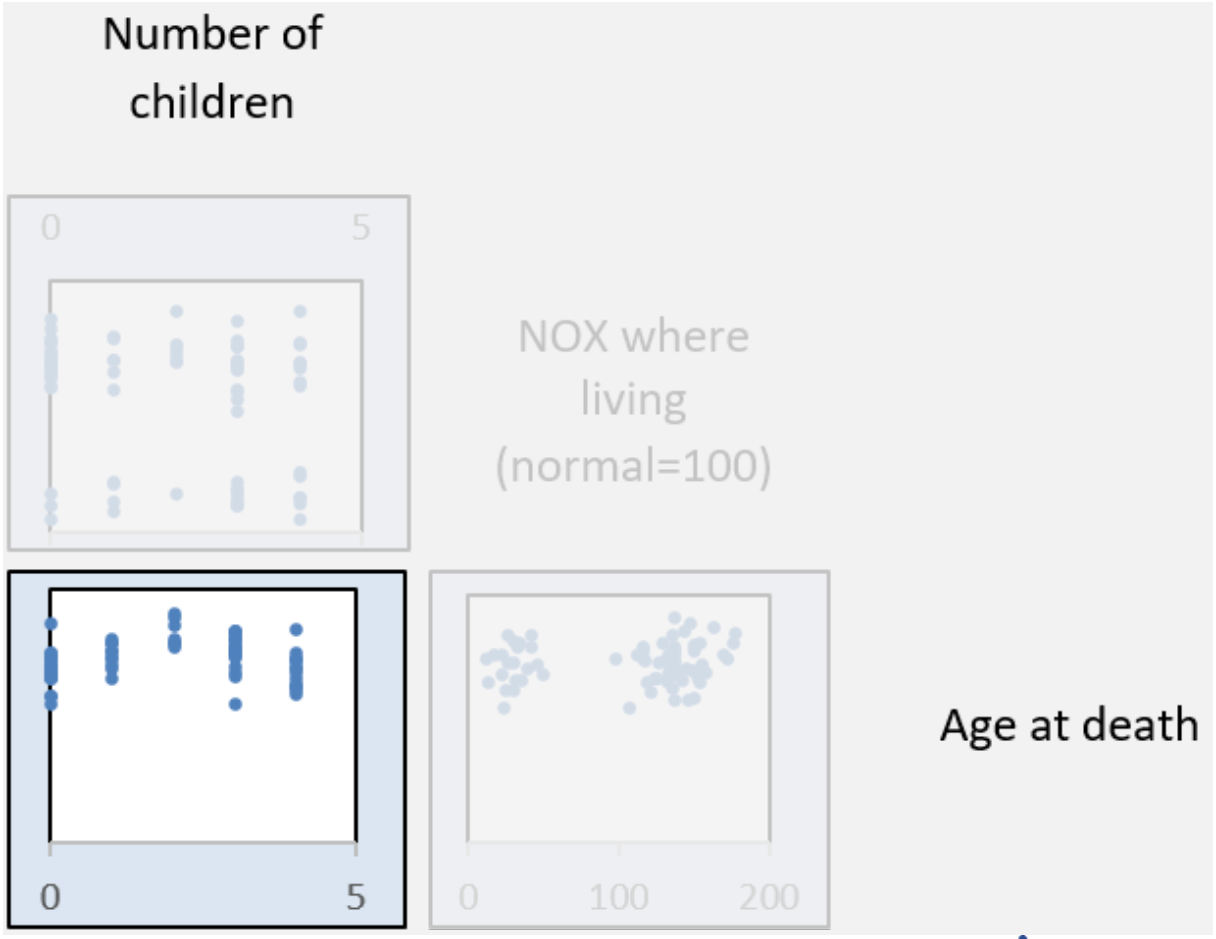


## Adding Squared Terms (Curvilinear regression)

It's always worth looking to see if there is a curved relationship between any of the Xs and the Y. If there are, we can add a squared term to model this.

- Reviewing the matrix plot, we notice that there may be a curved relationship between the number of children and age at death – people in this fictional study lived longer if they have two children than if they have 0 or 4. We have zoomed into the bottom-right corner of the matrix plot to highlight this.
- A straight line will not do much to predict age at death based on the number of children, but perhaps a curved line will.
- To add a squared term, all that is needed is to tick the relevant box under $X^2$, as shown.

| X | $X^2$ | Y | |
|---|---|---|---|
| ☐ | ☐ | ◉ | Age at death |
| ☑ | ☐ | ○ | Female |
| ☑ | ☐ | ○ | Father age at death |
| ☑ | ☐ | ○ | Mother age at death |
| ☑ | ☐ | ○ | Cigs per day |
| ☑ | ☐ | ○ | Units Alcohol per week |
| ☑ | ☐ | ○ | Portions fruit or veg per day |
| ☑ | ☑ | ○ | Number of children |
| ☑ | ☐ | ○ | NOX where living (normal=100) |
| ☐ | ☐ | ○ | Particulates where living (normal=100) |



# Multiple Regression



## Adding Squared Terms (Curvilinear regression) continued

The term ‘Number of children^2’ has now been added to the model, and our R-squared adjusted has jumped up to 59.57%, so this was a very helpful enhancement.

- The VIFs for Number of children and Number of children^2 have both jumped and are now over 17 – this is to be expected when you have a term which is the square of another term, and we can make an exception to the normal rule about removing terms with high VIFs in such cases.

**Regression Performance Summary for Age at death**

| s | R-sq | R-sq adj |
|---|---|---|
| 5.0734 | 64.49% | 59.57% |

**Regression Coefficients**

| Term | Coefficient | Std Error | T | P | VIF |
|---|---|---|---|---|---|
| Constant | 34.3 | 11 | 3.12 | 0.003 | |
| Female | 4.01 | 1.97 | 2.03 | 0.046 | 2.84 |
| Father age at death | 0.151 | 0.102 | 1.48 | 0.144 | 1.2 |
| Mother age at death | 0.254 | 0.092 | 2.76 | 0.008 | 1.14 |
| Cigs per day | -0.0972 | 0.0475 | -2.04 | 0.045 | 1.68 |
| Units Alcohol per week | -0.195 | 0.0887 | -2.20 | 0.031 | 1.72 |
| Portions fruit or veg pe | 0.0939 | 0.685 | 0.14 | 0.889 | 3.01 |
| Number of children | 12.1 | 1.6 | 7.59 | 0.000 | 17.7 |
| Number of children^2 | -2.95 | 0.414 | -7.14 | 0.000 | 17.9 |
| NOX where living (norn | 0.0125 | 0.0123 | 1.01 | 0.316 | 1.16 |

# Multiple Regression



## Reducing the model

We should now remove any terms that do not make a significant contribution to the model

- Review the column of P values. If the p value is above 0.05, that means this term is not statistically significant and should be removed.
- Start with the term with the highest p value (here, portions of fruit or veg) and take out the insignificant terms one at a time, checking the p-values each time to select the one with the highest p value to remove next.

**Regression Performance Summary for Age at death**

| s | R-sq | R-sq adj |
|---|---|---|
| 5.0734 | 64.49% | 59.57% |

**Regression Coefficients**

| Term | Coefficient | Std Error | T | P | VIF |
|---|---|---|---|---|---|
| Constant | 34.3 | 11 | 3.12 | 0.003 | |
| Female | 4.01 | 1.97 | 2.03 | 0.046 | 2.84 |
| Father age at death | 0.151 | 0.102 | 1.48 | 0.144 | 1.2 |
| Mother age at death | 0.254 | 0.092 | 2.76 | 0.008 | 1.14 |
| Cigs per day | -0.0972 | 0.0475 | -2.04 | 0.045 | 1.68 |
| Units Alcohol per week | -0.195 | 0.0887 | -2.20 | 0.031 | 1.72 |
| Portions fruit or veg pe | 0.0939 | 0.685 | 0.14 | 0.889 | 3.01 |
| Number of children | 12.1 | 1.6 | 7.59 | 0.000 | 17.7 |
| Number of children^2 | -2.95 | 0.414 | -7.14 | 0.000 | 17.9 |
| NOX where living (norn | 0.0125 | 0.0123 | 1.01 | 0.316 | 1.16 |

# Multiple Regression

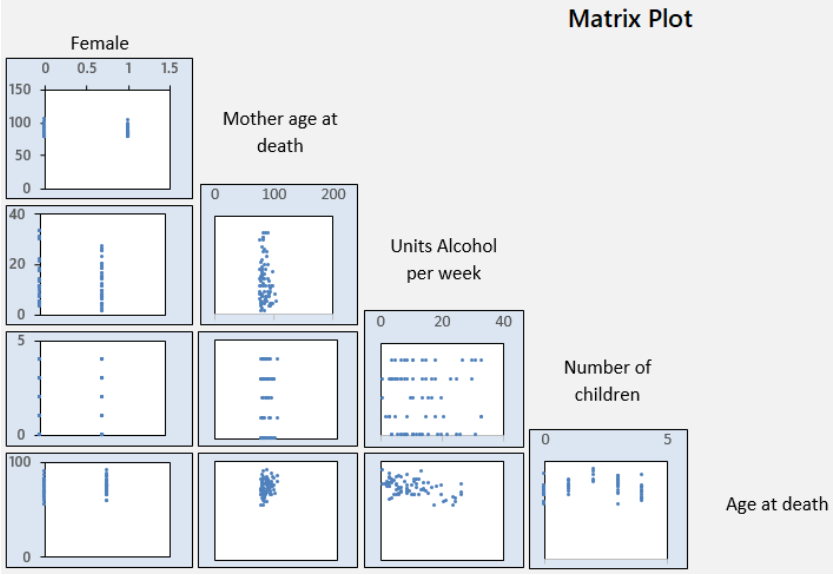


## Reducing the model, continued

After removing all the terms with p>0.05, starting with the highest p values, we have been able to reduce our model so that it contains only 5 factors (one of which is the squared term)*.

- The R-squared adjusted has fallen slightly to 57.24%, but it is worth it to have a simpler model.

- Returning to the Matrix Plot, there hardly seems to be any relationship with age of death, other than the curved relationship with children. Multiple regression enables us to find relationships from combinations of Xs.




**Regression Performance Summary for Age at death**

| s | R-sq | R-sq adj |
|---|---|---|
| 5.2176 | 60.13% | 57.24% |

**Regression Coefficients**

| Term | Coefficient | Std Error | T | P | VIF |
|---|---|---|---|---|---|
| Constant | 47 | 8.29 | 5.67 | 0.000 | |
| Female | 4.31 | 1.26 | 3.44 | 0.001 | 1.09 |
| Mother age at death | 0.261 | 0.0908 | 2.87 | 0.005 | 1.05 |
| Units Alcohol per week | -0.31 | 0.0717 | -4.32 | 0.000 | 1.07 |
| Number of children | 11.5 | 1.53 | 7.47 | 0.000 | 15.4 |
| Number of children^2 | -2.8 | 0.393 | -7.13 | 0.000 | 15.3 |

* Do not use the results of this made-up case study for health advice. Don't smoke, and eat your greens.

# Multiple Regression



## Residuals Plots

As with the simple linear regression model, if you scroll to the right you will find a graphical analysis of the residual errors.

### Normal Probability Plot (1)

- This should be a roughly straight line if the residuals are Normal – this one looks OK

### Histogram of Residuals (2)

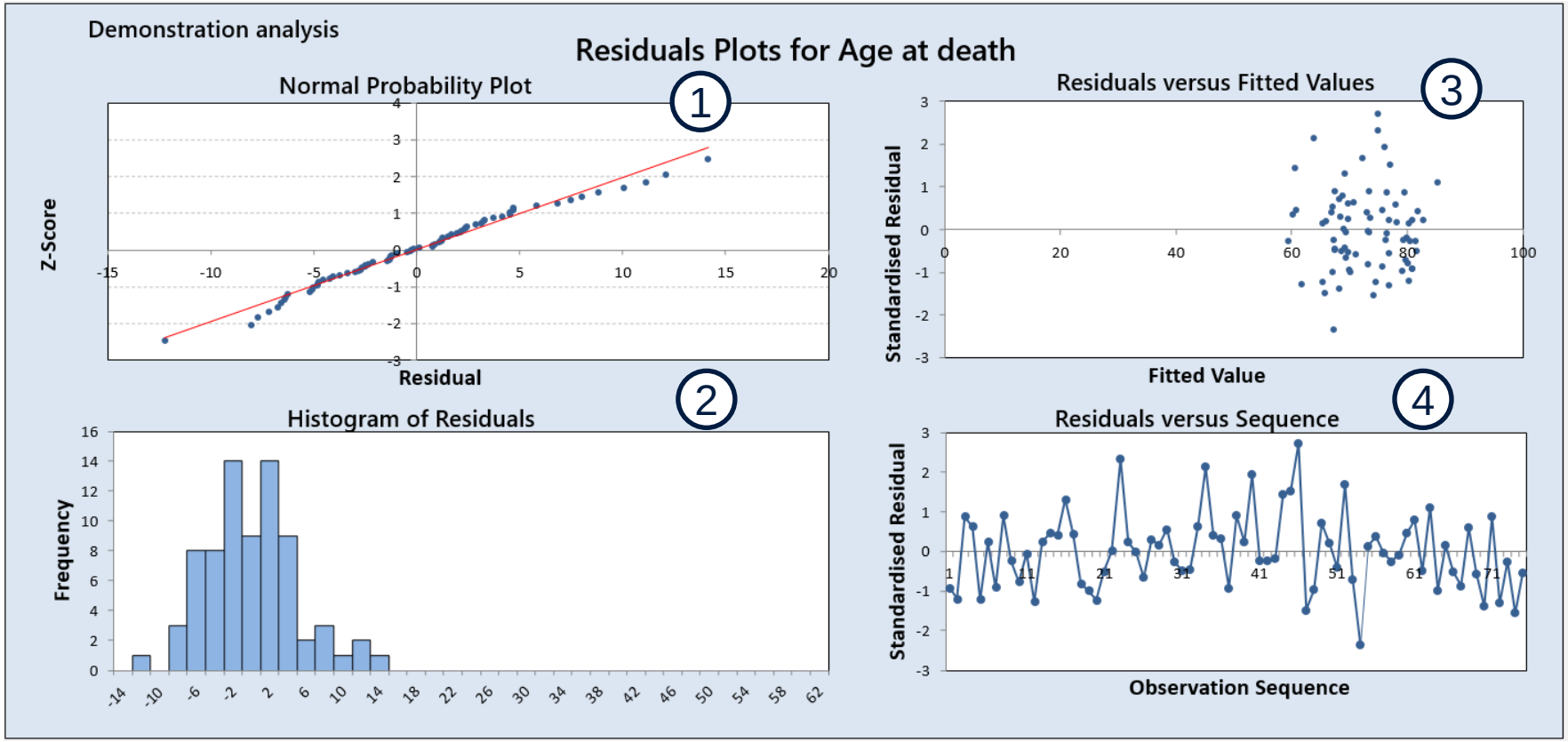
- This will have a bell shape if the residuals are Normal, and here it looks good

### Residuals versus Fitted Values (3)

- You would like the residual errors to be similar, regardless of the fitted values. This one looks fine.

### Residuals versus sequence (4)

- Nothing particularly stands out here, and in any case there is no time order to the data – it is just a collection of 75 individuals.



# Pareto Charts

## Determining priority areas within categorical data

# Pareto Charts

Pareto Chart

**Pareto Charts can be used to identify the dominant categories, to help with prioritisation.**

We shall perform a simple analysis of sales of financial products at a bank branch

- Delete all the data in the worksheet and copy-paste values the data from the worksheet ‘Bank Sales’
  - Ignore the data in the Pivot Tables, for now

We are looking at individual sales activities in a bank. The data shows:

- The date of the proposal
- The product type
- The customer type, based on occupation
  - ‘Retail’ means a person acting in a private capacity
  - ‘SME’ means a business customer - Small/Medium Enterprise
  - ‘Self-employed’ means a business customer – self-employed
  - ‘Student’ means someone in full-time education
- Whether the bank successfully made a sale or not

| 19 | | Exclusions | Time axis | Variable 1 | Variable 2 | Variable 3 |
|---|---|---|---|---|---|---|
| 20 | Row # | Exclude from Control Charts (only) | Date/Time | Product Type | Customer type | Sale? |
| 21 | 1 | | 30/05/2019 | Accident Cov | Retail | Yes |
| 22 | 2 | | 30/05/2019 | Accident Cov | SME | Yes |
| 23 | 3 | | 30/05/2019 | Accident Cov | Self-employe | No |
| 24 | 4 | | 31/05/2019 | Deposit | Self-employe | Yes |
| 25 | 5 | | 01/06/2019 | Accident Cov | Retail | No |
| 26 | 6 | | 02/06/2019 | Accident Cov | Student | No |
| 27 | 7 | | 02/06/2019 | Accident Cov | Retail | No |
| 28 | 8 | | 03/06/2019 | Deposit | Retail | Yes |
| 29 | 9 | | 03/06/2019 | Deposit | Self-employe | Yes |
| 30 | 10 | | 04/06/2019 | Deposit | Self-employe | No |
| 31 | 11 | | 05/06/2019 | Deposit | Student | No |
| 32 | 12 | | 05/06/2019 | Accident Cov | Student | No |
| 33 | 13 | | 06/06/2019 | Deposit | Retail | No |
| 34 | 14 | | 06/06/2019 | Accident cov | Self-employe | No |
| 35 | 15 | | 06/06/2019 | Deposit | Self-employe | No |
| 36 | 16 | | 06/06/2019 | House insura | Retail | Yes |
| 37 | 17 | | 07/06/2019 | Deposit | Student | No |
| 38 | 18 | | 07/06/2019 | Accident Cov | Retail | No |
| 39 | 19 | | 08/06/2019 | Deposit | Self-employe | Yes |

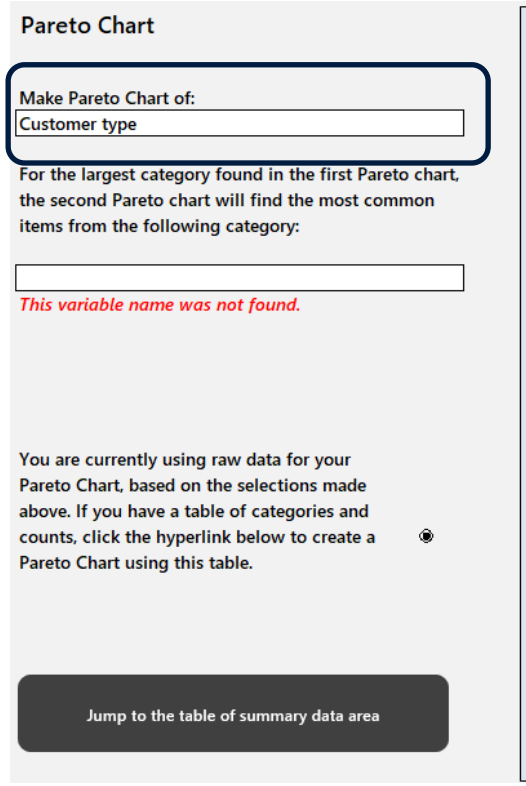
# Pareto Charts

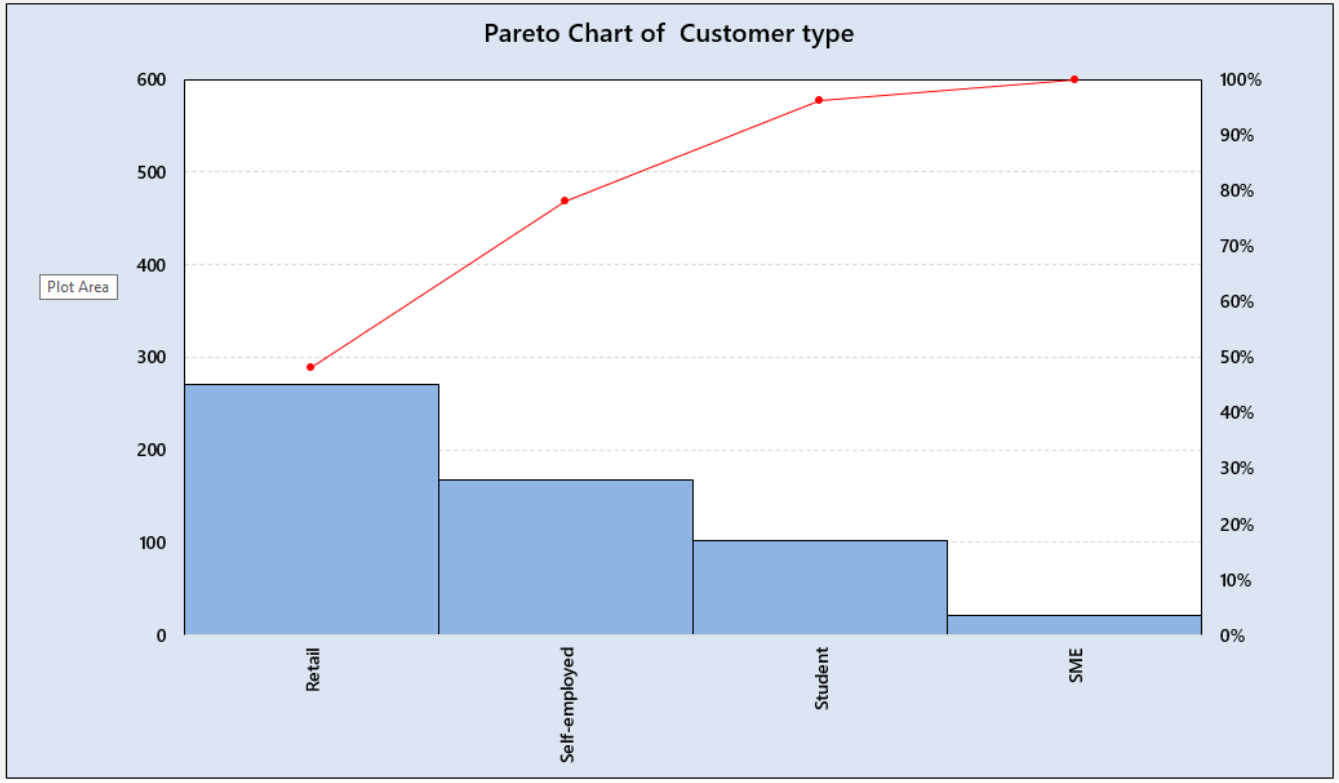
Pareto Chart

**Pareto Charts can be used to identify the dominant categories, to help with prioritisation.**

- Click the Pareto Chart Hyperlink in the normal way to begin

- To create a Pareto Chart, select a categorical factor from the drop-down list.
- Here, we have chosen Customer type, and the Pareto Chart shows that the most common Customer type in our sample is Retail
  - The are 270 of these - almost half our sample



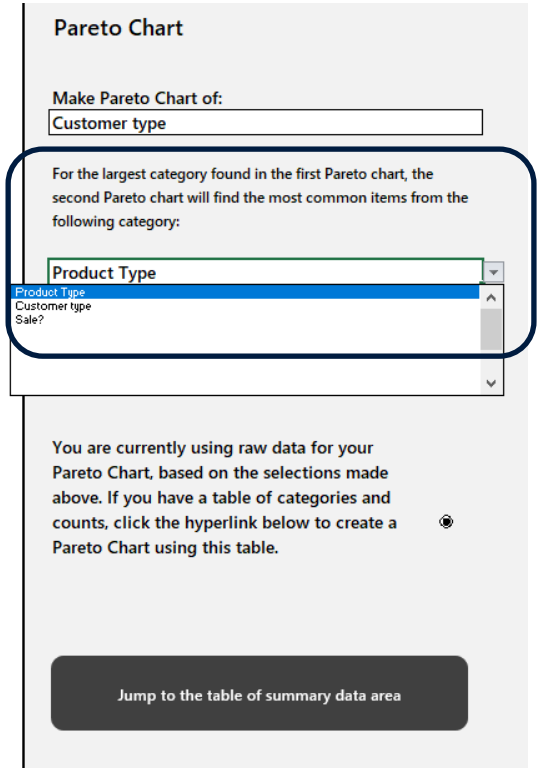



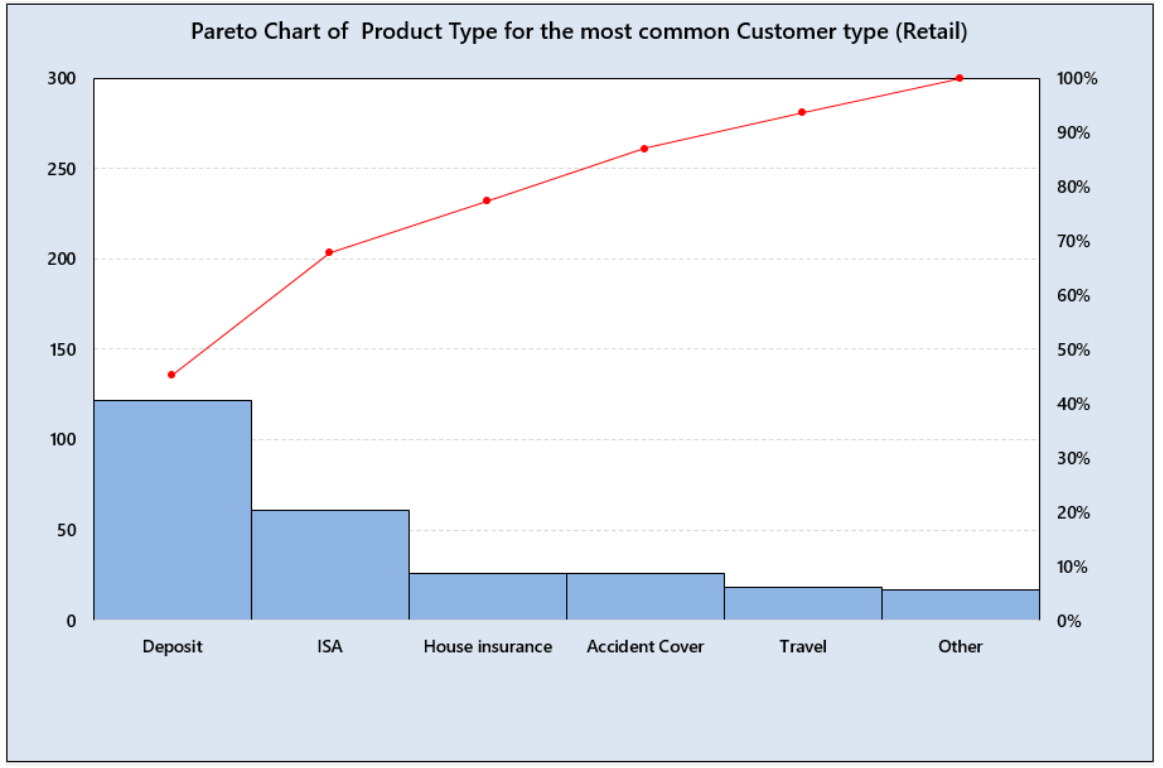
# Pareto Charts

Pareto Chart

## The Toolkit provides you the opportunity to investigate further the largest category (in this case, Retail)

- In the second drop-down box, select the factor that you would like to use to break down the top category
- Let's choose Product type
- Once you have done this, scroll to the right to see the second Pareto Chart
- The Retail customers mainly look for Deposit accounts and ISAs (tax-free Individual Savings Accounts)



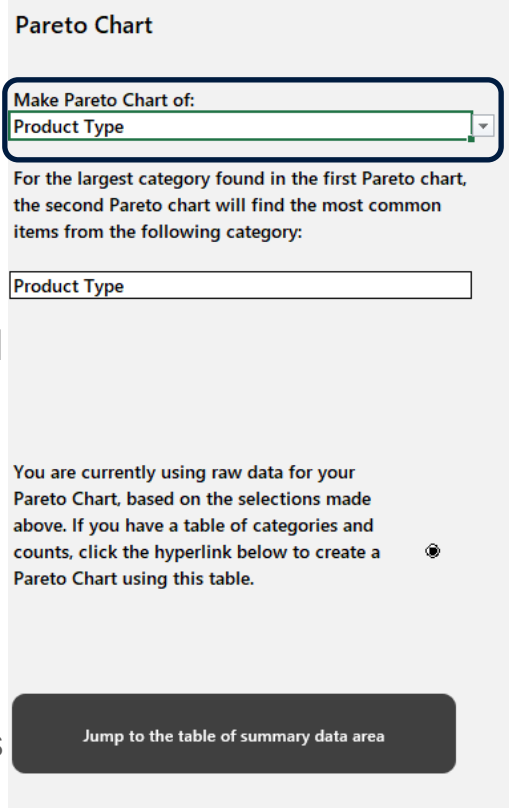



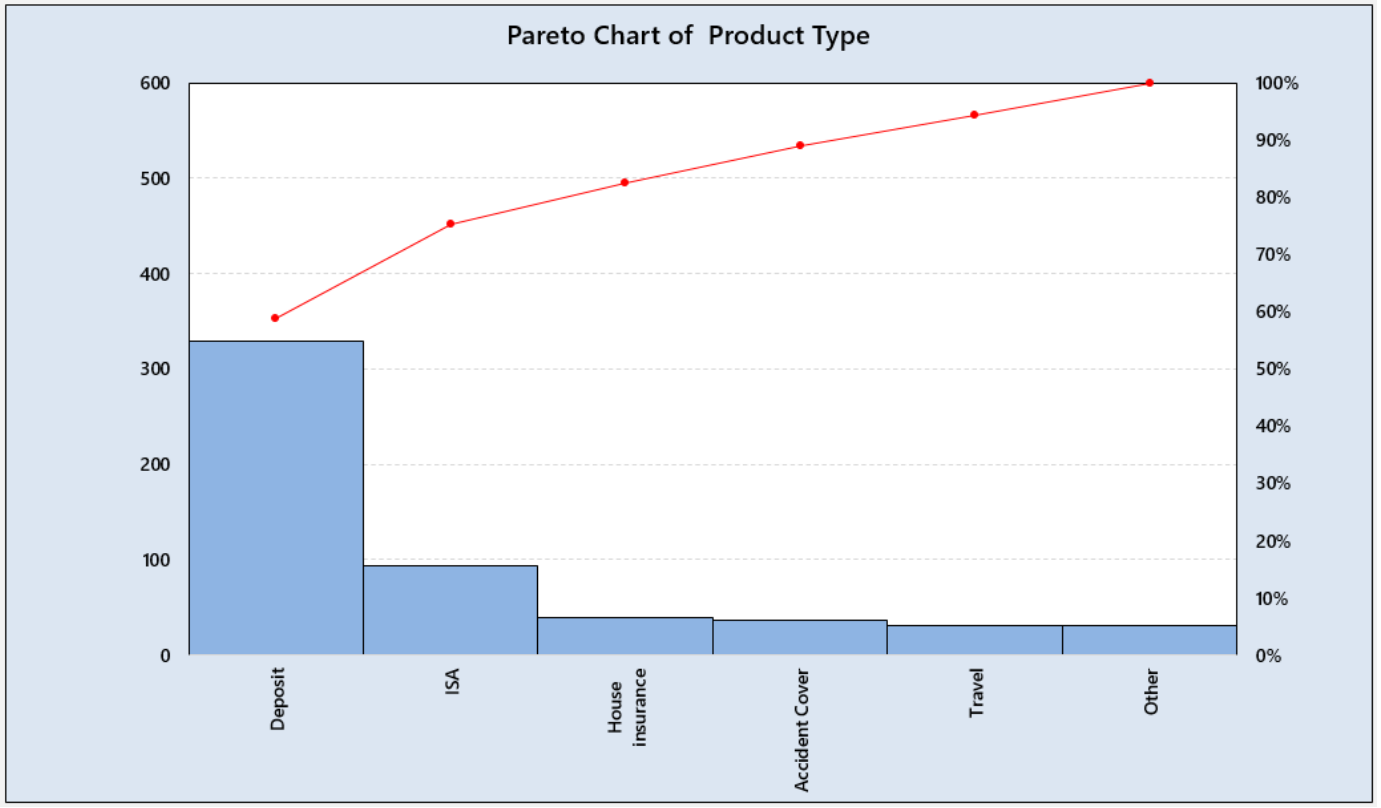
# Pareto Charts

Pareto Chart

## Let's try the analysis again, starting with Product Type

- Go back to the data selection area and make a Pareto Chart of Product type
- Again, Deposit is the top category – but now we can see that, when looking at all the proposals, Deposit really dominates – 59% of all proposals are in this category
  - You can see that figure by hovering your mouse over the red dot on Deposit. The scale for the red line is shown on the right-hand side of the graph.



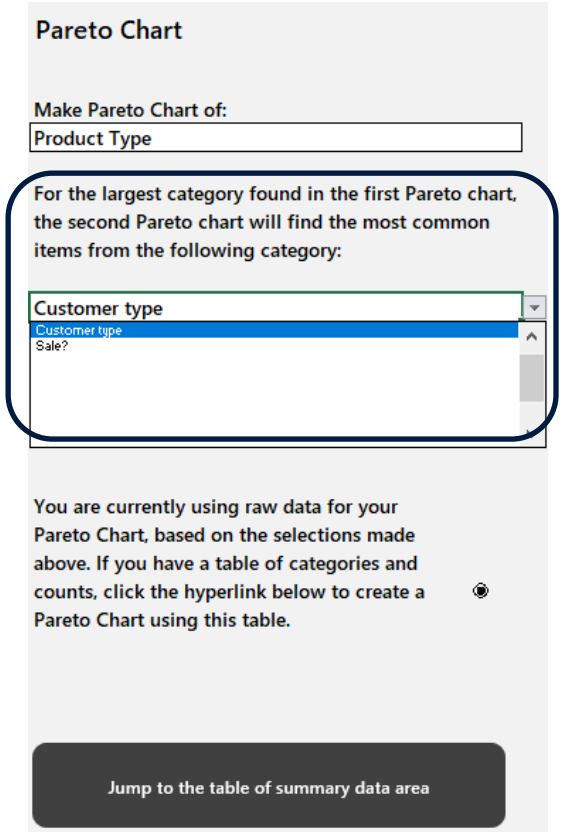



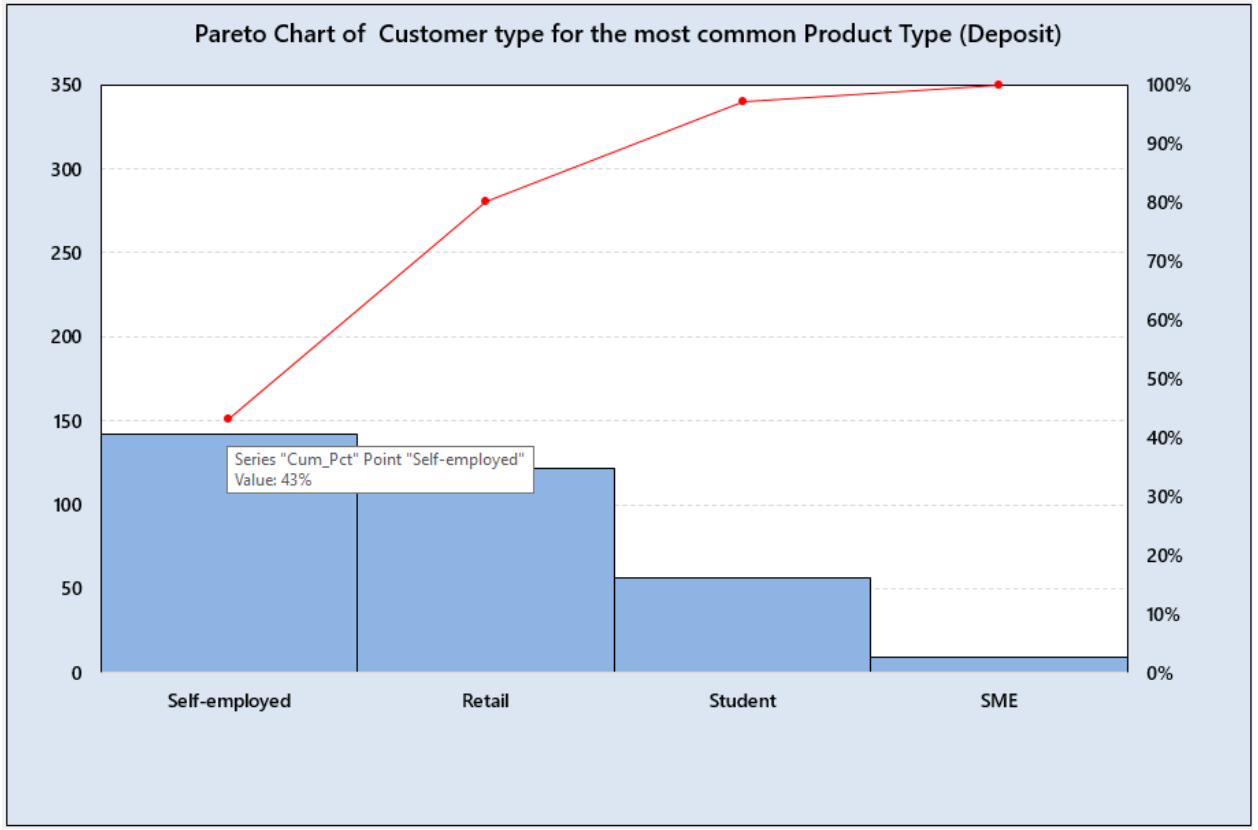
# Pareto Charts

Pareto Chart

## Let's look at who is interested in Deposits

- In the second drop-down box, select Customer type (to change the analysis in the second Pareto Chart)
- Once you have done this, scroll to the right to see the second Pareto Chart
- We can see (by hovering over the red dot) that 43% of prospective customers for Deposits are Self-Employed – so this product has much more appeal to self-employed people than retail investors.



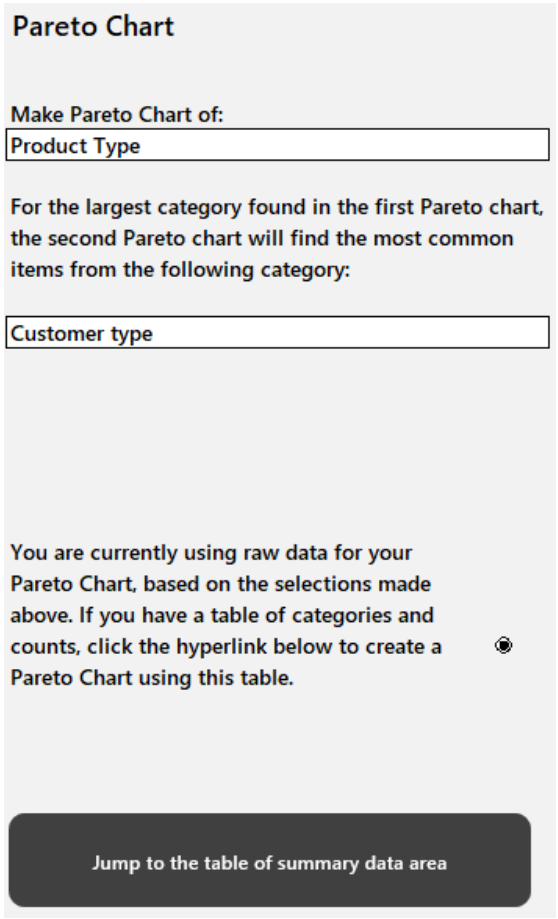



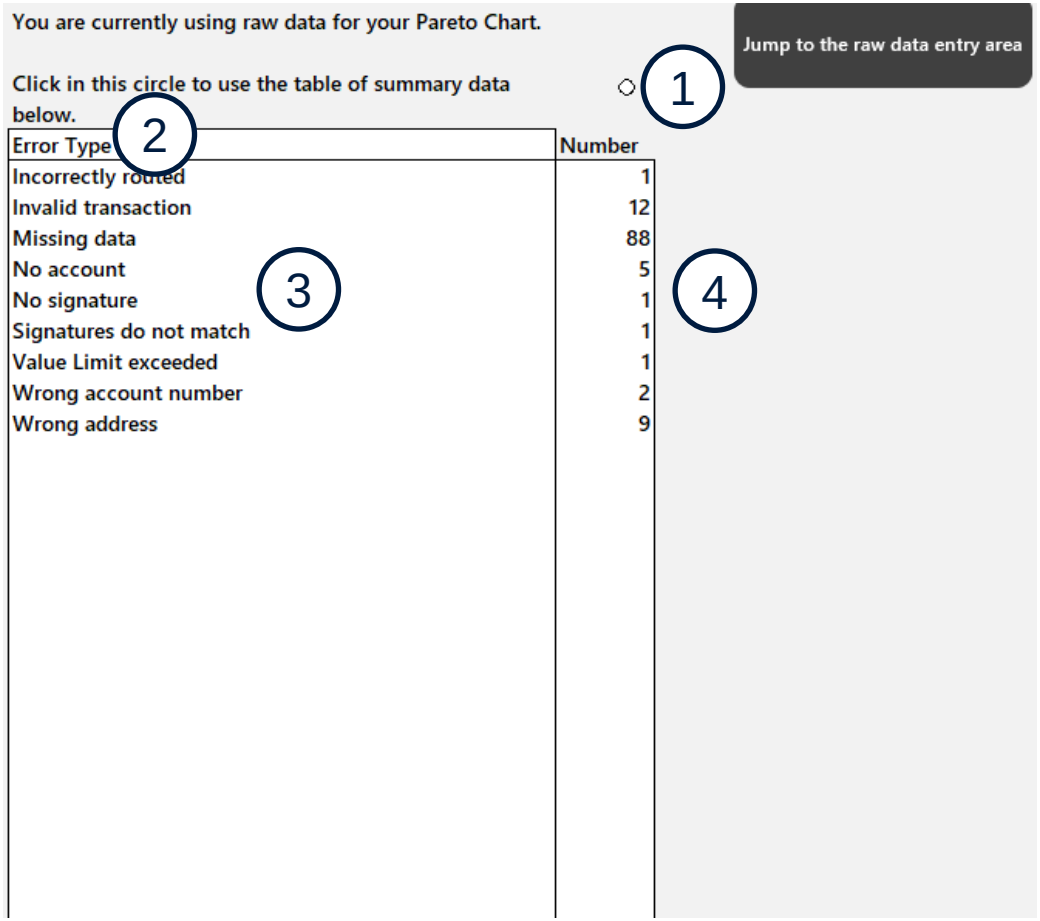
# Pareto Charts for Summary Data

Pareto Chart

## A second format for data entry is available: a table of data

- At the bottom of the data selection area you will see a black hyperlink that will take you to the summary data area. Click this hyperlink now.
- You will arrive at a table of data, positioned to the right of the secondary Pareto Chart.
- To use this table of data, click the radio button on the right (1)
- You can now enter (either directly, or by copying and pasting values from another source):
  - The title of the data set (2)
  - The names of the categories (3)
  - The counts for each category (4)
- A brief aside: this way of entering data might seem easier, but it does not permit more powerful analysis – if your data comes in a table like this, you cannot 'drill down' into secondary categories in the way that we just did for the bank's sales efforts



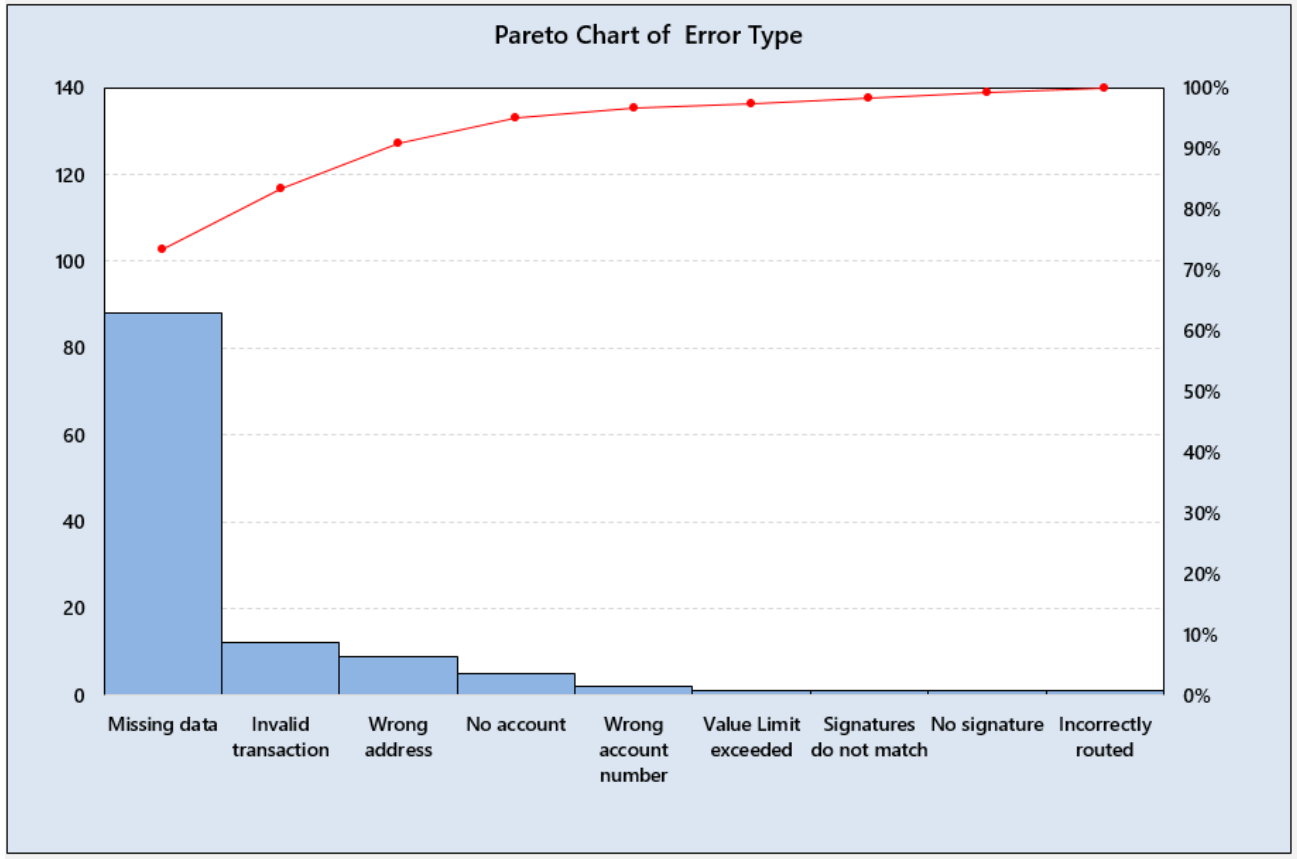



# Pareto Charts for Summary Data

Pareto Chart

## A second format for data entry is available: a table of data (continued)

- After clicking the radio button to select the table of data, you will see that the Pareto Charts now use this table instead of the raw data we used earlier
- To change back, click the black hyperlink “Jump to the raw data entry area” and click the radio button for raw data.




Data Entry

You are currently using a table of summary data (below) for the Pareto Chart.
There is no secondary analysis, so both Pareto Charts are the same.

| Error Type | Number |
|---|---|
| Incorrectly routed | 1 |
| Invalid transaction | 12 |
| Missing data | 88 |
| No account | 5 |
| No signature | 1 |
| Signatures do not match | 1 |
| Value Limit exceeded | 1 |
| Wrong account number | 2 |
| Wrong address | 9 |

Jump to the raw data entry area

If you have raw data, click this hyperlink to jump to the raw data section

# Pareto Charts for Summary Data

Pareto Chart

**We'll recreate one of our earlier Pareto Charts using a table of data**

- Start by deleting all the data (including the title) in the existing table of data
- In the Practice Data 'Bank Sales' worksheet, you will see two Pivot Tables
  - If you are not already familiar with Pivot Tables, take some time to learn them – they are a great tool and there are many pages and videos about them on the web.
- We shall use the top Pivot Table for this exercise.
- Copy the row labels (1) and paste-values into the category list in the toolkit
- Copy the totals for each customer type and paste-values into the Number column (2)
- Finally add the title, 'Customer Type' to the table (3).

**Pivot Table for customer type vs product type**

| Count of Custon | Column Labels | | | | | | |
|---|---|---|---|---|---|---|---|
| Row Labels | Accident Cover | Dep | House insura | ISA | Other | Travel | Grand Total |
| Retail | 26 | 122 | 26 | 61 | 17 | 18 | 270 |
| Self-employed | 4 | 142 | 4 | 9 | 5 | 3 | 167 |
| SME | 2 | 9 | 2 | 2 | 4 | 2 | 21 |
| Student | 4 | 56 | 8 | 21 | 5 | 8 | 102 |
| Grand Total | 36 | 329 | 40 | 93 | 31 | 31 | 560 |

(1) (2)

| Customer Type (3) | Number |
|---|---|
| Retail | 270 |
| Self-employed | 167 |
| SME | 21 |
| Student | 102 |

(1) (2)

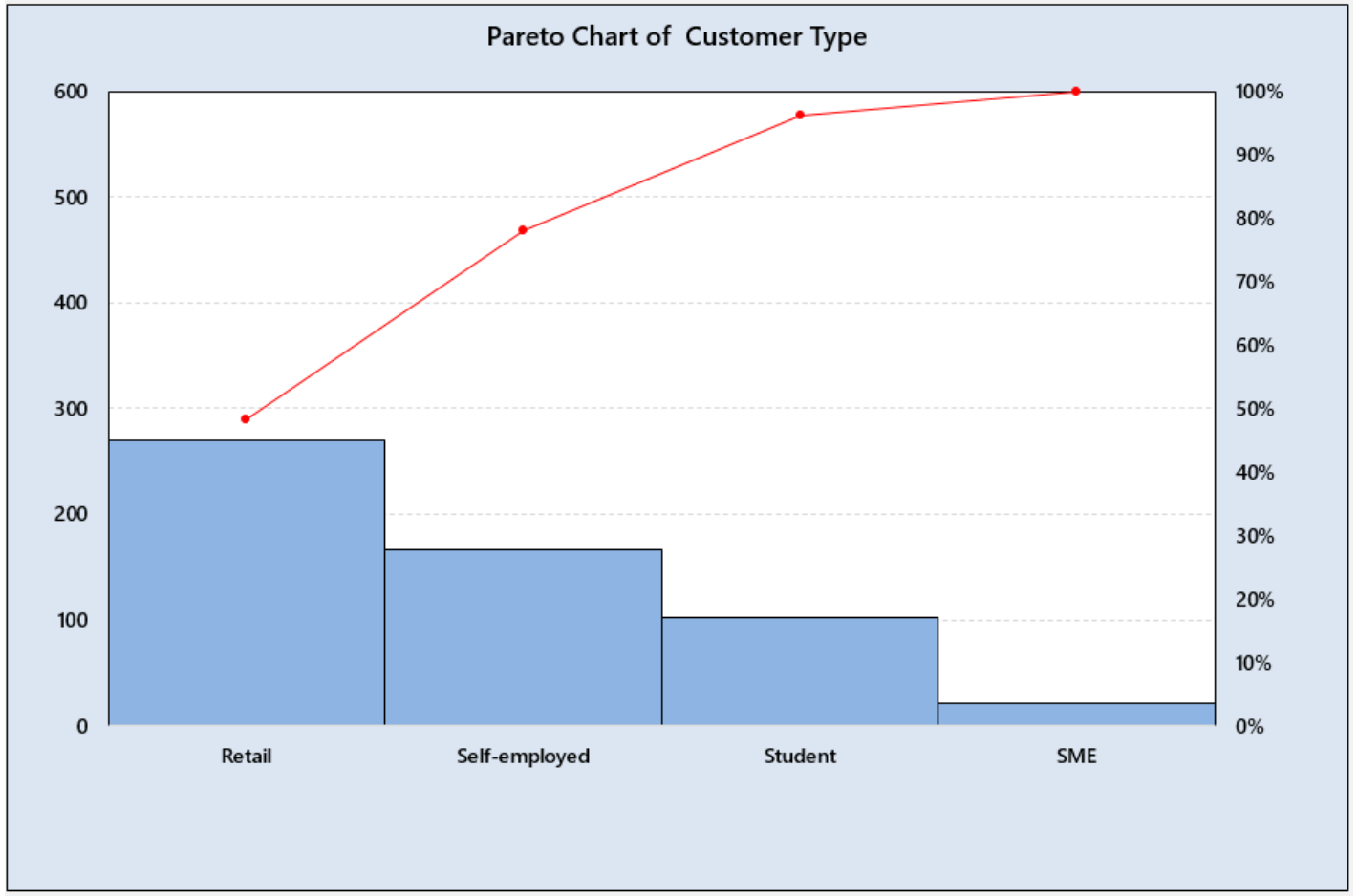
# Pareto Charts for Summary Data

Pareto Chart

**The Pareto made from the table of data is the same as the one originally made from the raw**

- As mentioned, this form of data entry does not permit a secondary analysis, so both Pareto Charts are the same.

| Customer Type | Number |
|---|---|
| Retail | 270 |
| Self-employed | 167 |
| SME | 21 |
| Student | 102 |



# 2-Proportions Test

**A hypothesis test to compare two proportions**

# 2-Proportions test

2-Proportions test

**2-proportions tests compare two proportions and see if it is possible to prove a difference**

Note: as with other hypothesis tests and advanced statistical analysis, this may be beyond the scope of what you have been taught. If you wish to learn about these tests, there are many books and online training courses available – this User Guide assumes you are familiar with the underlying principles.

We shall take the bank sales example from the Pareto Chart section as an illustration of this tool:

- The second pivot table shows the frequency of failure and success of the sales process for each customer type

| Count of Sale? | Column Labels | | |
|---|---|---|---|
| Row Labels | No | Yes | Grand Total |
| Retail | 200 | 70 | 270 |
| Self-employed | 83 | 84 | 167 |
| SME | 17 | 4 | 21 |
| Student | 83 | 19 | 102 |
| **Grand Total** | **383** | **177** | **560** |

As the 2-proportions test can only compare two proportions, we shall focus on the two biggest groups: Retail and Self-employed

The question is: is there a difference in the sales success rate between Retail and Self-Employed?

Click the 2-proportions test hyperlink to start.

# 2-Proportions test

2-Proportions test

## Enter the data from the Pivot Table into the table into the Two Proportions section

You can paste values or type them in directly, as you prefer

| Count of Sale? | Column Labels | | |
|---|---|---|---|
| Row Labels | No | Yes | Grand Total |
| Retail | 200 | 70 | 270 |
| Self-employed | 83 | 84 | 167 |
| SME | 17 | 4 | 21 |
| Student | 83 | 19 | 102 |
| **Grand Total** | **383** | **177** | **560** |

Notice that '#Events' means the number of successful sales (since that is what we are interested in' and '#Trials' means the total number of times we attempted a sale. The calculations show:

- The proportions in each group (sample) (1)
- The difference between the two sample proportions (2)
- The 95% confidence interval for the true difference (3)
- The p-value for the difference in proportions (4)

**Test and Confidence Interval for Two Proportions**

| Category | # Events | # Trials | Sample Proportion |
|---|---|---|---|
| Retail | 70 | 270 | 0.25926 |
| Self-employed | 84 | 167 | 0.50299 |

(1)

**Estimate for Difference**

| Difference (2) | 95% CI for Difference (3) | | P-Value (4) |
|---|---|---|---|
| -0.24373 | -0.3358 | -0.1516 | 0.000 |

Null hypothesis: averages are equal
Significance level: α=0.05

Alternative hypothesis: averages are not equal
Method: Normal approximation

Because the p value is below 0.05, you have proved there is a difference between the two groups

The comment at the bottom confirms that, as the p-value is below 0.05, this can be taken as statistical proof of a difference between the two groups: the proportion of successful sales is higher for self-employed people than for retail investors.

# Chi-Squared Test

**A hypothesis test to compare two or more proportions**

# Chi-squared test

Chi-squared test

**Chi-squared tests compare two or more proportions and see if it is possible to prove a difference**

Note: as with other hypothesis tests and advanced statistical analysis, this may be beyond the scope of what you have been taught. If you wish to learn about these tests, there are many books and online training courses available – this User Guide assumes you are familiar with the underlying principles.

We shall take the bank example from the Pareto Chart section as an illustration of this tool:

- This pivot table shows the frequency of failure and success of the sales process for each customer type

| Count of Sale? | Column Labels | | |
|---|---|---|---|
| Row Labels | No | Yes | Grand Total |
| Retail | 200 | 70 | 270 |
| Self-employed | 83 | 84 | 167 |
| SME | 17 | 4 | 21 |
| Student | 83 | 19 | 102 |
| **Grand Total** | **383** | **177** | **560** |

A Chi-squared test enables us to compare more than two proportions, so we shall look at all four groups together

The question is: is there a difference in the sales success rate between the four types of customers?*

Click the Chi-squared test hyperlink to start.

*admittedly, we have a pretty good idea what we will find, as we already proved a difference between Retail and self-employed… however, this will serve as an illustration of how the tool is used, and will highlight an important difference in the data that is entered.

# Chi-squared test



**Enter the data from the Pivot Table into the table into the Chi-squared analysis section**

You can paste values or type them in directly, as you prefer

| Count of Sale? | Column Labels | | |
|---|---|---|---|
| Row Labels | No | Yes | Grand Total |
| Retail | 200 | 70 | 270 |
| Self-employed | 83 | 84 | 167 |
| SME | 17 | 4 | 21 |
| Student | 83 | 19 | 102 |
| **Grand Total** | **383** | **177** | **560** |

| | | No | Yes |
|---|---|---|---|
| Retail | Observed | 200 | 70 |
| Self-employed | Observed | 83 | 84 |
| SME | Observed | 17 | 4 |
| Student | Observed | 83 | 19 |
| | Observed | | |
| | Observed | | |

Notice that, in contrast to the 2-proportions test, the Chi-squared test requires you to enter the number of cases in each category (here, 'No' and 'Yes' for Retail, Self-employed, SME and Student). Unlike with the 2-proportions test, you should *not* enter the totals – these are calculated automatically from the data you enter.

# Chi-squared test

Chi-squared test

## Here is the full Chi-Squared Analysis

- The data that you have just entered (1)
- The totals across the rows, automatically calculated (2)
- Calculations of the expected values (if all proportions were the same) and Chi-squared statistic for each cell (3)
- Total of the all the Chi-squared statistics (4)
- Based on the Chi-squared statistic and the number of rows and columns, the p-value is calculated

Here, the p-value is below 0.05, so we have (as expected) proved there is a difference between the groups

**Chi-Squared (χ2) Test for Association**

| | | No | Yes | All |
|---|---|---|---|---|
| Retail | Observed | 200 | 70 | 270 |
| Self-employed | Observed | 83 | 84 | 167 |
| SME | Observed | 17 | 4 | 21 |
| Student | Observed | 83 | 19 | 102 |
| | Observed | | | |
| | Observed | | | |
| | Observed | | | |
| | Observed | | | |
| | Observed | | | |
| | Observed | | | |
| | Observed | | | |
| Retail | Expected | 184.7 | 85.3 | |
| | *Chi-square (χ2)* | *1.3* | *2.8* | |
| Self-employed | Expected | 114.2 | 52.8 | |
| | *Chi-square (χ2)* | *8.5* | *18.5* | |
| SME | Expected | 14.4 | 6.6 | |
| | *Chi-square (χ2)* | *0.5* | *1.0* | |
| Student | Expected | 69.8 | 32.2 | |
| | *Chi-square (χ2)* | *2.5* | *5.4* | |
| | Expected | | | |
| | *Chi-square (χ2)* | | | |
| | Expected | | | |
| | *Chi-square (χ2)* | | | |
| | Expected | | | |
| | *Chi-square (χ2)* | | | |
| | Expected | | | |
| | *Chi-square (χ2)* | | | |
| | Expected | | | |
| | *Chi-square (χ2)* | | | |
| | Expected | | | |
| | *Chi-square (χ2)* | | | |
| | Expected | | | |
| | *Chi-square (χ2)* | | | |
| | All | 383 | 177 | 560 |

Data Entry

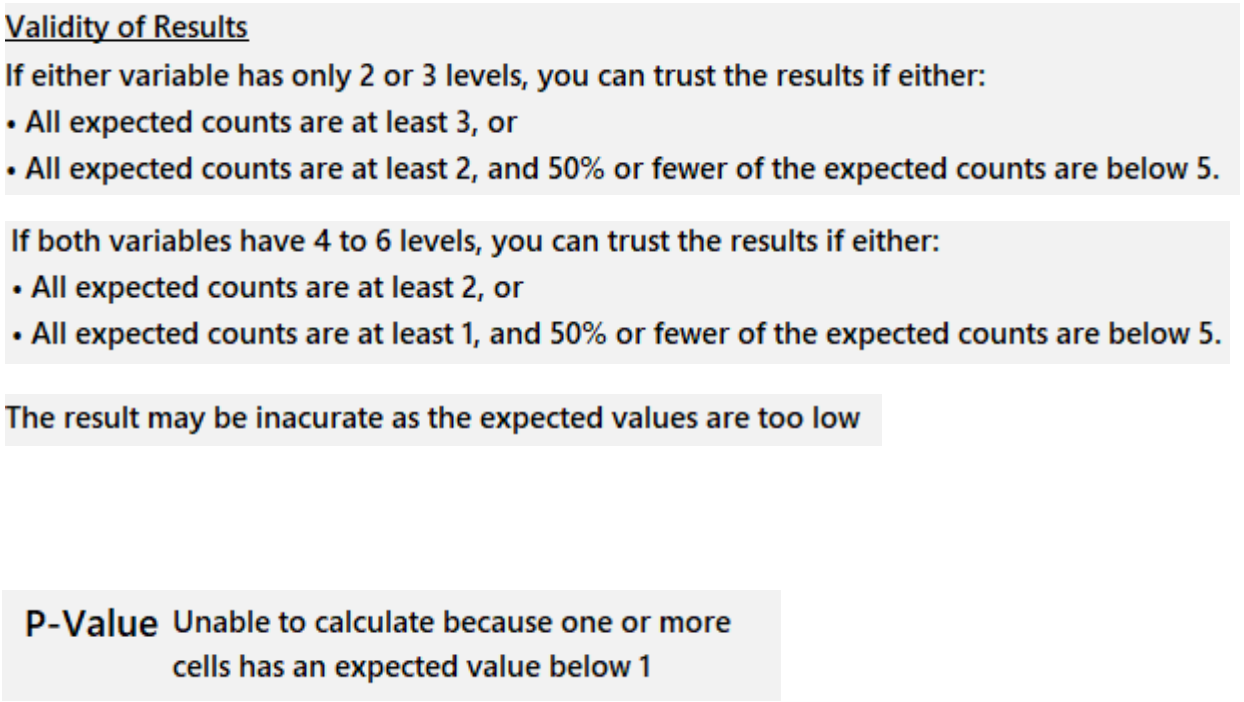
**Total Chi-squared (χ2) value:** 40.51 **P-Value** 0.000

# Chi-squared test

Chi-squared test

## Validity Check

- There are a number of requirements for the sample size which, if not met, reduce the accuracy of the results (or, in extreme cases, prevent the results from being calculated at all).
- If these requirements are not met, one or more of the messages shown here will appear to the right of the data entry table and next to the total Chi-Squared value
- The exact warning messages will depend on the number of factors and the seriousness of the problem incurred.
- The possible actions to take in these cases are:
  - Combine related categories, to avoid having cells with low expected values
  - Remove the categories will low expected values altogether
  - Collect more data



# Graphical and Statistical Analysis - Summary

**This worksheet contains all the analysis tools needed for typical variation reduction projects.**

The graphs can individually be selected with hyperlinks but and are arranged in a logical order for analysis:

Variable Data

- Time-ordered graphs first, to check for special causes
  - Time Series Plots plus I-MR, X Bar-R, C and NP Control Charts
- Histograms, to check the shape of the distribution – with follow-on Process Capability Analysis if needed including data transformation option
- Stratified and multiple column analysis, to determine the effect of a single discrete X on a variable Y
  - Time Series Plots and Box Plots (for both)
  - ANOVA (test for equal variance and test for equal means) is available in the stratified data section
  - 2-sample t-tests and Paired t-tests are available in the multiple data section
- Main Effects Plots and Multi-Vari Charts to determine the effect of up to three discrete Xs on a variable Y at the same time
- Scatter plot, with the option to stratify the data, to look for patterns in the relationship between a variable X and a variable Y
- Matrix Plot, to view the relationships between multiple Xs and a single Y at the same time
- Linear Regression to quantify the relationship between a single X and Y; Multiple Regression to examine several variable Xs at once, including the addition of $X^2$ terms to perform curvilinear regression. Both include residuals analysis.

Discrete Data

- Pareto Chart to enable you to prioritise amongst several categories, and conduct further investigation of the largest one
- 2-proportions test to compare two proportions statistically
- Chi-Squared test to compare more than two proportions statistically

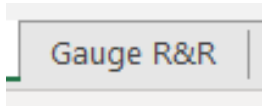
Data can be entered directly into the worksheet – there are no menus to learn. Up to 1000 rows are supported, with up to 12 variables plus Date/Time.

Helpful error messages are provided to politely point out the mistakes that are commonly made, to speed up the learning process.

# Gage Repeatability and Reproducibility

**Checking the precision of a measurement system for continuous data**

# Gage Repeatability and Reproducibility (GR&R)




**Gage R&R is a methodology for checking the precision of a measurement system**

- All measurement systems introduce some variation into the data they are measuring – if this is excessive, the measurement system may become unsuitable for inspection or diagnosing problems
- Gage R&R checks precision but it does not check calibration (which is normally performed by the company that made the measurement instrument

Gage R&R is an advanced topic

- If you wish to learn more about it, please check the training resources available in quality textbooks or on the web – this guide is intended to explain the use of the toolkit to people already familiar with Gage R&R.

Gage R&R has its own worksheet

- Data is pasted in the same way as with the Graphical and Statistical Analysis worksheet

# Gage Repeatability and Reproducibility (GR&R)

Gauge R&R

## Entering data into the Gage R&R Worksheet

- There are two places available for you to enter your data

**Gage R&R**

**Column format for data entry**

| Measurement Study Data | | |
|---|---|---|
| Part | Appraiser | Measurement |
| 6 | 1 | 1.000 |
| 8 | 1 | 0.800 |
| 7 | 1 | 0.950 |
| 5 | 1 | 0.450 |
| 4 | 1 | 0.950 |
| 1 | 1 | 0.600 |
| 10 | 1 | 0.700 |
| 9 | 1 | 1.000 |
| 3 | 1 | 0.800 |
| 2 | 1 | 1.000 |

- The column format area covers cells A21 to C110
- Separate columns are provided for Part, Appraiser and Measurement data

**Table Layout for data entry**

**Provided for compatibility with older GR&R Excel sheets**

| Appraiser | Trial # | Part 1 | Part 2 | Part 3 | Part 4 | Part 5 | Part 6 | Part 7 | Part 8 | Part 9 | Part 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| a | 1 | -0.79 | -1.09 | -0.14 | -1.23 | -0.97 | -0.75 | -0.23 | -1.24 | -0.14 | -0.86 |
| | 2 | -0.64 | -1.10 | -0.18 | -1.10 | -1.05 | -0.70 | -0.27 | -1.17 | -0.17 | -0.93 |
| | 3 | -0.54 | -1.08 | -0.17 | -1.17 | -0.94 | -0.49 | -0.31 | -1.10 | -0.20 | -0.99 |
| b | 1 | -0.80 | -1.21 | -0.14 | -1.23 | -0.97 | -0.75 | -0.23 | -1.24 | -0.14 | -0.86 |
| | 2 | -0.64 | -1.20 | -0.18 | -1.20 | -1.05 | -0.70 | -0.27 | -1.17 | -0.17 | -0.93 |
| | 3 | -0.54 | -1.22 | -0.20 | -1.17 | -0.94 | -0.49 | -0.31 | -1.10 | -0.20 | -0.99 |
| c | 1 | -0.62 | -1.09 | -0.06 | -1.16 | -0.89 | -0.71 | -0.16 | -1.16 | -0.03 | -0.78 |
| | 2 | -0.54 | -1.06 | -0.10 | -1.09 | -0.94 | -0.68 | -0.21 | -1.18 | -0.09 | -0.82 |
| | 3 | -0.53 | -1.10 | -0.14 | -1.13 | -0.87 | -0.51 | -0.24 | -1.09 | -0.11 | -0.91 |

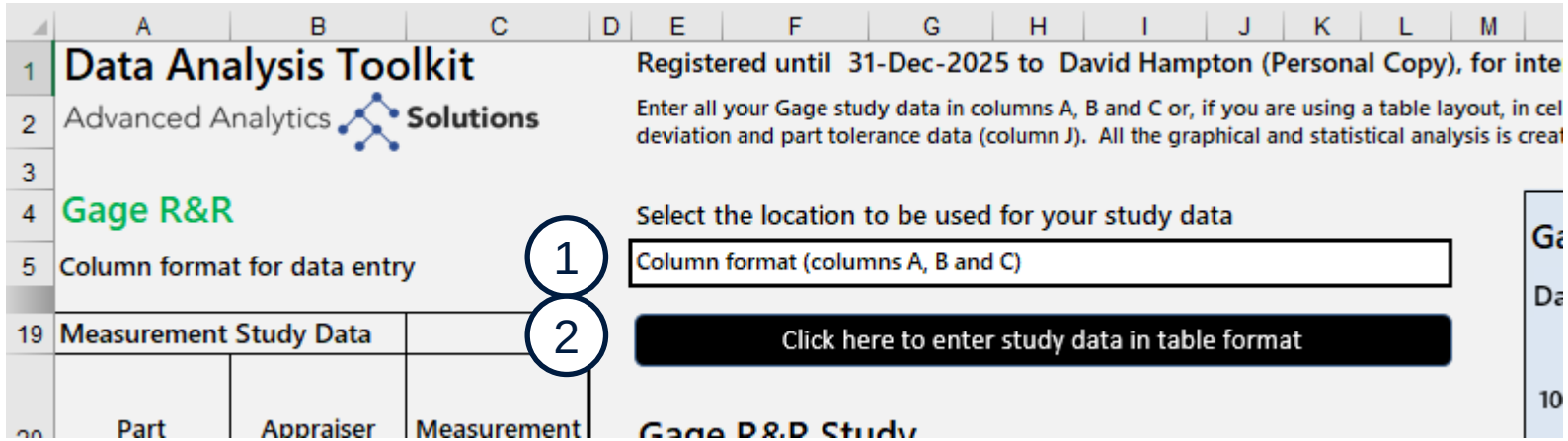
- The table format area covers cells O107 to Z119
- Separate sections are provided for each appraiser with separate columns for each part
- This layout is in common use in Excel-based analysis tools

# Gage Repeatability and Reproducibility (GR&R)

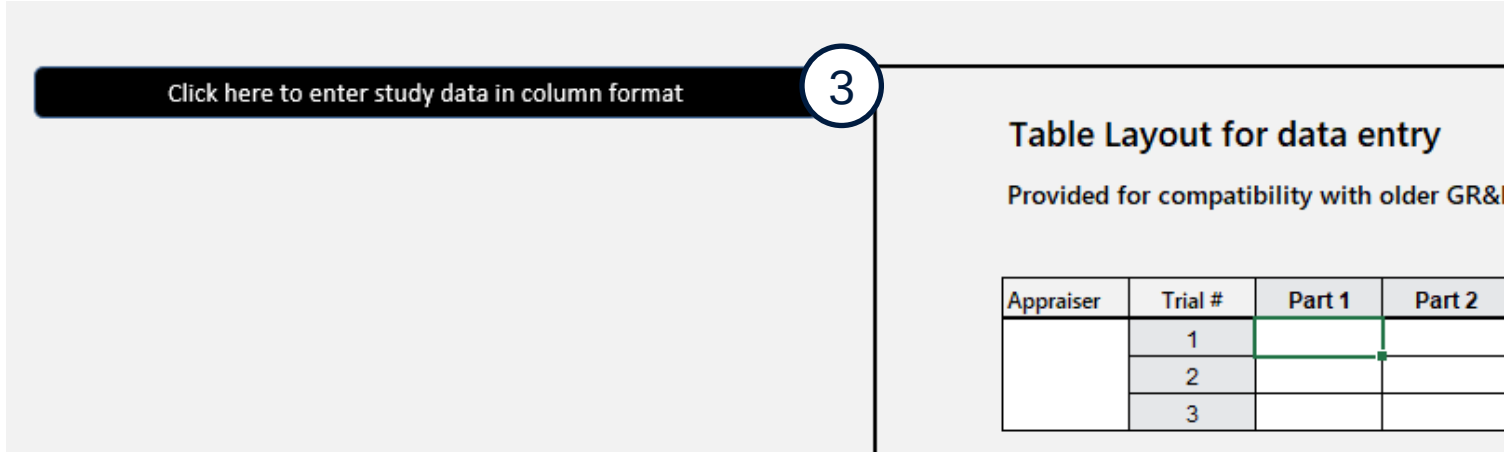


**Only one set of data can be used at once, of course…**

- Use cell E5 to select whether you will use the column format or table format section from the drop-down menu (1).




- Near the column format data entry section, there is a black hyperlink here to jump to the table data entry section (2)




- Another black hyperlink, near the table layout data entry section, will take you back to the column format section.
- So you can easily jump to the place where you will enter your data.

# Gage Repeatability and Reproducibility (GR&R)



**We shall look at the analysis of the data already in the toolkit in the column data section**

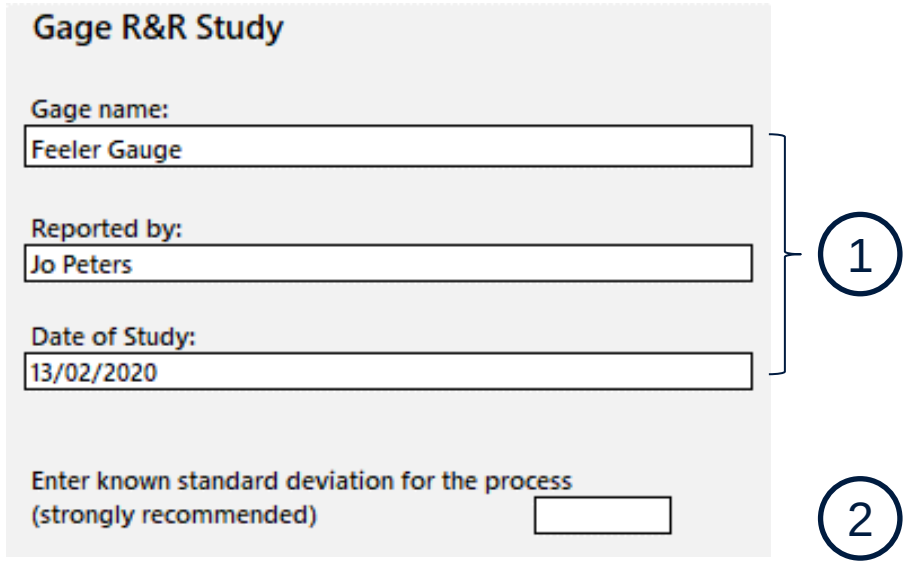
- There are 10 parts with 3 Appraisers (also referred to as Operators) and 2 Repeats

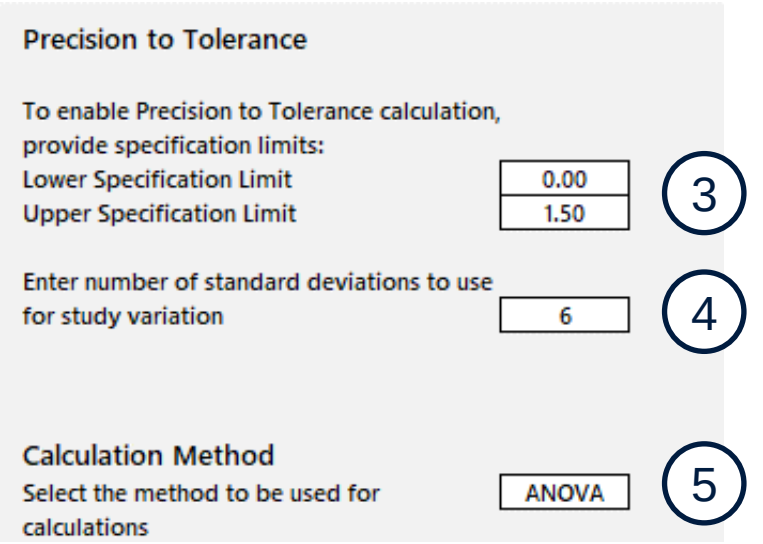
To the right of the data entry section you will find the section where you enter the general information about the study

- The name of the gage
- The person compiling the study
- The date of the study

} Basic administrative information, not used for calculations (1)

- The known standard deviation for the process, if available (2)
  - If this is provided, it will be used for $S_{total}$ – if not, the data from the study will be used
- For calculation of Precision to Tolerance, the upper and lower specification limits (3)
- The number of standard deviations to use for calculation of study variation (4)
  - This should be left at 6 unless you have a specific reason to change it
- The calculation method to be used (5)
  - This should be left as 'ANOVA' unless you have a specific reason to change it



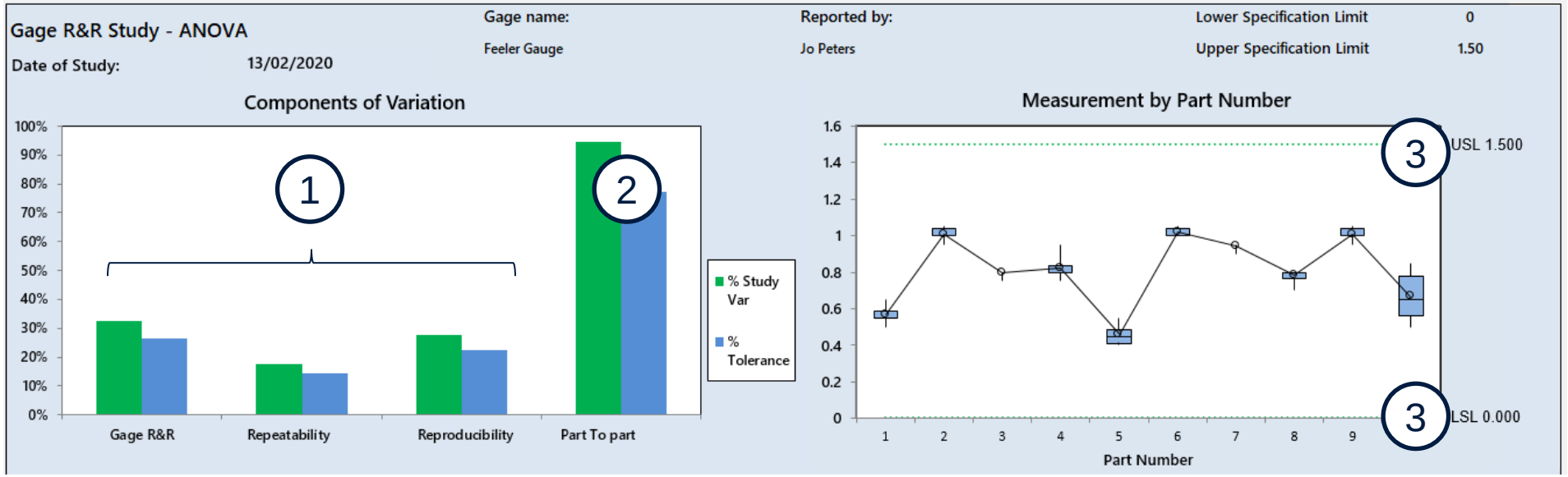



# Gage Repeatability and Reproducibility (GR&R)

Gauge R&R

**There are six graphs in a panel to the right of the data entry section - we shall look at them over the next few slides**

First, the top two graphs which cover information about the study results overall




- The bar char ‘Components of Variation’ shows:
  - Gage R&R (together with its Repeatability and Reproducibility components) (1)
  - Part to Part variation (2)
- The green bars show % Study variation (measurement variation relative to the variation seen in the study); the blue bars show % Tolerance (measurement variation relative to tolerance)
- Note, these are all based on Standard Deviation. Therefore the heights of the bars are not additive.

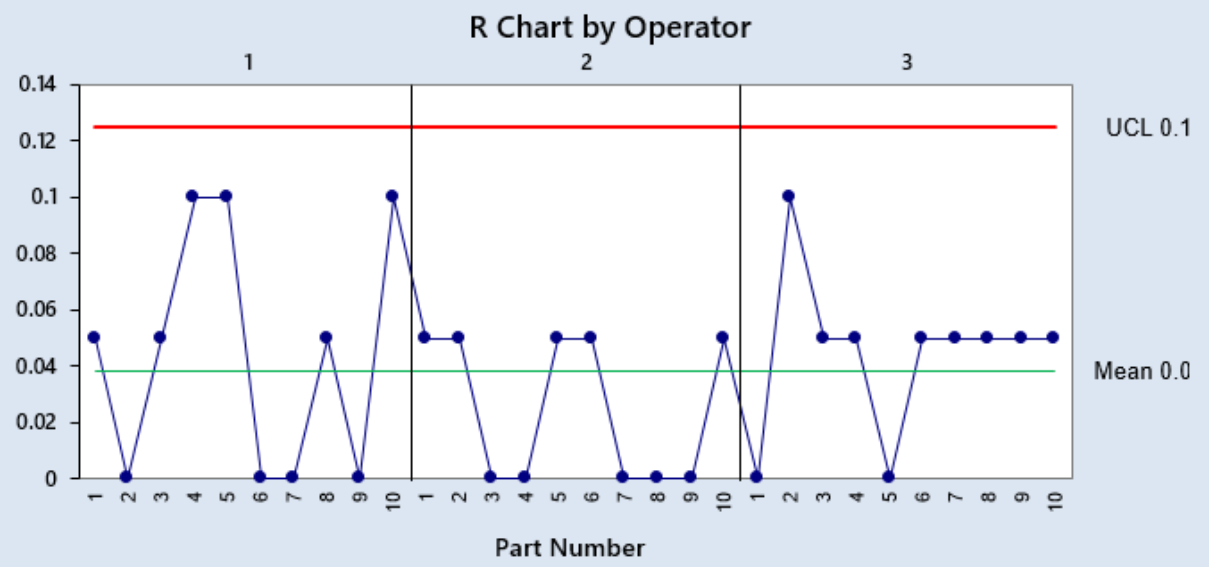
- The box plots ‘Measurements by Part Number’ show:
  - The variation of measurements for each part (represented by the size of the box plots)
  - The variation of mean measurements from part to part (represented by the line joining the box plots)
- The green dotted lines show the upper and lower specification limits (3).
  - Parts for a GR&R study should be selected at random; if any parts are close to the specification it may be a sign of unrepresentative part selection (‘fiddling’ the study)

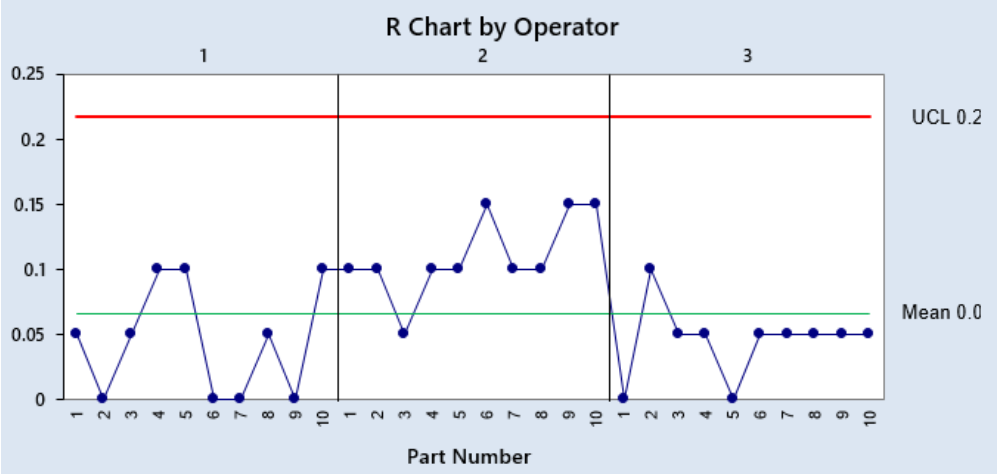
# Gage Repeatability and Reproducibility (GR&R)

Gauge R&R

**The middle graph on the left shows the range of measurements for each part, for each operator**

It is used to graphically represent the repeatability of the measurement system



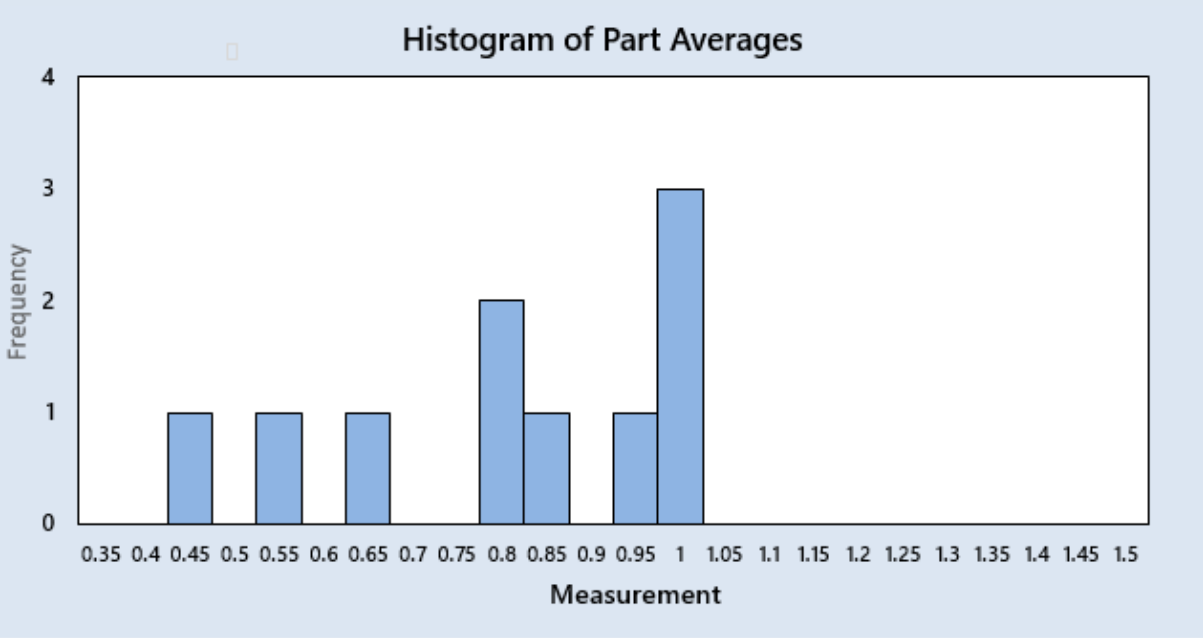




- There is a panel for each of the three operators
- Within each panel, the range is plotted for each of the 10 parts.
  - This is the difference between the smallest and largest measurements that operator made of that part
- The Upper Control Limit (UCL) shows the largest range that can be explained by chance variation, based on the this data
  - If a point goes above the UCL, this may be a sign of a problem in the study – a data entry error for example
- In addition to points outside the Upper Control Limit, you should look to see if there is a difference between the operators
- In the example above, we have altered the R chart for Operator 2, who now has a consistent pattern of higher ranges than the other two operators. Efforts should be made to train this operator as effectively as the other two

# Gage Repeatability and Reproducibility (GR&R)

Gauge R&R

**The bottom graph on the left shows a histogram of the average measurement for each part in the study**




Why have we provided this graph?

- The calculation of Gage R&R compares the measurement system variation to the total variation in the study
- The total variation in the study predominantly comes from the part variation
- Therefore, if the part variation is not representative of the true variation in the process, the GR&R calculations will be wrong
- In particular, if unusually large or small parts are chosen for the study, this will increase the calculated standard deviation for study variation, which will in turn cause an inaccurate (too low) figure for GR&R to be reported. Selection of unrepresentative parts can therefore give a misleadingly good impression of the measurement system.

The Histogram of Part Averages is used to help detect this problem

- If the parts have been selected randomly, they should follow a roughly Normal distribution
- An Anderson Darling test is applied to the distribution of the part averages, and if it shows that the parts are not Normally distributed, a red warning message will be displayed.

This histogram complements the use of specification limits in the top-right graph

# Gage Repeatability and Reproducibility (GR&R)



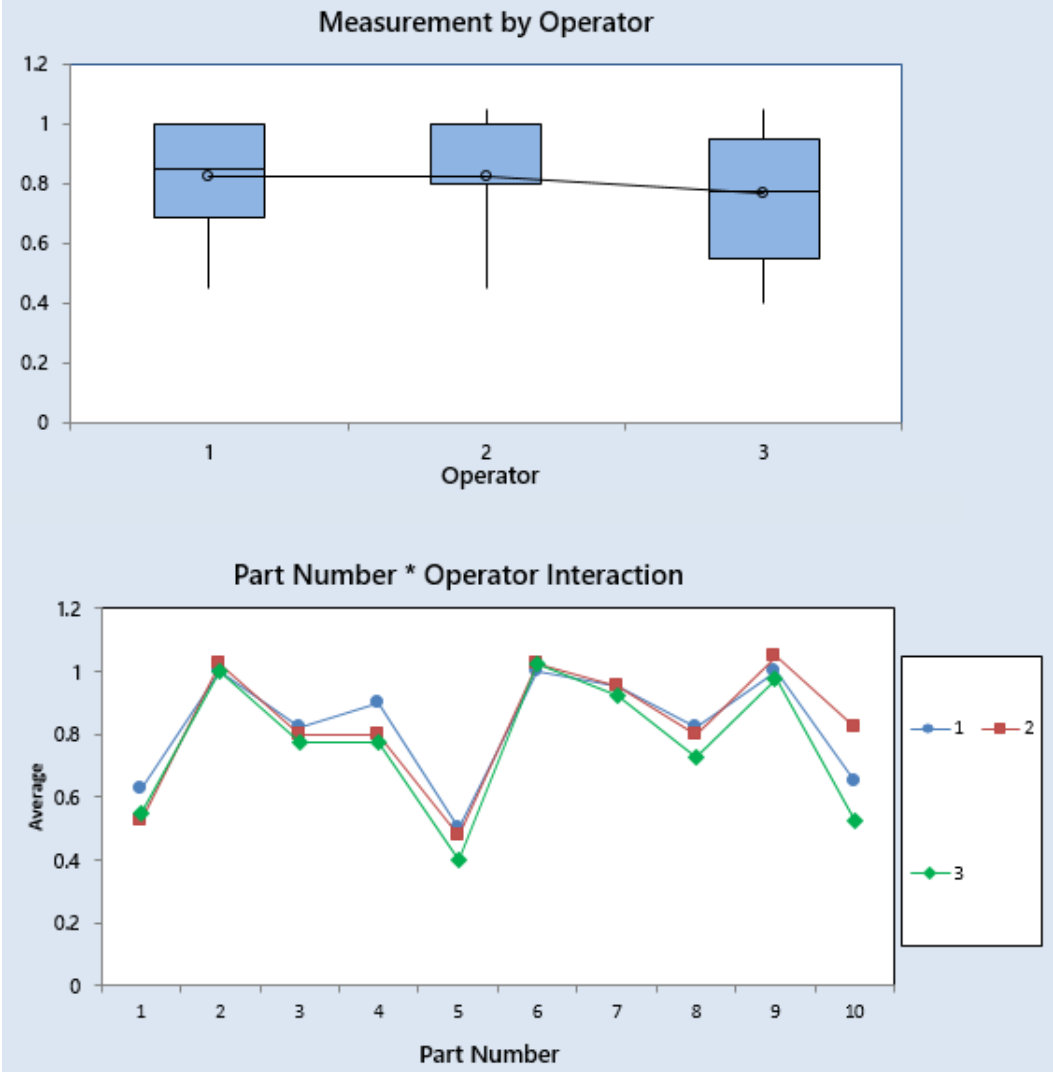
## The final two graphs examine Reproducibility

Measurement by Operator shows the operator-to-operator variation.

- The box plots show the range of variation for each operator's measurements
  - Most of this comes from the part-to-part variation, so you expect this to be large
- The line connecting the means of the box plots show the difference in overall averages of the three operators. You would want these to be the same, so the line should be horizontal. We can see from this graph that operator 2 has a lower mean than the other two, so this should be investigated

Part Number * Operator Interaction shows the operator*part interaction

- The lines should be roughly parallel. In this case, they diverge in a couple of places. This may be a sign of a problem in the way the study was conducted.



# Gage Repeatability and Reproducibility (GR&R)



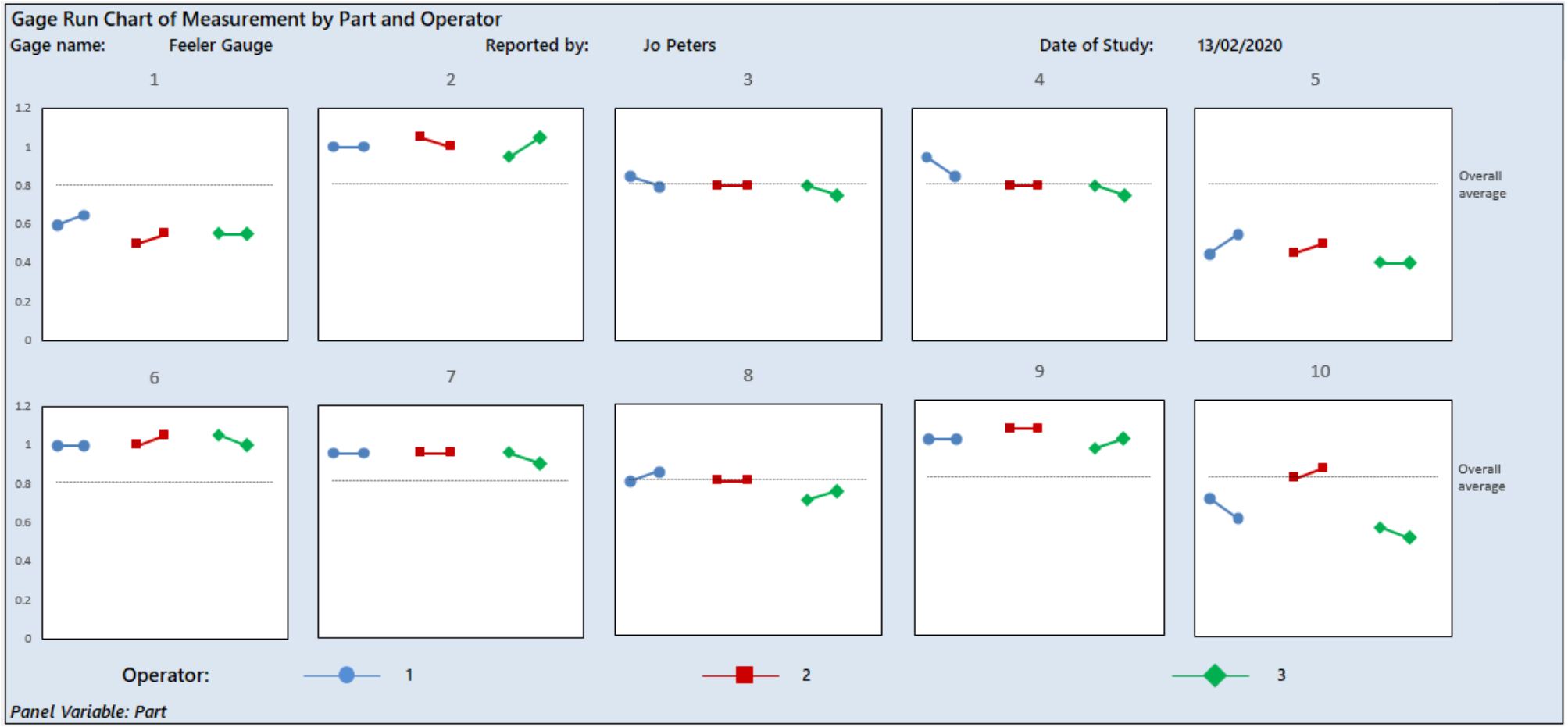
## Beneath the six-graph panel is the Gage Run Chart

This shows the individual measurements – one panel per part

Within each panel we see the repeated measurements of the individual operators

This is the most granular analysis of the part measurements and enables us to see any suspicious patterns.

Everything looks fine here.



# Gage Repeatability and Reproducibility (GR&R)

Gauge R&R

**The full statistical analysis is given to the right of the 6-panel graph**

Beneath some summary information we have the two-way ANOVA table (1)

- The different sources of variation are the parts, the operators, the operator*part interaction and repeatability
- For each of these, the DF (degrees of freedom), sum of squares (SS), mean square (MS), F ratio and P-value is given
  - Of note: the p-value tells you which factors are statistically significant

The first Gage R&R – ANOVA Method table shows the breakdown of study variation into variance components (2)

- Each source is listed along with its variance and its percentage contribution to the total

**Statistical Analysis: Gage R&R - ANOVA Method**

| | |
|---|---|
| Total number of inspections | 60 |
| Overall average | 0.8075 |

**Two-Way ANOVA Table With Interaction** (1)

| Source | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| Part Number | 9 | 2.0587 | 0.22875 | 39.718 | 0.000 |
| Operator | 2 | 0.048 | 0.024 | 4.1672 | 0.033 |
| Part Number * Operator | 18 | 0.10367 | 0.0057593 | 4.4588 | 0.000 |
| Repeatability | 30 | 0.03875 | 0.0012917 | | |
| Total | 59 | 2.2491 | | | |

**Gage R&R - ANOVA Method** (2)

| Source | Variance Component | % Contribution |
|---|---|---|
| Total Gage R&R | 0.00444 | 10.7% |
| Repeatability | 0.00129 | 3.1% |
| Reproducibility | 0.00315 | 7.6% |
| Operator | 0.00091 | 2.2% |
| Part Number * Operator | 0.00223 | 5.4% |
| Part To part | 0.03716 | 89.3% |
| Total Variation | 0.0416 | 100.0% |

# Gage Repeatability and Reproducibility (GR&R)

Gauge R&R

**The next table is based on the standard deviation of the variation sources**

- The standard deviation for each variation source is given (1) together with the study variation (2)
- The study variation as a percentage of the total is given (3) – the top figure in this column is the Gage R&R result, colour-coded based on the normal acceptance criteria
  - <10% - excellent, green
  - 10-30% - acceptable, blue
  - >30% - not acceptable, red
- If the specification limits are given, % Tolerance is quoted and again the top row is colour-coded as this figure is the declared % Tolerance result (4)
  - The acceptance levels and colour key are the same as for GR&R
- If the standard deviation of the process has been given (though not in this case), the GR&R figures are calculated here based on process variation rather than the parts in the study (5)

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Source | StdDev (SD) | Study Var (6 * SD) | %Study Var (%SV) | (Optional) % Tolerance (SV/Tol) | (Optional) % Process (SV/Proc) |
| Total Gage R&R | 0.06662 | 0.39969 | 32.66% | 26.65% | |
| Repeatability | 0.03594 | 0.21564 | 17.6% | 14.4% | |
| Reproducibility | 0.05609 | 0.33653 | 27.5% | 22.4% | |
| Operator | 0.0302 | 0.1812 | 14.8% | 12.1% | |
| Part Number * Operator | 0.04726 | 0.28358 | 23.2% | 18.9% | |
| Part To part | 0.19278 | 1.1567 | 94.5% | 77.1% | |
| Total Variation | 0.20397 | 1.2238 | 100.0% | 81.6% | |

Gage R&R as percentage of Study Variation (traditional approach) 32.66% ☹
This can be inaccurate as it is susceptible to non-random part selection
This is not good enough for process improvement
But see note below about data validity
% Tolerance: Precision/Tolerance Ratio 26.65% 😐
The measurement system is generally acceptable for inspection

For all these metrics: Less than 10% = Excellent (highted in green)
10-30% = Acceptable (highlighted in blue)
>30% = Unacceptable (highlighted in red)

Number of distinct categories = 4

# Gage Repeatability and Reproducibility (GR&R)

Gauge R&R

**Below the table you will find commentary to help you to interpret the data**

- Here you can see a warning message if there is evidence that the parts have not been sampled in a representative way (as indicated by non-Normality) (1)
- The headline Gauge R&R figure is repeated together with an explanation of what this means (2)
- The Precision to Tolerance ratio is also repeated together with an explanation of what this means (3)
- The criteria for acceptability are explained (4)
- The number of distinct categories is given (5) – this gives an intuitive feel for the meaning of Gage R&R (for example, a measurement system with 5 distinct categories could be thought of as being capable of grouping parts into Extra Small, Small, Medium, Large and Extra Large.
  - This must be at least 4, but you should rely on GR&R and Precision to Tolerance as your guide to whether or not the measurement system is acceptable.

| Source | StdDev (SD) | Study Var (6 * SD) | %Study Var (%SV) | (Optional) % Tolerance (SV/Tol) | (Optional) % Process (SV/Proc) |
|---|---|---|---|---|---|
| Total Gage R&R | 0.06661 | 0.3996874 | 32.66% | 26.65% | |
| Repeatability | 0.03594 | 0.2156386 | 17.6% | 14.4% | |
| Reproducibility | 0.05609 | 0.3365264 | 27.5% | 22.4% | |
| Operator | 0.0302 | 0.1811997 | 14.8% | 12.1% | |
| Part Number * Operator | 0.04726 | 0.2835783 | 23.2% | 18.9% | |
| Part To part | 0.19278 | 1.1566834 | 94.5% | 77.1% | |
| Total Variation | 0.20397 | 1.2237919 | 100.0% | 81.6% | |

WARNING: PARTS IN THIS STUDY ARE UNLIKELY TO HAVE BEEN SELECTED RANDOMLY. (1)
STUDY RESULTS ARE THEREFORE PROBABLY INVALID. SEEK HELP.

Gage R&R as percentage of Study Variation (traditional approach) 32.66% ☹
This can be inaccurate as it is susceptible to non-random part selection
This is not good enough for process improvement
But see note below about data validity
(2)

% Tolerance: Precision/Tolerance Ratio 26.65% 😐
The measurement system is generally acceptable for inspection
(3)

For all these metrics: Less than 10% = Excellent (highted in green)
10-30% = Acceptable (highlighted in blue)
>30% = Unacceptable (highlighted in red)
(4)

Number of distinct categories = 4
(5)

# Gage Repeatability and Reproducibility (GR&R)

Gauge R&R

## Below the previous analysis is a final thought about the validity of the data

***Is the data valid for Gage R&R?***

A rough estimate of process capability can be made from the tolerance and the parts in the study. The estimated Cp using the average measurements of these parts is 1.23

If this is much less than expected, it could be that the parts in this study do not truly represent the process variation. If that is the case, investigate the process by which parts were selected. Parts for a GR&R study should represent normal production; the best way to achieve this is to select them randomly.

Comment: typically, the acceptance standard for Cp is 2 or more. This warning appears because the CP is only 1.23.

- The traditional graphical and statistical analysis assumes that parts are representative of the current process and have been selected randomly… but most analysis packages fail to provide any check of this, even though it is critical to the reliability of the study. This means they do not detect manipulation of the GR&R score of someone only cares about achieving a 'pass'
- The Data Analysis Toolkit runs two checks:
  - Are the parts distributed Normally?
    - » We mentioned on the previous slide that red warnings appear when the toolkit detects that the parts are not Normally distributed. This warning is based on the Anderson-Darling Normality test giving a p-value below 0.05.
  - Do the parts exhibit acceptable Process Capability?
    - » This is the point of the text box shown above. If the process capability is poor, that means the part-to-part variation is high and that in turn makes the measurement system look good in comparison. For example, a crude set of supermarket scales that is not good enough to measure the variation in weight of boxes of cornflakes (low variability in weight) might still have a good GR&R score if used to compare pumpkins (high variability in weight)
    - » It's no good having a good GR&R but poor process capability – you need an acceptable GR&R AND an acceptable process capability.

# Gage Repeatability and Reproducibility - Summary

**This worksheet contains sophisticated analysis to evaluate a measurement system for variable data**

- You have the option to enter data in stacked columns or in a traditional Excel table layout
  - Up to 10 parts, 3 operators and 3 repeats are supported
- The graphical analysis features powerful diagnostic tools
  - Components of Variation, which gives a clear visualisation of the relative size of repeatability and reproducibility issues
  - Measurement by Part Number, which directly relates the overall measurement variation to the part variation
    - » This graph also shows how the parts compare to the specification limit, a useful visualisation to confirm that the parts are representative of normal production. The Data Analysis Toolkit is the only GR&R software that provides this information.
  - R Chart by Operator, which enables you to see differences in repeatability between operators and pick out special causes
  - Measurement by Operator, which enables you to see differences in the mean between operators (reproducibility)
  - Part Number*Operator Interaction, which graphically shows potential interactions
  - Histogram of Part Averages, which provides a visualisation of the distribution of the parts. This helps the user to identify parts that may have been 'cherry-picked' to increase the part variation and is another feature that is unique to the Data Analysis Toolkit.
  - A gage run chart to present a highly granular view of the individual measurements made
- A complete set of statistical analysis is also provided, including Gage R&R%, Precision to Tolerance%, %Process variation and all supporting ANOVA tables
  - The analysis is presented together with supporting explanation for easier interpretation
  - A unique feature of the Data Analysis Toolkit, warnings are given is there is evidence that the parts have not been randomly selected.
- Data can be entered directly into the worksheet – there are no menus to learn. Helpful error messages are provided to politely point out the mistakes that are commonly made, to speed up the learning process

# Attribute Agreement Analysis

**Checking the reliability of a discrete data measurement system**

# Attribute Agreement Analysis (AAA)

Attribute Agreement Analysis

**Attribute Agreement Analysis is a methodology for checking the reliability of a measurement system for discrete data.**

- These give measurement results as a category, not a number.  Examples might be:
  - Pass or Fail
  - (for a worrying lump): abscess, cyst, noncancerous lump, tumor
  - Perfect, minor defect, major defect, critical defect
- It is common for Attribute Agreement Analysis to involve a subjective visual assessment, and as a result calibration (“how strict should I be?”) is challenging. For this reason it is normally included in an Attribute Agreement Analysis.

Attribute Agreement Analysis is an advanced topic

- If you wish to learn more about it, please check the training resources available in quality textbooks or on the web – this guide is intended to explain the use of the toolkit to people already familiar with AAA.

Attribute Agreement Analysis has its own worksheet

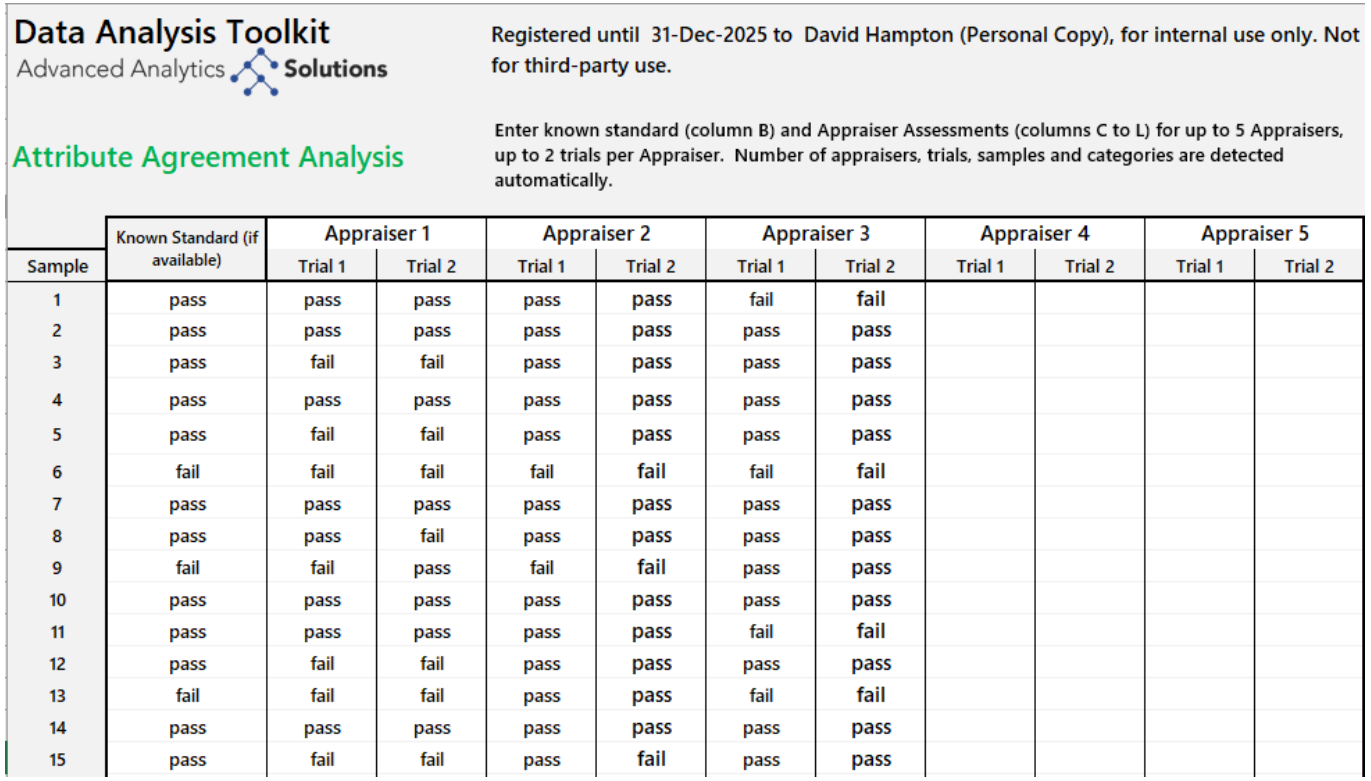
- Data is pasted in the same way as with the Graphical and Statistical Analysis worksheet

# Attribute Agreement Analysis (AAA)



## Entering data into the Attribute Agreement Analysis Worksheet

- The worksheet it configured for data to be entered in unstacked form. If your data is stacked, you can use the worksheet 'Unstack and Stack Data' to convert it.
- Enter the known standard (or expert opinion) in column B as shown with the example data (this is optional)
- Enter the results from up to 5 Appraisers, with one or two trials per appraiser, in columns C to L as shown with the example data.
  - You can enter the names of the appraisers in row 23 if required.
- Please note that the toolkit will handle a maximum of 7 different categories in the appraiser responses (if there are more than 7 you will see an error message)
  - In this case there are only two categories – "pass" and "fail".
  - It is very rare to need more than 7 categories.



Data Analysis Toolkit
Advanced Analytics Solutions

Registered until 31-Dec-2025 to David Hampton (Personal Copy), for internal use only. Not for third-party use.

Attribute Agreement Analysis

Enter known standard (column B) and Appraiser Assessments (columns C to L) for up to 5 Appraisers, up to 2 trials per Appraiser. Number of appraisers, trials, samples and categories are detected automatically.

| | Known Standard (if available) | Appraiser 1 | | Appraiser 2 | | Appraiser 3 | | Appraiser 4 | | Appraiser 5 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Sample | | Trial 1 | Trial 2 | Trial 1 | Trial 2 | Trial 1 | Trial 2 | Trial 1 | Trial 2 | Trial 1 | Trial 2 |
| 1 | pass | pass | pass | pass | pass | fail | fail | | | | |
| 2 | pass | pass | pass | pass | pass | pass | pass | | | | |
| 3 | pass | fail | fail | pass | pass | pass | pass | | | | |
| 4 | pass | pass | pass | pass | pass | pass | pass | | | | |
| 5 | pass | fail | fail | pass | pass | pass | pass | | | | |
| 6 | fail | fail | fail | fail | fail | fail | fail | | | | |
| 7 | pass | pass | pass | pass | pass | pass | pass | | | | |
| 8 | pass | pass | fail | pass | pass | pass | pass | | | | |
| 9 | fail | fail | pass | fail | fail | pass | pass | | | | |
| 10 | pass | pass | pass | pass | pass | pass | pass | | | | |
| 11 | pass | pass | pass | pass | pass | fail | fail | | | | |
| 12 | pass | fail | fail | pass | pass | pass | pass | | | | |
| 13 | fail | fail | fail | pass | pass | fail | fail | | | | |
| 14 | pass | pass | pass | pass | pass | pass | pass | | | | |
| 15 | pass | fail | fail | pass | fail | pass | pass | | | | |

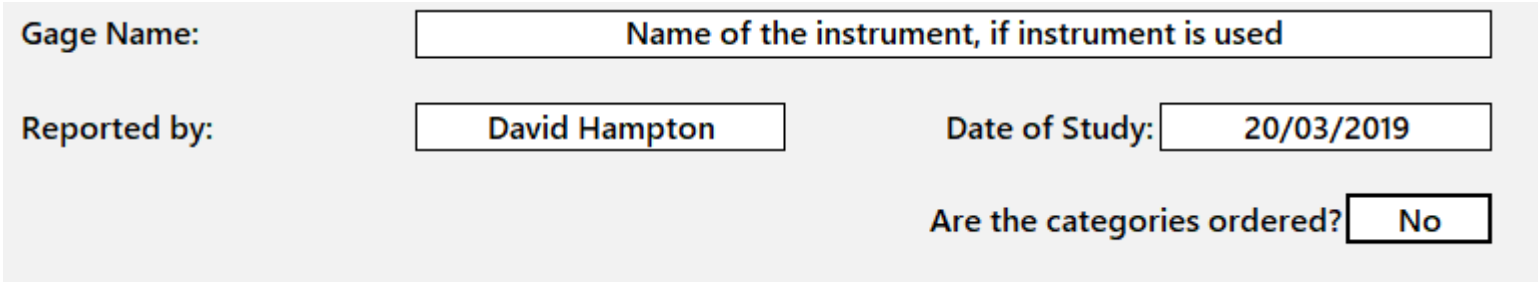
# Attribute Agreement Analysis (AAA)

Attribute Agreement Analysis

Enter general information about the study

This, and the rest of the AAA analysis, can be found to the right of the data entry section

- The name of the measurement instrument/gage, if applicable
- The person compiling the report
- The date of the study
- Are the categories ordered?
  - If the categories have a natural order (eg small-medium-large, A-B-C-D-E, perfect-minor damage-major damage), enter ‘Yes’ in this box
  - The calculations will change from Kappa values to Kendall’s coefficient if ‘Yes’ is selected. The difference is described in more detail here



# Attribute Agreement Analysis (AAA)

Attribute Agreement Analysis

## Within Appraiser analysis

For each appraiser, the table shows:

- The number of parts inspected
- The number matched (the appraiser gave consistent answers across the two trials)
- The percentage match
- The 95% confidence interval for the true percentage matched
- The Kappa value for within appraiser performance (opinions differ as to the thresholds for Kappa, AIAG (the Automotive Industry Action Group, which sets common standards for automotive manufacturers and their suppliers) recommend Kappa>0.75 and ideally >0.9).

The graph on the right shows the percentage agreement and confidence interval graphically.

Gage Name: The gauge being used

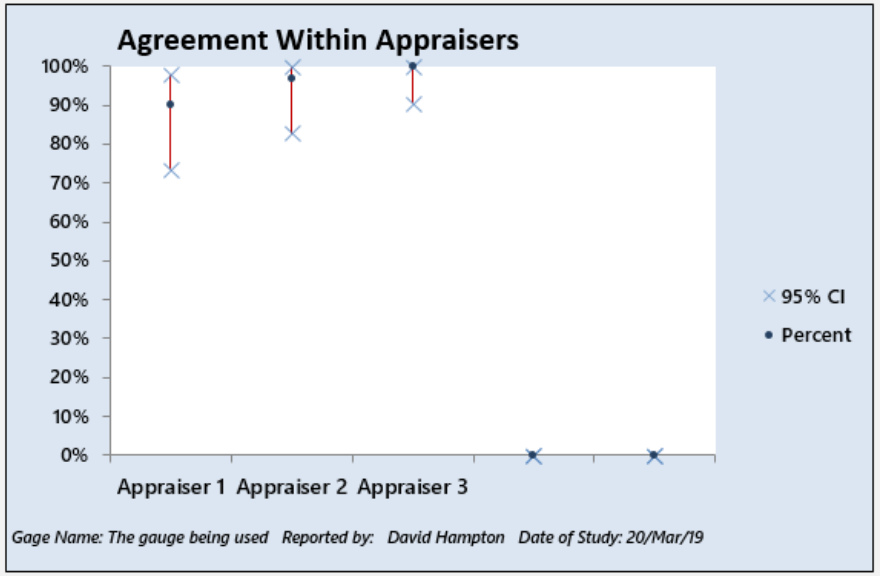
Reported by: David Hampton  Date of Study: 20/03/2019

Are the categories ordered? No

### Within Appraisers

| Appraiser | # Inspected | # Matched | Percent | 95% CI from | 95% CI to | Kappa Value |
|---|---|---|---|---|---|---|
| Appraiser 1 | 30 | 27 | 90.0% | 73.5% | 97.9% | 0.800 |
| Appraiser 2 | 30 | 29 | 96.7% | 82.8% | 99.9% | 0.923 |
| Appraiser 3 | 30 | 30 | 100.0% | 90.5% | 100.0% | 1.000 |

*# Matched: appraiser agrees with him/herself across trials*



# Attribute Agreement Analysis (AAA)



## Each Appraiser versus Standard

For each appraiser, the table shows:

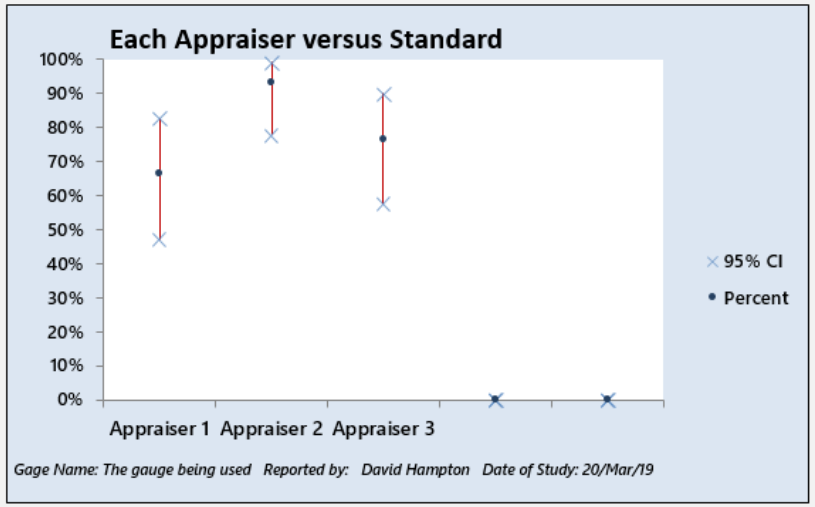
- The number of parts inspected
- The number where the appraiser consistently agreed with the expert/standard
- The percentage match
- The 95% confidence interval for the true percentage matched
- The Kappa value for Appraiser versus Standard

The graph on the right shows the percentage agreement and confidence interval graphically.

Covered in this slide

**Each Appraiser versus Standard**

| Appraiser | # Inspected | # Matched | Percent | 95% CI from | 95% CI to | Kappa Value |
|---|---|---|---|---|---|---|
| Appraiser 1 | 30 | 20 | 66.7% | 47.2% | 82.7% | 0.421 |
| Appraiser 2 | 30 | 28 | 93.3% | 77.9% | 99.2% | 0.886 |
| Appraiser 3 | 30 | 23 | 76.7% | 57.7% | 90.1% | 0.461 |




See next slide

**Assessment Disagreement**

| Appraiser | # pass/fail | Percent | # fail/pass | Percent | # Mixed | Percent |
|---|---|---|---|---|---|---|
| Appraiser 1 | 1 | 10% | 6 | 30% | 3 | 10% |
| Appraiser 2 | 1 | 10% | 0 | 0% | 1 | 3% |
| Appraiser 3 | 4 | 40% | 3 | 15% | 0 | 0% |

*# pass/fail: Assessments across trials = pass / standard = fail*
*# fail/pass: Assessments across trials = fail / standard = pass*
*# Mixed: Assessments across trials are not identical*

# Attribute Agreement Analysis (AAA)



## Assessment Disagreement

This information will only be presented if there are just two categories, which are usually pass/fail, good/bad, OK/NOK etc. It is used to help you see if the problem is that the appraisers are too strict or too lenient. The data shown for each appraiser is:

- The number of times they reported 'pass' for parts which should have been a fail (note, the two category names will depend on your data) and the percentage this represents
- The number of times they reported 'fail' for parts which should have been a pass (or whatever your two categories are) and the percentage
- The number of times they had a mix of the two categories for the same part ('Mixed') and the percentage

See previous slide

Each Appraiser versus Standard

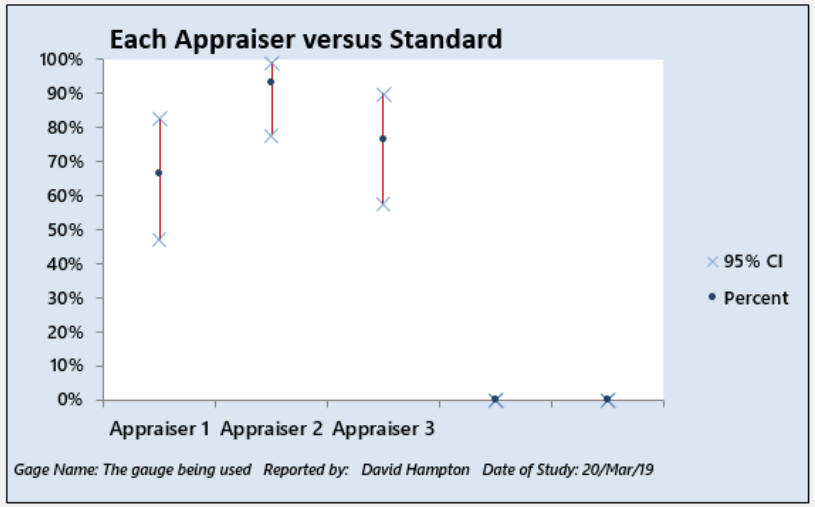
| Appraiser | # Inspected | # Matched | Percent | 95% CI from | 95% CI to | Kappa Value |
|---|---|---|---|---|---|---|
| Appraiser 1 | 30 | 20 | 66.7% | 47.2% | 82.7% | 0.421 |
| Appraiser 2 | 30 | 28 | 93.3% | 77.9% | 99.2% | 0.886 |
| Appraiser 3 | 30 | 23 | 76.7% | 57.7% | 90.1% | 0.461 |

Covered in this slide

**Assessment Disagreement**

| Appraiser | # pass/fail | Percent | # fail/pass | Percent | # Mixed | Percent |
|---|---|---|---|---|---|---|
| Appraiser 1 | 1 | 10% | 6 | 30% | 3 | 10% |
| Appraiser 2 | 1 | 10% | 0 | 0% | 1 | 3% |
| Appraiser 3 | 4 | 40% | 3 | 15% | 0 | 0% |

*# pass/fail: Assessments across trials = pass / standard = fail*
*# fail/pass: Assessments across trials = fail / standard = pass*
*# Mixed: Assessments across trials are not identical*



# Attribute Agreement Analysis (AAA)

Attribute Agreement Analysis

## Between Appraisers

This section shows you agreement between the appraisers (whether or not they agree with the standard)

- All the appraisers have to give consistent answers themselves, and agree with each other, to count as an agreement between appraisers
- The percentage agreement, 95% confidence interval and Kappa values are given as in the previous sections.

**Between Appraisers**

| | | | 95% CI | | |
|---|---|---|---|---|---|
| # Inspected | # Matched | Percent | from | to | Kappa Value |
| 30 | 14 | 46.7% | 28.3% | 65.7% | 0.395 |

# Attribute Agreement Analysis (AAA)

Attribute Agreement Analysis

## All Appraisers versus Standard

This final section provides the overall performance: how often did the assessors all get the right answer for both trials?

- The percentage agreement, 95% confidence interval and Kappa values are given as in the previous sections.

Finally, there is link to an article on acceptance levels for Kappa.

**All Appraisers versus Standard**

| # Inspected | # Matched | Percent | 95% CI from | 95% CI to | Kappa Value |
|---|---|---|---|---|---|
| 30 | 14 | 46.7% | 28.3% | 65.7% | 0.589 |

**Interpreting Kappa Value and Kendall Coefficient**

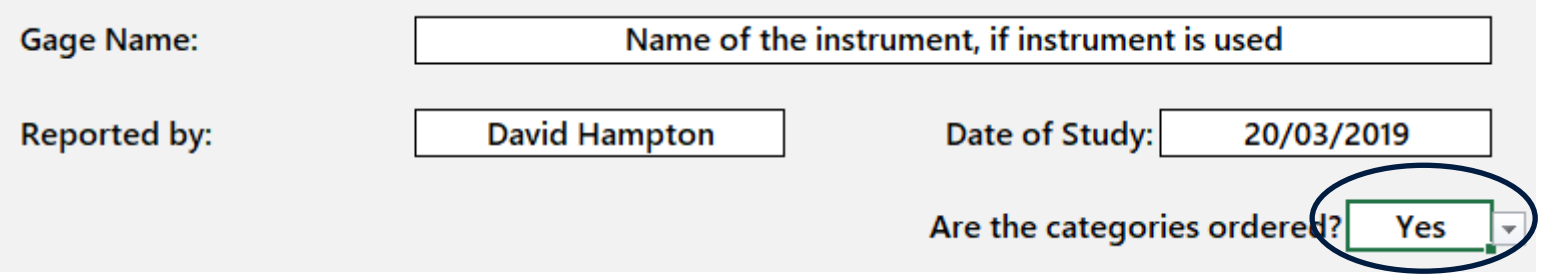
https://support.minitab.com/en-us/minitab/18/help-and-how-to/quality-and-process-improvement/measurement-system-analysis/supporting-topics/attribute-agreement-analysis/kappa-statistics-and-kendall-s-coefficients/

# Attribute Agreement Analysis (AAA)

Attribute Agreement Analysis

## AAA with ordered data

As previously mentioned, in cases where there are more than two different categories and these have a natural order (such as no defect, minor defect, major defect and critical defect), you should select ‘Yes’ in response to the question ‘Are the categories ordered?’.




Kendall’s coefficient does not apply when there are only two categories, so we’ll need a different dataset to show you how it works

- Delete all the data in the data entry area (cells B25 – H54)
- Copy the data from the Practice Data worksheet ‘AAA – ordered’ (cells B3 – H62) and paste the values into the Toolkit, starting at cell B25

# Attribute Agreement Analysis (AAA)

Attribute Agreement Analysis

**AAA with ordered data**

This data set has 6 categories (a, b, c, d, e, f).

- The toolkit will treat the natural order of the categories as being alphabetical, which works in this case.

If you need a different order, give your categories names such as

  - 1 – XS
  - 2 – S
  - 3 – M
  - 4 – L
  - 5 – XL
  - 6 – XXL
- By using a numeric prefix you will ensure that the sequence is interpreted correctly.

Attribute Agreement Analysis

Enter known standard (column B) and Appraiser Assessments (columns C to L) for up to 5 Appraisers, up to 2 trials per Appraiser. Number of appraisers, trials, samples and categories are detected automatically.

| Sample | Known Standard (if available) | Appraiser 1 | | Appraiser 2 | | Appraiser 3 | | Appraiser 4 | | Appraiser 5 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Trial 1 | Trial 2 | Trial 1 | Trial 2 | Trial 1 | Trial 2 | Trial 1 | Trial 2 | Trial 1 | Trial 2 |
| 1 | a | a | b | b | b | b | b | | | | |
| 2 | a | a | a | a | a | a | a | | | | |
| 3 | a | a | a | a | d | a | a | | | | |
| 4 | a | a | d | d | d | d | d | | | | |
| 5 | e | e | c | c | c | c | c | | | | |
| 6 | f | f | f | f | e | f | e | | | | |
| 7 | e | e | e | e | e | e | e | | | | |
| 8 | b | b | a | c | c | c | c | | | | |
| 9 | c | c | c | c | c | c | c | | | | |
| 10 | f | f | f | e | f | e | f | | | | |
| 11 | a | a | b | b | b | b | b | | | | |
| 12 | a | a | a | a | a | a | a | | | | |
| 13 | a | a | e | e | d | e | d | | | | |
| 14 | a | a | d | d | d | d | d | | | | |
| 15 | e | e | c | c | c | c | c | | | | |
| 16 | f | f | f | f | e | f | e | | | | |
| 17 | e | e | e | e | e | e | e | | | | |
| 18 | b | b | a | c | c | c | c | | | | |

# Attribute Agreement Analysis (AAA)

Attribute Agreement Analysis

## AAA with ordered data - Within Appraiser analysis

For each appraiser, the table shows:

- The number of parts inspected
- The number matched (the appraiser gave consistent answers across the two trials)
- The percentage match
- The 95% confidence interval for the true percentage matched
- The Kendall's coefficient for within appraiser performance – this is the only change to the output you will see when the data is not ordered

The graph on the right shows the percentage agreement and confidence interval graphically, as before.

Gage Name: The gauge being used

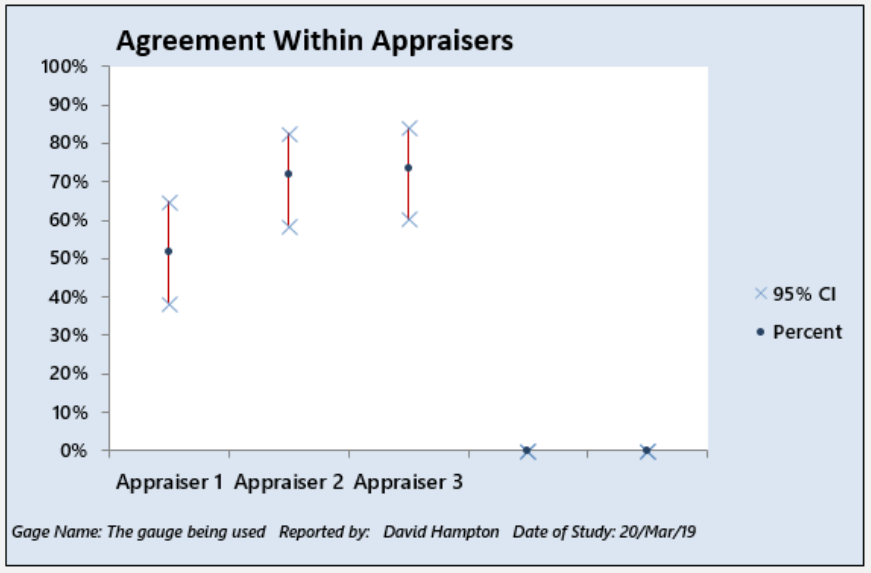
Reported by: David Hampton  Date of Study: 20/03/2019

Are the categories ordered? Yes

**Within Appraisers**

| Appraiser | # Inspected | # Matched | Percent | 95% CI from | 95% CI to | Kendall's Coefficient |
|---|---|---|---|---|---|---|
| Appraiser 1 | 60 | 31 | 51.7% | 38.4% | 64.8% | 0.816 |
| Appraiser 2 | 60 | 43 | 71.7% | 58.6% | 82.5% | 0.960 |
| Appraiser 3 | 60 | 44 | 73.3% | 60.3% | 83.9% | 0.977 |

*# Matched: appraiser agrees with him/herself across trials*



# Attribute Agreement Analysis (AAA)



## AAA with ordered data - Each Appraiser versus Standard

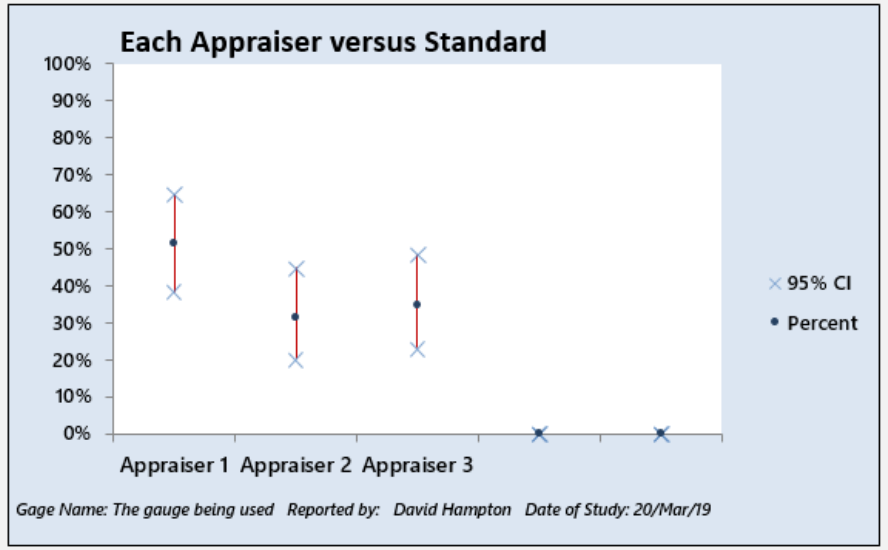
For each appraiser, the table shows:

- The number of parts inspected
- The number where the appraiser consistently agreed with the expert/standard
- The percentage match
- The 95% confidence interval for the true percentage matched
- The Kendall's coefficient for within appraiser performance – this is the only change to the output you will see when the data is not ordered

The remainder is blank as there are more than 2 categories

**Each Appraiser versus Standard**

| Appraiser | # Inspected | # Matched | Percent | 95% CI from | 95% CI to | Kendall's Coefficient |
|---|---|---|---|---|---|---|
| Appraiser 1 | 60 | 31 | 51.7% | 38.4% | 64.8% | 0.770 |
| Appraiser 2 | 60 | 19 | 31.7% | 20.3% | 45.0% | 0.563 |
| Appraiser 3 | 60 | 21 | 35.0% | 23.1% | 48.4% | 0.581 |



# Attribute Agreement Analysis (AAA)

Attribute Agreement Analysis

## AAA with ordered data - Between Appraisers and All Appraisers versus Standard

These sections continue the approach used for non-ordered data, with the Kappa values again being replaced by Kendall's coefficients.

As a reminder, there is a link to a helpful article on acceptable values for Kendall and Kappa. A range of values have been suggested for 'just good enough' for these coefficients – most references consider 0.7 or 0.75 to be the 'pass' mark.

### Between Appraisers

| | | | 95% CI | | Kendall's |
|---|---|---|---|---|---|
| # Inspected | # Matched | Percent | from | to | Coefficient |
| 60 | 18 | 30.0% | 18.8% | 43.2% | 0.866 |

### All Appraisers versus Standard

| | | | 95% CI | | Kendall's |
|---|---|---|---|---|---|
| # Inspected | # Matched | Percent | from | to | Coefficient |
| 60 | 18 | 30.0% | 18.8% | 43.2% | 0.638 |

**Interpreting Kappa Value and Kendall Coefficient**

https://support.minitab.com/en-us/minitab/18/help-and-how-to/quality-and-process-improvement/measurement-system-analysis/supporting-topics/attribute-agreement-analysis/kappa-statistics-and-kendall-s-coefficients/

# Attribute Agreement Analysis - Summary

**This worksheet contains detailed analysis of a discrete data measurement system**

- Within Appraisers
  - Includes supporting graphical analysis to compare percentage agreement and confidence intervals
- Appraisers versus Standard
  - Includes supporting graphical analysis to compare percentage agreement and confidence intervals
  - Including Assessment Disagreement analysis in cases where there are two categories (eg Pass/Fail)
- Between Appraisers
- All appraisers versus Standard

Kappa values are calculated for non-ordered data; Kendall's coefficients are calculated for ordered data.

Up to 250 parts, 5 appraisers, two trials and seven different attribute variations (eg XXS, XS, S, M, L, XL, XXL) can be accommodated.

Data can be entered directly into the worksheet – there are no menus to learn.  Helpful error messages are provided to politely point out the mistakes that are commonly made, to speed up the learning process

# Design of Experiments

**Investigating main effects and interactions in a structured experimental design**

# Design of Experiments (DOE)



**Design of Experiments is a means to simultaneously explore the effect of several Xs on the response**
The structured design also makes it possible to investigate interactions between two or more factors.

Design of Experiments is an advanced topic

- If you wish to learn more about it, please check the training resources available in quality textbooks or on the web – this guide is intended to explain the use of the toolkit to people already familiar with DOE.

Design of Experiments has its own worksheet

- Data is pasted in the same way as with the Graphical and Statistical Analysis worksheet

All experimental designs are $2^k$ full factorial

- That means that the number of runs for each replicate is $2^k$, where k is the number of factors

The worksheet is in three sections, which all work in the same way:

- DOE with 2 factors – rows 20 to 43 – offers 1, 2 or 4 replicates (a maximum of 16 runs)
- DOE with 3 factors – rows 60 to 83 – offers 1 or 2 replicates (a maximum of 16 runs)
- DOE with 4 factors – rows 100 to 129 – offers 1 replicate with 16 runs.

This User Guide will take you through the 3 factor section; the 2-factor and 4-factor sections work in the same way.

# Design of Experiments (DOE)



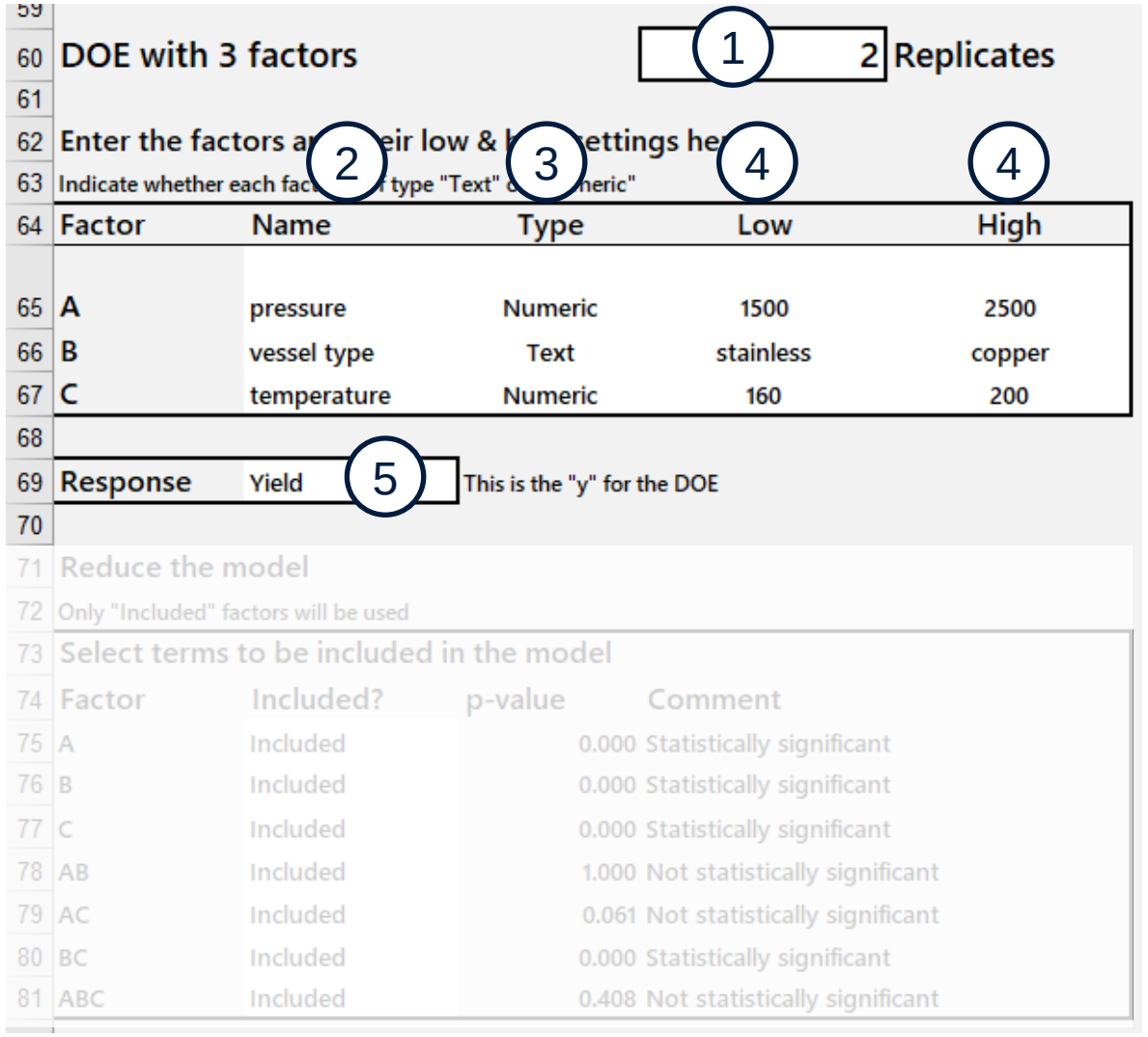
## Design of Experiments with 3 Factors

Scroll down to row 60 to find the 3-factor DOE section.

Experimental design

- Specify the number of replicates you plan to use (1)
- Enter the names of the three factors (2)
- Specify whether the factors are Text or Numeric (3)
  - Your choice will affect the prediction equation in uncoded units
- Enter the High and Low settings for each factor (4)
- Enter the name of the Response (5)

We shall explain this section later



# Design of Experiments (DOE)

DOE

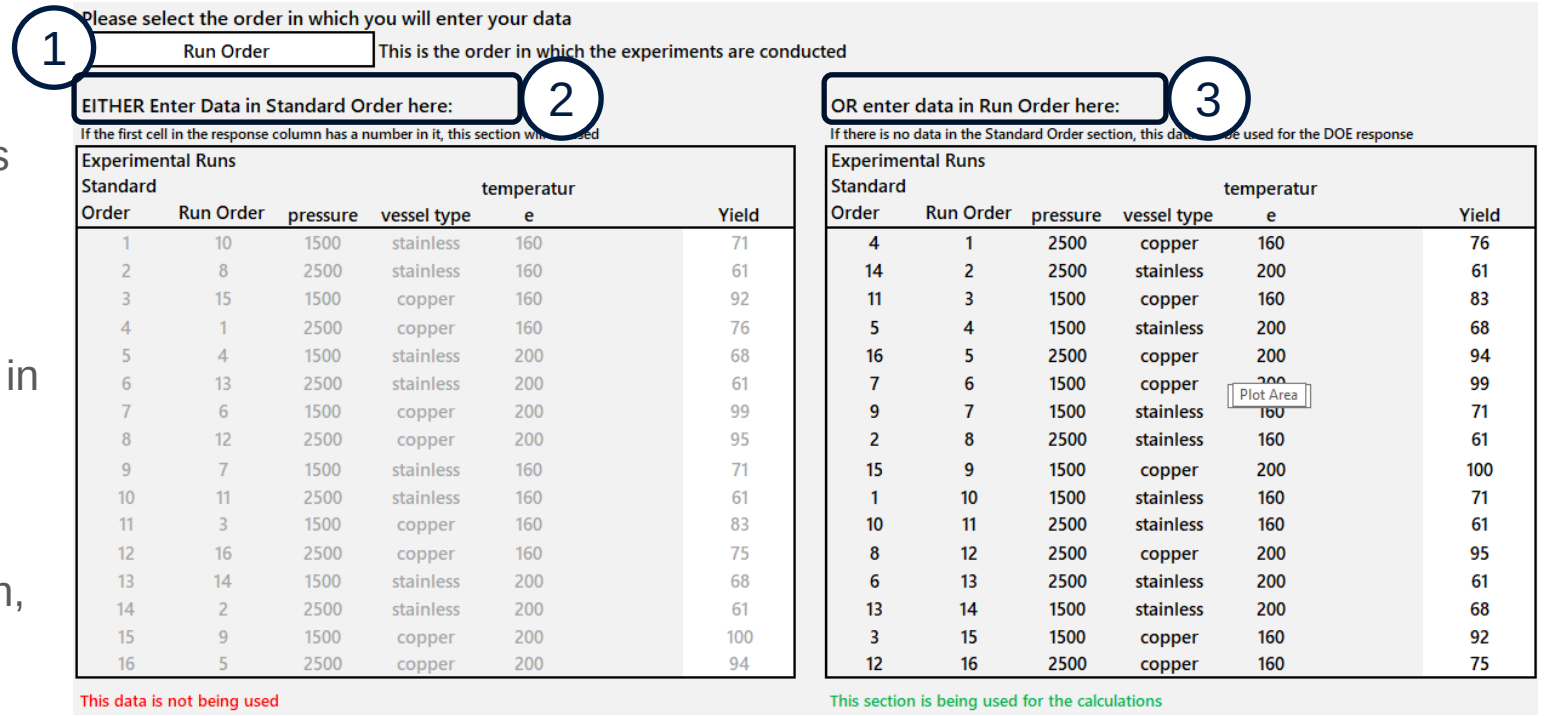
## Design of Experiments with 3 Factors

If you choose one replicate, there will be 8 rows for data entry; here we have chosen two replicates, so there are 16.

In both cases, you have a choice of two places in which to enter the experimental data, and the choice is made at the top of this section (1):

- If you have collected your data in the standardised order of 2k experimental design, select ‘Standard Order’ at (1) and enter the data in the left-hand section (2)
- Typically, however, experimental runs are randomised. If you select ‘Run Order’ at (1), as we have here, you will use the right-hand section, where the Toolkit offers a randomised sequence.

We shall use the demonstration data in the toolkit to explore the means of analysis.




Whichever order you use, the experimental data needs to follow the sequence of experimental settings listed in the appropriate section.

# Design of Experiments (DOE)

## Prediction Equation, Model Summary and Pareto Chart

The Prediction Equation and Model Summary sections give details of the terms that have been used in the model:

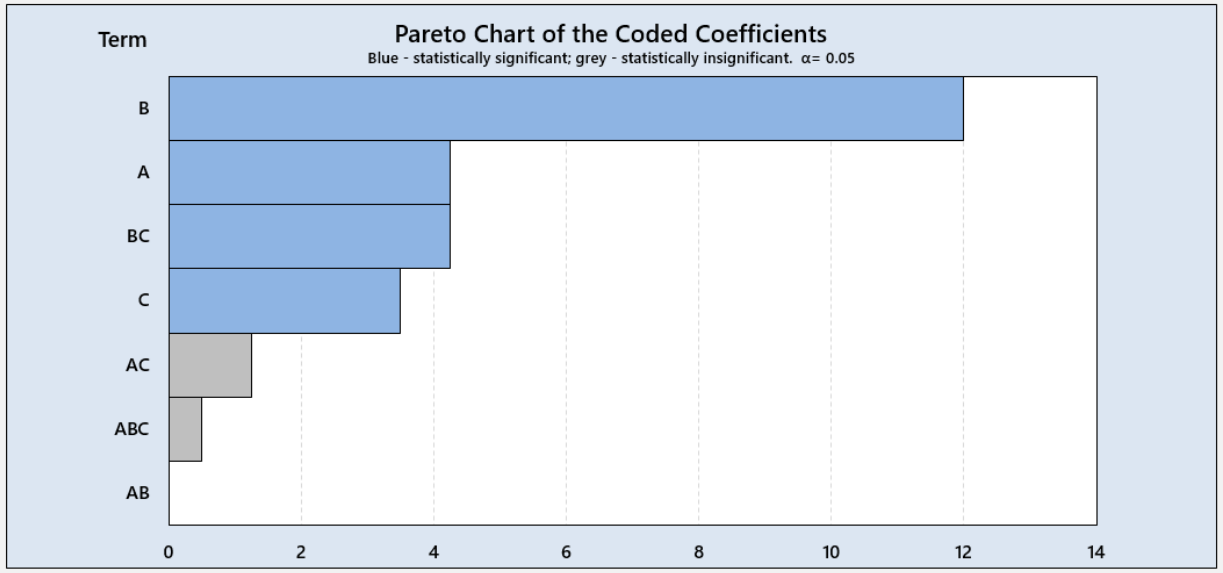
- A list of the terms that have been selected (1)
- The coefficients of these terms in the prediction equation – both with coded units and uncoded units (2)
  - 'Coded units' maps Low settings to -1 and High settings to +1
  - 'Uncoded units' uses the actual numerical values (if the values are Text then -1 and +1 are used instead)
- The p-value for each term (3)
- The standard deviation of the residual errors (4)
- The R-squared and R-squared adjusted measures of correlation (5)

**Prediction equation**

| Term | Model coefficients: Coded units | Uncoded units | p-value |
|---|---|---|---|
| Constant | 77.3 | 108 | 0.000 |
| pressure | -4.25 | -0.031 | 0.000 |
| vessel type | 12 | -8.25 | 0.000 |
| temperature | 3.5 | -0.075 | 0.000 |
| pressure * vessel type | 0 | -0.009 | 1.000 |
| pressure * temperature | 1.25 | 0.000125 | 0.061 |
| vessel type * temperature | 4.25 | 0.112 | 0.000 |
| pressure * vessel type * temperature | 0.5 | 0.00005 | 0.408 |

**Model Summary**

| | |
|---|---|
| s = | 2.291 |
| R-Sq = | 98.7% |
| R-sq (adj) = | 97.5% |




The Pareto Chart shows the relative importance of each term

- Statistically significant terms are coloured blue
- Insignificant terms are grey

It is handy to look at the Pareto when you are reducing the model – which will be described shortly

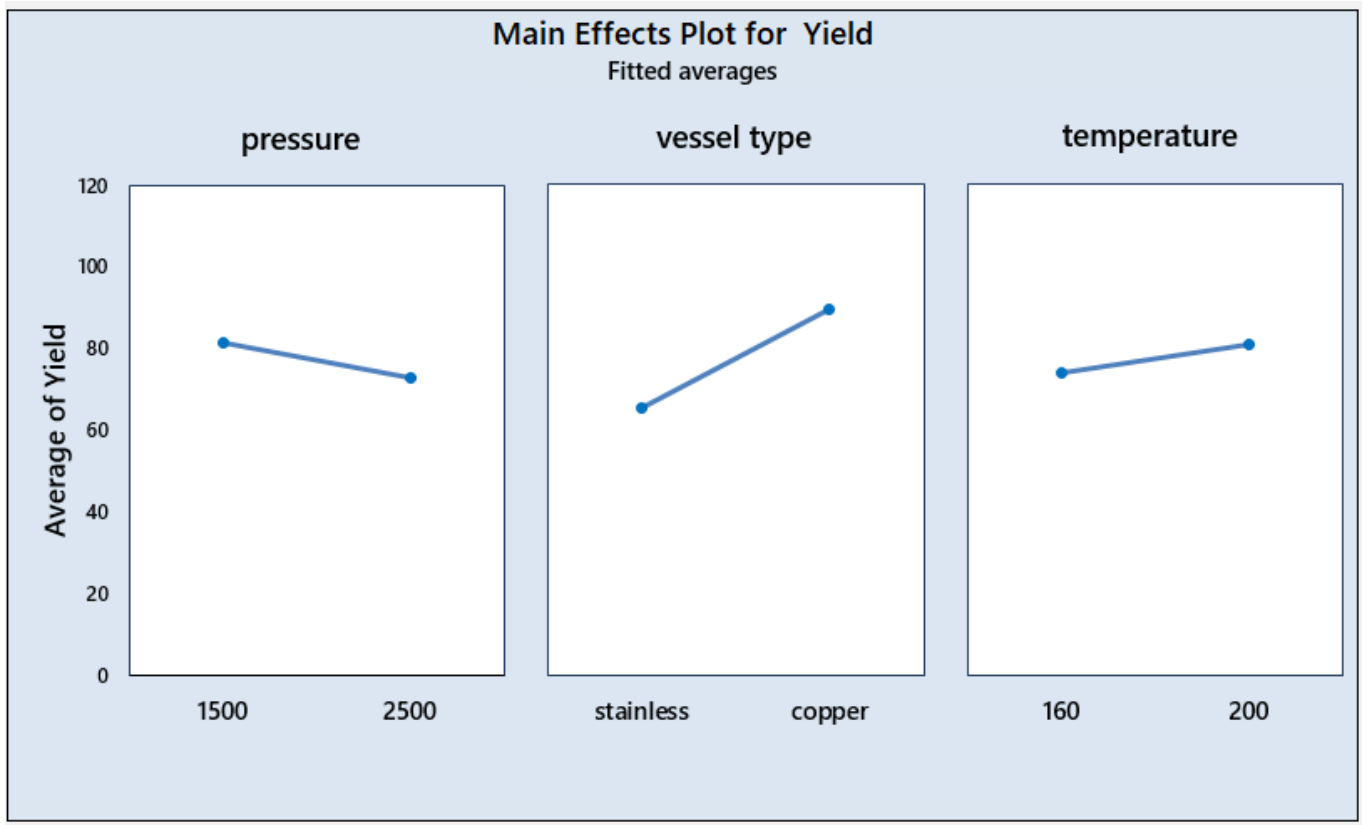
# Design of Experiments (DOE)



## Main Effects Plots

Scroll further to the right to find the Main Effects Plots

- These plots graphically show the effect of changing the setting for each of the main effects



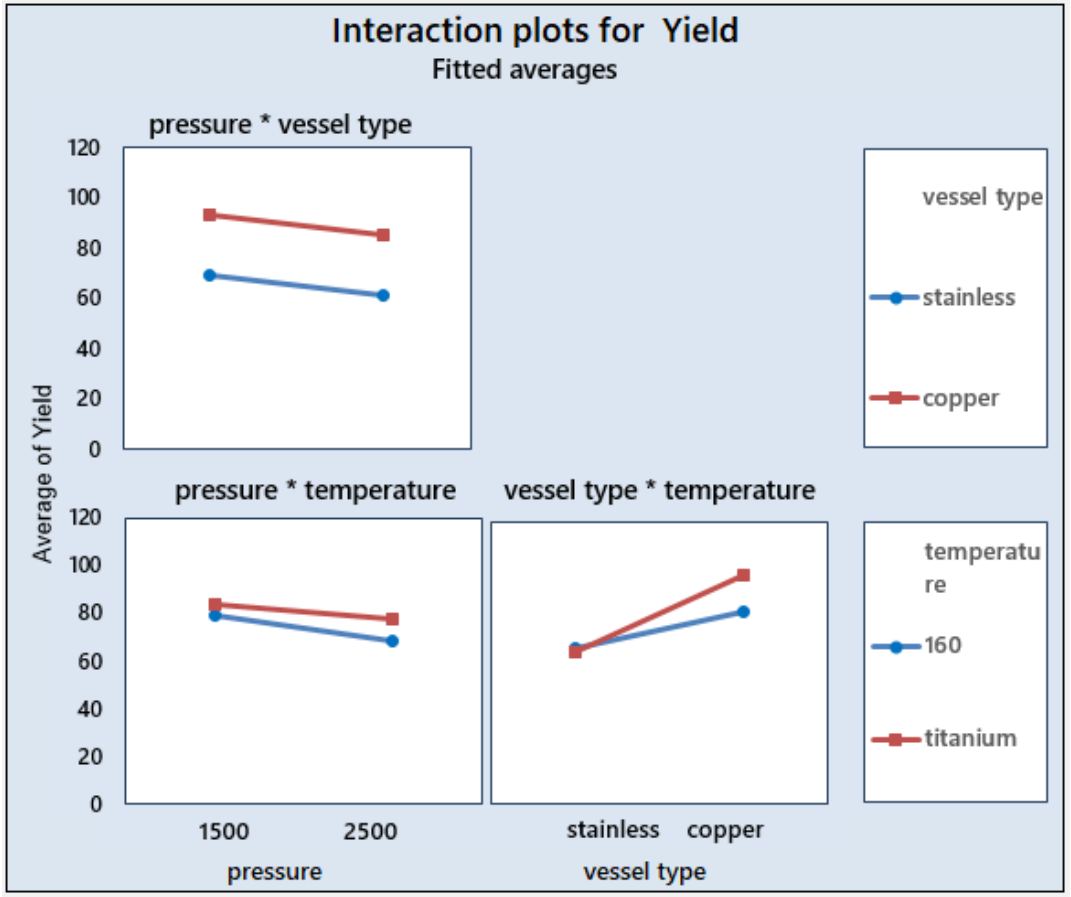
# Design of Experiments (DOE)



## Interaction Plots

Scroll further to the right to find the Interaction Plots

- These plots graphically show the interactions between each pair of two main effects
- Only the interactions included in the model are shown (we'll cover reducing the model shortly).



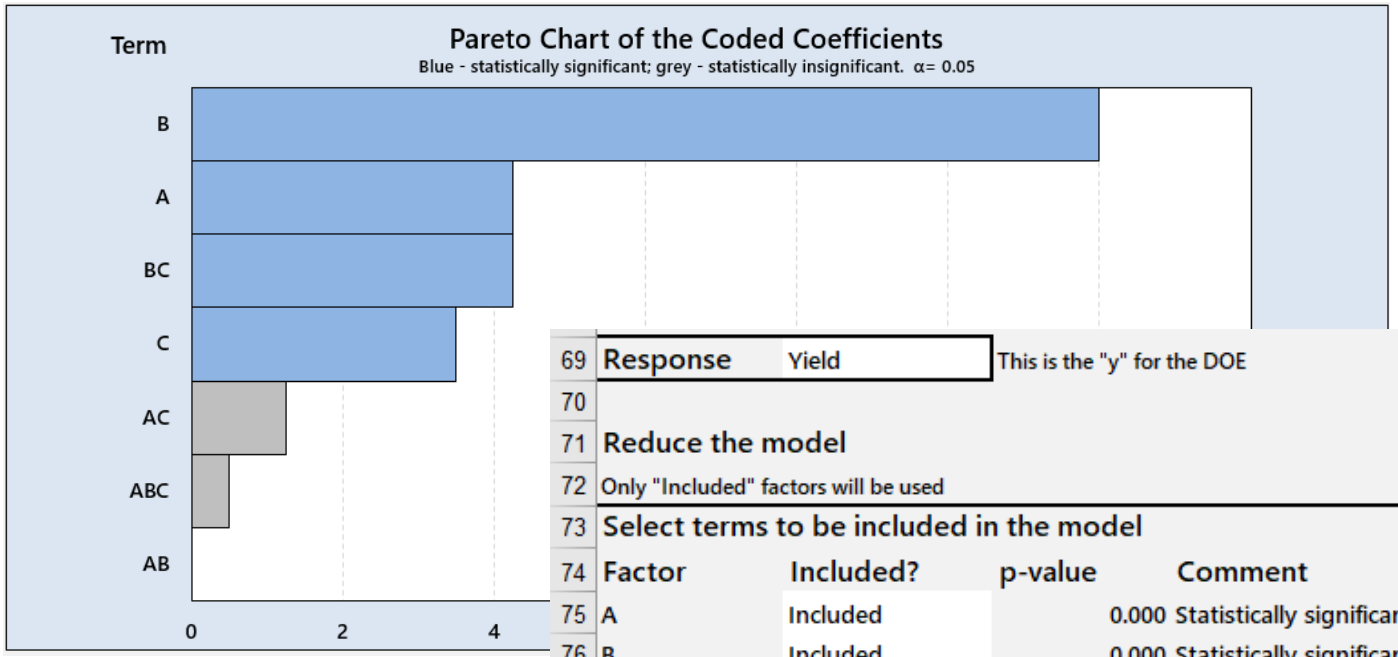
# Design of Experiments (DOE)



## Reducing the Model

Now that we have seen the output generated by the DOE tab in the Toolkit, we'll look at the usual next step – reducing the model in order to focus only on the statistically significant factors and interactions.

- The Pareto chart enables us to see that the AB interaction is zero and ABC is very small – let's take both of these out.
- Scroll all the way back to the left and you will see the 'Select terms to be included in the model' section
- Simply delete the word 'Included' from opposite the relevant terms, as indicated here
- All the rest of the calculations and graphs will update automatically (though it may take a couple of seconds to refresh)




| | | | | |
|---|---|---|---|---|
| 69 | Response | Yield | | This is the "y" for the DOE |
| 70 | | | | |
| 71 | Reduce the model | | | |
| 72 | Only "Included" factors will be used | | | |
| 73 | Select terms to be included in the model | | | |
| 74 | Factor | Included? | p-value | Comment |
| 75 | A | Included | 0.000 | Statistically significant |
| 76 | B | Included | 0.000 | Statistically significant |
| 77 | C | Included | 0.000 | Statistically significant |
| 78 | AB | Included | 1.000 | Not statistically significant |
| 79 | AC | Included | 0.061 | Not statistically significant |
| 80 | BC | Included | 0.000 | Statistically significant |
| 81 | ABC | Included | 0.408 | Not statistically significant |

Delete this

Delete this

# Design of Experiments (DOE)

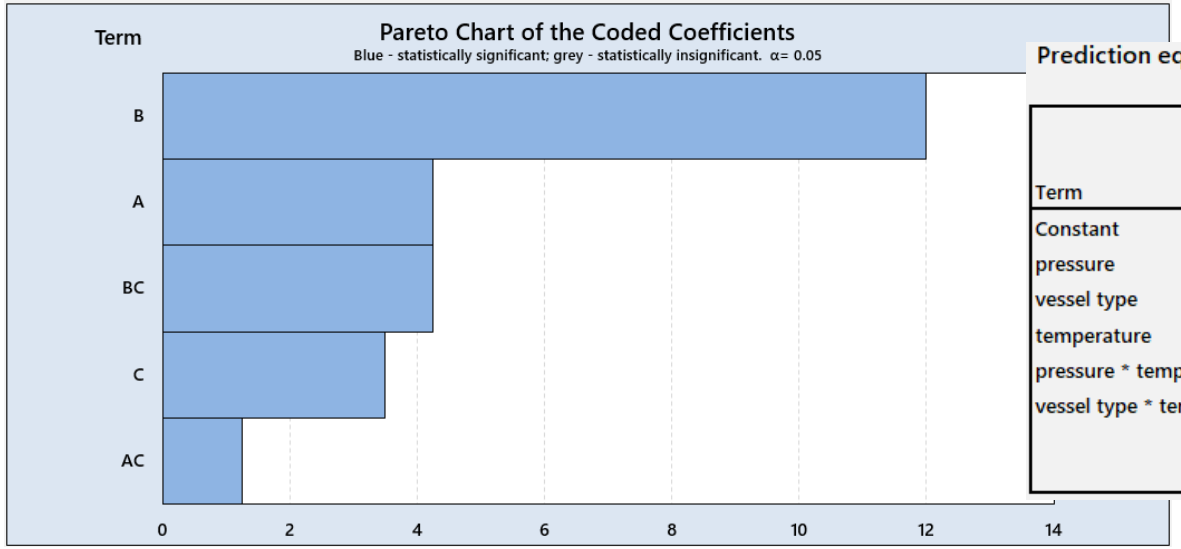
## Reducing the Model

After you remove the AB and ABC terms, they no longer appear in the Pareto Chart or the Prediction Equation.

You'll continue removing terms, prioritising the smallest insignificant ones.

The theory of Designed Experiments requires that you keep the Main Effects in the model if they are involved in any of the interactions.

Lets' break that rule and remove A – delete the word 'Included' next to A




**Prediction equation**

| Term | Model coefficients Coded units | Uncoded units | p-value |
|---|---|---|---|
| Constant | 77.3 | 108 | 0.000 |
| pressure | -4.25 | -0.031 | 0.000 |
| vessel type | 12 | -26.2 | 0.000 |
| temperature | 3.5 | -0.075 | 0.000 |
| pressure * temperature | 1.25 | 0.000125 | 0.042 |
| vessel type * temperature | 4.25 | 0.213 | 0.000 |

**Reduce the model**

Only "Included" factors will be used

**Select terms to be included in the model**

| Factor | Included? | p-value | Comment |
|---|---|---|---|
| A | Included | | Delete this stically significant |
| B | Included | 0.000 | Statistically significant |
| C | Included | 0.000 | Statistically significant |
| AB | | | |
| AC | Included | 0.042 | Statistically significant |
| BC | Included | 0.000 | Statistically significant |
| ABC | | | |

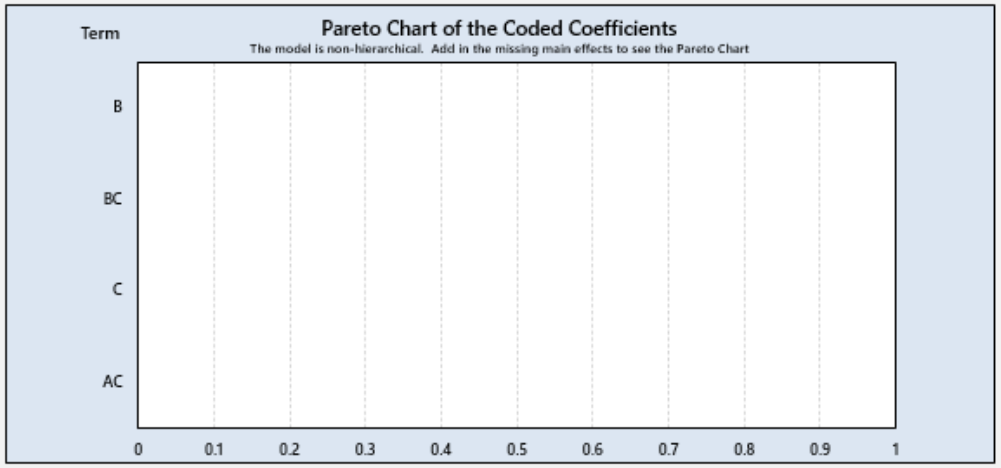
# Design of Experiments (DOE)

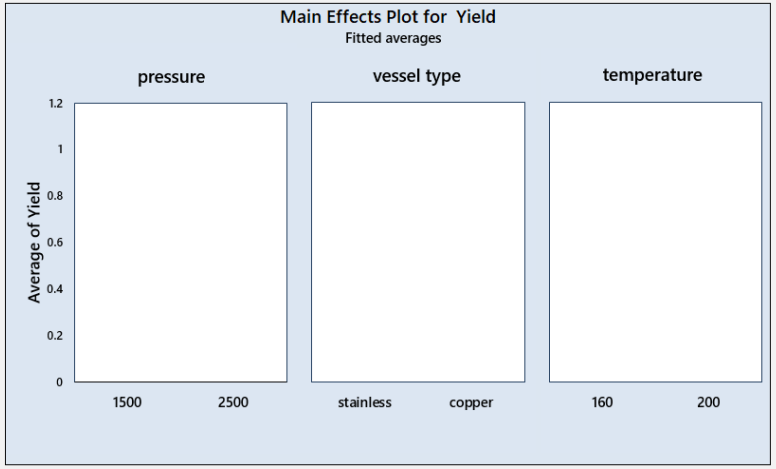
## Reducing the Model

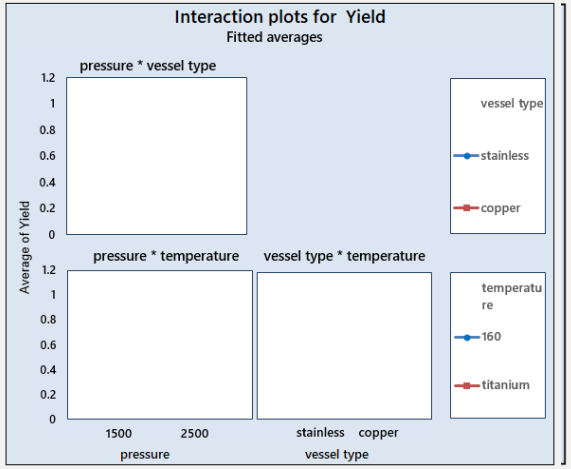
Removing A when this Main Effect is included in one of the interactions is not allowed – doing so will trigger red warning messages and the graphs will sulk.

Change A back to ‘Included’ to restore the graphs.

| | | | | |
|---|---|---|---|---|
| 71 | **Reduce the model** | | | |
| 72 | Only "Included" factors will be used | | | |
| 73 | **Select terms to be included in the model** | | | |
| 74 | **Factor** | **Included?** | **p-value** | **Comment** |
| 75 | A | | | Invalid non-hierarchical model |
| 76 | B | Included | 0.000 | Invalid non-hierarchical model |
| 77 | C | Included | 0.028 | Invalid non-hierarchical model |
| 78 | AB | | | Invalid non-hierarchical model |
| 79 | AC | Included | 0.384 | Invalid non-hierarchical model |
| 80 | BC | Included | 0.010 | Invalid non-hierarchical model |
| 81 | ABC | | | Invalid non-hierarchical model |
| 82 | The model is non-hierarchical because interactions are present that include a main effect that has been excluded from the model. Remove all involved interactions before removing the main effect. | | | |



# Design of Experiments (DOE)

## Final Reduced Model

After restoring A to 'Included' again, we can see that all terms are significant and so on this occasion we shall consider this to be our final reduced model.

We can now review the Prediction Equation and Model Summary for the final reduced model. Recall that the response for this experiment is Yield.

**The prediction equation shown here in coded units is:**

Yield = 77.3 - 4.25*pressure + 12*vessel type + 3.5*temperature + 1.25*pressure*temperature + 4.25*vessel type*temperature

**The prediction equation shown here in uncoded units is:**

Yield = 108 – 0.031*pressure – 26.2*vessel type – 0.075*temperature + 0.000125*pressure*temperature + 0.213*vessel type*temperature

Finally, the standard deviation of the residual errors is 2.145 and the R-Squared adjusted is 98.5%.

**Prediction equation**

| | Model coefficients | | |
|---|---|---|---|
| Term | Coded units | Uncoded units | p-value |
| Constant | 77.3 | 108 | 0.000 |
| pressure | -4.25 | -0.031 | 0.000 |
| vessel type | 12 | -26.2 | 0.000 |
| temperature | 3.5 | -0.075 | 0.000 |
| pressure * temperature | 1.25 | 0.000125 | 0.042 |
| vessel type * temperature | 4.25 | 0.213 | 0.000 |

**Model Summary**

| | |
|---|---|
| s = | 2.145 |
| R-Sq = | 98.5% |
| R-sq (adj) = | 97.8% |

# Design of Experiments - Summary

**This worksheet contains detailed analysis of Designed Experiments**

- 2, 3 and 4 factor designs are available, with up to 16 runs in each case.
- All experiments are $2^k$ full factorial designs.
- Data, factor names, types and settings can be directly typed into the worksheet rather than having to work through a series of menus, making the worksheet completely intuitive to use.
- Data can be entered in either standard order or in a randomised run order
- Pareto analysis shows significant and insignificant factors clearly
- Main Effect Plots and Interaction Plots enable the effects to be visualised and interpreted
- Statistical analysis of model coefficients (both coded and uncoded) is provided, plus p-values, standard deviation and correlation coefficient
- The model can be reduced in a quick and simple way marking factors as included or not – taking a fraction of the time needed by other analysis packages.

There are no menus to learn.

Helpful error messages are provided to politely point out the mistakes that are commonly made, to speed up the learning process

# Report

Free space for you to record your work

# Report


**The Report sheet is intended for you to record your work**

There are no restrictions on what you can do with it.

- You can paste snips of analysis of graphs that you have created
- You can use it as an occasional temporary storage for data that you wish to manipulate in ways that are not possible with the other toolkit sheets
  - You can use it to paste and sort data – the restrictions in the other sheets, which are designed to prevent unintentional corruption, mean that you cannot sort data within them
  - You can use it to create Pivot Tables, so that you can create summary tables of data held in other sheets

# Unstack and Stack Data

# Unstack and Stack Data



**Data can be presented in either stacked or unstacked format.**

As explained in the section on Multiple plots and t-tests, stacked data is generally more useful than unstacked data

- Stacked data enables us to stratify data in as many ways as we like… just add extra columns. In the example below, we are investigating the causes of variation of impurity content and we can easily add columns for day of the week, say, or production supervisor.
- Stacked data also enables us to use date and time – one date/time value applies to the whole of each row

There may be times when, despite this, you prefer your data to be unstacked – for example:

- Preparing data for a report that has already been configured as unstacked
- Preparing data for a paired t-test, which needs unstacked data
- Improving the appearance of a Time Series Plot with several lines

- Stacked data used for Stratified Plots

| Date/Time | Impurity Content | Batch | Shift | Machine |
|---|---|---|---|---|
| 26/11/2019 | 3.21 | a | Day | 110 |
| 26/11/2019 | 3.01 | a | Night | 180 |
| 26/11/2019 | 2.78 | a | Day | 220 |
| 27/11/2019 | 2.94 | a | Night | 110 |
| 28/11/2019 | 2.97 | a | Day | 180 |
| 29/11/2019 | 2.95 | a | Night | 220 |

- Unstacked Data used for Multiple Plots

| Impurity Content_Day | Impurity Content_Night |
|---|---|
| 3.21 | 3.01 |
| 2.78 | 2.94 |
| 2.97 | 2.95 |
| 2.95 | 3.3 |
| 3.38 | 3.05 |

The Unstack and Stack Data provides a simple way to convert between these two layouts

# Unstack and Stack Data

## Unstacking data

Delete all the data in the  Graphs and Statistical Analysis worksheet and then select the worksheet ‘Unstack and Stack Data’.

- This data contains lap times of four people who have competed against each other in two races.

The original stacked data is given in columns B and C – this is where you will enter your own data.

The Data Unstacker has automatically created four columns, one for each driver, by unstacking the source data.

Note that the date is not used here because it will anyway be lost during the unstacking process.

For the purpose of illustration, paste *both* the original and unstacked data into the Graphs and Statistical Analysis worksheet – we’ll compare what we can do with them.

Data Unstacker

Jump to data stacker

Unstacked data is normally best for multiple Time Series Plots where you want the datasets to appear in parallel (if you use unstacked data th
provided with unstacked data and ANOVA is provided with stacked data - so if you have two data sets and wish to compare them with a t-tes

| Stacked Data (input table) | |
|---|---|
| Variable | Subscript |
| Laptime | Driver |
| 34.449 | Rob |
| 35.258 | Rob |
| 33.997 | Rob |
| 42.493 | Rob |
| 34.617 | Rob |
| 34.617 | Rob |
| 36.373 | Rob |
| 34.534 | Rob |
| 34.739 | Rob |
| 34.328 | Rob |
| 34.534 | Rob |
| 35.836 | Alex |
| 36.105 | Alex |
| 36.064 | Alex |
| 38.462 | Alex |
| 36.457 | Alex |
| 35.319 | Alex |
| 36.994 | Alex |
| 35.485 | Alex |
| 34.436 | Alex |
| 36.597 | Alex |

| Unstacked Data suitable for Multiple Plots and T-Tests | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Laptime_Rob | Laptime_Alex | Laptime_David | Laptime_Hayley | | | | | |
| 34.449 | 35.836 | 36.291 | 48.016 | | | | | |
| 35.258 | 36.105 | 35.774 | 45.66 | | | | | |
| 33.997 | 36.064 | 35.837 | 46.691 | | | | | |
| 42.493 | 38.462 | 37.676 | 43.715 | | | | | |
| 34.617 | 36.457 | 35.774 | 43.674 | | | | | |
| 34.617 | 35.319 | 35.692 | 44.563 | | | | | |
| 36.373 | 36.994 | 35.65 | 42.494 | | | | | |
| 34.534 | 35.485 | 35.795 | 43.881 | | | | | |
| 34.739 | 34.436 | 35.443 | 42.144 | | | | | |
| 34.328 | 36.597 | 36.829 | 40.489 | | | | | |
| 34.534 | 35.402 | 35.319 | 40.903 | | | | | |
| 36.251 | 41.596 | 35.008 | 41.192 | | | | | |
| 35.444 | 34.563 | 34.099 | 40.179 | | | | | |
| 34.966 | 34.803 | 34.348 | 40.261 | | | | | |
| 35.486 | 41.421 | 34.326 | 39.352 | | | | | |
| 34.863 | 34.139 | 34.286 | 40.448 | | | | | |
| 35.115 | 34.203 | 34.658 | 40.571 | | | | | |
| 38.668 | 34.719 | 34.284 | 39.206 | | | | | |
| 35.196 | 35.113 | 35.051 | 39.289 | | | | | |
| 36.477 | 36.044 | 34.555 | 38.545 | | | | | |

# Unstack and Stack Data



## Unstacking data

Here is the data, as pasted into the Graphical and Statistical Analysis worksheet.

You would not normally paste both stacked and unstacked data – we have done so purely to be able to compare and contrast.

| Exclusions | Time axis | Variable 1 | Variable 2 | Variable 3 | Variable 4 | Variable 5 | Variable 6 |
|---|---|---|---|---|---|---|---|
| Exclude from Control Charts (only) | | Laptime | Driver | Laptime_Rob | Laptime_Alex | Laptime_David | Laptime_Hayley |
| | | 34.449 | Rob | 34.449 | 35.836 | 36.291 | 48.016 |
| | | 35.258 | Rob | 35.258 | 36.105 | 35.774 | 45.66 |
| | | 33.997 | Rob | 33.997 | 36.064 | 35.837 | 46.691 |
| | | 42.493 | Rob | 42.493 | 38.462 | 37.676 | 43.715 |
| | | 34.617 | Rob | 34.617 | 36.457 | 35.774 | 43.674 |
| | | 34.617 | Rob | 34.617 | 35.319 | 35.692 | 44.563 |
| | | 36.373 | Rob | 36.373 | 36.994 | 35.65 | 42.494 |
| | | 34.534 | Rob | 34.534 | 35.485 | 35.795 | 43.881 |
| | | 34.739 | Rob | 34.739 | 34.436 | 35.443 | 42.144 |
| | | 34.328 | Rob | 34.328 | 36.597 | 36.829 | 40.489 |
| | | 34.534 | Rob | 34.534 | 35.402 | 35.319 | 40.903 |
| | | 35.836 | Alex | 36.251 | 41.596 | 35.008 | 41.192 |
| | | 36.105 | Alex | 35.444 | 34.563 | 34.099 | 40.179 |
| | | 36.064 | Alex | 34.966 | 34.803 | 34.348 | 40.261 |
| | | 38.462 | Alex | 35.486 | 41.421 | 34.326 | 39.352 |
| | | 36.457 | Alex | 34.863 | 34.139 | 34.286 | 40.448 |
| | | 35.319 | Alex | 35.115 | 34.203 | 34.658 | 40.571 |
| | | 36.994 | Alex | 38.668 | 34.719 | 34.284 | 39.206 |
| | | 35.485 | Alex | 35.196 | 35.113 | 35.051 | 39.289 |

# Unstack and Stack Data



## Unstacking data

Here is a stratified Time Series Plot that we made with the first two columns.

We will let you decide how helpful this presentation is

**Stratified Time Series Plot, Box Plots and ANOVA**

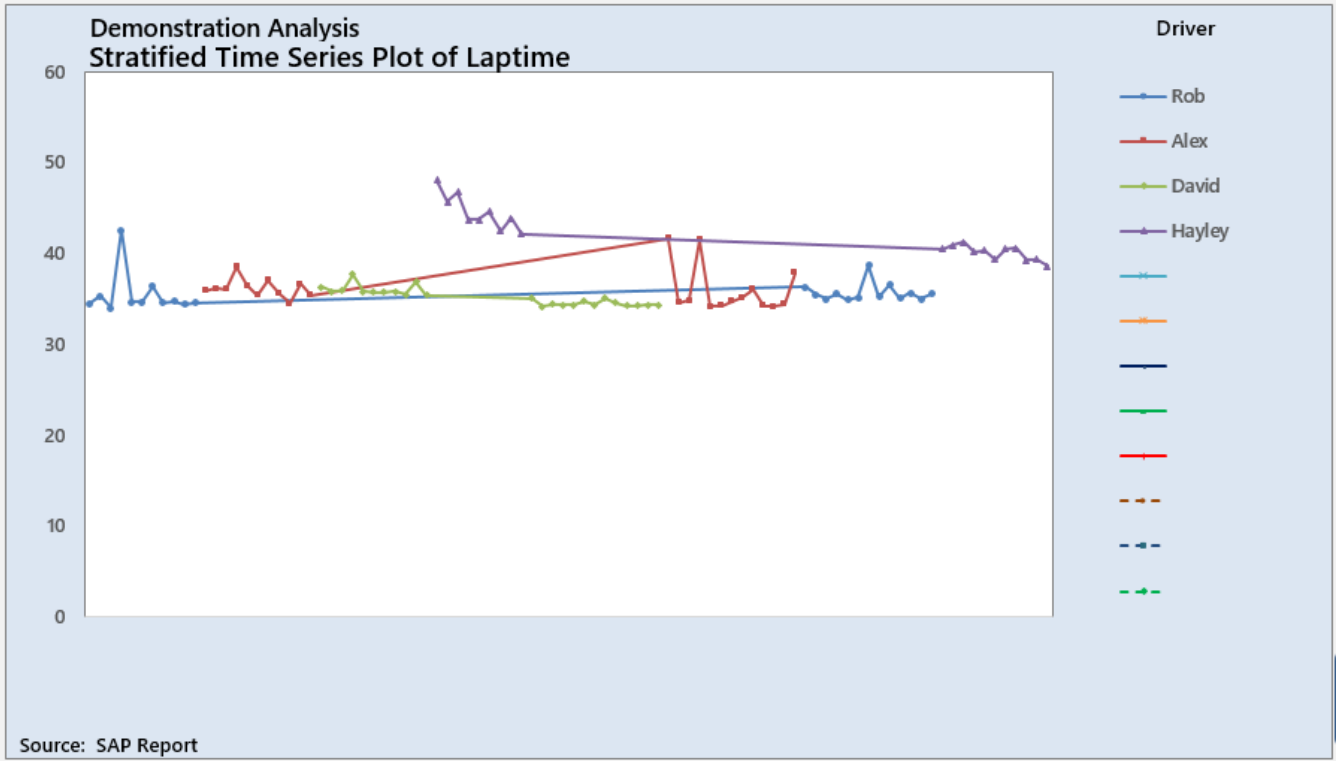
Use these graphs if you have a categorical column that can be used to stratify your data (up to a maximum of 10 lines for the Time Series Plot and 32 groups in the Box Plot)

Which Column to plot?

Laptime

Which column to use to stratify the data?

Driver



# Unstack and Stack Data

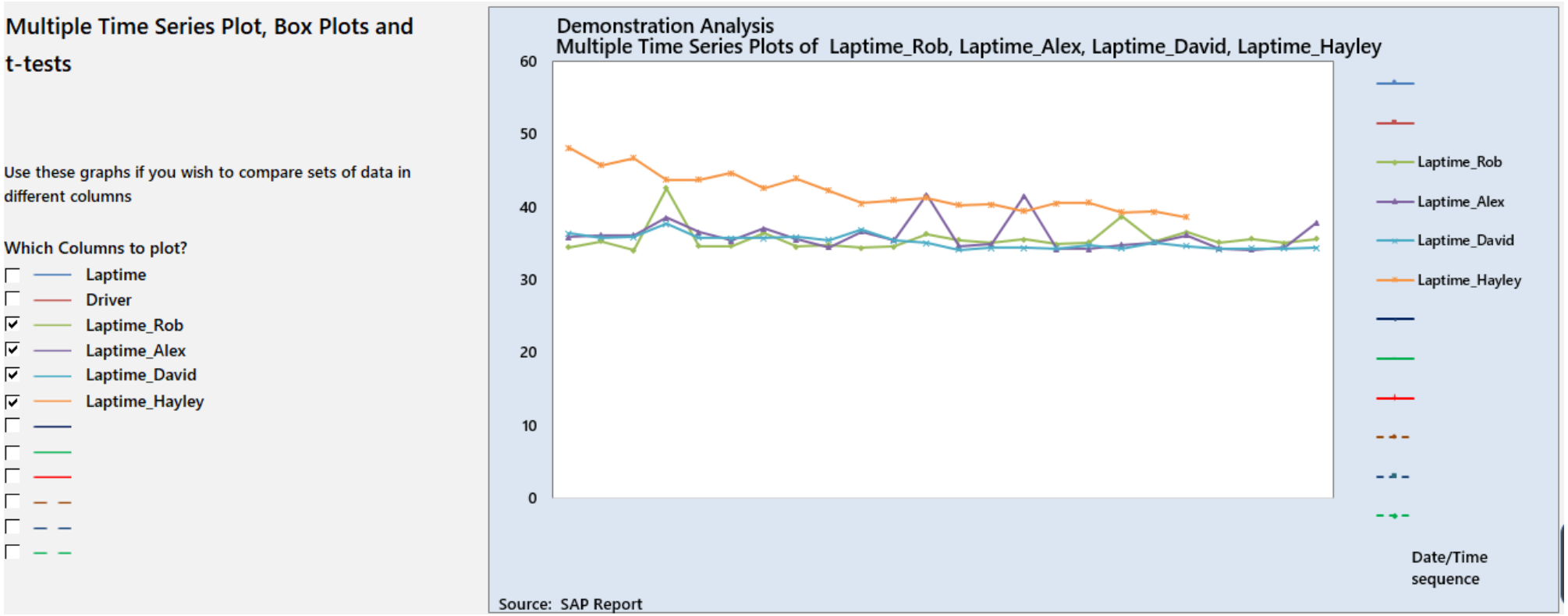


## Unstacking data

Here is the same data, but plotted using the four columns that were created after we unstacked it.

It is probably easier to see the key features with this version of the graph:

- Alex, Rob and David have very similar times, but Rob and Alex suffer from occasional slow laps
- Hayley was slower at first, but her lap times are improving rapidly



# Unstack and Stack Data

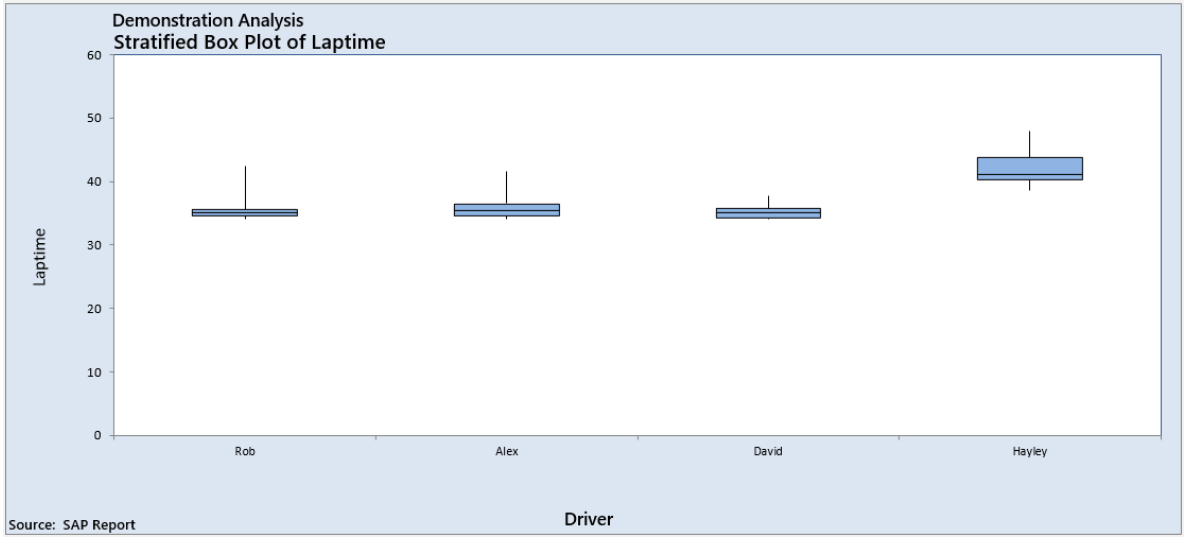
## Unstacking data

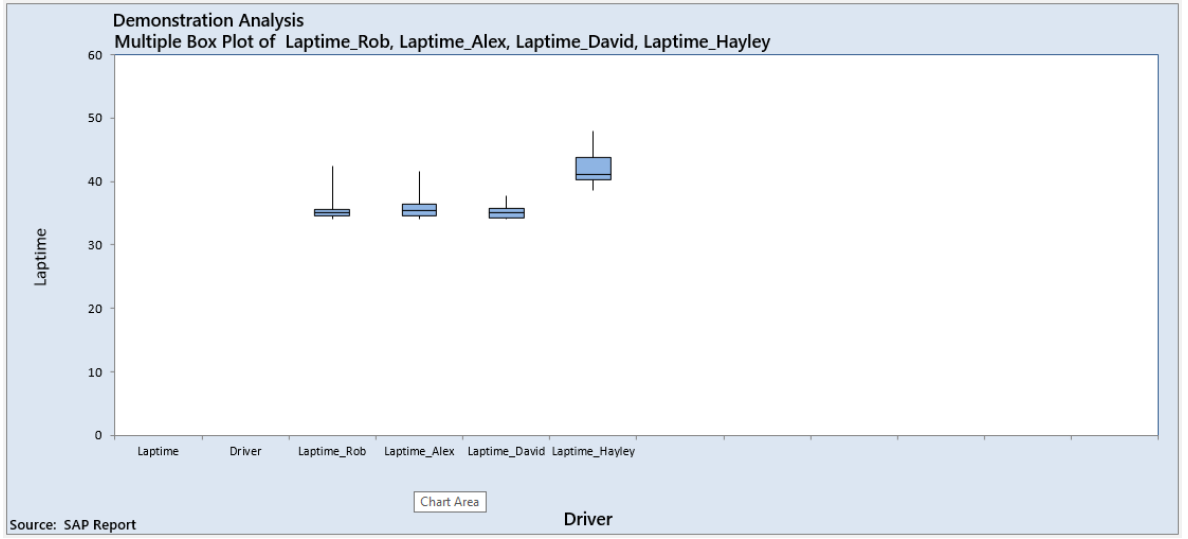
The box plots are the same, as you would expect

- The only difference between these two graphs is the change in shape and layout from stratified to multiple plots

The ANOVA analysis is available in the same rows as the stratified data

- There is no comparable t-test analysis because there are more than two groups



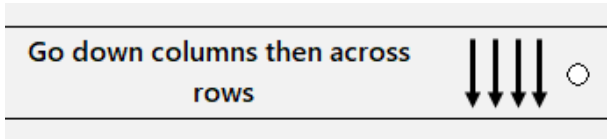



# Unstack and Stack Data

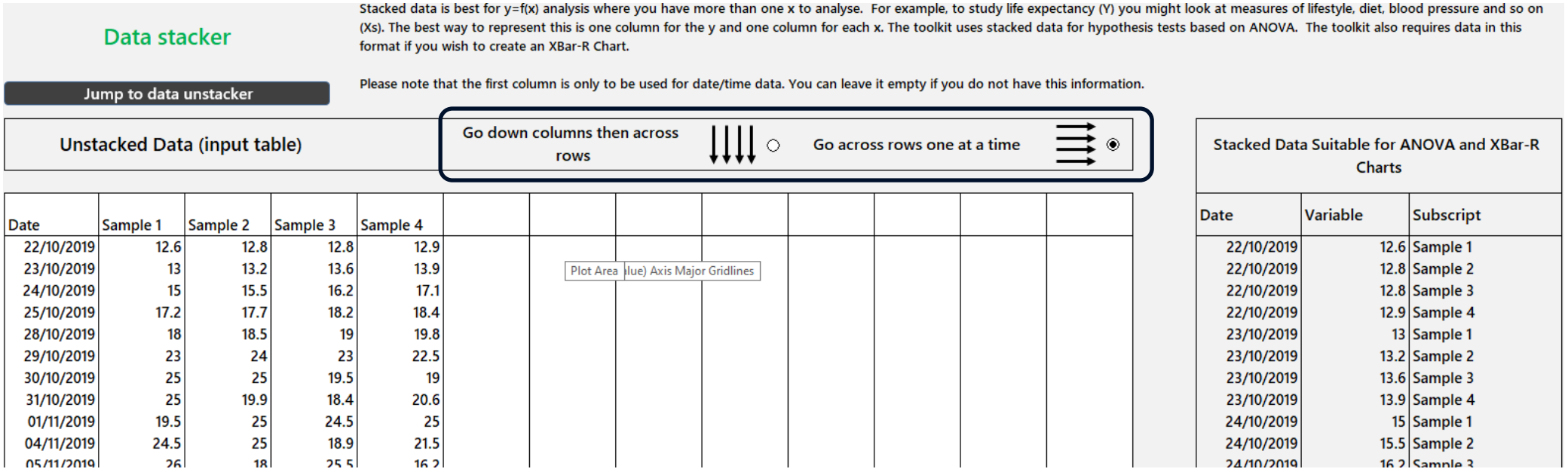


## Stacking Data

- We have completed our study of the lap time data, so delete all the data in the Graphical and Statistical Analysis worksheet.
- In the Unstack and Stack Data worksheet, scroll to the right and you will find the Data stacker.
- The same basic idea applies as before: paste your data into the white cells and it will be processed automatically – though in this case, the data will be stacked.
- You will notice there is a choice of method:




or


| Date | Sample 1 | Sample 2 | Sample 3 | Sample 4 |
|---|---|---|---|---|
| 22/10/2019 | 12.6 | 12.8 | 12.8 | 12.9 |
| 23/10/2019 | 13 | 13.2 | 13.6 | 13.9 |
| 24/10/2019 | 15 | 15.5 | 16.2 | 17.1 |
| 25/10/2019 | 17.2 | 17.7 | 18.2 | 18.4 |
| 28/10/2019 | 18 | 18.5 | 19 | 19.8 |
| 29/10/2019 | 23 | 24 | 23 | 22.5 |
| 30/10/2019 | 25 | 25 | 19.5 | 19 |
| 31/10/2019 | 25 | 19.9 | 18.4 | 20.6 |
| 01/11/2019 | 19.5 | 25 | 24.5 | 25 |
| 04/11/2019 | 24.5 | 25 | 18.9 | 21.5 |
| 05/11/2019 | 26 | 18 | 25.5 | 16.2 |

| Date | Variable | Subscript |
|---|---|---|
| 22/10/2019 | 12.6 | Sample 1 |
| 22/10/2019 | 12.8 | Sample 2 |
| 22/10/2019 | 12.8 | Sample 3 |
| 22/10/2019 | 12.9 | Sample 4 |
| 23/10/2019 | 13 | Sample 1 |
| 23/10/2019 | 13.2 | Sample 2 |
| 23/10/2019 | 13.6 | Sample 3 |
| 23/10/2019 | 13.9 | Sample 4 |
| 24/10/2019 | 15 | Sample 1 |
| 24/10/2019 | 15.5 | Sample 2 |
| 24/10/2019 | 16.2 | Sample 3 |

# Unstack and Stack Data

## Stacking Data

Here are the two configurations of the stacked data you can use, depending on whether you choose the columns-first or rows-first method.

- The second option (rows first) is usually preferable, but it makes less sense when the columns are unequal length.
- Your choice will ultimately depend on what the data represents and what you want to do with it.

**Unstacked Data (input table)** — Go down columns then across rows ◉ — Go across rows one at a time ○

| Date | Sample 1 | Sample 2 | Sample 3 | Sample 4 |
|---|---|---|---|---|
| 22/10/2019 | 12.6 (1) | 12.8 | 12.8 | 12.9 |
| 23/10/2019 | 13 (2) | 13.2 | 13.6 | 13.9 |
| 24/10/2019 | 15 (3) | 15.5 | 16.2 | 17.1 |
| 25/10/2019 | 17.2 (4) | 17.7 | 18.2 | 18.4 |
| 28/10/2019 | 18 (5) | 18.5 | 19 | 19.8 |
| 29/10/2019 | 23 (6) | 24 | 23 | 22.5 |
| 30/10/2019 | 25 | 25 | 19.5 | 19 |

**Stacked Data Suitable for ANOVA and XBar-R Charts**

| Date | Variable | Subscript |
|---|---|---|
| 22/10/2019 | 12.6 | Sample 1 (1) |
| 23/10/2019 | 13 | Sample 1 (2) |
| 24/10/2019 | 15 | Sample 1 (3) |
| 25/10/2019 | 17.2 | Sample 1 (4) |
| 28/10/2019 | 18 | Sample 1 (5) |
| 29/10/2019 | 23 | Sample 1 (6) |
| 30/10/2019 | 25 | Sample 1 |

- All the data from the first column, then the second, and so on

**Unstacked Data (input table)** — Go down columns then across rows ○ — Go across rows one at a time ◉

| Date | Sample 1 | Sample 2 | Sample 3 | Sample 4 |
|---|---|---|---|---|
| 22/10/2019 | 12.6 (1) | 12.8 (2) | 12.8 (3) | 12.9 (4) |
| 23/10/2019 | 13 (5) | 13.2 (6) | 13.6 | 13.9 |
| 24/10/2019 | 15 | 15.5 | 16.2 | 17.1 |
| 25/10/2019 | 17.2 | 17.7 | 18.2 | 18.4 |
| 28/10/2019 | 18 | 18.5 | 19 | 19.8 |
| 29/10/2019 | 23 | 24 | 23 | 22.5 |

**Stacked Data Suitable for ANOVA and XBar-R Charts**

| Date | Variable | Subscript |
|---|---|---|
| 22/10/2019 | 12.6 | Sample 1 (1) |
| 22/10/2019 | 12.8 | Sample 2 (2) |
| 22/10/2019 | 12.8 | Sample 3 (3) |
| 22/10/2019 | 12.9 | Sample 4 (4) |
| 23/10/2019 | 13 | Sample 1 (5) |
| 23/10/2019 | 13.2 | Sample 2 (6) |

- All the data from the first row, then the second, and so on

# Unstack and Stack Data

## Stacking Data for X Bar-R Control Charts

One specific (and very common) purpose of stacking data is to prepare an Xbar-R Control Chart

- You may have one column for each sample in a subgroup – by the Toolkit requires stacked data for the X-Bar R Control Chart
- In this case, you need to choose the “Go across rows one at a time” option

**Unstacked Data (input table)** — Go down columns then across rows ○ — Go across rows one at a time ◉

| Date | Sample 1 | Sample 2 | Sample 3 | Sample 4 |
|---|---|---|---|---|
| 22/10/2019 | 12.6 | 12.8 | 12.8 | 12.9 |
| 23/10/2019 | 13 | 13.2 | 13.6 | 13.9 |
| 24/10/2019 | 15 | 15.5 | 16.2 | 17.1 |
| 25/10/2019 | 17.2 | 17.7 | 18.2 | 18.4 |
| 28/10/2019 | 18 | 18.5 | 19 | 19.8 |
| 29/10/2019 | 23 | 24 | 23 | 22.5 |

Plot Area lue) Axis Major Gridlines

**Stacked Data Suitable for ANOVA and XBar-R Charts**

| Date | Variable | Subscript |
|---|---|---|
| 22/10/2019 | 12.6 | Sample 1 |
| 22/10/2019 | 12.8 | Sample 2 |
| 22/10/2019 | 12.8 | Sample 3 |
| 22/10/2019 | 12.9 | Sample 4 |
| 23/10/2019 | 13 | Sample 1 |
| 23/10/2019 | 13.2 | Sample 2 |

- With this option chosen, copy and paste all the data from the right-hand section into the Graphs and Statistical Analysis worksheet.

| | Exclusions | Time axis | Variable 1 | Variable 2 |
|---|---|---|---|---|
| Row # | Exclude from Control Charts (only) | Date | Variable | Subscript |
| 1 | | 22/10/2019 | 12.6 | Sample 1 |
| 2 | | 22/10/2019 | 12.8 | Sample 2 |
| 3 | | 22/10/2019 | 12.8 | Sample 3 |
| 4 | | 22/10/2019 | 12.9 | Sample 4 |
| 5 | | 23/10/2019 | 13 | Sample 1 |
| 6 | | 23/10/2019 | 13.2 | Sample 2 |

# Unstack and Stack Data

## Stacking Data for X Bar-R Control Charts

With the data now pasted:

- Select the Control Charts hyperlink
- Select the variable ‘Variable’
- Select the subgroup size – 4, in this case

Scroll to the right and you will see the Xbar-R chart made with your data

**Control Charts**

Tick here to use the same data
as the Time Series Plot ☐

I-MR (individual points) on the left
Xbar-R (with subgroups) on the right
Only the relevant chart will be active

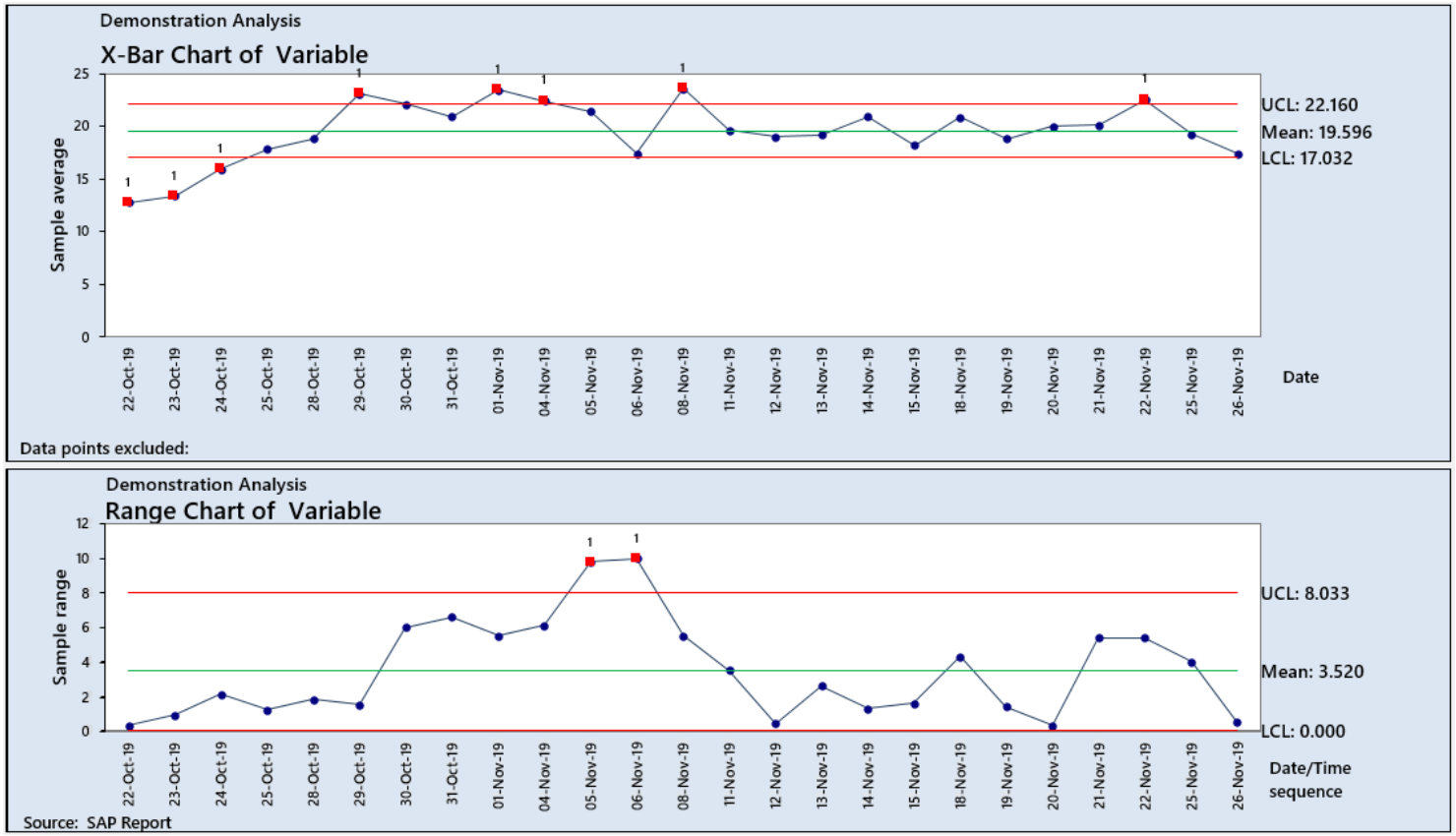
Use column B to identify rows to be excluded from control limit calcs

Make a control chart of:
Variable

***Subgroup Size*** 4
(leave empty for no subgroups)
as your data is in subgroups,
only an X bar-R chart is shown

***# 'Before' Points***
(to create a Before/After chart)

***Control Chart with fixed limits***
Enter required values here:
Upper Control Limit
Average
Lower Control Limit



# Unstack and Stack Data - Summary

**This worksheet contains a utility to perform two tasks which are otherwise fiddly to do in Excel**

The Data Unstacker data separate columns for each level of the stratification factor (up to 12 different columns)

- Potential applications
  - Preparing stacked data to be used for the toolkit's 2-sample t-test
  - Preparing stacked data to be used in a multiple time series plot, to visualise the effect of different settings alongside each other.

The Data Stacker recombines unstacked columns (up to 12 in total) into stacked columns

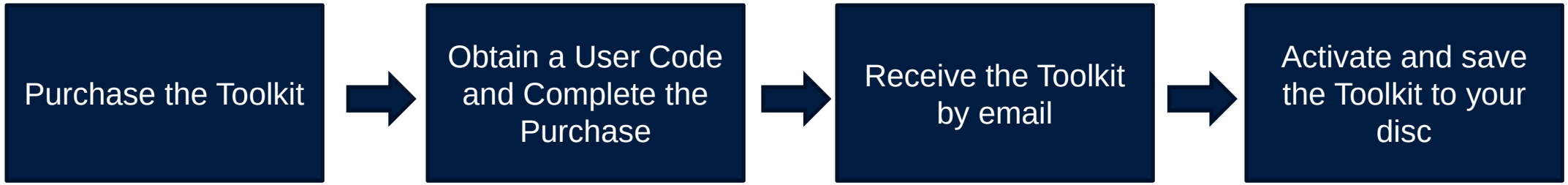
- Potential applications
  - Preparing data so that it is ready for additional variables to be added to increase the sophistication of analysis
  - Preparing data for an X-Bar R control chart

# Troubleshooting

# Troubleshooting – Getting Started

**The Data Analysis Toolkit is very easy to use once you have it properly saved on your laptop or PC.**
Such problems as there are generally occur before you get to use the toolkit. This section will explain what to do in the event of a problem.

**The steps to purchasing and using your toolkit are:**



# Troubleshooting

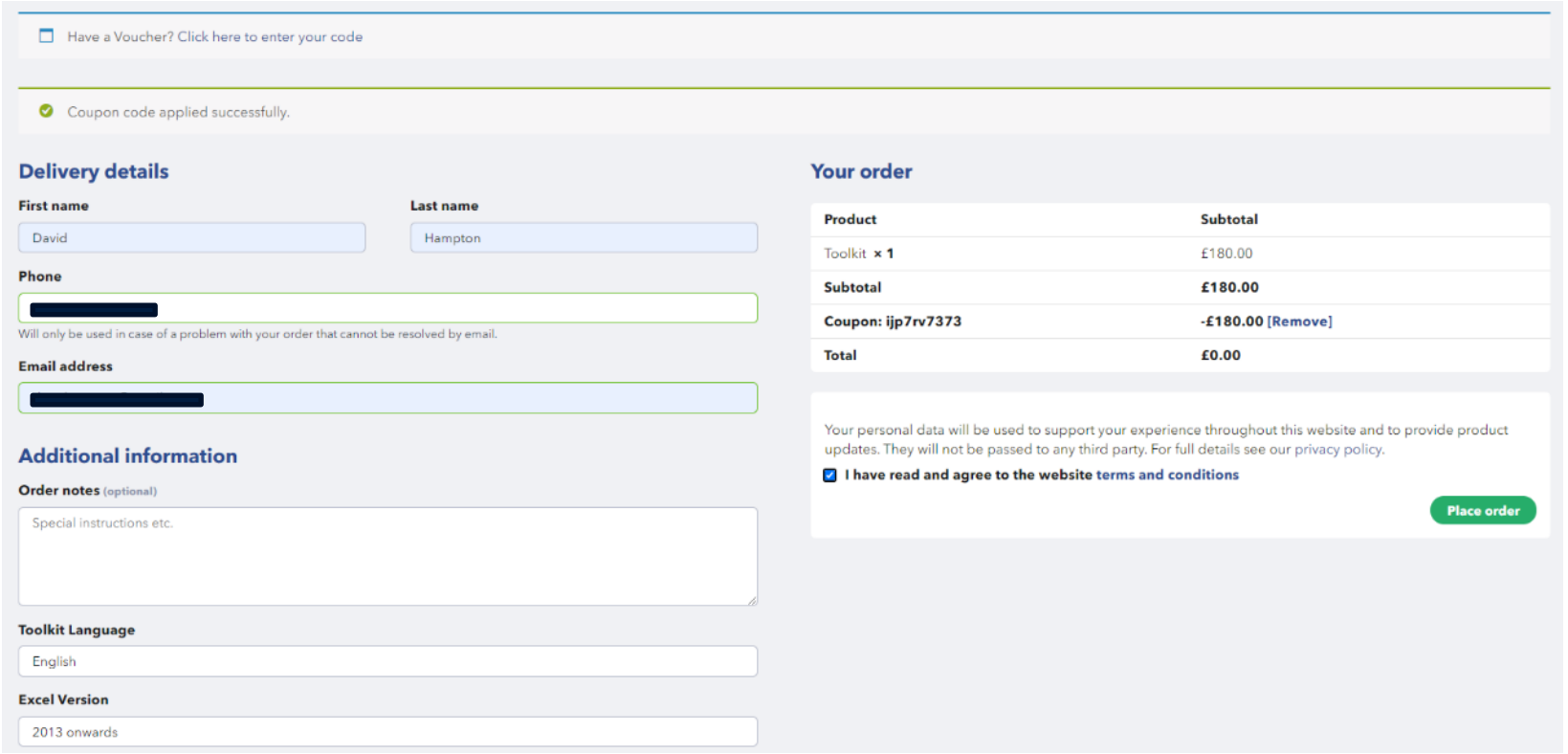
Purchase the Toolkit

## Purchasing the Toolkit

The toolkit can be purchased by following the link at the bottom of this page: https://advancedanalyticssolutions.co.uk/data-analysis-toolkit/

The web form to complete your transaction is conventional. Note:

- We need your phone number as a back-up in case of problems contacting you by email
- The screen shot shows an example where the buyer has used a voucher (this typically occurs when the Toolkit is bought through a reseller)




### What can go wrong here

- You must be using the desktop version of Excel, 2013 or later, or the Toolkit will not work.
  - Not Excel 2010 or earlier
  - Not the web version of Excel
  - Not Libre Office or Apple numbers

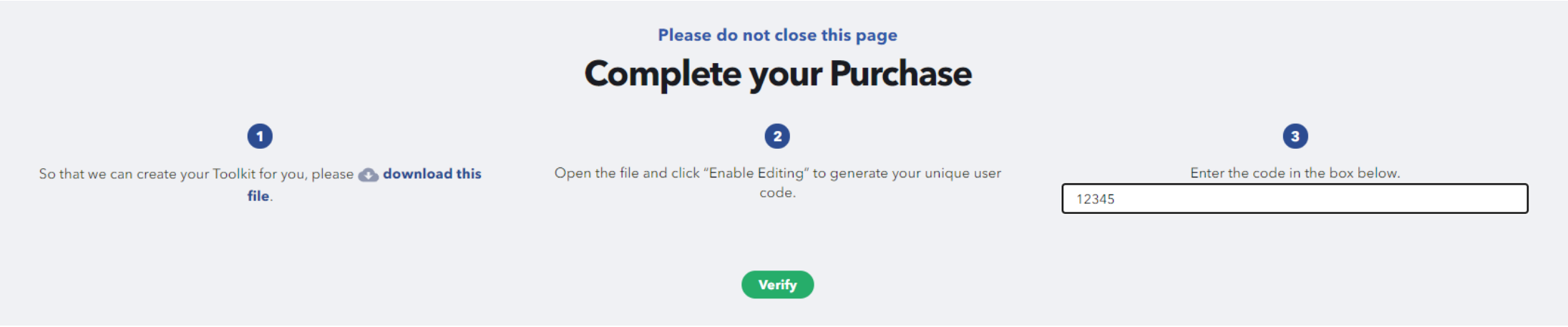
# Troubleshooting



## Obtain a User Code and Complete the Purchase

To prevent unauthorized copying, the Toolkit includes copy protection that locks it to the specific laptop or PC that uses it

1. This page requires you to download a small Excel file, “User Code Generator”
2. When you open this file and click ‘Enable Editing’, you will see a 5-digit code
   – This code is created from your laptop or PC (but cannot be used to identify anything specific about it) – the only purpose of the code is to prevent the toolkit from working on somebody else’s laptop
3. Enter the 5-digit code in the box in step 3 (where we have entered ‘12345’) and click Verify to complete your purchase




## What can go wrong here

- If you use a different laptop for the purchase to the one you will save the toolkit to, the toolkit won’t work (it’s copy-protected)
- If you do not click ‘Enable Editing’ you will see a reminder message instead of your 5-digit code
- If you have any problems obtaining your User Code, you can get help by emailing toolkit@advancedanalyticssolutions.co.uk

# Troubleshooting

Receive the Toolkit by email

## Receive the Toolkit by email

Once you have completed your purchase you will receive an automated email with the subject line “Your Advanced Analytics Solutions order has been received!” to confirm the transaction.

The generation of the toolkit itself involves an offline customisation process and there will be a lead time of a few hours (never more than 24 hours) before you receive your toolkit by email.

## What can go wrong here

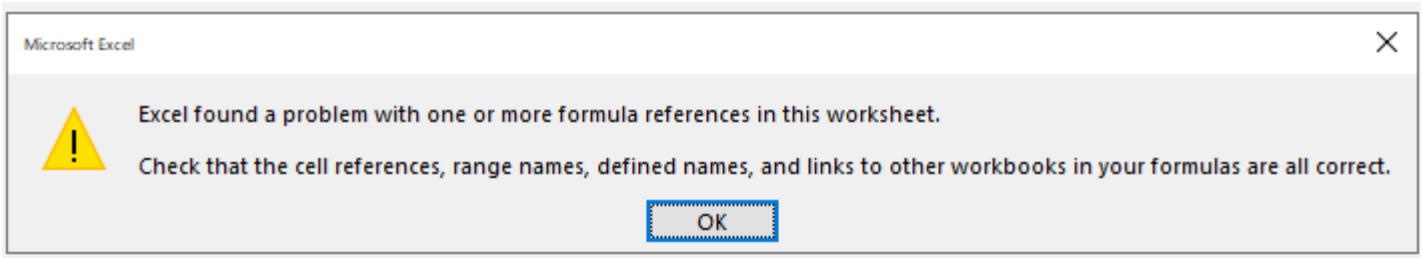
- It is possible that the email with your toolkit could be sent to your spam folder.  If you have not received the toolkit within 24 hours:
  - Check your spam folder
  - If the Toolkit is not in there, email toolkit@advancedanalyticssolutions.co.uk to chase it up

# Troubleshooting

Activate and save the Toolkit to your disc

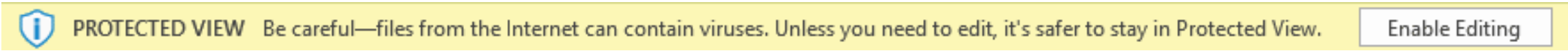
## Activate and save the Toolkit to your disc

- We recommend that you save the toolkit to a folder within your ‘My Documents’ master folder
- The first time you open the toolkit, you will most likely see an error message such as:




These errors are perfectly normal the first time you open the toolkit.  You need to activate it as follows:

- Click ‘Enable Editing’ in the banner at the top of the sheet




- Now save the toolkit to the folder on your laptop where you want to keep it
- In most cases, this will clear the error messages and you will not see them again

## What can go wrong here

- See details on the next slides

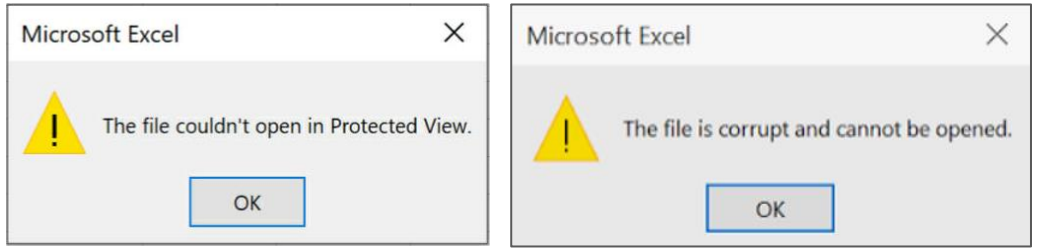
# Troubleshooting

Activate and save the Toolkit to your disc

## Activate and save the Toolkit to your disc – What can go wrong here

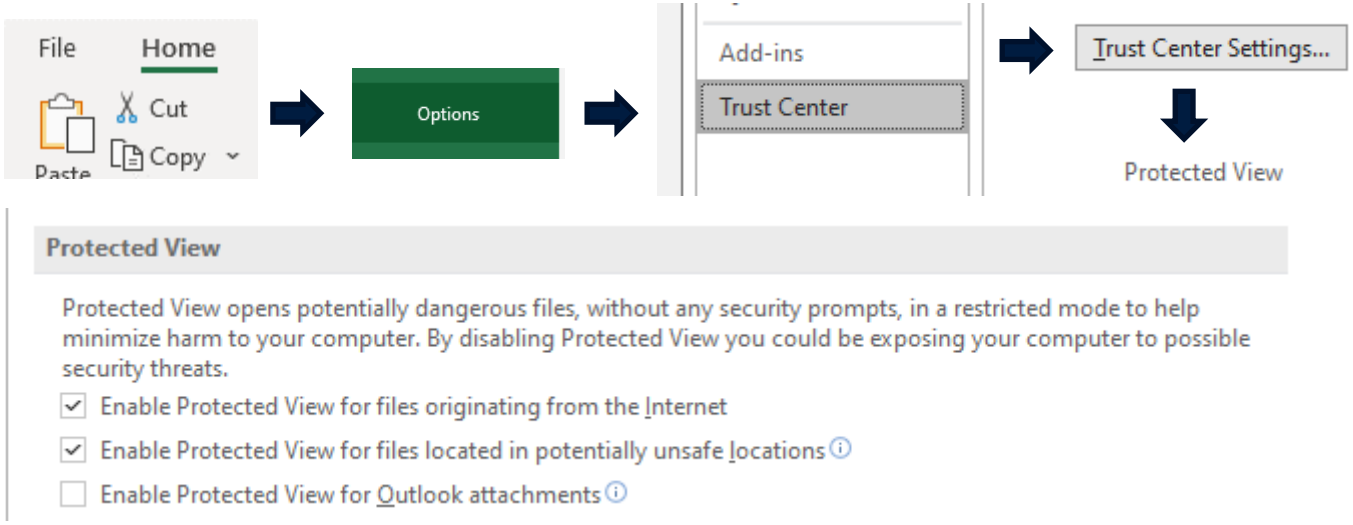
Problem 1: The Toolkit file will not open at first

- On a few computers, the Toolkit will not open at all the first time, and you will see one of these two error messages.
- Save the file to the folder where you will keep it. Open Excel first, and then open the Toolkit file
- Proceed as shown in the previous slide




Problem 2: over-strict security settings

- You can check your security settings in Excel by selecting File > Options > Trust Center > Trust Center Settings > Protected View
- Normal settings are shown here – if you have a tick in the box 'Enable Protected View for outlook attachments' you may be able to solve the problem by unticking this box
- This is only needed for the first opening of the toolkit, so you can change it back afterwards



# Troubleshooting

Activate and save the Toolkit to your disc

## Save the Toolkit to your disc – What can go wrong here

Problem 3: Mixed filing systems

- Some corporate laptops use OneDrive or a similar cloud storage system to store your working files, and they do not reside on your laptop at all
- In principle this is not a problem but you will have a problem if:
  - You save your toolkit in the company filing system
  - You save your User Code Generator in your Desktop (which, even on laptops that use cloud storage, will be on your C drive)
  - Or, the other way round
- The Toolkit must be saved to the same drive as the User Code Generator
  - If your User Code Generator was saved to your hard drive, the Toolkit will need to be saved to your hard drive as well
  - If your User Code Generator was saved to the cloud, the Toolkit will need to be saved to the cloud as well
- If you must save your toolkit to the cloud but unfortunately you generated your User Code from your desktop, we can easily generate a new User Code for you – please contact toolkit@advancedanalyticssolutions.co.uk and we'll take care of it.

# Troubleshooting

Activate and save the Toolkit to your disc

## Save the Toolkit to your disc – What can go wrong here

Problem 3: Mixed filing systems

- If you are using Parallels on an Apple computer, this section is for you (if not, skip it)
- Parallels Desktop for Mac is software providing hardware virtualisation for Macintosh computers with Intel processors.
- The issue here is that Parallels creates two parallel filing systems:
  - A Mac-like filing system that looks something like this: /Users/username/Library/whatever
  - A Windows-like filing system that looks something like this: C:\Users\username\Documents\whatever
- The Data Analysis Toolkit has security features that are designed to prevent it being pirated by copying it to another computer – and these two parallel partitions look like different computers
  - So, by design, the toolkit will not work on both

- The issue is that the toolkit will say that its license is invalid and will not work
- The cause of the issue is usually that the user has created the User Code from the User Code Generator while it was saved in one filing system…
- … but has then saved the Toolkit into the other filing system

- The solution is simply to save the Toolkit into the same partition that you were using when you obtained your User Code

# Troubleshooting

Activate and save the Toolkit to your disc

## Save the Toolkit to your disc – What can go wrong here

Problem 3: Second laptop issues

- The purchase price of the Toolkit covers installation on one laptop/PC only.
- If have two laptops/PCs, the toolkit will only work on one of them.
- If you generated your User Code on a different laptop to the one that you plan to use the Toolkit on, contact toolkit@advancedanalyticssolutions.co.uk within 30 days of your purchase and we'll help you to solve the problem
- Please bear in mind that you will need to purchase another toolkit if you purchase a new laptop/PC

# Troubleshooting

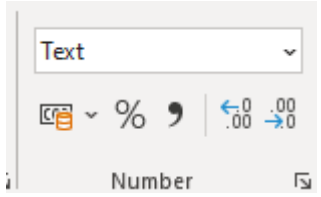
Activate and save the Toolkit to your disc

## Save the Toolkit to your disc – What can go wrong here

Problem 4: License Key problems

- A few users have reported issues with the License Key cell (L8 in the Instructions and License worksheet)
- The license key is a very long number which presents potential risks
  - If you type it in directly, you could get a digit wrong
  - If you copy and paste it, you could accidentally include a spurious space
- This should not be an issue when you first purchase the toolkit
  - The correct license code is already entered for you
- You might have a problem when upgrading the toolkit
  - We publish updates every 6-12 months and these are sent to existing users free of charge; you will need to type your details into the new sheet for yourself

- We recommend that you:
  - Ensure that the format used for your License Key is 'Text'




  - Manually type your license key when updating the toolkit, to avoid any risk of spurious characters
  - Ensure that all the digits can be seen - you don't want to see engineering notation like this:

# Troubleshooting

## Graphs and Statistical Analysis

As there are no menus, this worksheet (and all the others) rarely presents users with problems.

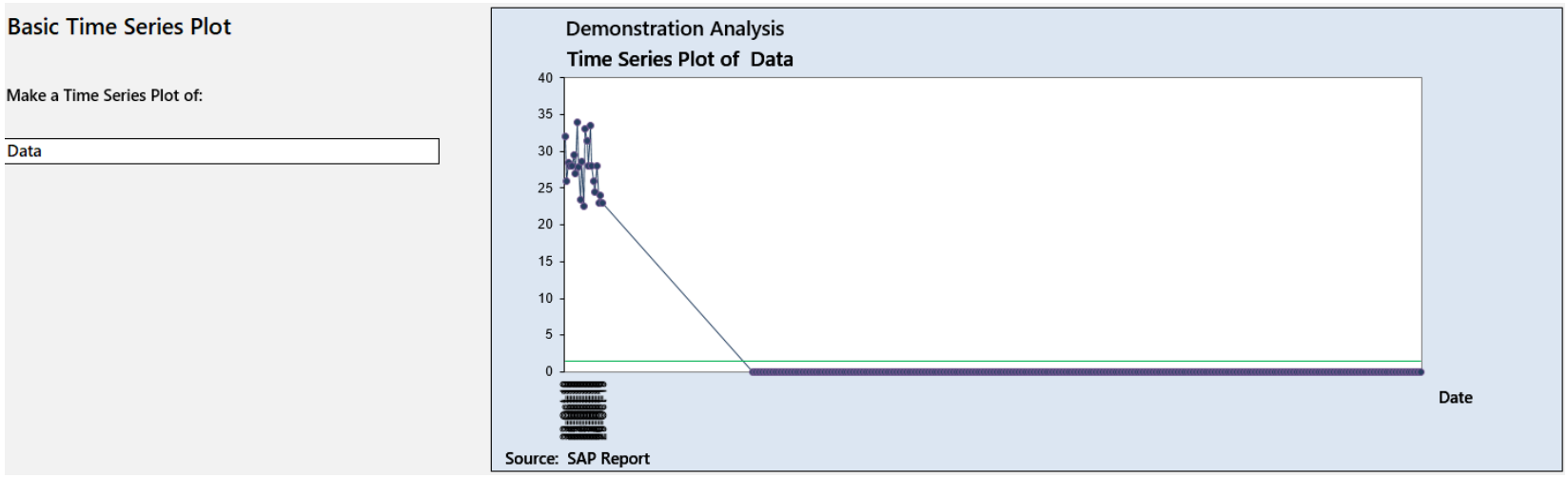
There is, however, one specific issue that we see from time to time

**Our Data**

| 19 | | Exclusions | Time axis | Variable 1 |
|---|---|---|---|---|
| 20 | Row # | Exclude from Control Charts (only) | Date | Data |
| 21 | 1 | | 22/10/2019 | 25 |
| 22 | 2 | | 22/10/2019 | 26 |
| 23 | 3 | | 22/10/2019 | 30.5 |
| 24 | 4 | | 22/10/2019 | 26 |
| 25 | 5 | | 23/10/2019 | 27 |
| 26 | 6 | | 23/10/2019 | 24.5 |
| 27 | 7 | | 23/10/2019 | 26 |
| 28 | 8 | | 23/10/2019 | 31 |
| 29 | 9 | | 24/10/2019 | 26.9 |
| 30 | 10 | | 24/10/2019 | 19.4 |
| 31 | 11 | | 24/10/2019 | 23.6 |
| 32 | 12 | | 24/10/2019 | 20.5 |
| 33 | 13 | | 25/10/2019 | 30 |
| 34 | 14 | | 25/10/2019 | 33.5 |
| 35 | 15 | | 25/10/2019 | 29 |
| 36 | 16 | | 25/10/2019 | 28.5 |
| 37 | 17 | | 28/10/2019 | 32 |
| 38 | 18 | | 28/10/2019 | 19.9 |
| 39 | 19 | | 28/10/2019 | 24.5 |
| 40 | 20 | | 28/10/2019 | 28 |
| 41 | 21 | | 29/10/2019 | 23 |

**Time Series Plot**

(or any other graph)




**What could be causing this strange, corrupted appearance?**

# Troubleshooting

## Graphs and Statistical Analysis

Watch out for the remains of previous data!

There appears to be 23 rows of data here…

| | Exclusions | Time axis | Variable 1 |
|---|---|---|---|
| Row # | Exclude from Control Charts (only) | Date | Data |
| 1 | | 22/10/2019 | 32 |
| 2 | | 22/10/2019 | 26 |
| 3 | | 22/10/2019 | 28.5 |
| 4 | | 22/10/2019 | 28 |
| 5 | | 23/10/2019 | 28 |
| 6 | | 23/10/2019 | 29.5 |
| 7 | | 23/10/2019 | 27 |
| 8 | | 23/10/2019 | 34 |
| 9 | | 24/10/2019 | 27.9 |
| 10 | | 24/10/2019 | 23.4 |
| 11 | | 24/10/2019 | 28.6 |
| 12 | | 24/10/2019 | 22.5 |
| 13 | | 25/10/2019 | 33 |
| 14 | | 25/10/2019 | 31.5 |
| 15 | | 25/10/2019 | 28 |
| 16 | | 25/10/2019 | 33.5 |
| 17 | | 28/10/2019 | 28 |
| 18 | | 28/10/2019 | 25.9 |
| 19 | | 28/10/2019 | 24.5 |
| 20 | | 28/10/2019 | 28 |
| 21 | | 29/10/2019 | 23 |
| 22 | | 29/10/2019 | 24 |
| 23 | | 29/10/2019 | 23 |
| 24 | | | |
| 25 | | | |
| 26 | | | |

… But scroll further down and we find the remains of some previous data, which the toolkit has found and has attempted to plot

| | |
|---|---|
| 107 | |
| 108 | |
| 109 | |
| 110 | |
| 111 | Day |
| 112 | Night |
| 113 | Day |
| 114 | Night |
| 115 | Day |
| 116 | Night |
| 117 | Day |
| 118 | Night |
| 119 | Day |
| 120 | Night |
| 121 | Day |
| 122 | Night |
| 123 | Day |
| 124 | Night |
| 125 | Day |
| 126 | Night |
| 127 | Day |
| 128 | Night |
| 129 | Day |
| 130 | Night |
| 131 | Day |
| 132 | Night |
| 133 | Day |
| 134 | Night |
| 135 | Day |
| 136 | Night |

When a graph looks completely wrong, check that you really have removed all your old data from the columns involved.


https://advancedanalyticssolutions.co.uk