
PLURALSIGHT

M&T Bank

Data Academy

Week 2

Tarek Atwan
Instructor, Pluralsight

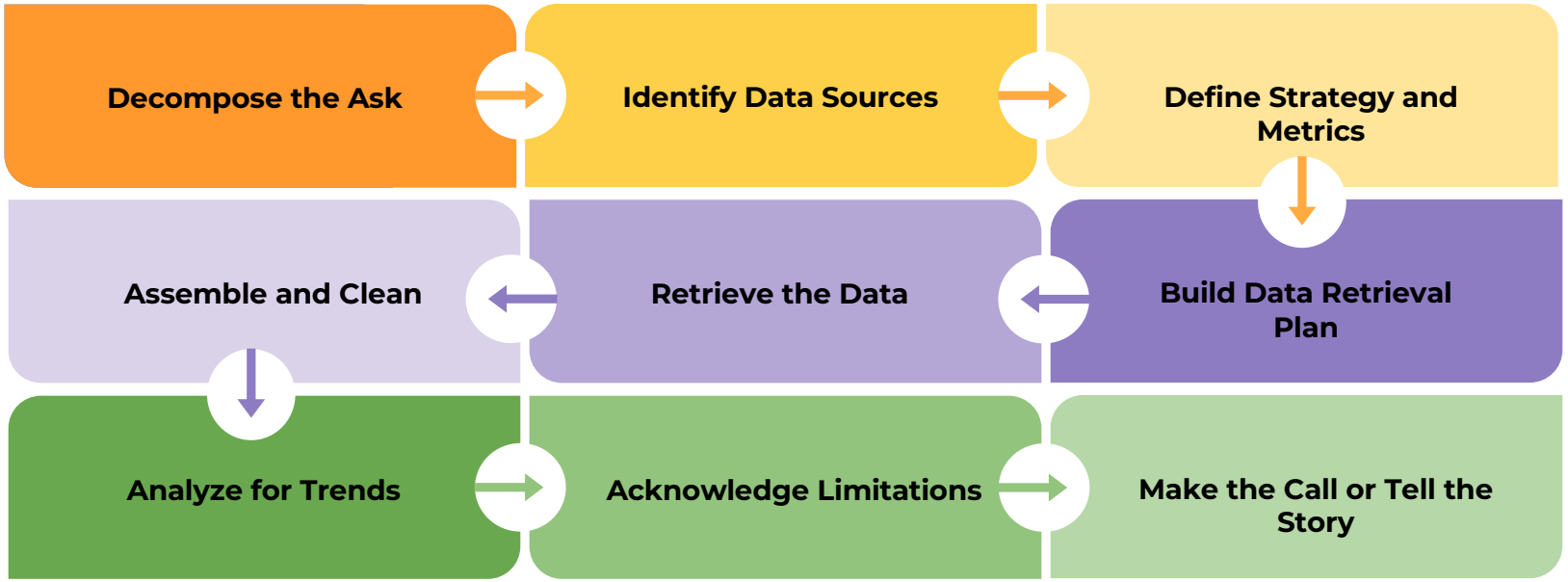
PLURALSIGHT

What is an area you are struggling with today?

An Analytics Paradigm

Analytics Paradigm

Regardless of type or industry, this paradigm provides a repeatable pathway for effective data problem-solving.



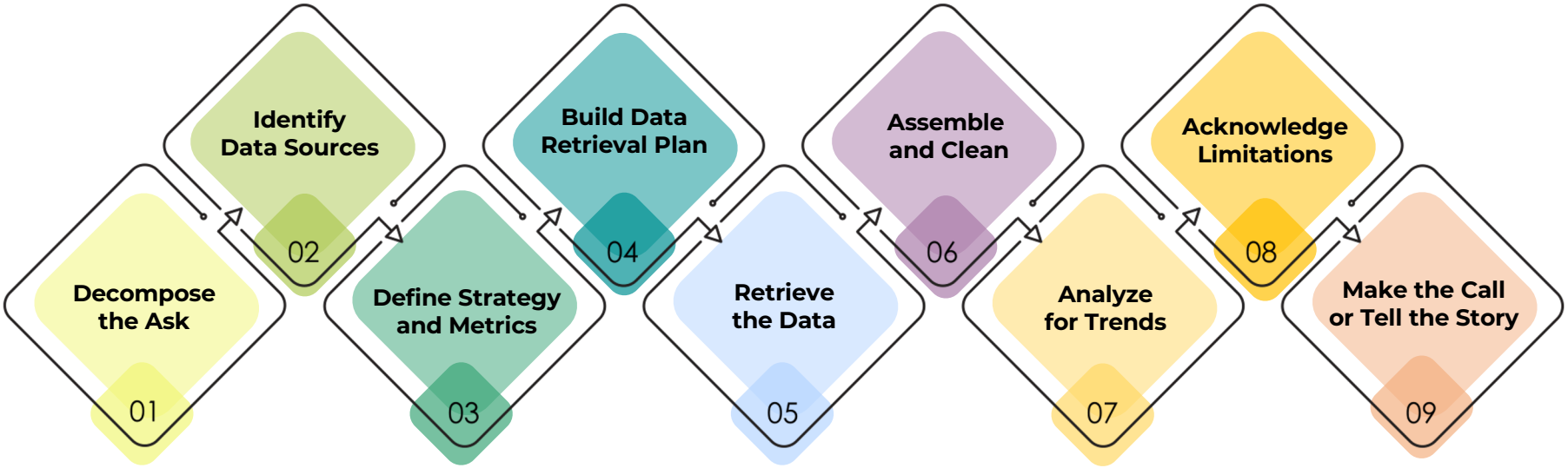
Data analytics is about what two things?

Fundamentally, data analytics
is about storytelling and truth-telling.

What are the steps in the Analytics Paradigm?

Analytics Paradigm

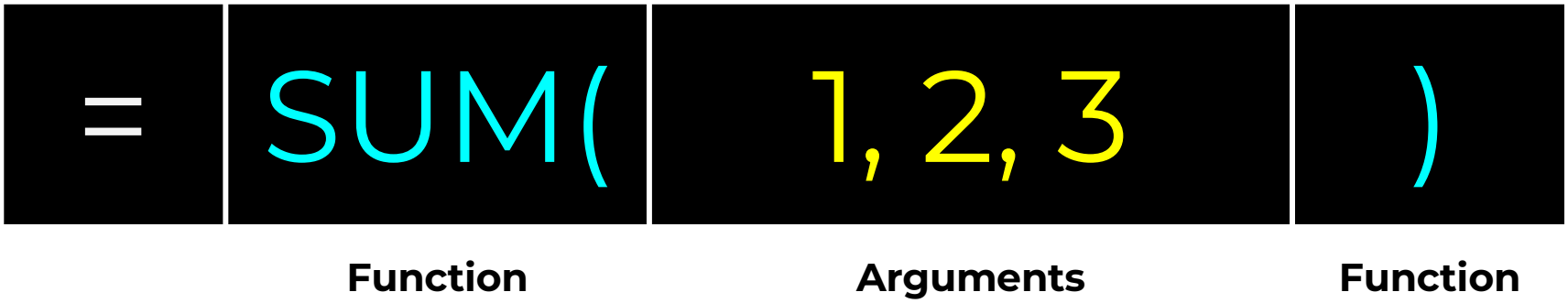
Regardless of type or industry, this paradigm provides a repeatable pathway for effective data problem solving.



Formulas

Ooh...Coding! (Sort Of)

Excel has introduced you to a sort of proto-programming. When you write scripts, you will rely on functions (methods) that do something to or with arguments.



Ooh…Coding! (Sort Of)

When we reference a range or a set of ranges, Excel is given a set of variable inputs. Excel will determine the actual values of these inputs prior to executing the function.

Ooh...Coding! (Sort Of)

What about this example?
Which is the function?
Which are the arguments?

```
= SUM( AVG(F4:F6), AVG(G4:G6) )
```

Ooh...Coding! (Sort Of)

What about this example?
Which is the function?
Which are the arguments?

The **AVG functions** take the provided ranges as their arguments.

```
= SUM( AVG(F4:F6), AVG(G4:G6) )
```

Ooh...Coding! (Sort Of)

What about this example?
Which is the function?
Which are the arguments?

This is a **nested function**. We'll be doing plenty of complex nests in this class.

```
= SUM( AVG(F4:F6), AVG(G4:G6) )
```

There are multiple ways to select data in a formula

Most of us learned to select a range of cells to input into a function

```
=AVG(A1:A10)
```

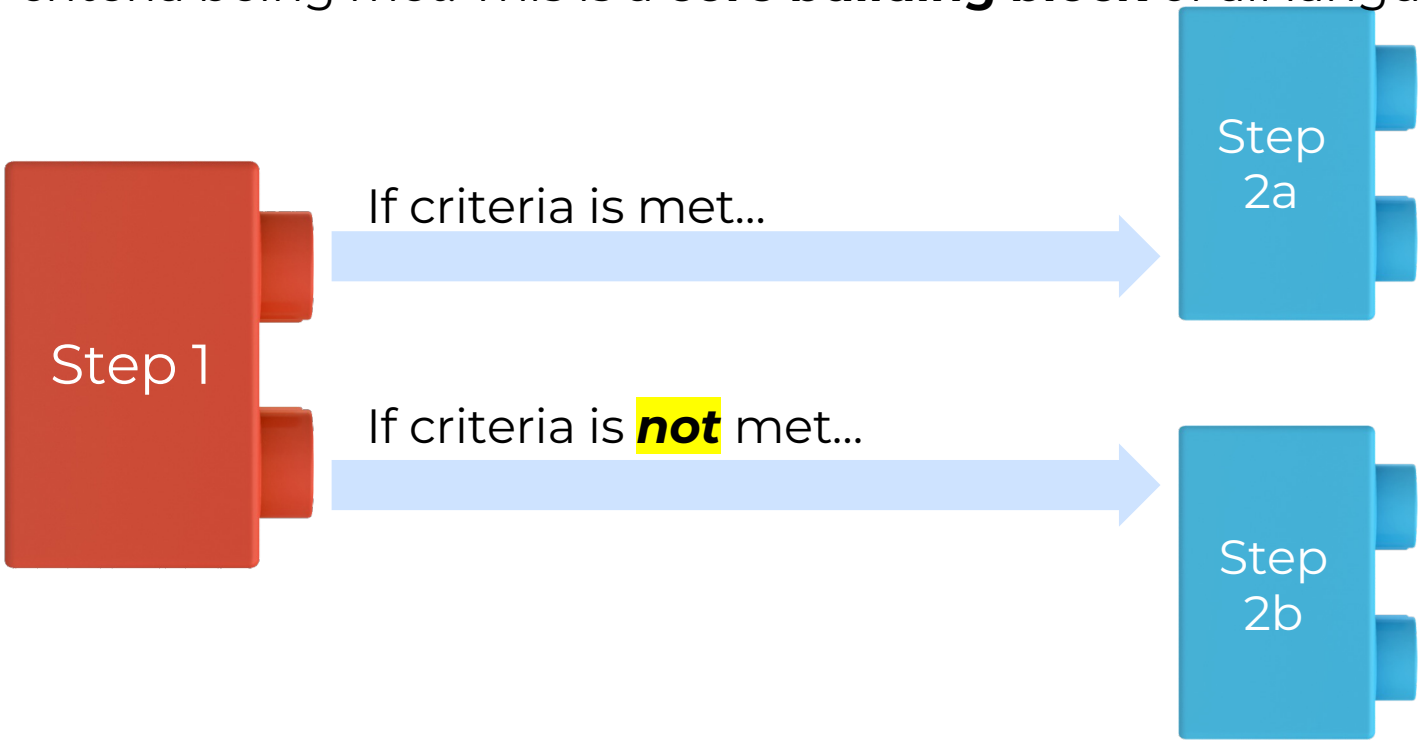
There are multiple ways to select data in a formula

But we can name a range of values to make interpreting formulas easier!

=AVG(A1:A10)  =AVG(prices)

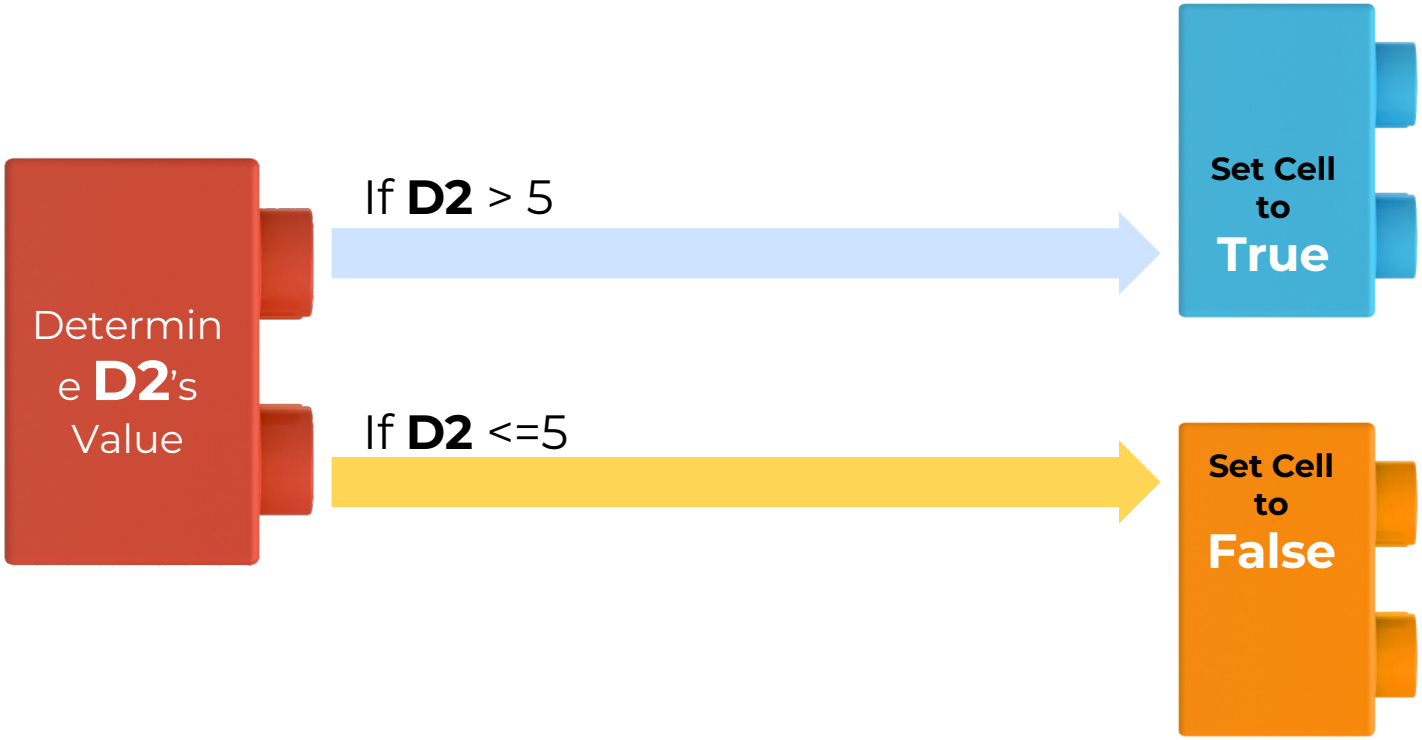
Conditionals: If This, Then That

Conditionals present a way to control the flow of logic based on certain criteria being met. This is a **core building block** of all languages.



Conditionals: If This, Then That

=IF(D2>5,TRUE,FALSE)



But what if we want to combine conditions?

AND, NOT, OR

Ooh...Coding! (Sort Of)

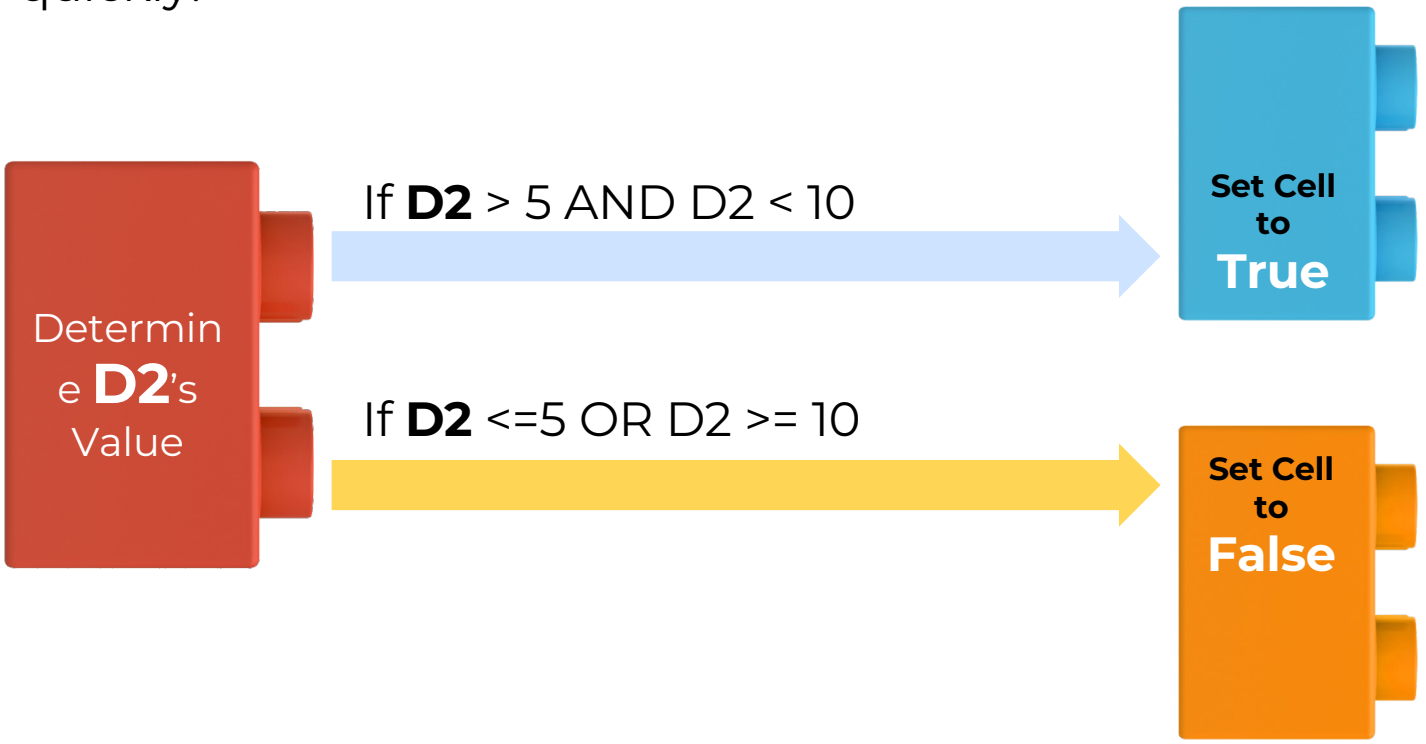
But what if we want
to **combine** conditions?

AND, NOT, OR

```
=IF(AND(D2>5, D2<10),TRUE,FALSE)
```

Conditionals: If This, Then That

Nesting conditionals are powerful, but can become convoluted very quickly!



Activity: Gradebook

Create a formula that calculates the final grade for a student based on their previous exams and papers.

Suggested Time:

15 minutes

Activity: Gradebook

To do	• Create a formula which calculates the final grade for a student based upon their previous exams and papers.
When making this calculation	• Consider every paper and exam to be equal in weight; each should comprise one-fourth of the overall grade. • Round the result to the nearest integer. • Using conditionals, create a formula that returns `PASS` if a student's final grade is greater than or equal to 60. If a student's final grade is below 60, the formula should return `FAIL`.
Bonus	Create a nested `IF()` formula that returns a letter grade based on a student's final grade. • Greater than or equal to 90 = A • Greater than or equal to 80 and less than 90 = B • Greater than or equal to 70 and less than 80 = C • Greater than or equal to 60 and less than 70 = D • Anything less than 60 = F

What are “measures of central tendency”?

Values used to describe the center of a data set.

Central Tendency

Three most common measures of central tendency:

Mean

The "arithmetic" average

To calculate: The sum of all values, divided by the number of values

Median

The middle value of a data set

To calculate: Sort the data set and find the center

Mode

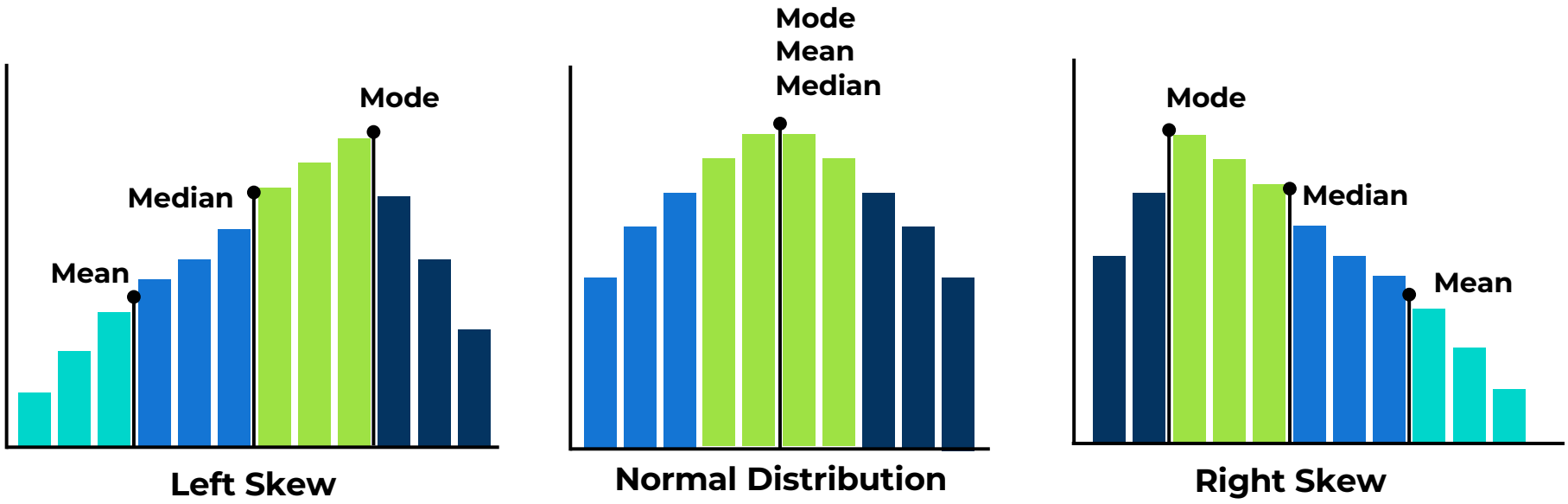
The most frequent value of a data set

To calculate: Count the frequency of each value in a data set, determine the most frequent value

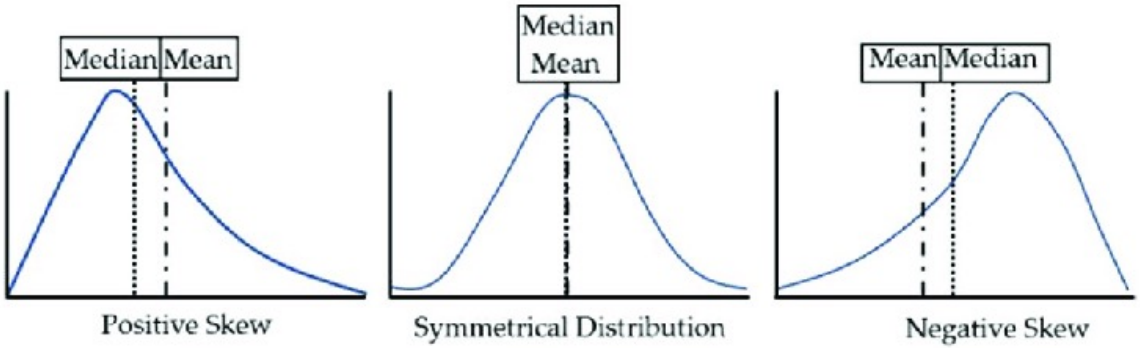
What are the three measures of central tendency?

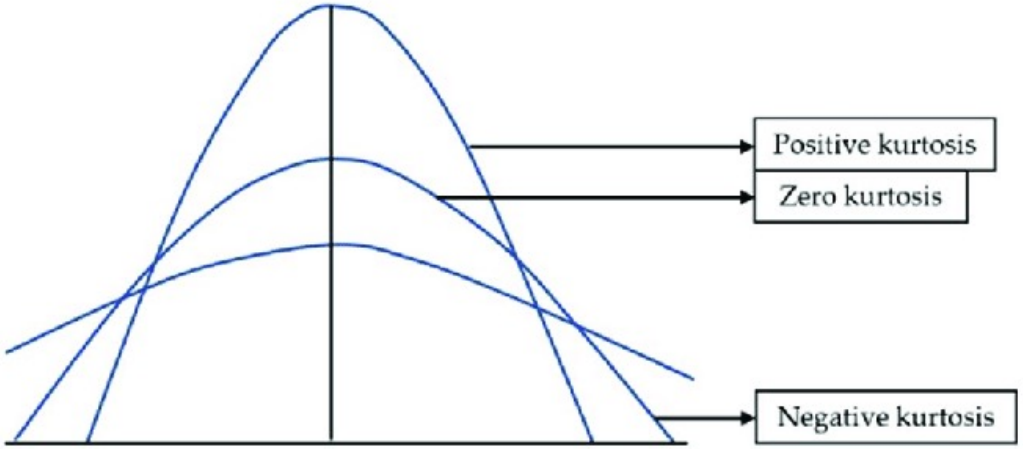
The mean, median and mode.

The mean, median and mode.



Skewness and Kurtosis





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?

Instructor Demonstration

Formatting

Formatting in Excel falls into two categories

Data Formatting

- Changes the way a value is represented in a cell.
- Used to help with interpretation or to add context to the range of values

Examples

- Date and Time
- Currency
- Percentage
- Scientific Notation

Style Formatting

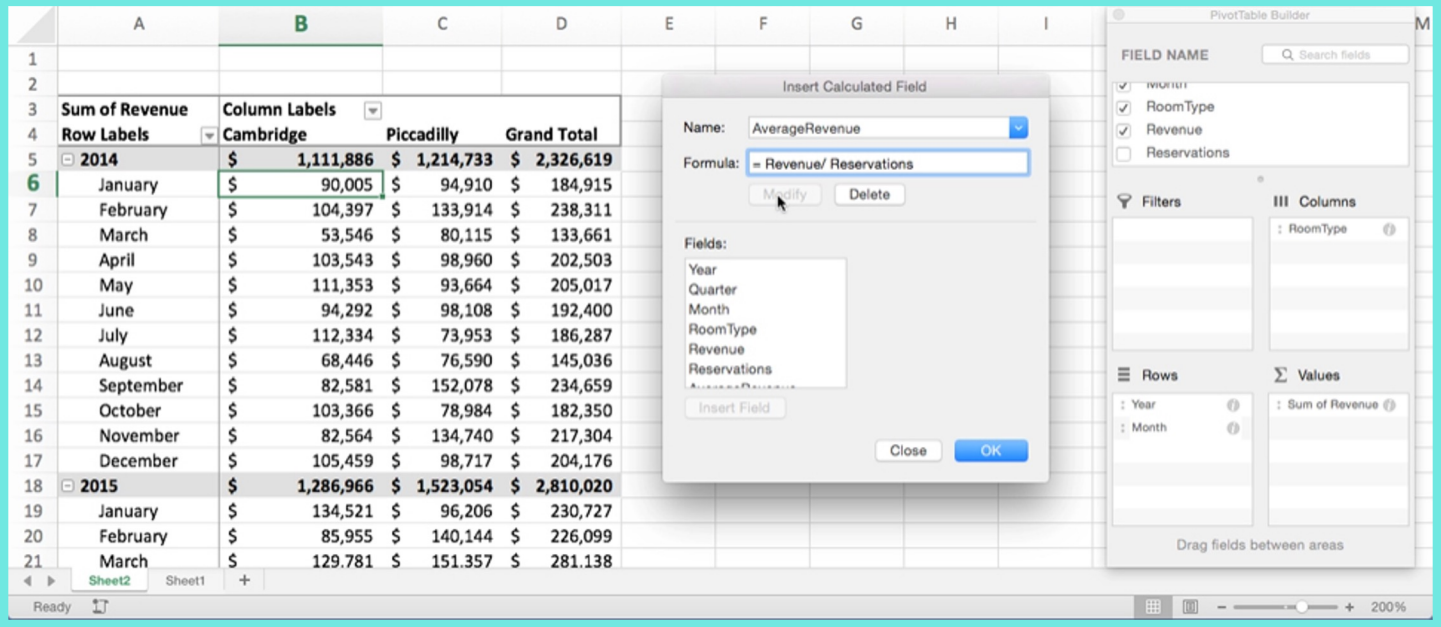
- Changes the way the cell and text are viewed
- Can include font color, cell highlighting, borders, etc.
- Can be performed manually or using formulas/logic (conditional formatting)

Instructor Demonstration

Pivot Tables

Get Pivot With It

Pivot tables are one of the most important data visualization concepts to master in this class. (Don't worry. They are a cinch to deal with.)



Get Pivot With It

In essence, a pivot table is a **summative** analytic tool that allows us to perform aggregate functions that allow any combination of fields. (The term *pivot table* comes from the fact that we are pivoting along a data axis).

Seller	Qty. Sold	Date
Joseph	$42.50	1/1/17
Jacob	$65.00	1/3/17
Jacob	$5.25	1/6/17
Joseph	$125.00	1/6/17
Jacob	$3.50	1/7/17
Matt	$32.00	1/9/17

Seller	Total Sold
Joseph	$167.50
Jacob	$73.75
Matt	$32.00

Word to the Wise: Keep It Flat!

Modern Business Intelligence (BI) tools like Tableau, Sisense, and Salesforce work best if data is stored in flat CSVs—meaning column headers represent fields (vertically) on the spreadsheet. This is largely because all of these technologies heavily utilize pivot tables as a tool for their visualizations. **Don't try to confuse this simplicity. "Spreadsheet magic" is a nightmare to analyze.**

B	C	D	E	F	G	H
DateTime	Week #	Section?	Pace	Academic Support	Self-Mastery	Instructor Er
2016-09-11T04:00:00.000Z	18	RCB0503FSF - CCC	3	5	5	4
2016-09-11T05:00:00.000Z	6	UT0726FSF	3	5	3	4
2016-09-12T04:00:00.000Z	11	UCF062016FSF	4	4	3	5
2016-09-12T04:00:00.000Z	23	UCF0329FSF	2	4	5	1
2016-09-12T04:00:00.000Z	9	UNC0712FSF	3	4	4	3
2016-09-12T04:00:00.000Z	23	UCF0328FSF	4	3	2	3
2016-09-12T04:00:00.000Z	6	RUT0725FSF-NB	5	4	4	5
2016-09-12T04:00:00.000Z	6	RUT0725FSF-NB	5	5	4	5
2016-09-12T04:00:00.000Z	6	RUT0725FSF-NB	2	4	4	4
2016-09-12T04:00:00.000Z	11	UCF062016FSF	4	5	4	5
2016-09-12T04:00:00.000Z	13	UCF061416FSF	4	5	1	5

Activity: Top Songs Pivot Table

In this activity, you will use a 5000 row spreadsheet containing data for the top 5000 songs from 1901 onward. Using pivot tables, you will uncover which artists have the most songs in the top 5000, the song titles, and the year each song was released.

Suggested Time:

17 minutes

Top Songs Pivot Table Instructions

- Select all of the data in your worksheet and create a new pivot table.
- Make a pivot table that can be filtered by year and contains two rows: *Artist* and *Name*.
- All of an artist's songs should be listed below their name.

Update your pivot table to contain values for:

- How many songs an artist has in the top 5000
- The sum of the final_score of their songs.

- Sort your pivot table by descending sum of the final_score.

Instructor Demonstration

Lookups

Look It Up with Lookups

Assume this table is gigantic. How would we **retrieve** the population of a specific planet for use in another formula?

Planet	Population
Zeelo	5020
Merinoa	380
Cardboard Box	2
...	...
Asteroid 9	95

Look It Up with Lookups

Assume this table is gigantic. How would we **retrieve** the population of a specific planet for use in another formula?

=vlookup(<value>, <full table>,
<column to retrieve>,<match parameter>)

Planet	Population
Zeelo	5020
Merinoa	380
Cardboard Box	2
...	...
Asteroid 9	95

Look It Up with Lookups

What will this yield?
=vlookup(“Asteroid 9”, Planets, 3, FALSE)

Planet	Population	Species
Zeelo	5020	Zoltans
Merinoa	380	Murphies
Cardboard Box	2	Hambones
…	…	
Asteroid 9	95	Asterisks

Look It Up with Lookups

What will this yield?
=vlookup(“Asteroid 9”, Planets, 3, FALSE)

Planet	Population	Species
Zeelo	5020	Zoltans
Merinoa	380	Murphies
Cardboard Box	2	Hambones
…	…	
Asteroid 9	95	Asterisks

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).

Instructor Demonstration

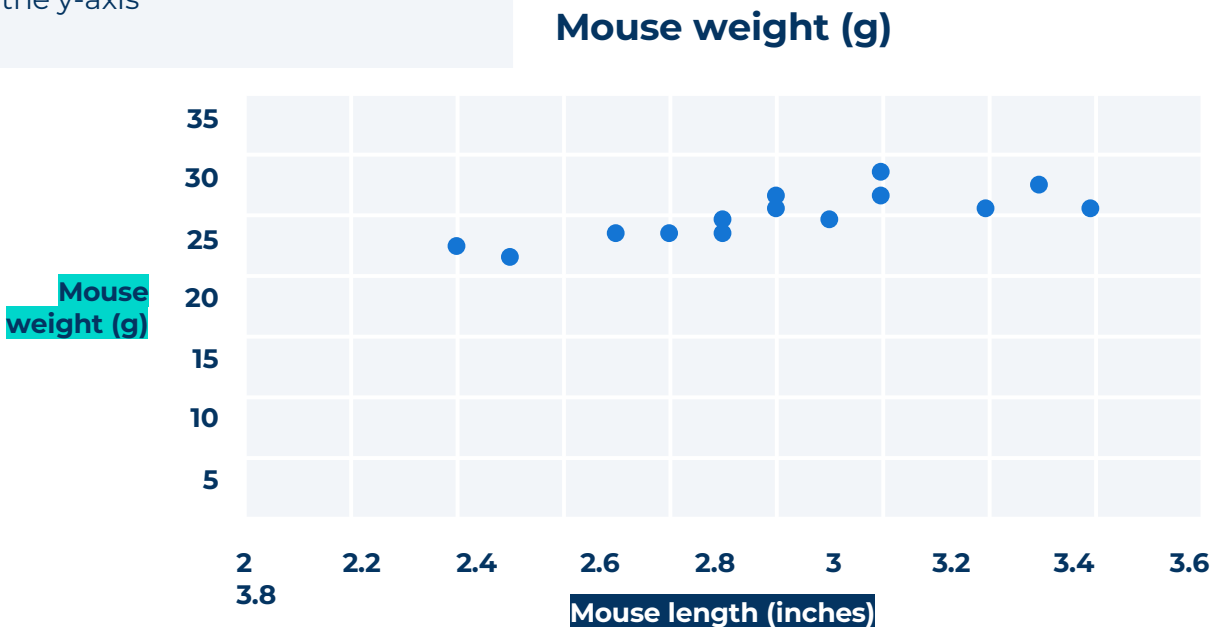
Scatter Plots and Trend Lines

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!



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	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

- 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

Instructor Demonstration

Variance, Standard Deviation and Z-Score

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

The mean value of all observations

Sample variance

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

The number of observations

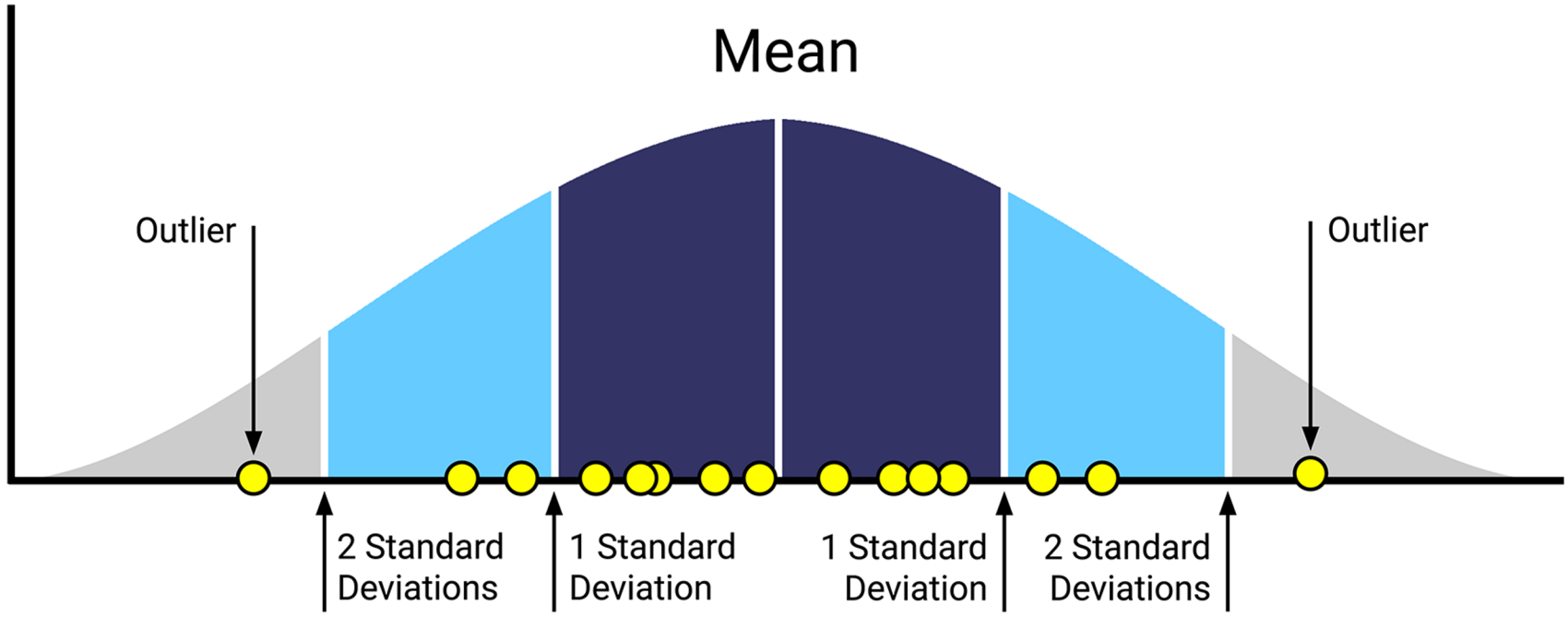
Standard Deviation

- 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

A single value X

$$z = \frac{X - \mu}{\sigma}$$

The mean of the dataset μ

The standard deviation of the dataset σ

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.

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?

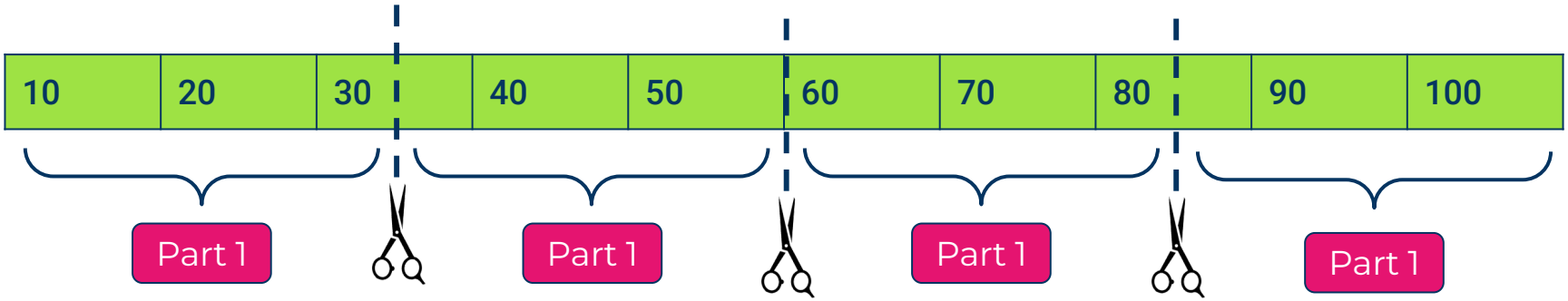
Quantiles: Used to Describe Segments of a Dataset

Quantiles separate a sorted dataset into equally sized fragments.

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

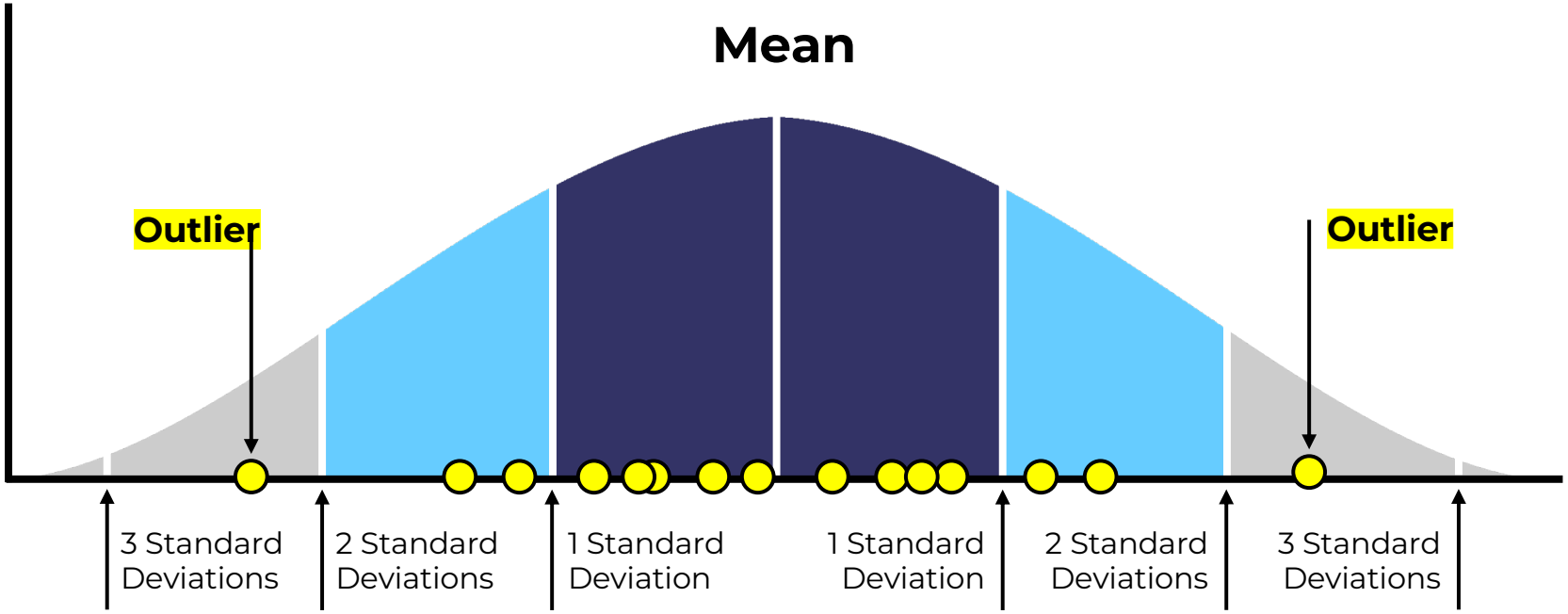
01 **Quartiles** divide the dataset into four equally sized parts.

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



Outliers

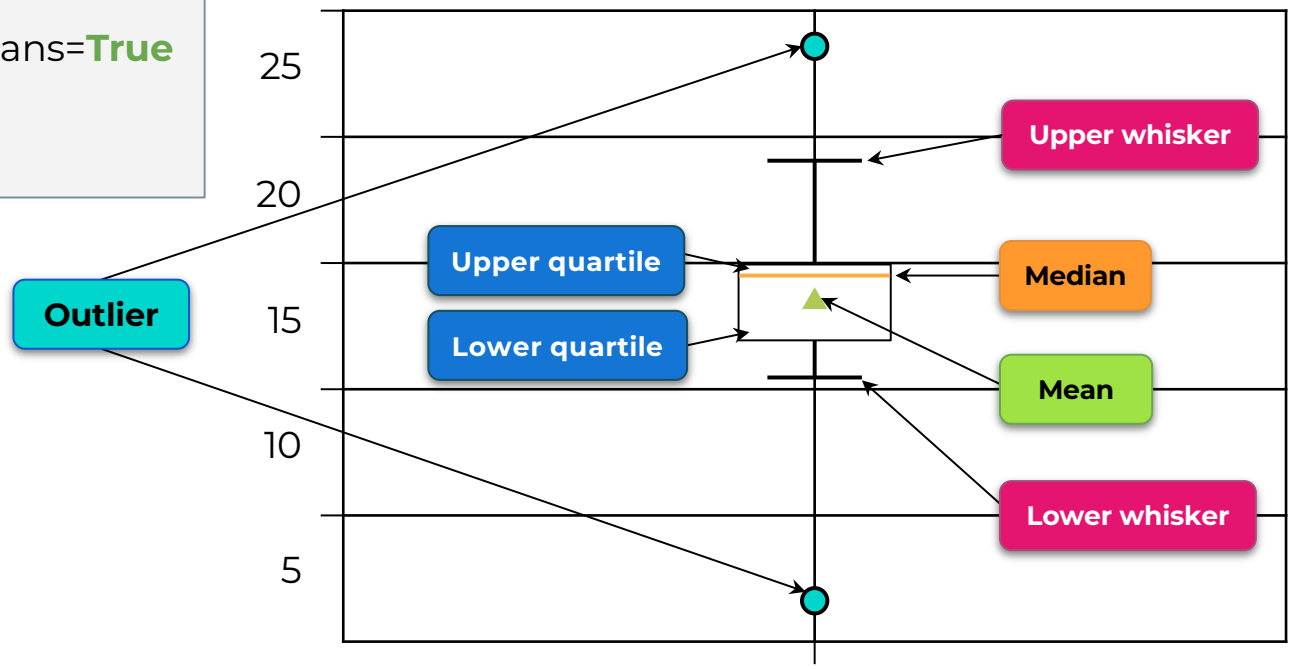
Suspicious values are called potential outliers. An outlier is a data point that differs from the rest of a data set. Outliers can inaccurately skew a data set.



Qualitatively

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

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



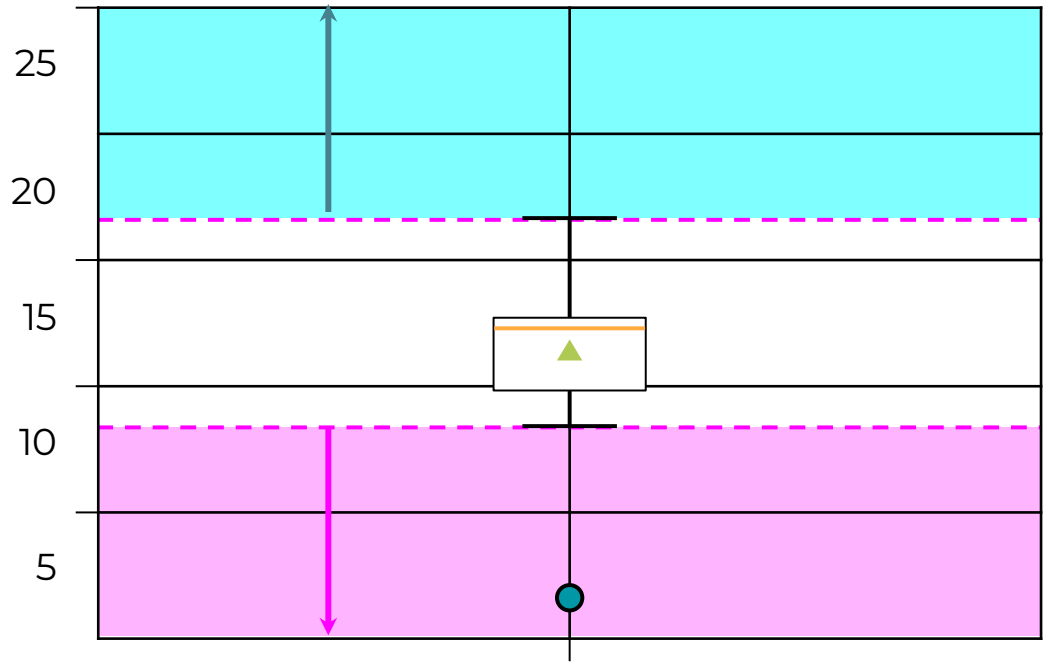
Quantitatively

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.

Instructor Demonstration

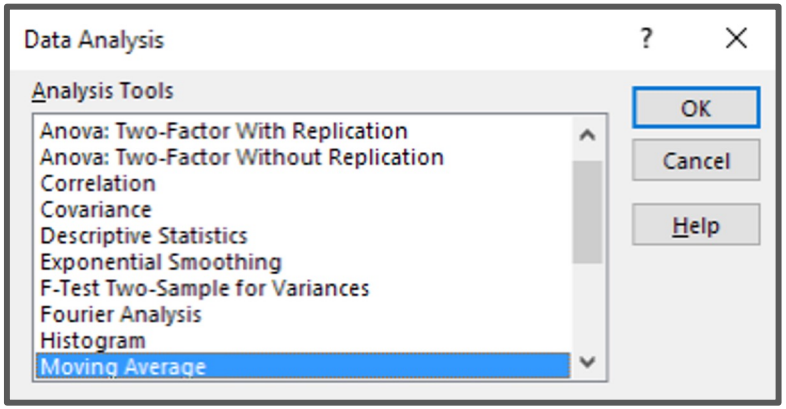
Excel's Statistics Add-On

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

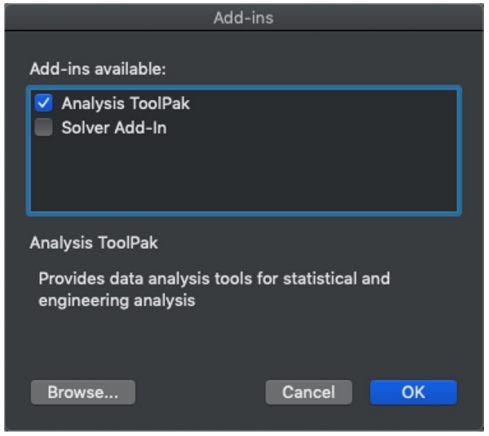
To Install:

01 Go to the “Tools” menu in Excel.

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

03 Enable the “Analysis ToolPak” option.

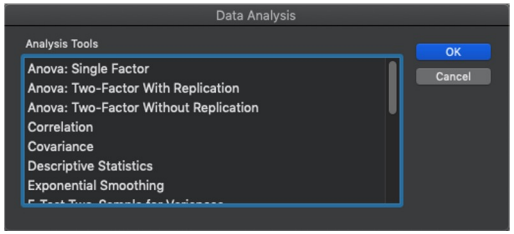
04 Press “OK”.



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:

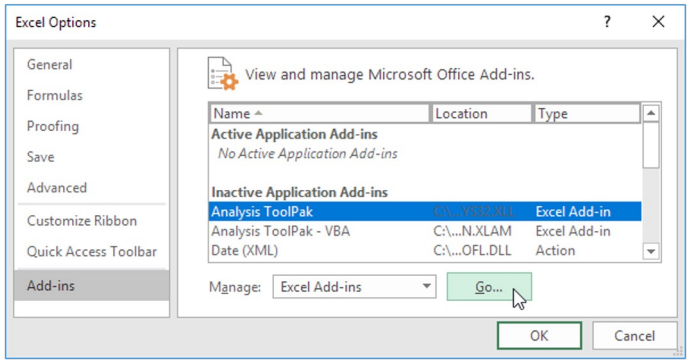
01 Click the File tab

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.