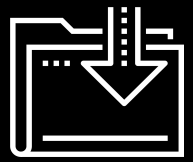
Charting a New Course with Excel

Data Boot Camp

Lesson 1.3




WELCOME

Instructor Demonstration

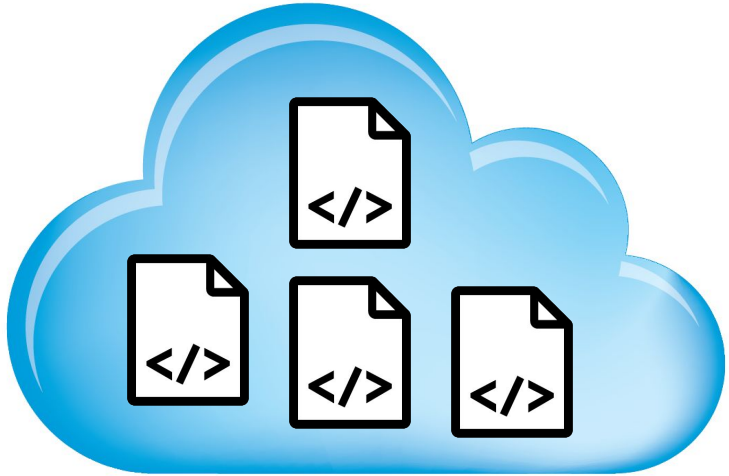
Adding Files to GitHub

GitHub Is a Hosting Service for Source Code

GitHub is a web interface for Git.

Git is version control software that can:

- Track source code history.
- Allow for collaboration on the same code files across a team or organisation.
- Easily update and roll back software versions.



GitHub is used by over 4 million organisations.

Proficiency in Git and GitHub are highly desired skills in many industries.

Git and Github

We will use Git and Github throughout the curriculum.

You will submit your Challenge assignments by using GitHub.

You will version control your individual project work by using Git.

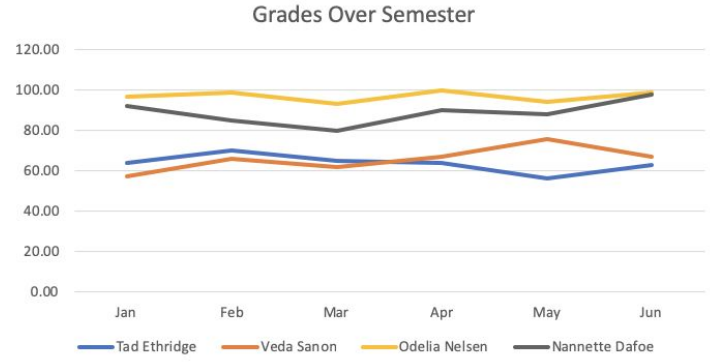
You will collaborate with teammates by using GitHub.

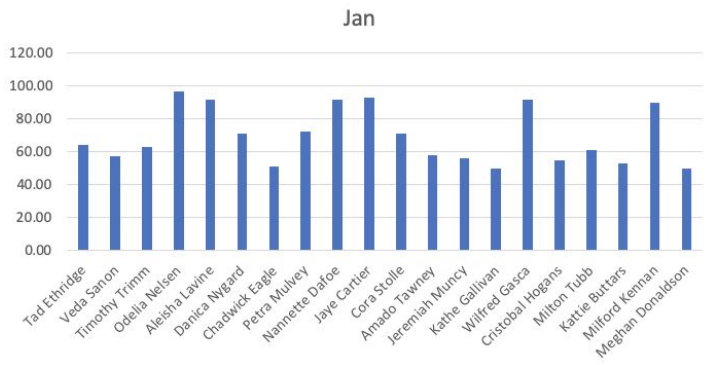
You should become proficient with the basic Git and GitHub functionality by the end of the curriculum.

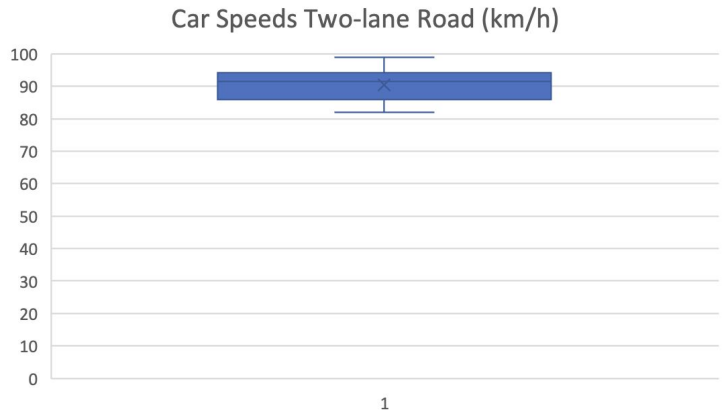
Instructor Demonstration

Basic Charting

Excel Visualisations





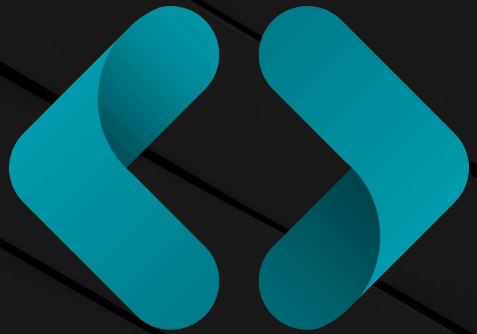


Excel Visualisations: Examples and Use Cases

In this activity, we will:

- Examine an example dataset.
- Select some data of interest.
- Visualise the selected data.
- Add labels and titles to the visualization.

Do not hesitate to ask questions.
The TAs will slack out images for each operating system.



Time to <code>

Activity: The Line and Bar Grades

For this activity, you'll take on the role of the teacher as you create bar and line graphs to visualise the grades of your class over a semester.

Suggested Time:

15 minutes

Activity: Line and Bar Grades

Instructions

- Create a series of bar graphs that visualise the grades of all the students in the class, with one graph for every month.
- Create a line graph by using all the data that can be used to compare students' grades across the semester.
- When creating the line graph, use filtering to drill down to an individual student’s performance.

Hint

When duplicating bar graphs, it helps to get the formatting and style of the chart as you want for the first graph (that is, for January). Then copy that chart, and reselect the data for each subsequent copy. That is, keep the style and format but change the included data.

Time’s Up! Let’s Review.

Instructor Demonstration

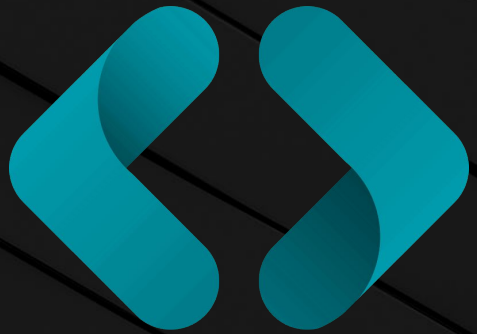
Scatter Plots and Trend Lines

The Scatter Plot: A Powerful Visualisation Tool

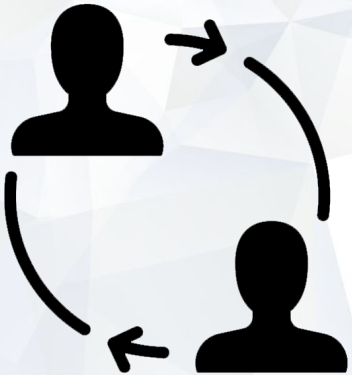
Visualises the comparison between two variables:

One variable	is located on the x-axis.
Another variable	is plotted on the y-axis.

- Each data point represents a pair of measurements.
- The measurements on a scatter plot are independent.
- A scatter plot can help us identify a positive or negative relationship between two variables.
- Adding a trend line to a scatter plot can further help us visualise this relationship.



Time to <code>



Partner Activity: Home Sales

For this activity, you will work in pairs to create a series of scatter plots that compare home prices against home attributes in the Adelaide, SA region.

Suggested Time:

10 minutes

Partner Activity: Home Sales

Instructions:

- Create a scatter plot that compares the price of the home with the square metres of the home (`sqm_living`). Make sure to add in axis titles, a chart title, and a trend line.
- Create a scatter plot that compares the price of the home with the number of bedrooms. Make sure to add in axis titles, a chart title, and a trend line.
- Create a scatter plot that compares the price of the home with the number of bathrooms. Make sure to add in axis titles, a chart title, and a trend line.
- Go back into each of your charts, and modify the value range on each axis so that they are consistent across charts.

We want the axes to match so the data is conveyed in a consistent, truthful manner.

Time’s Up! Let’s Review.

Instructor Demonstration

The Need to Filter

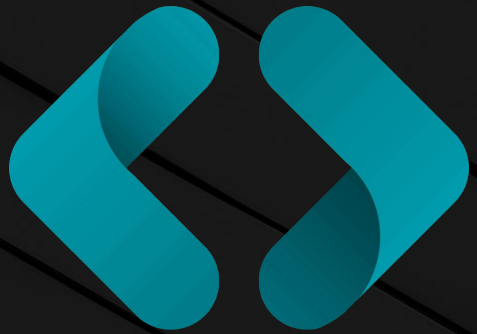
Do You Notice Anything About the Following Data?

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	GroupAN	GroupId	Year	DateTimeStart	DateTimeEnd	Latitude	Longitude	Observer	IceConcentra	IceForm	DistanceToGroup	FlightDistance	ApproachDirection	GroupSize	GroupSizingMethod	MMPAtake	Observation
2	4	NM-2013-06	2013	6/6/13 17:29	6/6/13 18:31	62.47	-168.78	John Citta, Mary Cody, Chadwick Jay, Mark Nelson, Lori Quakenbush	0.3	Medium Floe	400	280	315	42	Count	42	NM
3	5	NM-2013-06	2013	6/6/13 18:34	6/6/13 19:10	62.47	-168.78	John Citta, Mary Cody, Chadwick Jay, Mark Nelson, Lori Quakenbush	0.3	Medium Floe	223	200	315	29	Count	29	NM
4	6	NM-2013-06	2013	6/6/13 19:10	6/6/13 19:10	62.47	-168.78	John Citta, Mary Cody, Chadwick Jay, Mark Nelson, Lori Quakenbush	0.3	Medium Floe	200		315	2	Count	0	NM
5	7	NM-2013-06	2013	6/6/13 21:43	6/6/13 21:50	62.52	-168.76	John Citta, Mary Cody, Chadwick Jay, Mark Nelson, Lori Quakenbush	0.1	Small Floe	218	218	70	2	Count	1	NM
6	8	NM-2013-06	2013	6/6/13 21:43	6/6/13 21:50	62.52	-168.76	John Citta, Mary Cody, Chadwick Jay, Mark Nelson, Lori Quakenbush	0.1	Small Floe	218	218	70	2	Count	1	NM
7	9	NM-2013-06	2013	6/6/13 22:32	6/6/13 22:53	62.51	-168.75	John Citta, Mary Cody, Chadwick Jay, Mark Nelson, Lori Quakenbush	0.1	Small Floe	200	30	209	14	Count	14	NM
8	12	NM-2013-06	2013	6/7/13 14:12	6/7/13 15:15	62.54	-168.3	John Citta, Mary Cody, Chadwick Jay, Mark Nelson, Lori Quakenbush	0.6	Small Floe		100	183	11	Count	2	NM
9	13	NM-2013-06	2013	6/7/13 14:12	6/7/13 15:15	62.54	-168.3	John Citta, Mary Cody, Chadwick Jay, Mark Nelson, Lori Quakenbush	0.6	Small Floe		100	183	5	Count	1	NM
10	14	S2-2013-06-	2013	6/6/13 16:19		62.45	-168.87	Geoffrey Cook, Jason Everett, Joel Garlich-Miller	0.3	Ice Cake	20	20		1	Count	1	S2
11	15	NM-2013-06	2013	6/7/13 14:12	6/7/13 15:15	62.54	-168.3	John Citta, Mary Cody, Chadwick Jay, Mark Nelson, Lori Quakenbush	0.6	Small Floe		100	183	8	Count	2	NM
12	16	NM-2013-06	2013	6/7/13 14:12	6/7/13 15:15	62.54	-168.3	John Citta, Mary Cody, Chadwick Jay, Mark Nelson, Lori Quakenbush	0.6	Small Floe		100	183	10	Count	3	NM
13	17	NM-2013-06	2013	6/7/13 16:35	6/7/13 17:11	62.53	-168.31	John Citta, Mary Cody, Chadwick Jay, Mark Nelson, Lori Quakenbush	0.4	Ice Cake	400	200	138	16	Count	16	NM
14	18	NM-2013-06	2013	6/7/13 16:35	6/7/13 17:11	62.53	-168.31	John Citta, Mary Cody, Chadwick Jay, Mark Nelson, Lori Quakenbush	0.4	Ice Cake	400	200	138	11	Count	9	NM
15	19	NM-2013-06	2013	6/7/13 18:00	6/7/13 18:05	62.53	-168.34	John Citta, Mary Cody, Chadwick Jay, Mark Nelson, Lori Quakenbush	0.4	Small Floe	450		300	2	Count	0	NM
16	20	NM-2013-06	2013	6/7/13 18:50	6/7/13 18:53	62.53	-168.35	John Citta, Mary Cody, Chadwick Jay, Mark Nelson, Lori Quakenbush	0.2	Ice Cake	300	300	342	5	Count	1	NM
17	21	NM-2013-06	2013	6/7/13 19:31	6/7/13 19:46	62.52	-168.36	John Citta, Mary Cody, Chadwick Jay, Mark Nelson, Lori Quakenbush	0.3	Ice Cake	400	182	236	8	Count	8	NM
18	22	NM-2013-06	2013	6/7/13 19:50	6/7/13 20:29	62.35	-168.37	John Citta, Mary Cody, Chadwick Jay, Mark Nelson, Lori Quakenbush	0.3	Ice Cake	250	250	103	3	Count	3	NM
19	23	NM-2013-06	2013	6/7/13 19:50	6/7/13 20:29	62.35	-168.37	John Citta, Mary Cody, Chadwick Jay, Mark Nelson, Lori Quakenbush	0.3	Ice Cake	250	200	103	8	Count	8	NM
20	24	NM-2013-06	2013	6/7/13 19:50	6/7/13 20:29	62.35	-168.37	John Citta, Mary Cody, Chadwick Jay, Mark Nelson, Lori Quakenbush	0.3	Ice Cake	250	103	103	16	Count	16	NM
21	25	NM-2013-06	2013	6/7/13 19:50	6/7/13 20:29	62.35	-168.37	John Citta, Mary Cody, Chadwick Jay, Mark Nelson, Lori Quakenbush	0.3	Ice Cake	250	103	103	28	Count	28	NM
22	26	NM-2013-06	2013	6/7/13 20:34	6/7/13 20:39	62.52	-168.36	John Citta, Mary Cody, Chadwick Jay, Mark Nelson, Lori Quakenbush	0.3	Ice Cake	400		182	2	Count	0	NM
23	27	NM-2013-06	2013	6/7/13 20:41	6/7/13 21:05	62.52	-168.36	John Citta, Mary Cody, Chadwick Jay, Mark Nelson, Lori Quakenbush	0.3	Ice Cake	300	150	310	9	Count	4	NM
24	28	NM-2013-06	2013	6/7/13 20:41	6/7/13 21:05	62.52	-168.36	John Citta, Mary Cody, Chadwick Jay, Mark Nelson, Lori Quakenbush	0.3	Ice Cake	300	150	310	3	Count	0	NM
2078	2176	S3-2015-06-	2015	6/20/15 18:23		70.99	-165.23	Alexi, Yura Burkanov, Maxim, Z Sergei						4		4	S3
2079	2177	S3-2015-06-	2015	6/20/15 18:54		70.99	-165.24	Alexi, Yura Burkanov, Maxim, Z Sergei						2		2	S3
2080	2178	S3-2015-06-	2015	6/20/15 19:07		70.99	-165.24	Alexi, Yura Burkanov, Maxim, Z Sergei						2		2	S3
2081	2179	S3-2015-06-	2015	6/20/15 10:26		70.99	-165.23	Alexi, Yura Burkanov, Maxim, Z Sergei						5		5	S3
2082	2180	S3-2015-06-	2015	6/6/15 0:00				Alexi, Yura Burkanov, Maxim, Z Sergei						10		10	S3
2083	2181	S3-2015-05-	2015	5/30/15 23:45				Alexi, Yura Burkanov, Maxim, Z Sergei						2		2	S3

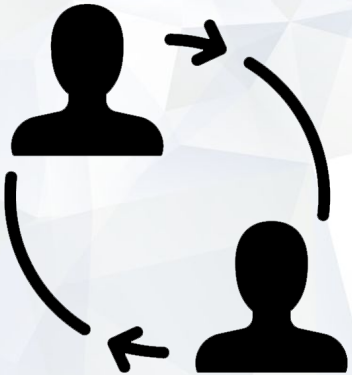
There Is Lots of Missing and Unneeded Data

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	GroupAN	GroupId	Year	DateTimeStart	DateTimeEnd	Latitude	Longitude	Observer	IceConcentr	IceForm	DistanceToGroup	FlightDistance	ApproachDirection	GroupSize	GroupSizingMethod	MMPAtake	Observation
2	4	NM-2013-0€	2013	6/6/13 17:29	6/6/13 18:31	62.47	-168.78	John Citta, Mary Cody, Chadwick Jay, Mark Nelson, Lori Quakenbush	0.3	Medium Floe	400	280	315	42	Count	42	NM
3	5	NM-2013-0€	2013	6/6/13 18:34	6/6/13 19:10	62.47	-168.78	John Citta, Mary Cody, Chadwick Jay, Mark Nelson, Lori Quakenbush	0.3	Medium Floe	223	200	315	29	Count	29	NM
4	6	NM-2013-0€	2013	6/6/13 19:10	6/6/13 19:10	62.47	-168.78	John Citta, Mary Cody, Chadwick Jay, Mark Nelson, Lori Quakenbush	0.3	Medium Floe	200		315	2	Count	0	NM
5	7	NM-2013-0€	2013	6/6/13 21:43	6/6/13 21:50	62.52	-168.76	John Citta, Mary Cody, Chadwick Jay, Mark Nelson, Lori Quakenbush	0.1	Small Floe	218	218	70	2	Count	1	NM
6	8	NM-2013-0€	2013	6/6/13 21:43	6/6/13 21:50	62.52	-168.76	John Citta, Mary Cody, Chadwick Jay, Mark Nelson, Lori Quakenbush	0.1	Small Floe	218	218	70	2	Count	1	NM
7	9	NM-2013-0€	2013	6/6/13 22:32	6/6/13 22:53	62.51	-168.75	John Citta, Mary Cody, Chadwick Jay, Mark Nelson, Lori Quakenbush	0.1	Small Floe	200	30	209	14	Count	14	NM
8	12	NM-2013-0€	2013	6/7/13 14:12	6/7/13 15:15	62.54	-168.3	John Citta, Mary Cody, Chadwick Jay, Mark Nelson, Lori Quakenbush	0.6	Small Floe		100	183	11	Count	2	NM
9	13	NM-2013-0€	2013	6/7/13 14:12	6/7/13 15:15	62.54	-168.3	John Citta, Mary Cody, Chadwick Jay, Mark Nelson, Lori Quakenbush	0.6	Small Floe		100	183	5	Count	1	NM

- Most datasets contain multiple variables and factors.
- When exploring a dataset, it can be hard to determine which data is useful.
- It can be hard to find the data of interest.
- So, we need to filter our data.



Time to <code>



Partner Activity: Filtering Home Sales

For this activity, you'll create a filtered chart that visualises the increases in waterfront properties over time in the Adelaide area.

Suggested Time:

20 minutes

Partner Activity: Filtering Home Sales

In this activity, you'll work in pairs to create a filtered chart that visualises the increases in waterfront properties over time in the Adelaide area.

Instructions:

- Use the Adelaide Home Sales dataset provided.
- Examine the data and check out the available columns.
- Create a line graph that shows the price trend of waterfront homes in Adelaide by the age of the home.

Time’s Up! Let’s Review.


Break

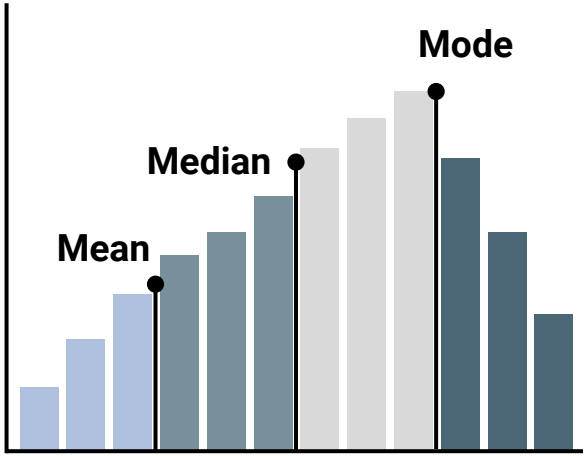
Instructor Demonstration

Variance, Standard Deviation and Z-Score

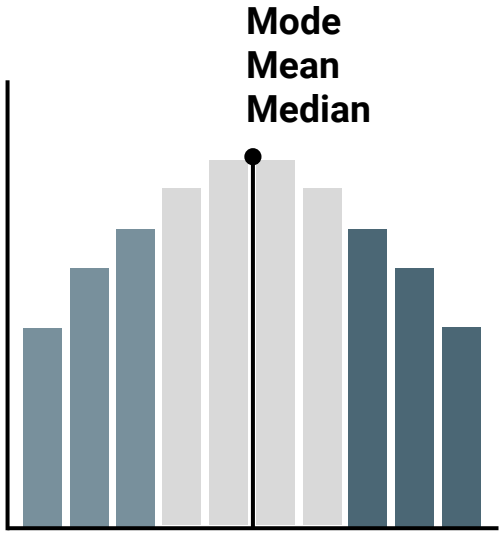
Quick Refresher

What are the three measures of central tendency?

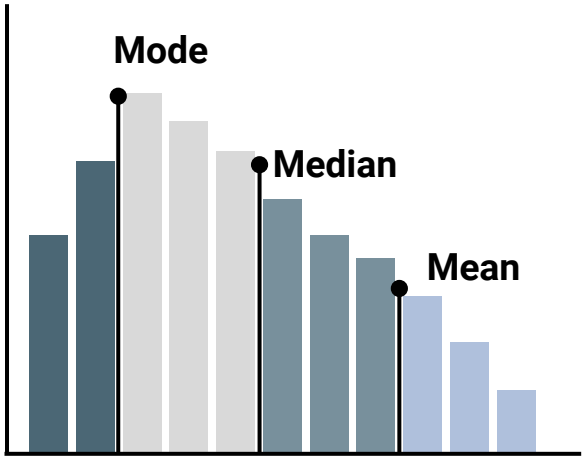
Mean, Median and Mode



Left Skew



Normal Distribution



Right Skew

What are the measures of central tendency used for?

To describe the centre of a dataset

How do we describe the variability of a dataset?

Variability of a Dataset

The three summary statistics metrics for describing variability:

01 Variance

02 Standard deviation

03 Z-score

Variance

- Describes how far values in the dataset are from the mean.
- Describes how much variation exists in the data.
- Considers the distance of each value in the dataset from the centre of the data.

The value of the one observation x_i

The mean value of all observations $\bar{x}$

Sample variance $S^2 = \dfrac{\sum (x_i - \bar{x})^2}{n - 1}$

The number of observations n

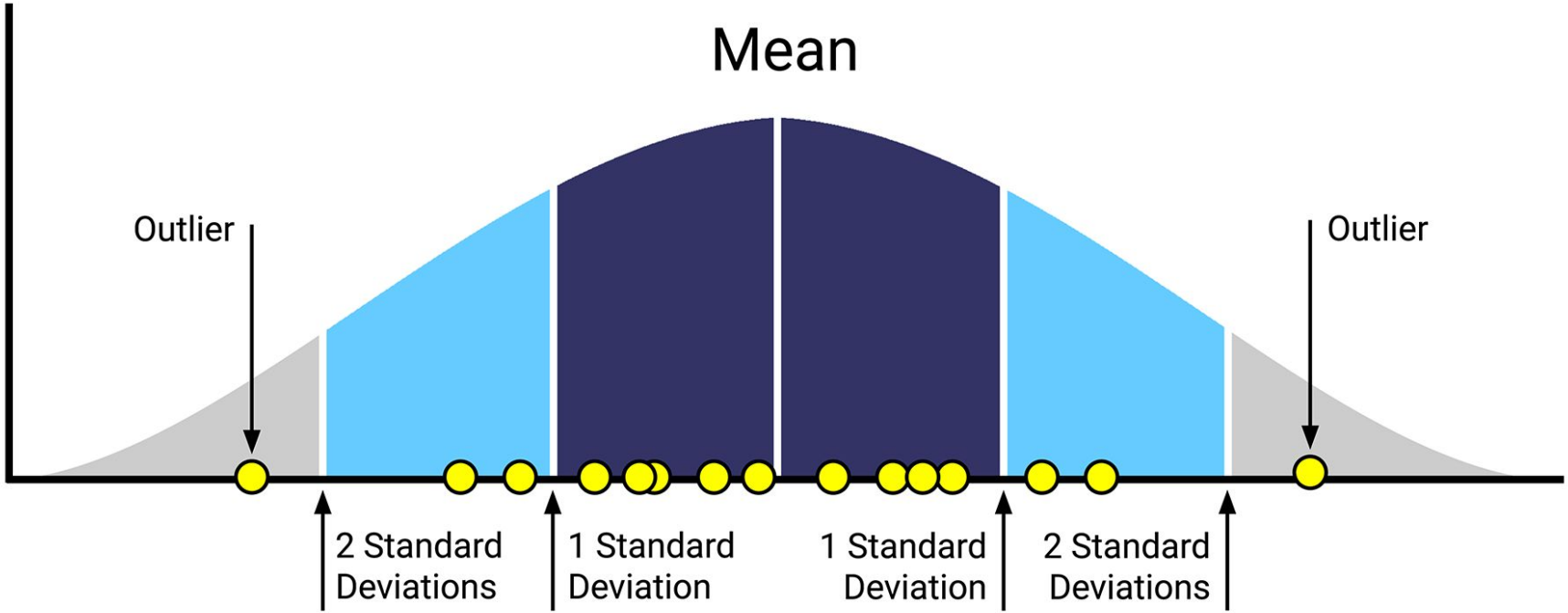
Standard Deviation

- Describes how spread out the data is from the mean.
- Gets calculated from the square root of the variance.
- Uses the same unit of measurement as the mean.

Standard deviation $\sigma = \sqrt{S^2}$ **Variance**

Standard Deviation

The standard deviation is the square root of the variance and a measure that quantifies the dispersion of a set of observations.

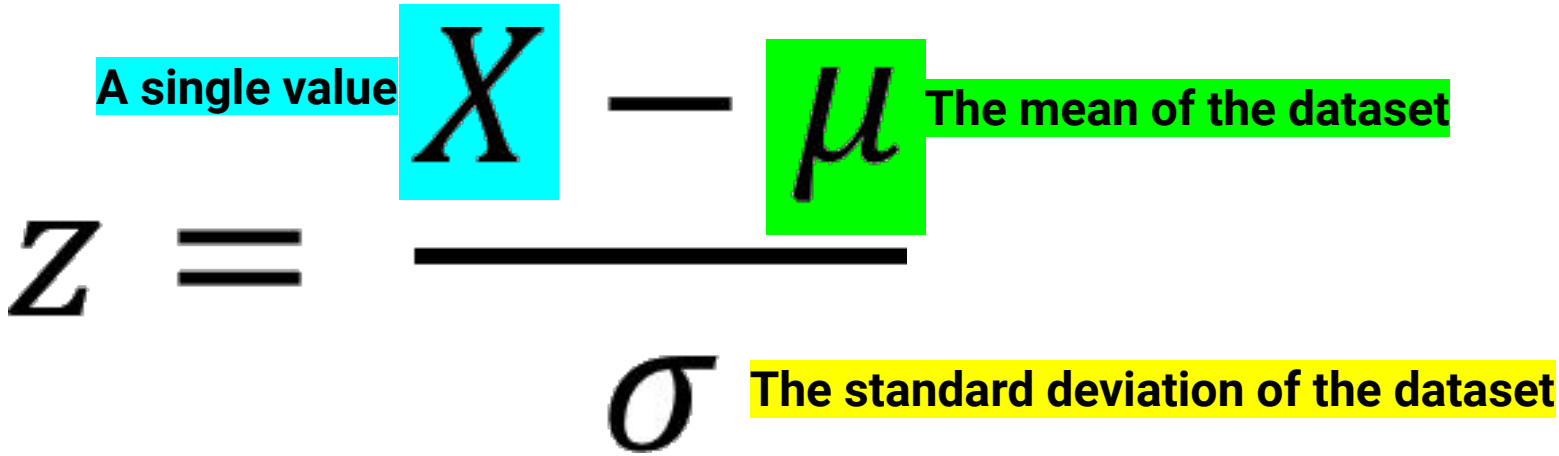


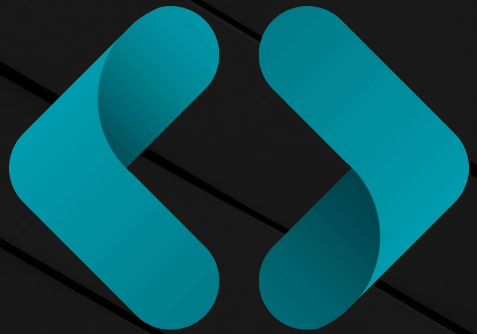
Z-Score

The z-score describes the distance of a single value from the mean of the dataset. This distance is in standard deviations and can be positive or negative.

If negative	The value is less than the mean.
If positive	The value is greater than the mean. .

The smaller the z-score, the closer the value is to the mean.





Time to <code>

Activity: Variance, Standard Deviation, and Z-Score Review

In this activity, you will practice summarising the variability of a dataset by using employment data from the Australian Bureau of Statistics.

Suggested Time:

15 minutes

Activity: Variance, Standard Deviation, and Z-Score Review

Instructions:

- Open the variance_review.xlsx workbook that contains your raw data.
- Make a copy of the worksheet. This way, if you make any mistakes, you will have a backup of the original dataset.
- Create a new sheet in the workbook, and name the sheet "Summary Table".
    - If you are uncertain of how to make a new sheet in an Excel workbook, refer to the Insert or delete a worksheet Microsoft Office support page.
- Within the new sheet, create a State column, which contains the following states: New South Wales, Victoria, Queensland, South Australia, Western Australia
- For each state, determine the mean, variance and standard deviation for the overall median income.
- Based upon your calculated summary statistics, determine which state had the highest average median income. What was the median income?
- Based upon your calculated summary statistics, determine which state had the greatest difference in median income across all of its statistical areas.
- Based upon your calculated summary statistics, determine which state had the lowest variance in median income. What was the median income?
- Create a new sheet in the workbook, and name the sheet "Western Australia Z-Scores".
- Within this new sheet, copy over the Statistical Area and Median Income columns from the raw data for only the state Western Australia.
- Calculate the z-score for the overall median income by statistical area across the whole state.
- Based upon your calculated z-scores, determine which statistical area had the largest difference in median income from the mean of the state.

Time’s Up! Let’s Review.

Instructor Demonstration

Quantiles, Outliers, and Box Plots

Real-World Data

Be careful when describing real-world data:

- Real-world data can contain extreme values.
- Some summary statistics, such as the mean, take into account all the values of a dataset.
- Extreme values can skew these statistics.

How can we summarise real-world data?

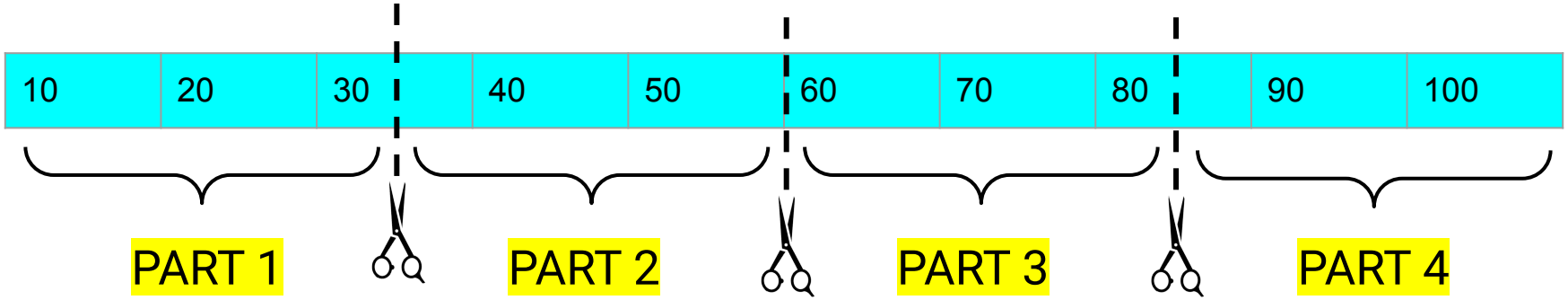
Quantiles: Used to Describe Segments of a Dataset

Quantiles separate a sorted dataset into equally sized fragments.

The two most popular types of quantiles are **quartiles** and **percentiles.**

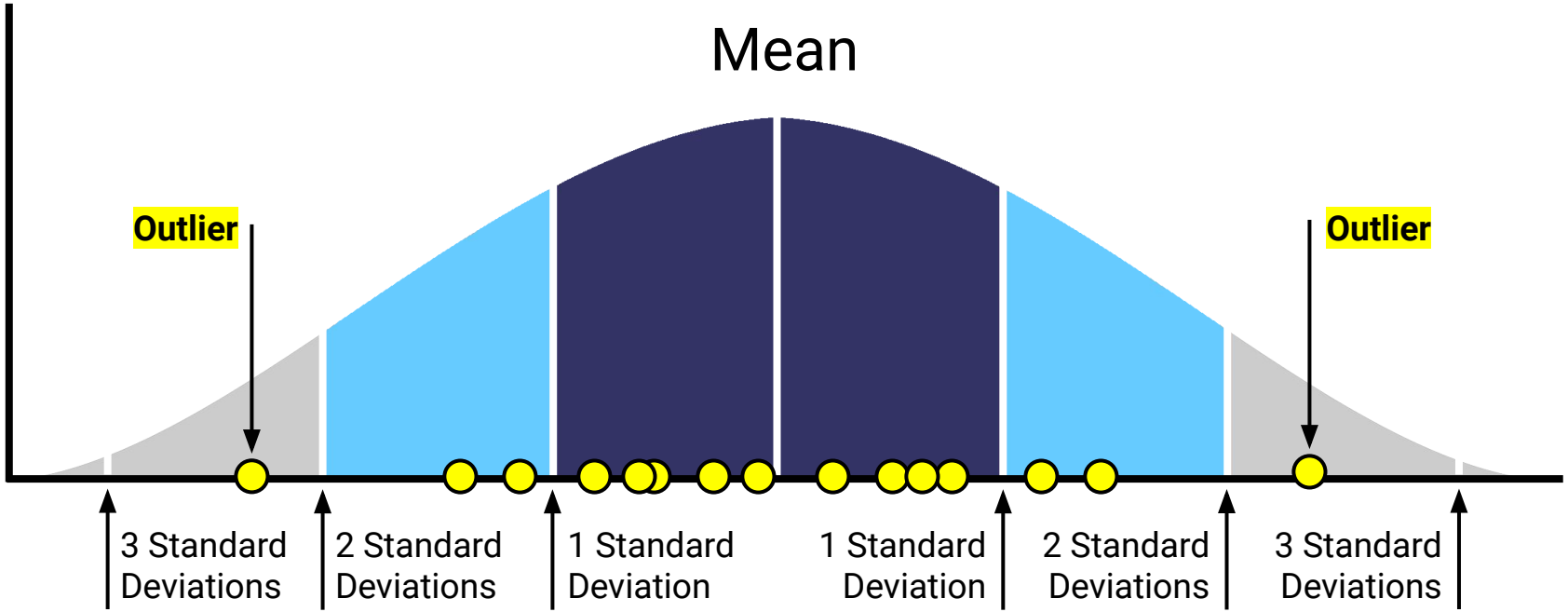
01 **Quartiles** divide the dataset into four equally sized parts.

02 **Percentiles** divide the dataset into 100 equally sized parts.



Outliers

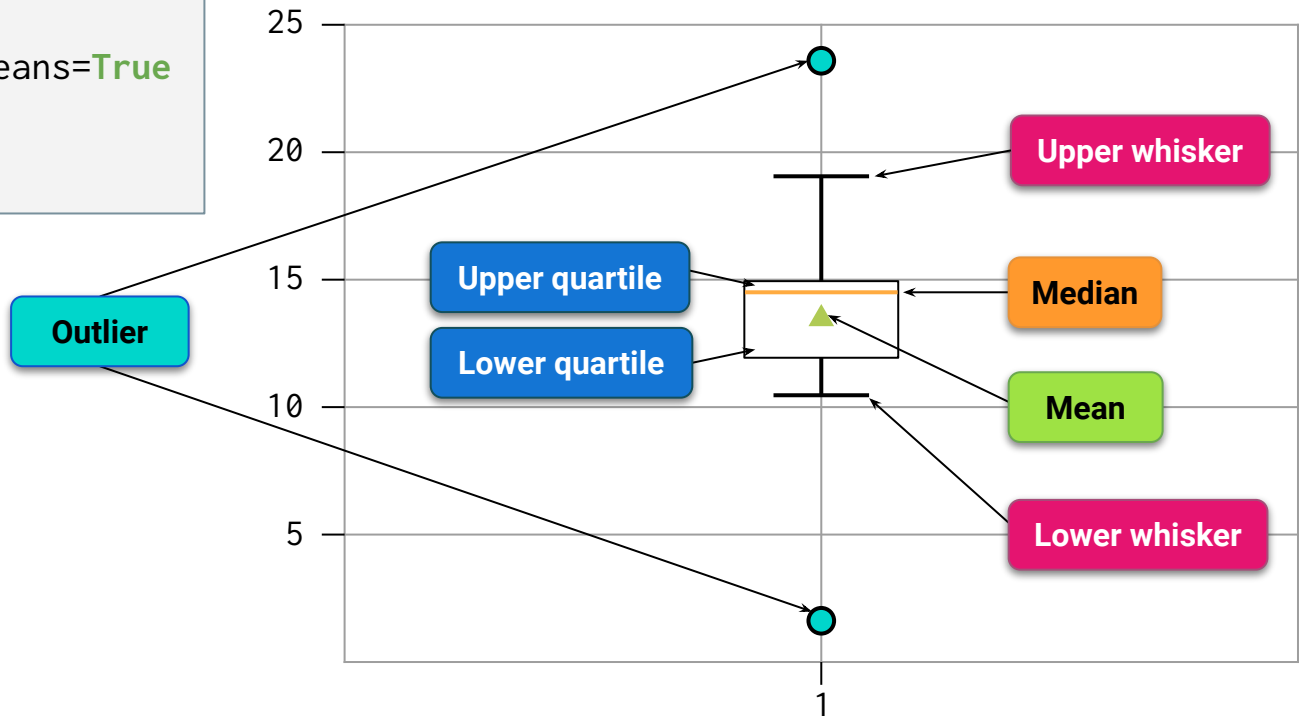
Suspicious values are called potential outliers. An outlier is a data point that differs from the rest of a dataset. Outliers can inaccurately skew a dataset.



Qualitatively

Use **box-and-whisker plots** to visually identify potential outliers.

```
# Create box plot
plt.boxplot(arr, showmeans=True
plt.grid()
plt.show()
```



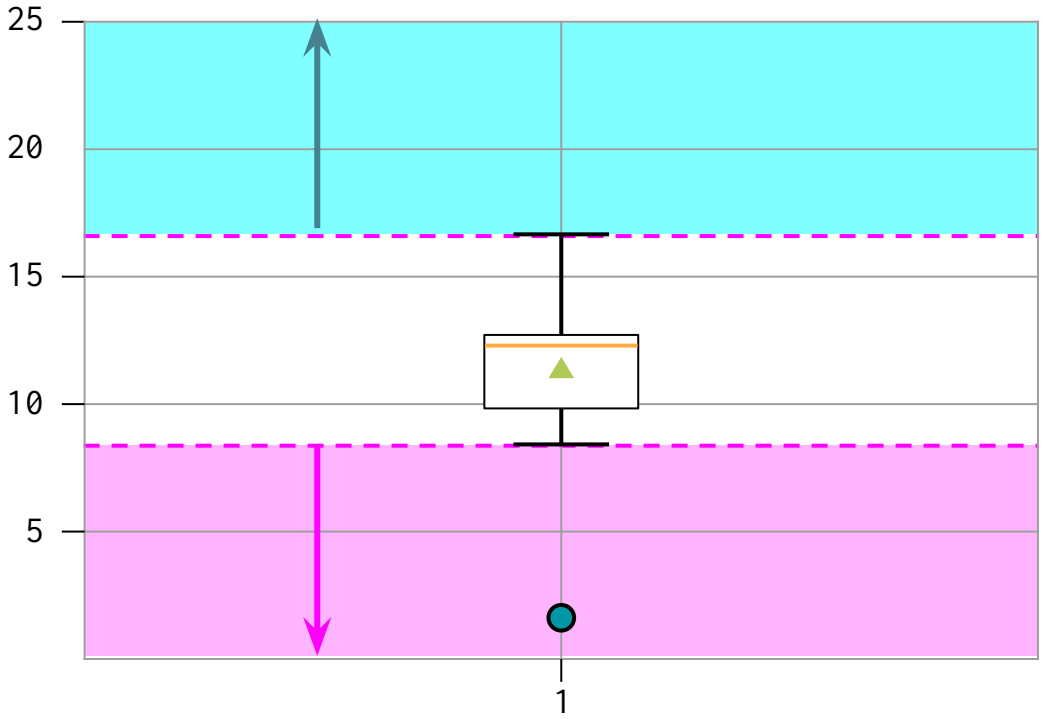
Quantitatively

Determine the outlier boundaries in a dataset by using the **1.5 × IQR rule**.

The IQR is the range between the first and the third quartile.

Anything less than, or below, Quartile 1 − (1.5 × IQR) might be an outlier.

Anything greater than, or above, Quartile 3 + (1.5 × IQR) might be an outlier.



Activity: Cereal Outliers

In this activity, you will be investigating data from a dataset called 80 Cereals. Your task is to search through the ratings of each product and determine if there are any potential outliers in the dataset.

Suggested Time:

10 minutes

Activity: Cereal Outliers

Instructions:

- Open up the activity workbook, and familiarise yourself with the raw data.
  - File: Unsolved/Outliers_Activity_Unsolved.xlsx
- Create a new worksheet, and name it "Outlier Testing".
- In the "Outlier Testing" worksheet, create a summary statistics table of the “rating” for the following statistics:
  - Mean
  - Median
  - Minimum value
  - Maximum value
  - First quartile
  - Third quartile
  - Interquartile Range
- Using the calculations from the table, determine the lower and upper boundaries of the 1.5*IQR rule.
- Determine if there are any products whose rating falls outside of the 1.5*IQR boundaries. List those products and their rating on the worksheet.
- Create a box plot of the rating for all products.
  - **Note**: Be sure to add a title and label your *y*-axis.

Time's Up! Let's Review.

Instructor Demonstration

Excel's Statistics Add-On

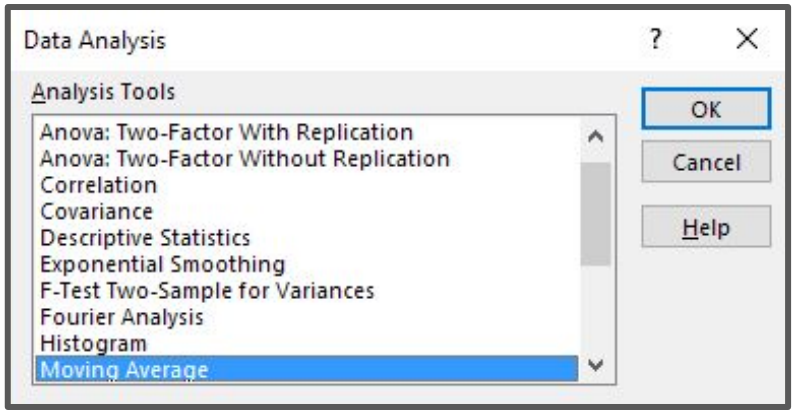
Excel: A Great Foundational Tool

Up to this point, we have covered only the summary statistics.

We Can Use Excel for Even More Statistics

The Excel Analysis ToolPak contains:

- T-tests
- Correlation tests
- Regression tests
- ANOVA



We will cover all of these functions throughout the course.

Analysis ToolPak: Not Designed for In-Depth Data Analytics

Excel struggles with medium to large datasets:

- >200 columns or >100,000 rows
- Depends on the machine

Excel does not automatically record parameters for statistical tests.

Excel's Analysis ToolPak should be used for:

- Gut checks
- One-off analyses

Install and Use the Analysis ToolPak: Mac

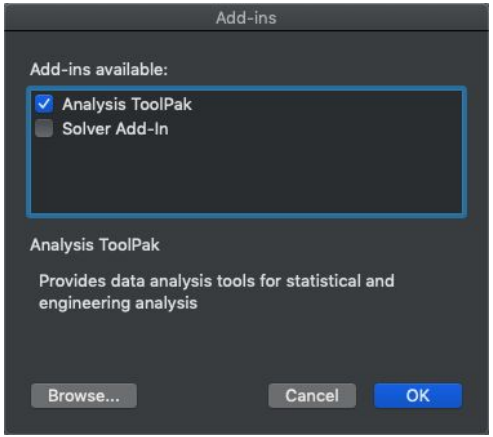
To install:

01 In Excel, go to the Tools menu.

02 Select the Excel Add-Ins option.

03 Enable the Analysis ToolPak option.

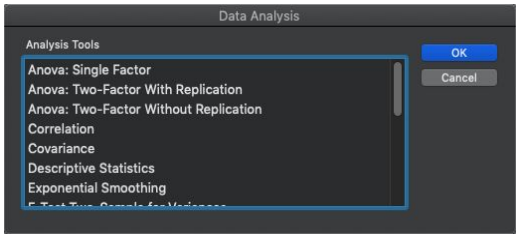
04 Click OK.



To use:

01 In Excel, go to the Data menu.

02 Select the Data Analysis option.



Install and Use the Analysis ToolPak: PC

To install:

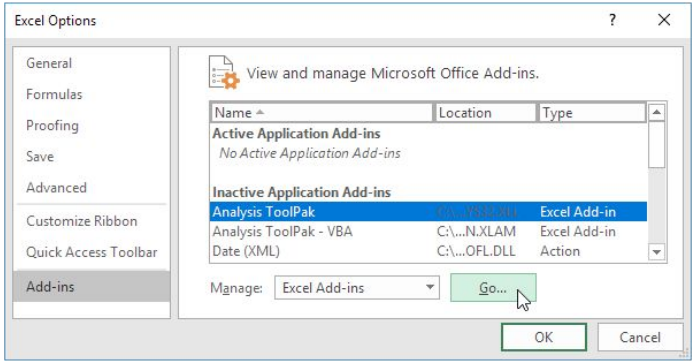
01 Click the File tab.

02 Go to Options.

03 Select the Add-Ins category.

04 In the Manage box, select Excel Add-ins, and then click Go.

05 In the Add-Ins box, enable the Analysis ToolPak and click OK.

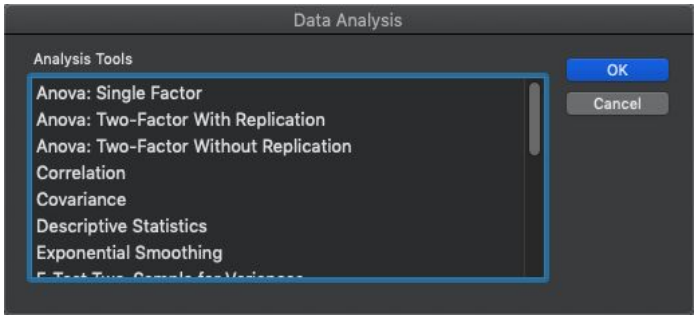


To use:

01 In Excel, go to the Data menu.

02 Go to the Analyze section.

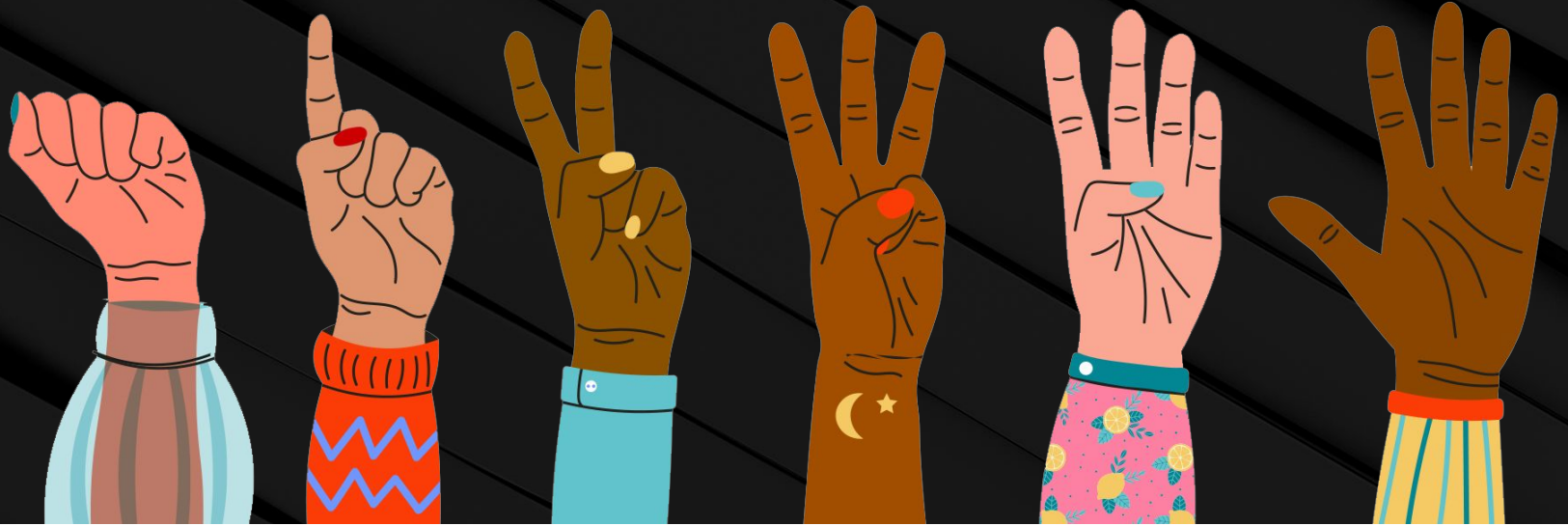
03 Select the Data Analysis option.



Questions?

FIST TO FIVE:

Who feels comfortable with plotting figures in Excel?



FIST TO FIVE:

Who feels comfortable calculating summary statistics in Excel?