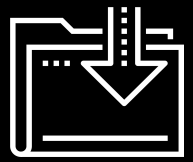
Excel Plotting

Data Boot Camp

Lesson 1.3




WELCOME

We are off to the races!


This will be you at the end of class.

Instructor Demonstration

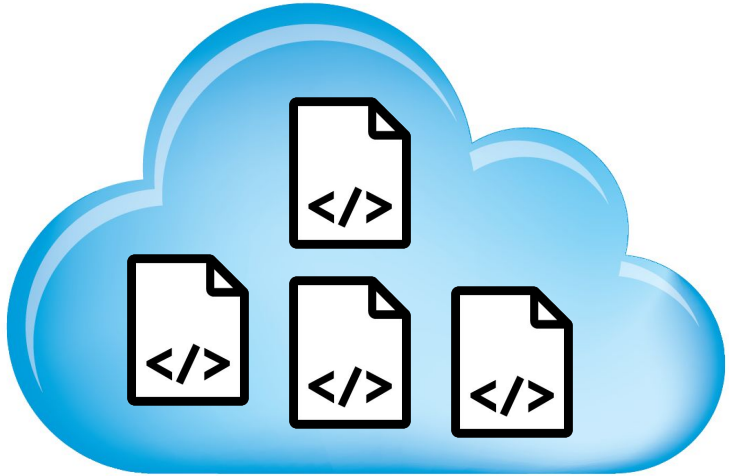
Adding Files to Github

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 organization
- Easily update and roll back software versions



GitHub is used by over 4 million organizations.

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

Git and Github

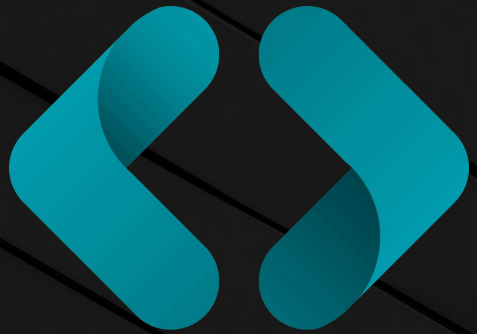
We will use Git and Github throughout the curriculum

You will submit your homework assignments using Github.

Your individual project work will be version controlled using Git.

You will be collaborating with teammates using Github.

By the end of the curriculum, you should be proficient with the basic Git and Github functionality.

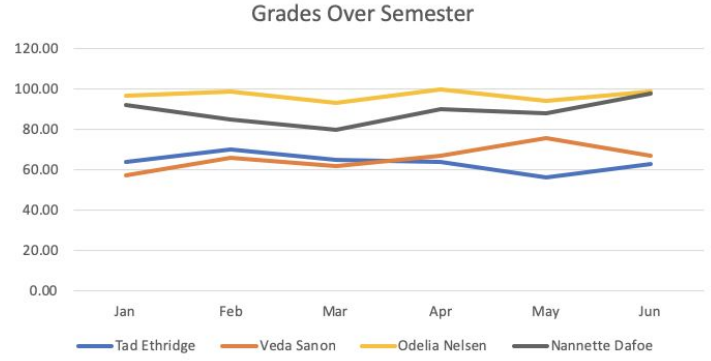


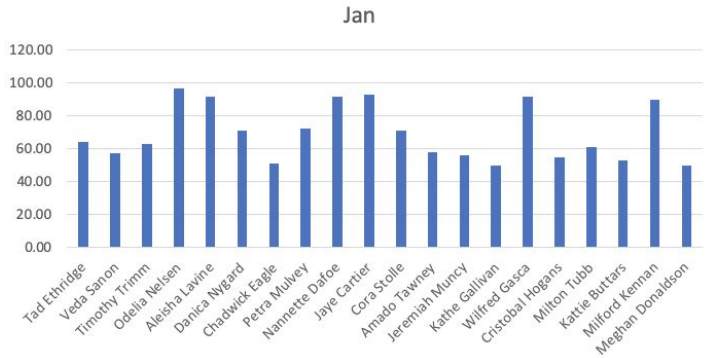
Time to <code>

Instructor Demonstration

Basic Charting

It is time to learn Excel visualizations!



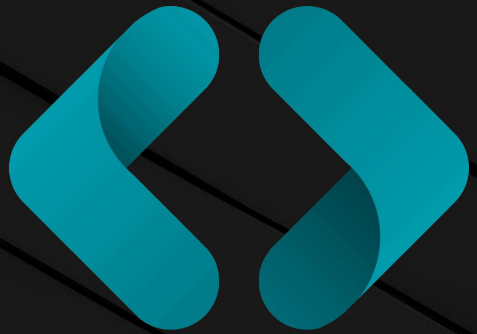


We will look at a few examples and use cases

In this activity, we will:

- Look at an example data set
- Select data of interest
- Visualize selected data
- Add labels and titles to our visualization

Do not hesitate to ask questions.
Our 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 visualize your class's grades over a semester.

Suggested Time:

15 minutes

Activity: Line and Bar Grades

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

Instructions:

- Create a series of bar graphs that visualize the grades of all students in the class, with one graph for every month.
- Create a line graph using all of the data that can be used to compare students' grades across the semester.
- Use filtering in the line graph to allow you to drill down to a specific student's progress throughout the semester.

Hint:

When duplicating bar graphs, it pays to get the formatting and look of the chart where you want it for the first graph (e.g., for January), and to then copy that chart and re-select the data for the subsequent copies (keeping the style and format, but just changing the data).

Time's Up! Let's Review.

Instructor Demonstration

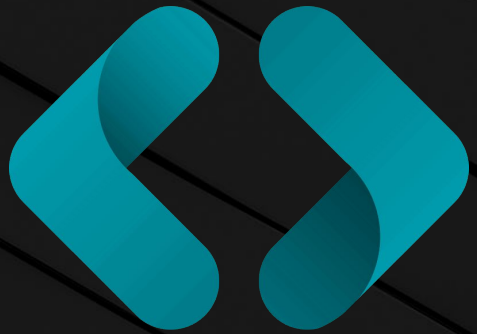
Scatter Plots and Trend Lines

Scatter plots are a powerful visualization tool!

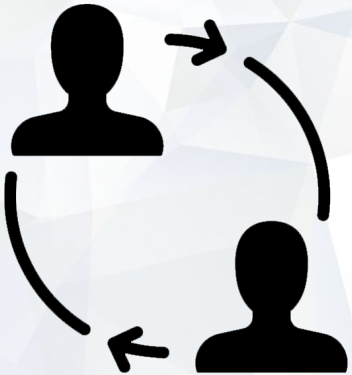
Visualizes the comparison between two variables:

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

- Each data point represents a pair of measurements
- Measurements on a scatter plot are independent
- Scatter plots can help to identify positive or negative relationships between two variables
- Adding a trend line to a scatterplot can visualize this relationship even easier!



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 in the St. Louis, MO, region.

Suggested Time:

15 minutes

Partner Activity: Home Sales

Instructions:

- Create a scatter plot that compares the price of the home with the square feet of the home (`sqft_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

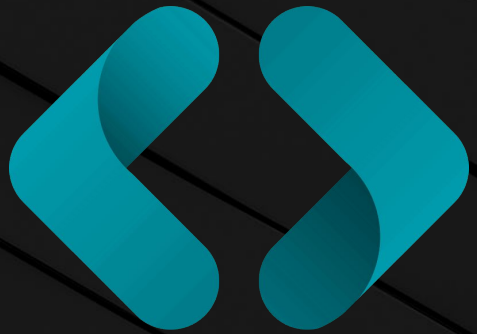
Do you notice anything about the following data?

	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

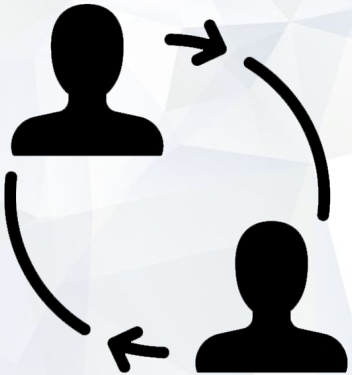
There is a LOT of missing and unneeded data

	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 data sets contain multiple variables and factors
- It can be difficult to determine what data is useful when exploring a data set
- It can be hard to locate data of interest
- We need to filter our data



Time to <code>



Partner Activity: Filtering Home Sales

For this activity, you'll create a filtered chart that visualizes increases in waterfront properties over time in the St. Louis Area.

Suggested Time:

15 minutes

Partner Activity: Filtering Home Sales

In this activity, you will pair up with one of your classmates in order to create a filtered chart that visualizes increases in waterfront properties over time in the St. Louis Area.

Instructions:

- Use the St. Louis 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 St. Louis by the age of the home.

Time's Up! Let's Review.


Break

Instructor Demonstration

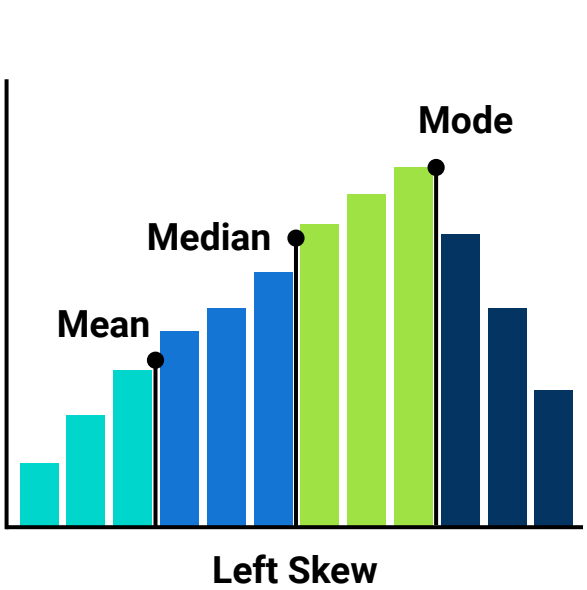
Variance, Standard Deviation and Z-Score

Quick Refresher

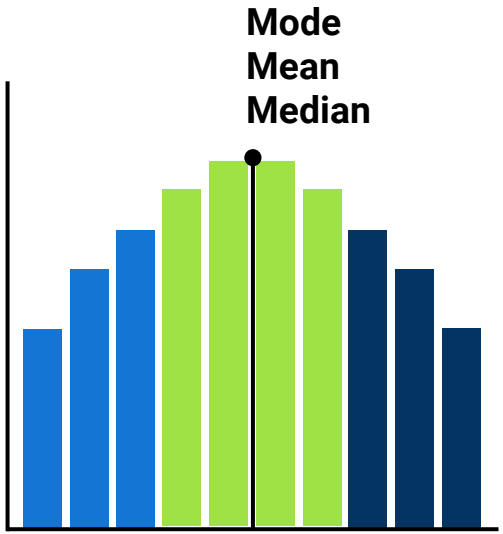
What are the three measures of central tendency?

The mean, median and mode.

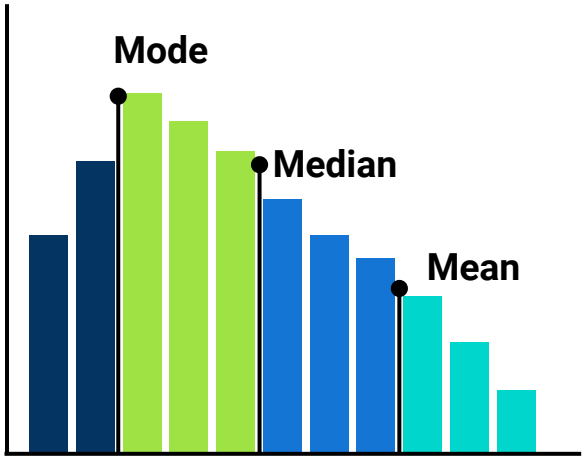
The mean, median and mode.



Left Skew



Normal Distribution



Right Skew

What are the measures of central tendency used for?

Metrics used to describe the center of a data set.

How do you describe the variability of a data set?

Variability of a Data Set

Three summary statistics metrics for describing variability:

01 Variance

02 Standard Deviation

03 Z-Score

Variance

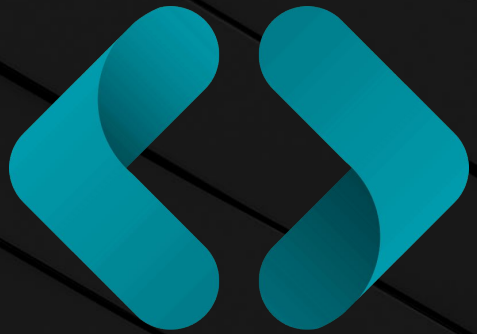
- Used to describe how far values in the data set are from the mean
- Describes how much variation exists in the data
- Variance considers the distance of each value in the data set from the center of the data

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



Time to <code>

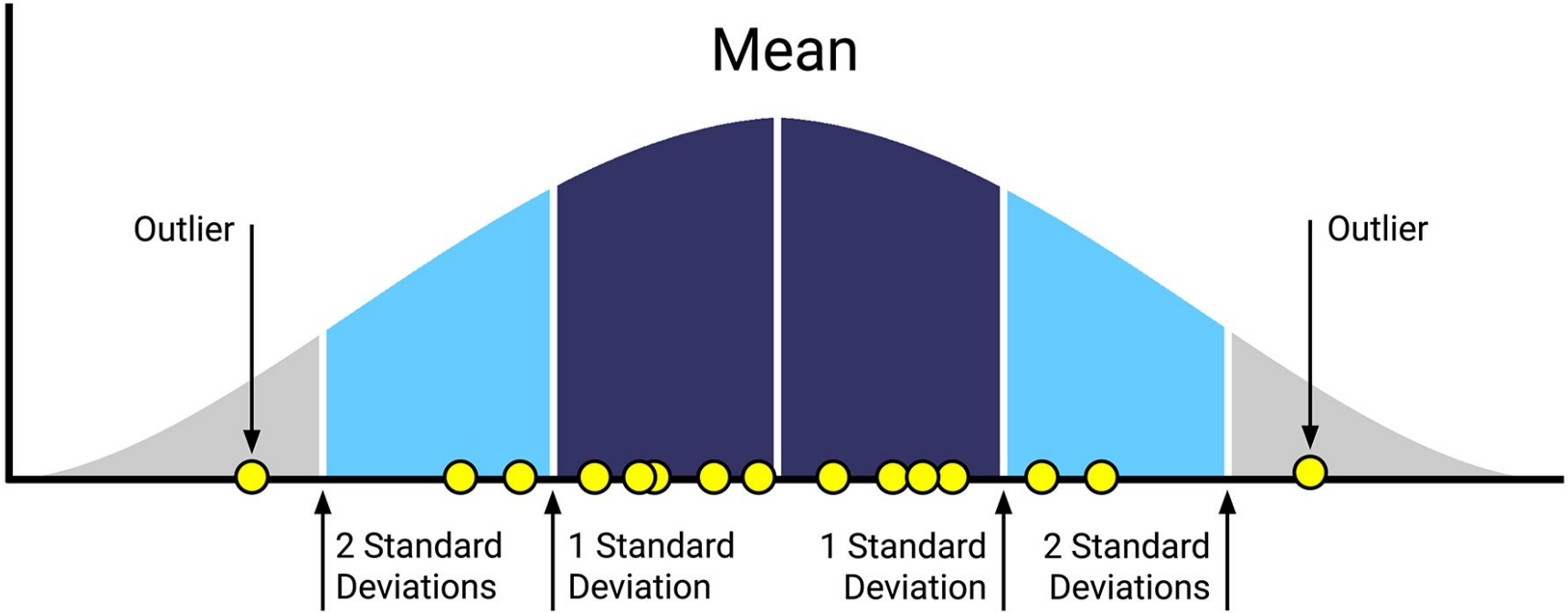
Standard Deviation

- Describes how spread out the data is from the mean
- Calculated from the square root of the variance
- In the same units of measurement as the mean

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

Standard Deviation

Square root of the variance; a measure used to quantify the dispersion of a set of observations.

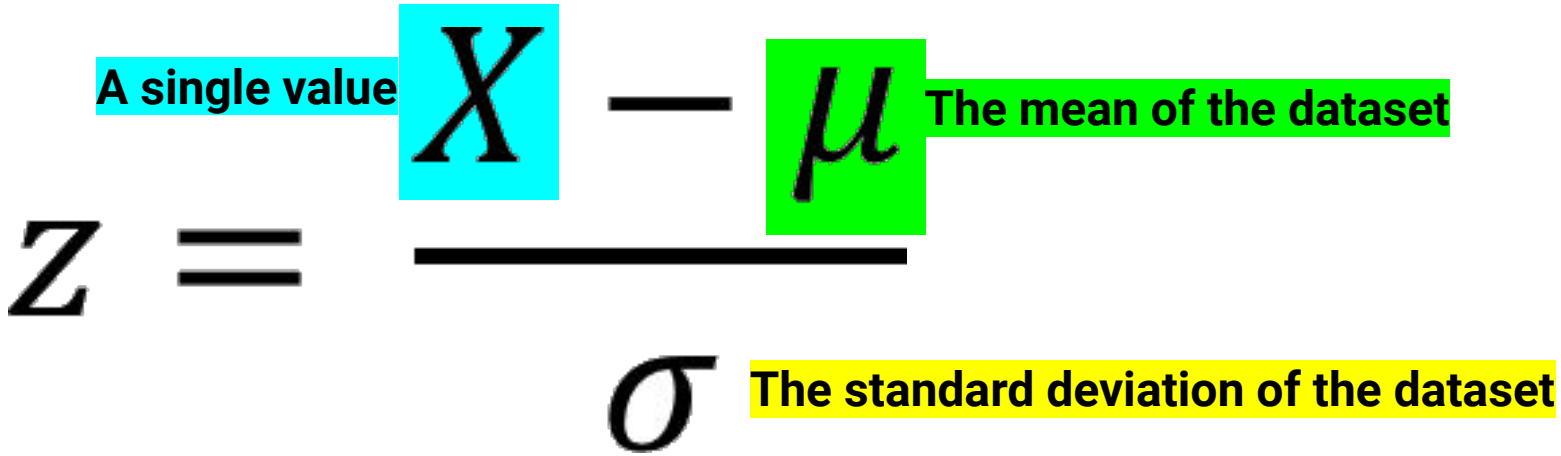


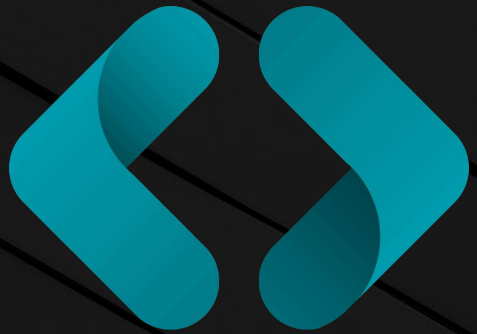
Z-Score

Z-Score describes a single value's distance from the mean of the data set
The distance is in terms of standard deviations. 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

It is now your turn to practice summarizing the variability of a data set using heart disease death rate data from the CDC.

Suggested Time:

15 minutes

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

Open the variance_review.xlsx workbook that contains your raw data Then clean up the dataset as follows:

- Rename the **Data_Value** column to **Death Rate Per 100,000**.
- This column contains missing data, so add a filter to the column that displays all rows except **(blanks)**.
- Rename the **Stratification1** and **Stratification2** columns to **Gender** and **Race/Ethnicity**, respectively.
- Rename **LocationAbbr** to **State**.
- Filter the **GeographicLevel** column so that **State** and **county** values are not compared together.
- Create a new sheet in the workbook named **Summary Table** that has a **State** column containing the following values: **AR** - Arkansas , **CA** - California, **FL** - Florida, **ME** - Maine, **MS** - Mississippi, **OR** - Oregon
- For each state, determine the **mean**, **variance**, and **standard deviation** for the overall death rate.
- Based on your calculated summary statistics determine which state had the greatest difference in death rate across all its counties and which state had the lowest variance in death rate. What was the death rate?
- Create a new sheet in the workbook named **Oregon Z-Scores**. Within this new sheet, copy over the **LocationDesc** (renamed to **County**) and **Death Rate Per 100,000** columns from the raw data for *only* the state **OR** where **Gender** is **Overall**.
- Calculate the **z-score** for the overall death rate by county across the whole state and use those values to determine which county had the largest difference in death rate from the mean of the state.
- Based upon your calculated z-scores, determine which county had the largest difference in death rate from the mean of the state.

Time's Up! Let's Review.

Instructor Demonstration

Quantiles, Outliers and Boxplots

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 values of a data set
- Extreme values can skew these statistics!

But how can we summarize real-world data?

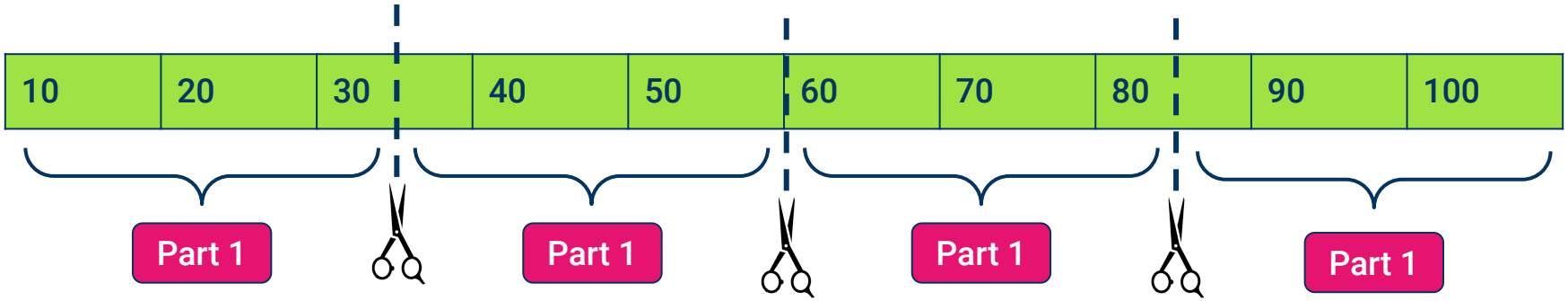
Quantiles: Used to Describe Segments of a Dataset

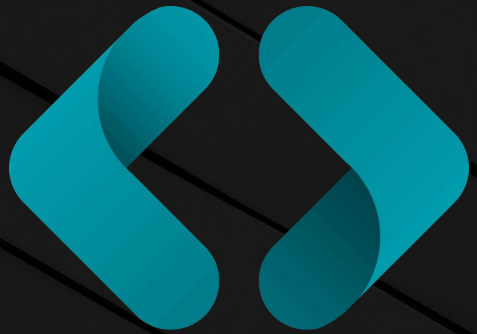
Quantiles separate a sorted dataset into equally sized fragments.

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.

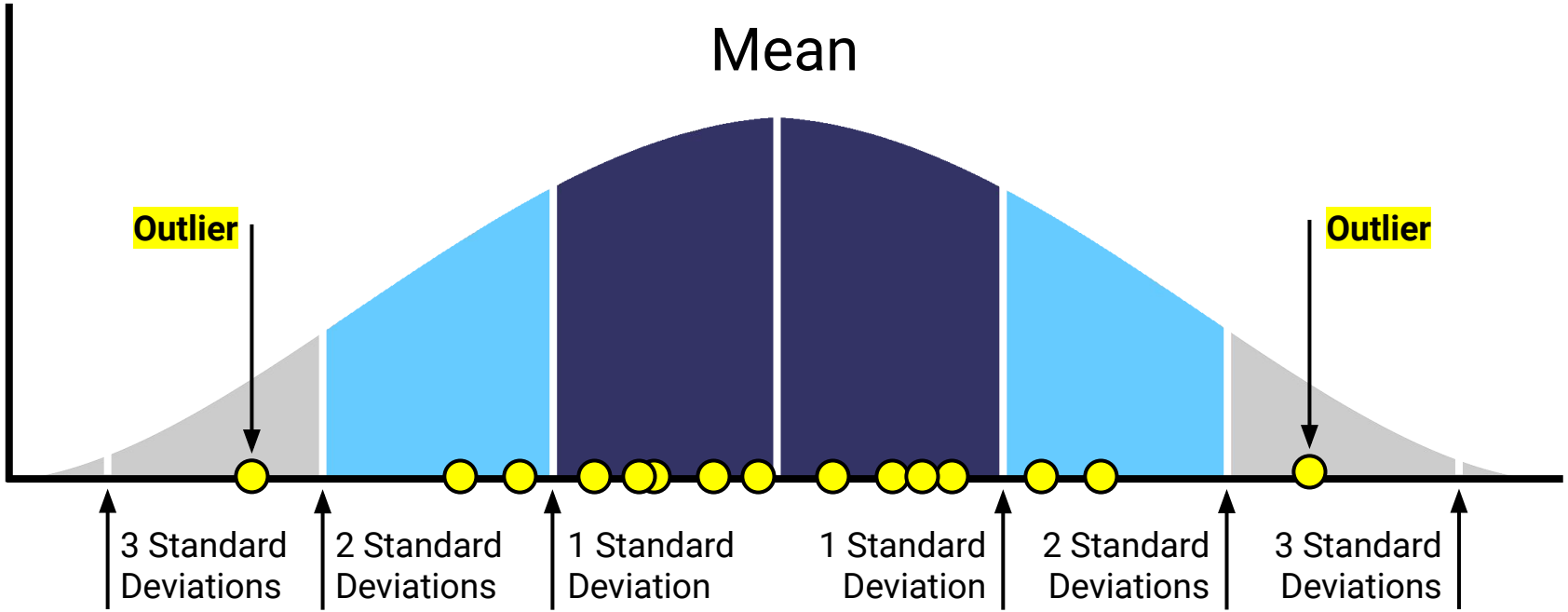




Time to <code>

Outliers

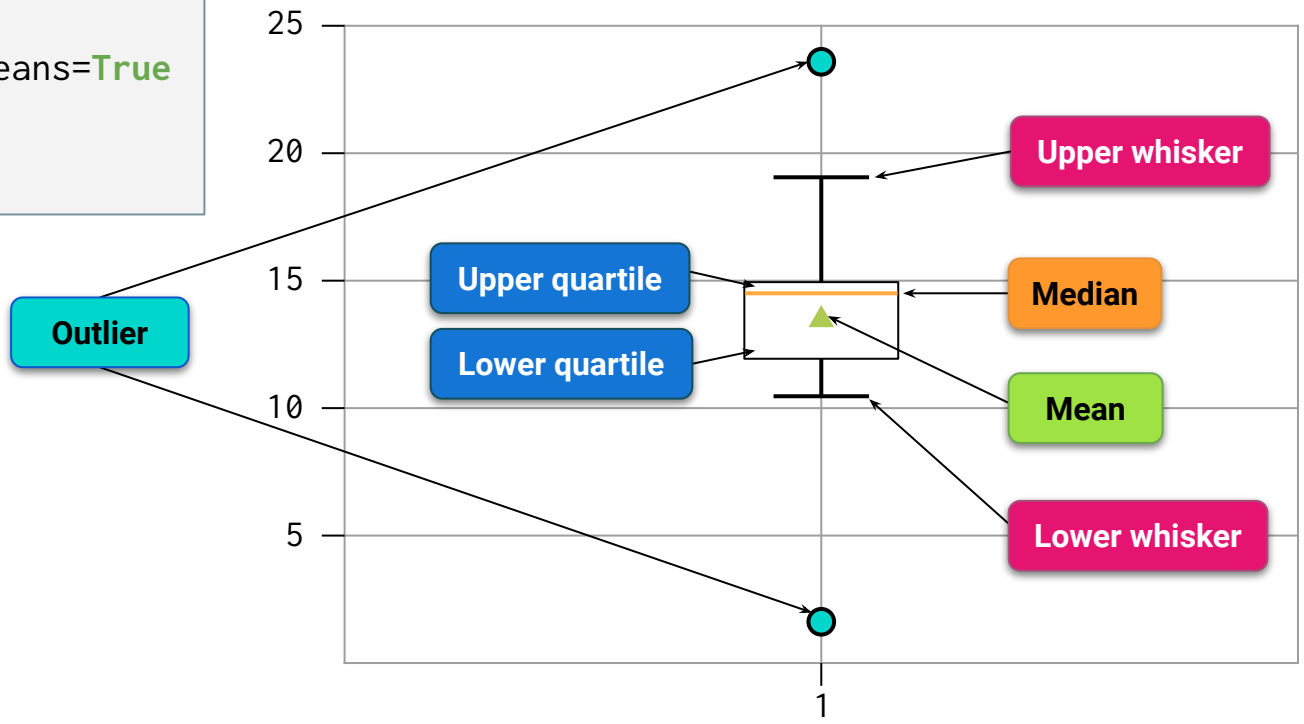
Suspicious values are called potential outliers. An outlier is a data point that differs from the rest of a data set. Outliers can inaccurately skew a data set.



Qualitatively

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

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



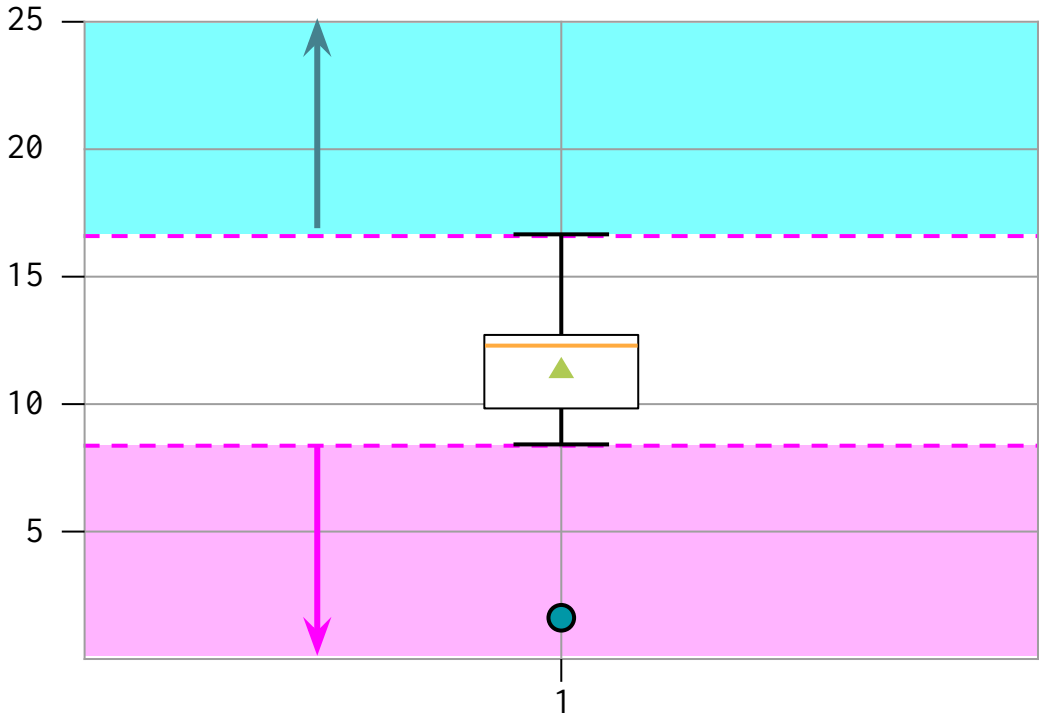
Quantitatively

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 familiarize 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 Antioxidant_content_in_mmol_100g 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 Antioxidant_content_in_mmol_100g falls outside of the 1.5*IQR boundaries. List those products and their antioxidant content on the worksheet.
- Create a box plot of the Antioxidant_content_in_mmol_100g 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

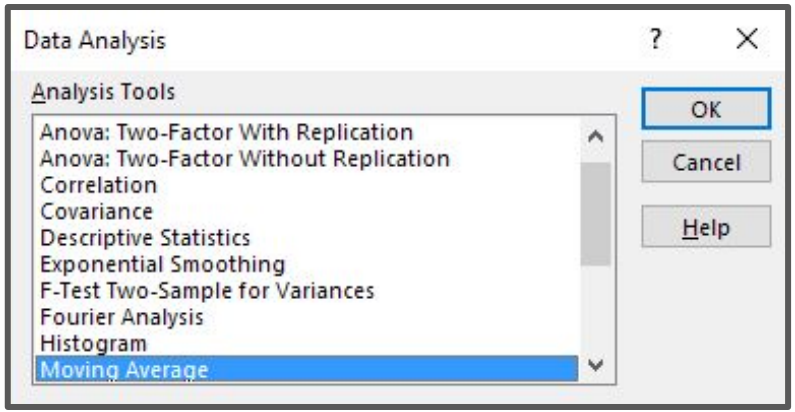
Excel is a great foundational tool

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

But Excel can be used for even MORE statistics!

The Excel Analysis ToolPak contains:

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



All of these functions we will cover throughout the course!

Analysis ToolPak is not designed for in-depth data analytics

Excel struggles with medium to large data sets:

- >200 columns or >100000 rows
- Depends on machine

Excel does not automatically record parameters for statistical tests

Excel's Analysis ToolPak ***should*** be used

- Gut-checks
- One-off analysis

How to install and use the Excel Analysis ToolPak: Mac

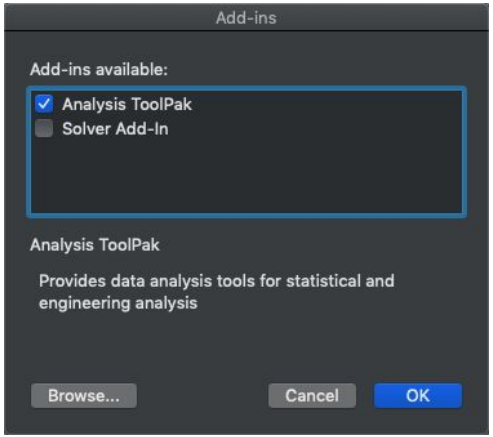
To Install:

01 Go to the “Tools” menu in Excel.

02 Select the “Excel Add-Ins…” option.

03 Enable the “Analysis ToolPak” option.

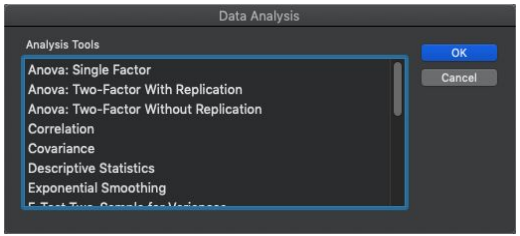
04 Press “OK”.



To Use:

01 Go to the “Data” menu in Excel.

02 Select the “Data Analysis” option.



How to install and use the Excel Analysis ToolPak: PC

To Install:

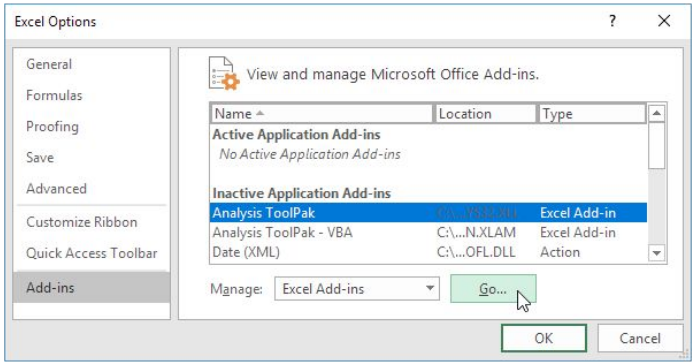
01 Click the File tab

02 Go to Options

03 Select the Add-Ins category

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

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

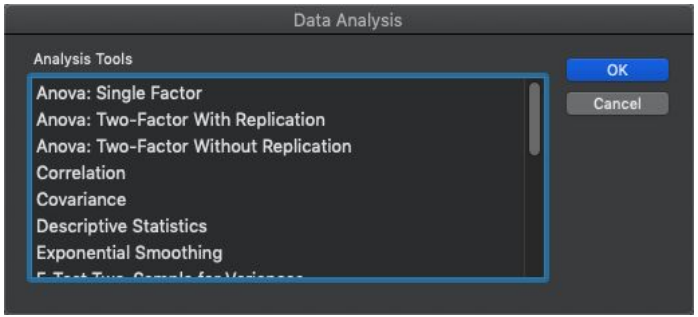


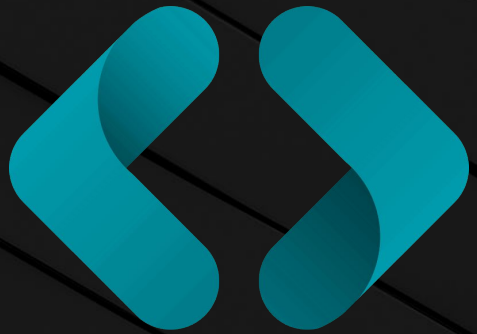
To Use:

01 Go to the “Data” menu in Excel.

02 Go to the “Analyze” section.

03 Select the “Data Analysis” option.

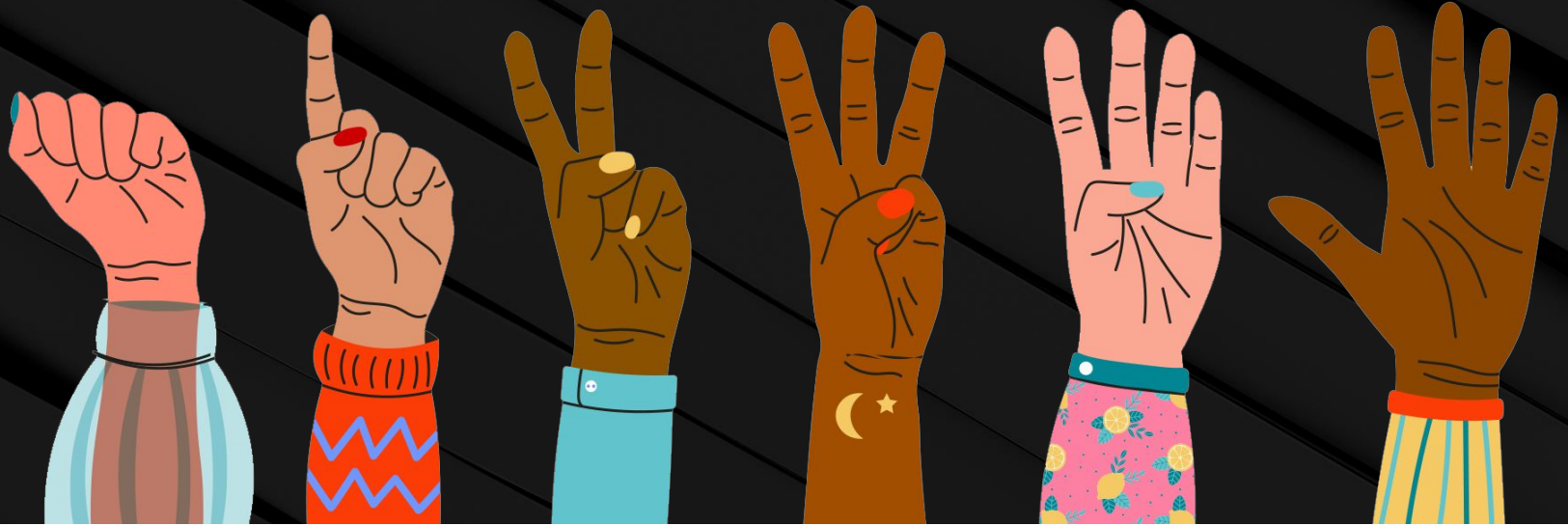




Time to <code>

FIST TO FIVE:

Who feels comfortable with plotting figures in Excel?



FIST TO FIVE:

Who feels comfortable calculating summary statistics in Excel?