Descriptive Statistics

is useful in many different jobs, and activities. Having a good understanding of descriptive statistics will help anyone working in:

- Business Analytics
- Data Analysis
- Data Engineering
- Product Management

Data is defined as distinct pieces of information and it can come in many forms. From numbers in a spreadsheet, text to video and databases, to images and audio recordings, utilizing data in its different forms is the new way of the world.

Data is used to understand and improve nearly every facet of our lives. So, no matter what field you are in, you can utilize data to make better decisions and accomplish your goals.

two data types are introduced: **Quantitative** and **Categorical**.

Quantitative data takes on numeric values that allow us to perform mathematical operations (like the number of dogs).

Categorical is used to label a group or set of items (like dog breeds - Collies, Labs, Poodles, etc.).

divide categorical data further into two types: **Ordinal** and **Nominal**.

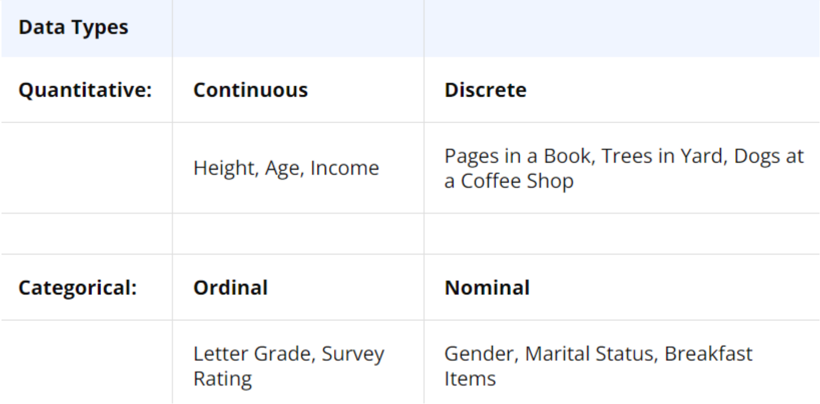
Categorical Ordinal data take on a ranked ordering (like a ranked interaction on a scale from `Very Poor` to `Very Good` with the dogs).

Categorical Nominal data do not have an order or ranking (like the breeds of the dog).

quantitative data as being either **continuous** or **discrete**.

Continuous data can be split into smaller and smaller units, and still a smaller unit exists. An example of this is the age of the dog - we can measure the units of the age in years, months, days, hours, seconds, but there are still smaller units that could be associated with the age.

Discrete data only takes on countable values. The number of dogs we interact with is an example of a discrete data type.



Data Types		
Quantitative:	**Continuous**	**Discrete**
	Height, Age, Income	Pages in a Book, Trees in Yard, Dogs at a Coffee Shop
Categorical:	**Ordinal**	**Nominal**
	Letter Grade, Survey Rating	Gender, Marital Status, Breakfast Items

To break down our data types, there are two main blocks:

Quantitative and **Categorical**

Quantitative can be further divided into `Continuous` or `Discrete`.

Categorical data can be divided into `Ordinal` or `Nominal`.

Continuous vs. Discrete

To consider if we have continuous or discrete data, we should see if we can split our data into smaller and smaller units. Consider time - we could measure an event in years, months, days, hours, minutes, or seconds, and even at seconds we know there are smaller units we could measure time in. Therefore, we know this data type is continuous. **Height**, **age**, and **income** are all examples of `continuous data`. Alternatively, the **number of pages in a book**, **dogs I count outside a coffee shop**, or **trees in a yard** are `discrete data`. We would not want to split our dogs in half.

Ordinal vs. Nominal

In looking at categorical variables, we found **Gender**, **Marital Status**, **Zip Code**, and your **Breakfast items** are `nominal variables` where there is no order ranking associated with this type of data. Whether you ate cereal, toast, eggs, or only coffee for breakfast; there is no rank-ordering associated with your breakfast.

Alternatively, the **Letter Grade** or **Survey Ratings** have a rank ordering associated with it, as `ordinal data`. If you receive an A, this is higher than an A-. An A- is ranked higher than a B+, and so on... Ordinal variables frequently occur on rating scales from very poor to very good. In many cases, we turn these ordinal variables into numbers, as we can more easily analyze them, but more on this later!

Analyzing Quantitative Data

Four Aspects for Quantitative Data

There are four main aspects to analyzing **Quantitative** data.

- Measures of `Center`
- Measures of `Spread`
- The `Shape` of the data.
- `Outliers`

Analyzing Categorical Data

Though not discussed in the video, analyzing categorical data has fewer parts to consider. **Categorical** data is analyzed usually by looking at the counts or proportion of individuals that fall into each group. For example, if we were looking at the breeds of the dogs, we would care about how many dogs are of each breed, or what proportion of dogs are of each breed type.

Measures of Center

There are three measures of center:

- `Mean`
- `Median`
- `Mode`

The mean is often called the average or the **expected value** in mathematics. We calculate the mean by adding all of our values together and dividing by the number of values in our dataset.

The Median

The **median** splits our data so that 50% of our values are lower and 50% are higher. We found in this video that how we calculate the median depends on if we have an even number of observations or an odd number of observations.

Median for Odd Values

If we have an **odd** number of observations, the **median** is simply the number in the **direct middle**. For example, if we have 7 observations, the median is the fourth value when our numbers are ordered from smallest to largest. If we have 9 observations, the median is the fifth value.

Median for Even Values

If we have an **even** number of observations, the **median** is the **average of the two values in the middle**. For example, if we have 8 observations, we average the fourth and fifth values together when our numbers are ordered from smallest to largest.

In order to compute the median, we MUST sort our values first.

Whether we use the mean or median to describe a dataset is largely dependent on the **shape** of our dataset and if there are any **outliers**. We will talk about this in just a bit!

The Mode

The **mode** is the most frequently observed value in our dataset.

There might be multiple modes for a particular dataset or no mode at all.

No Mode

If all observations in our dataset are observed with the same frequency, there is no mode. If we have the dataset:

1, 1, 2, 2, 3, 3, 4, 4

There is no mode because all observations occur the same number of times.

Many Modes

If two (or more) numbers share the maximum value, then there is more than one mode. If we have the dataset:

1, 2, 3, 3, 3, 4, 5, 6, 6, 6, 7, 8, 9

There are two modes 3 and 6, because these values share the maximum frequencies at 3 times, while all other values only appear once.

Notation is a common language used to communicate mathematical ideas. **Think of notation as a universal language used by academic and industry professionals to convey mathematical ideas.**

You likely already know some notation. Plus, minus, multiply, division, and equal signs

If you aren't familiar with spreadsheets, this will be covered in detail in future lessons. Spreadsheets are a common way to hold data. They are composed of rows and columns. Rows run horizontally, while columns run vertically. Each column in a spreadsheet commonly holds a specific **variable**, while each row is commonly called an **instance** or **individual**.

A **random variable** is a placeholder for the possible values of some process (mostly... the term 'some process' is a bit ambiguous). As was stated before, notation is useful in that it helps us take complex ideas and simplify (often to a single letter or single symbol). We see random variables represented by capital letters (**X**, **Y**, or **Z** are common ways to represent a random variable).

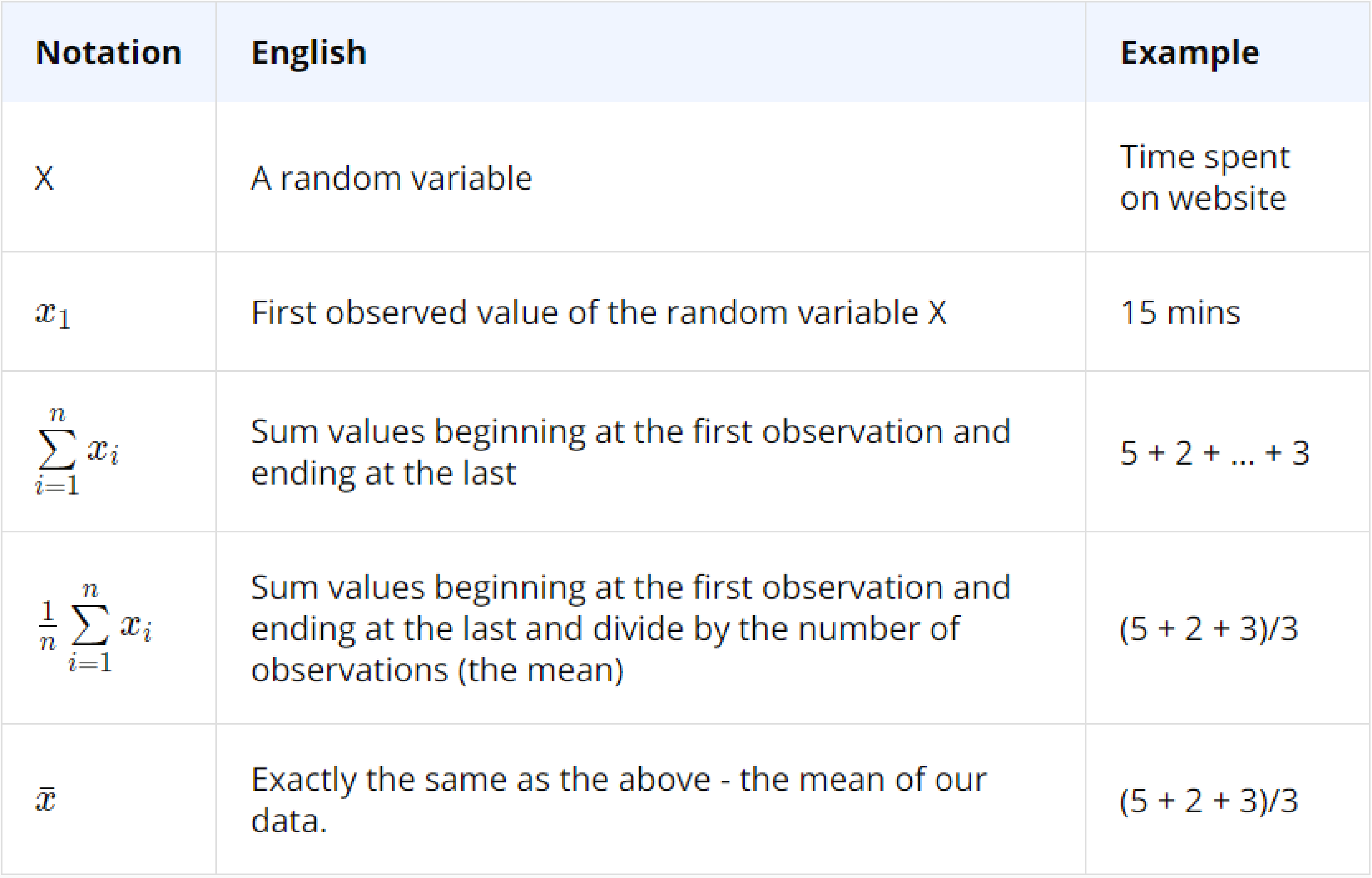
We might have the random variable **X**, which is a holder for the possible values of the amount of time someone spends on our site. Or the random variable **Y**, which is a holder for the possible values of whether or not an individual purchases a product.

X is 'a holder' of the values that could possibly occur for the amount of time spent on our website. Any number from 0 to infinity really.

Random variables are represented by capital letters. Once we observe an outcome of these random variables, we notate it as a lower case of the same letter.

An **aggregation** is a way to turn multiple numbers into fewer numbers (commonly one number).

Summation is a common aggregation. The notation used to sum our values is a greek symbol called sigma Σ



Notation	English	Example
X	A random variable	Time spent on website
x_1	First observed value of the random variable X	15 mins
$\sum_{i=1}^{n} x_i$	Sum values beginning at the first observation and ending at the last	5 + 2 + ... + 3
$\frac{1}{n}\sum_{i=1}^{n} x_i$	Sum values beginning at the first observation and ending at the last and divide by the number of observations (the mean)	(5 + 2 + 3)/3
$\bar{x}$	Exactly the same as the above - the mean of our data.	(5 + 2 + 3)/3

- Evaluate measures of spread
- **Range**
- **Interquartile Range (IQR)**
- **Standard Deviation**
- **Variance**

Histograms

Histograms are super useful for understanding the different aspects of data and they are the most common visual used for quantitative data. In the upcoming concepts, you will see histograms used all the time to help you understand the four aspects we outlined earlier regarding a quantitative variable:

- **center**
- **spread**
- **shape**
- **outliers**

How are Histograms constructed?

First, we need to bin our data. Each **bin** represents a range of values in a dataset. The number of values that fall in the range of each bin determines the height of each histogram bar. As shown in the video above, changing the range of our bins can result in slightly different visuals. However, there is no right or wrong answer in choosing how to bin, and in most cases, the software you use will choose the appropriate bins for you.

Calculating the 5 Number Summary

The five-number summary consist of 5 values:

- **Minimum:** The smallest number in the dataset.
- $Q1$ The value such that 25% of the data fall below.
- $Q2$ The value such that 50% of the data fall below.
- $Q3$ The value such that 75% of the data fall below.
- **Maximum:** The largest value in the dataset.

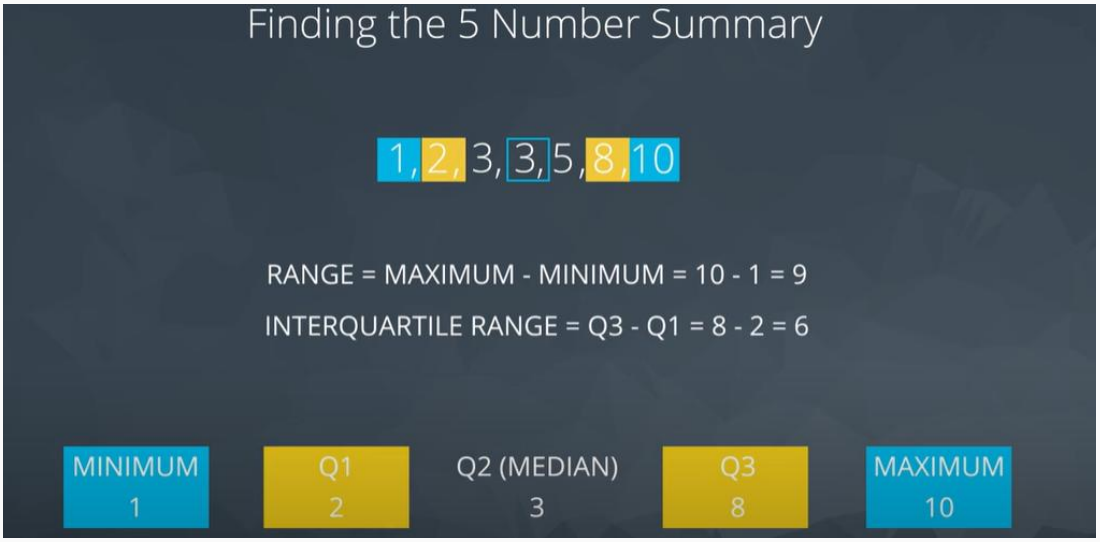
In the above video, we saw that calculating each of these values was essentially just finding the median of a bunch of different datasets. Because we are essentially calculating a bunch of medians, the calculation depends on whether we have an odd or even number of values.

Range

The **range** is then calculated as the difference between the **maximum** and the **minimum**.

IQR

The **interquartile range** is calculated as the difference between Q3 and Q1



Finding the 5 Number Summary

1, 2, 3, 3, 5, 8, 10

RANGE = MAXIMUM - MINIMUM = 10 - 1 = 9

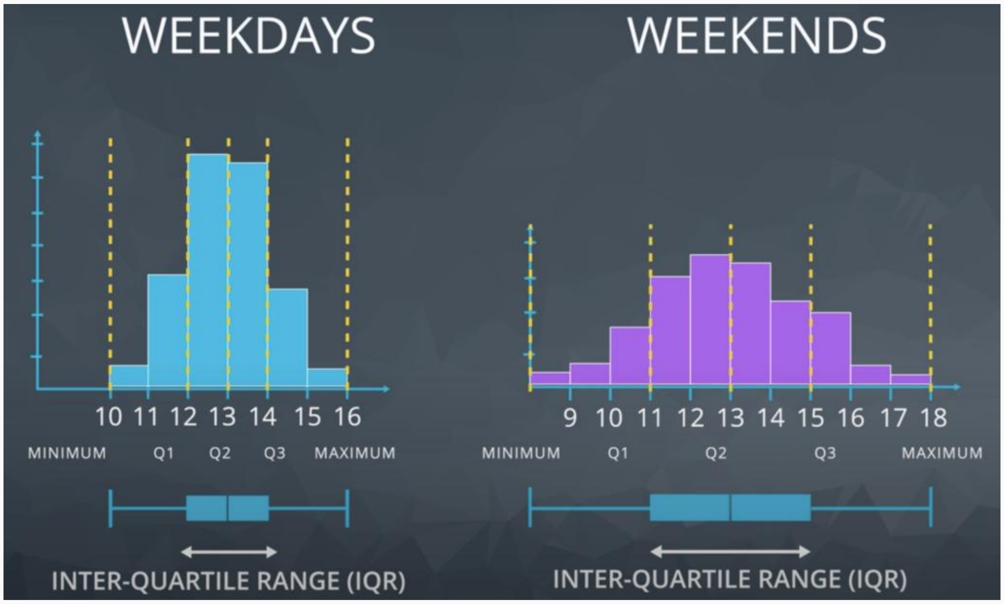
INTERQUARTILE RANGE = Q3 - Q1 = 8 - 2 = 6

MINIMUM	Q1	Q2 (MEDIAN)	Q3	MAXIMUM
1	2	3	8	10

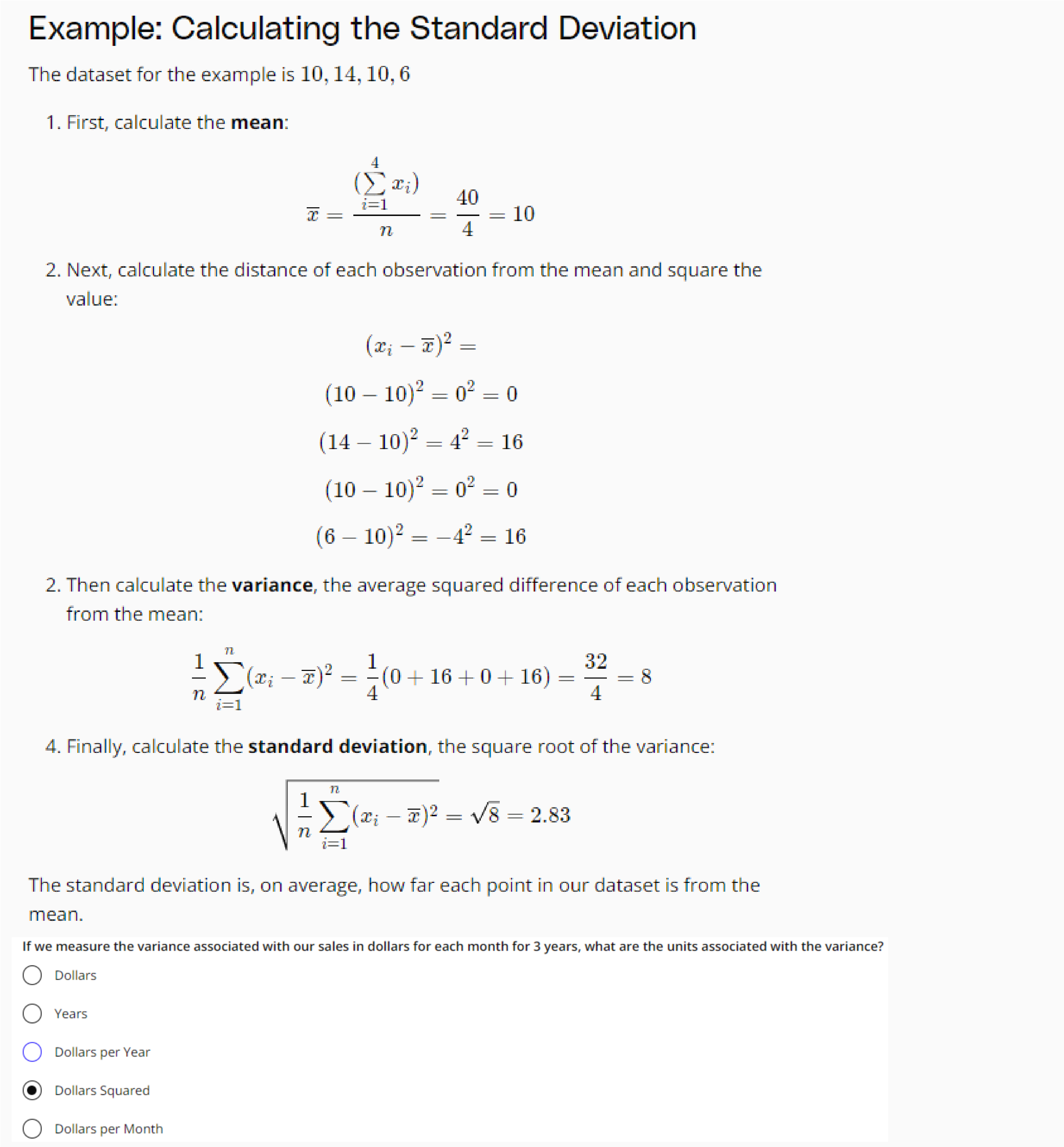
- **Box plots** are useful for quickly comparing the spread of two data sets across some key metrics, like quartiles, maximum, and minimum.

How do we create the box plot?

- The beginning of the line to the left of the box and the end of the line to the right of the box represent the minimum and maximum values in a dataset.
- The visual distance between these markings is an indication of the range of the values.
- The box itself represents the IQR. The box begins at the Q1 value, ends at the Q3 value, and Q2, or the median, is represented by a line within the box.



The **standard deviation** is one of the most common measures for talking about the spread of data. It is defined as **the average distance of each observation from the mean**.



Example: Calculating the Standard Deviation

The dataset for the example is $10, 14, 10, 6$

1. First, calculate the **mean**:

$$\overline{x} = \frac{(\sum_{i=1}^{4} x_i)}{n} = \frac{40}{4} = 10$$

2. Next, calculate the distance of each observation from the mean and square the value:

$$(x_i - \overline{x})^2 =$$

$$(10 - 10)^2 = 0^2 = 0$$

$$(14 - 10)^2 = 4^2 = 16$$

$$(10 - 10)^2 = 0^2 = 0$$

$$(6 - 10)^2 = -4^2 = 16$$

2. Then calculate the **variance**, the average squared difference of each observation from the mean:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{4}(0 + 16 + 0 + 16) = \frac{32}{4} = 8$$

4. Finally, calculate the **standard deviation**, the square root of the variance:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{8} = 2.83$$

The standard deviation is, on average, how far each point in our dataset is from the mean.

If we measure the variance associated with our sales in dollars for each month for 3 years, what are the units associated with the variance?

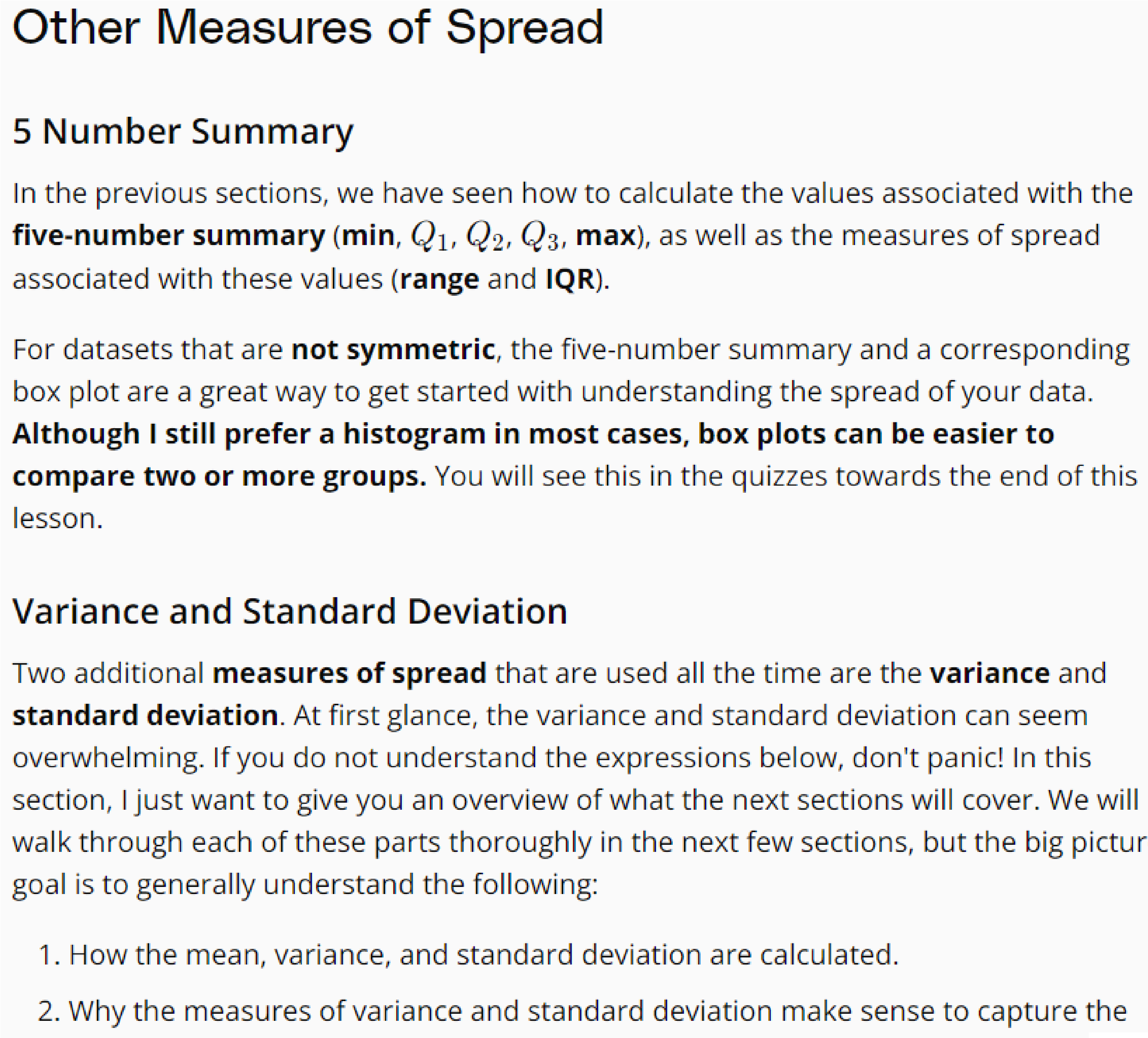
○ Dollars

○ Years

○ Dollars per Year

● Dollars Squared

○ Dollars per Month



Other Measures of Spread

5 Number Summary

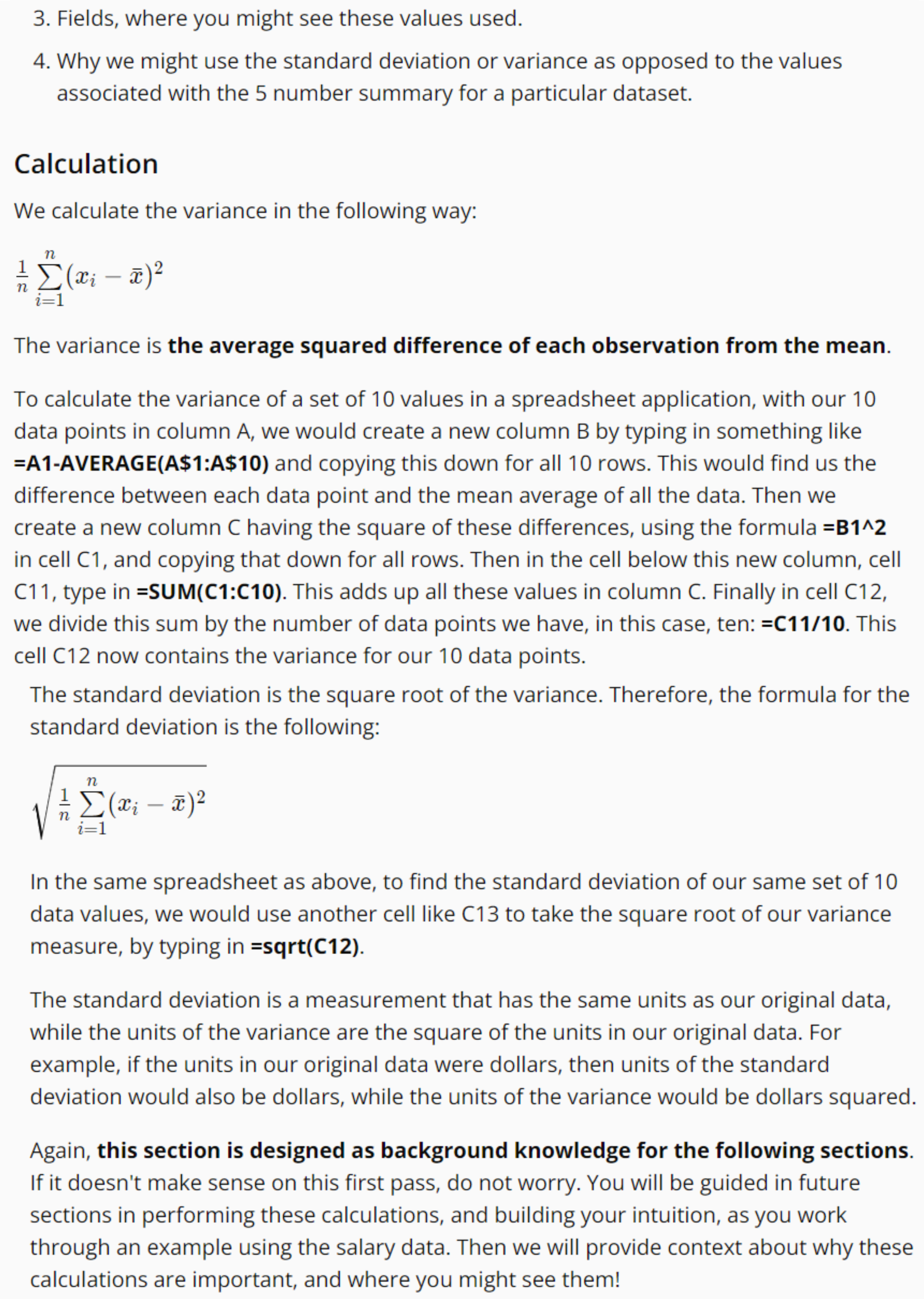
In the previous sections, we have seen how to calculate the values associated with the **five-number summary** (**min**, Q_1, Q_2, Q_3, **max**), as well as the measures of spread associated with these values (**range** and **IQR**).

For datasets that are **not symmetric**, the five-number summary and a corresponding box plot are a great way to get started with understanding the spread of your data. **Although I still prefer a histogram in most cases, box plots can be easier to compare two or more groups.** You will see this in the quizzes towards the end of this lesson.

Variance and Standard Deviation

Two additional **measures of spread** that are used all the time are the **variance** and **standard deviation**. At first glance, the variance and standard deviation can seem overwhelming. If you do not understand the expressions below, don't panic! In this section, I just want to give you an overview of what the next sections will cover. We will walk through each of these parts thoroughly in the next few sections, but the big pictur goal is to generally understand the following:

1. How the mean, variance, and standard deviation are calculated.
2. Why the measures of variance and standard deviation make sense to capture the



3. Fields, where you might see these values used.
4. Why we might use the standard deviation or variance as opposed to the values associated with the 5 number summary for a particular dataset.

Calculation

We calculate the variance in the following way:

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

The variance is **the average squared difference of each observation from the mean**.

To calculate the variance of a set of 10 values in a spreadsheet application, with our 10 data points in column A, we would create a new column B by typing in something like **=A1-AVERAGE(A$1:A$10)** and copying this down for all 10 rows. This would find us the difference between each data point and the mean average of all the data. Then we create a new column C having the square of these differences, using the formula **=B1^2** in cell C1, and copying that down for all rows. Then in the cell below this new column, cell C11, type in **=SUM(C1:C10)**. This adds up all these values in column C. Finally in cell C12, we divide this sum by the number of data points we have, in this case, ten: **=C11/10**. This cell C12 now contains the variance for our 10 data points.

The standard deviation is the square root of the variance. Therefore, the formula for the standard deviation is the following:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

In the same spreadsheet as above, to find the standard deviation of our same set of 10 data values, we would use another cell like C13 to take the square root of our variance measure, by typing in **=sqrt(C12)**.

The standard deviation is a measurement that has the same units as our original data, while the units of the variance are the square of the units in our original data. For example, if the units in our original data were dollars, then units of the standard deviation would also be dollars, while the units of the variance would be dollars squared.

Again, **this section is designed as background knowledge for the following sections**. If it doesn't make sense on this first pass, do not worry. You will be guided in future sections in performing these calculations, and building your intuition, as you work through an example using the salary data. Then we will provide context about why these calculations are important, and where you might see them!

Standard deviation is a common metric used to compare the spread of two datasets. The benefits of using a single metric instead of the 5 number summary are:

- It simplifies the amount of information needed to give a measure of spread
- It is useful for inferential statistics

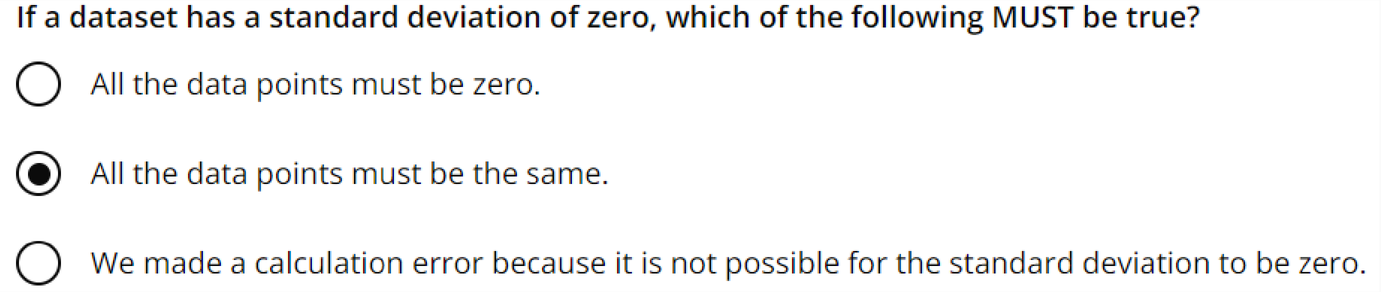
Important Final Points

- The variance is used to compare the spread of two different groups. A set of data with higher variance is more spread out than a dataset with lower variance. Be careful though, there might just be an outlier (or outliers) that is increasing the variance when most of the data are actually very close.
- When comparing the spread between two datasets, the units of each must be the same.
- When data are related to money or the economy, higher variance (or standard deviation) is associated with higher risk.
- The standard deviation is used more often in practice than the variance because it shares the units of the original dataset.

Use in the World

The standard deviation is associated with risk in finance, assists in determining the significance of drugs in medical studies, and measures the error of our results for predicting anything from the amount of rainfall we can expect tomorrow to your predicted commute time tomorrow.

These applications are beyond the scope of this lesson as they pertain to specific fields, but know that understanding the spread of a particular set of data is extremely important to many areas. In this lesson, you mastered the calculation of the most common measures of spread.

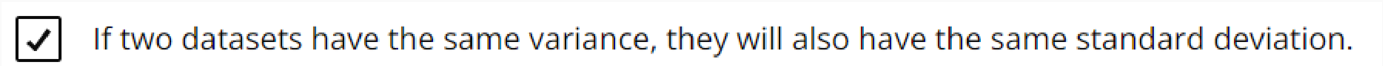


If a dataset has a standard deviation of zero, which of the following MUST be true?

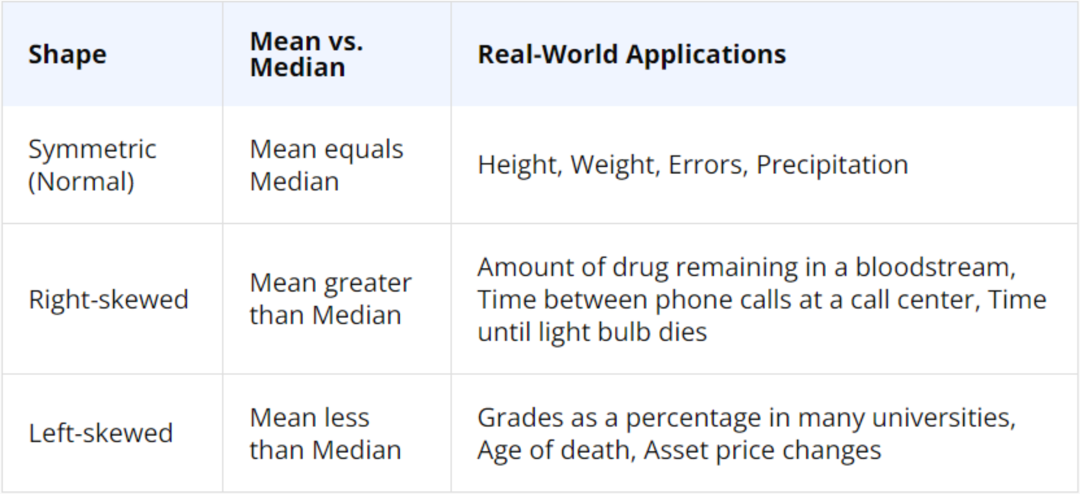
○ All the data points must be zero.

● All the data points must be the same.

○ We made a calculation error because it is not possible for the standard deviation to be zero.



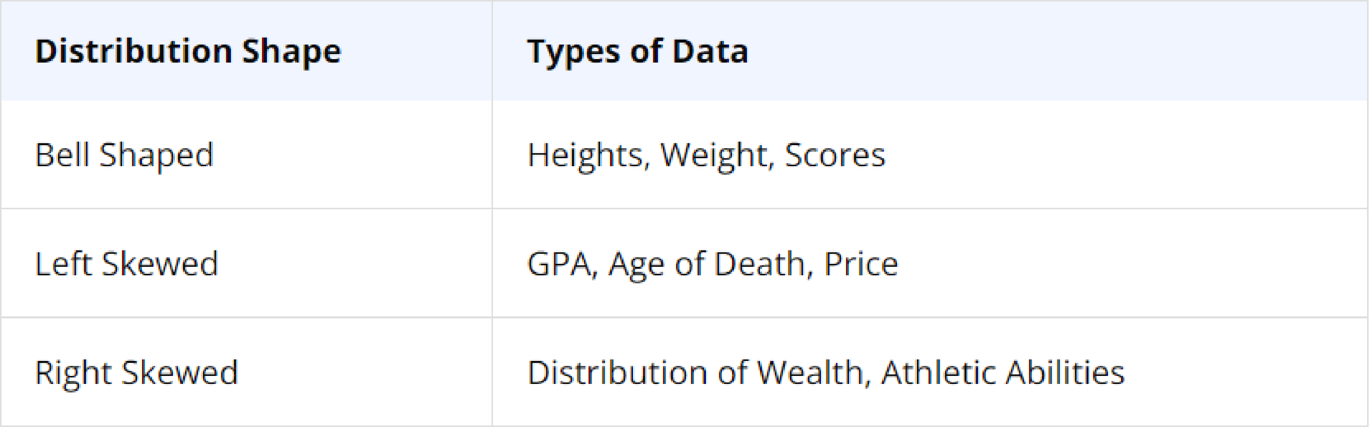
☑ If two datasets have the same variance, they will also have the same standard deviation.



Shape	Mean vs. Median	Real-World Applications
Symmetric (Normal)	Mean equals Median	Height, Weight, Errors, Precipitation
Right-skewed	Mean greater than Median	Amount of drug remaining in a bloodstream, Time between phone calls at a call center, Time until light bulb dies
Left-skewed	Mean less than Median	Grades as a percentage in many universities, Age of death, Asset price changes

Histograms

We learned how to build a **histogram** in this video, as this is the most popular visual for quantitative data.



Distribution Shape	Types of Data
Bell Shaped	Heights, Weight, Scores
Left Skewed	GPA, Age of Death, Price
Right Skewed	Distribution of Wealth, Athletic Abilities

outliers are points that fall very far from the rest of our data points. This influences measures like the mean and standard deviation much more than measures associated with the five-number summary.

Common Techniques

When outliers are present we should consider the following points.

1. Noting they exist and the impact on summary statistics.

2. If typo - remove or fix

3. Understanding why they exist, and the impact on questions we are trying to answer about our data.

4. Reporting the 5 number summary values is often a better indication than measures like the mean and standard deviation when we have outliers.

5. Be careful in reporting. Know how to ask the right questions.

`Descriptive statistics` **is about describing our collected data**.

`Inferential Statistics` **is about using our collected data to draw conclusions about a larger population**.

We looked at specific examples that allowed us to identify the

- **Population** - our entire group of interest.
- **Parameter** - numeric summary about a population
- **Sample** - a subset of the population
- **Statistic** numeric summary about a sample

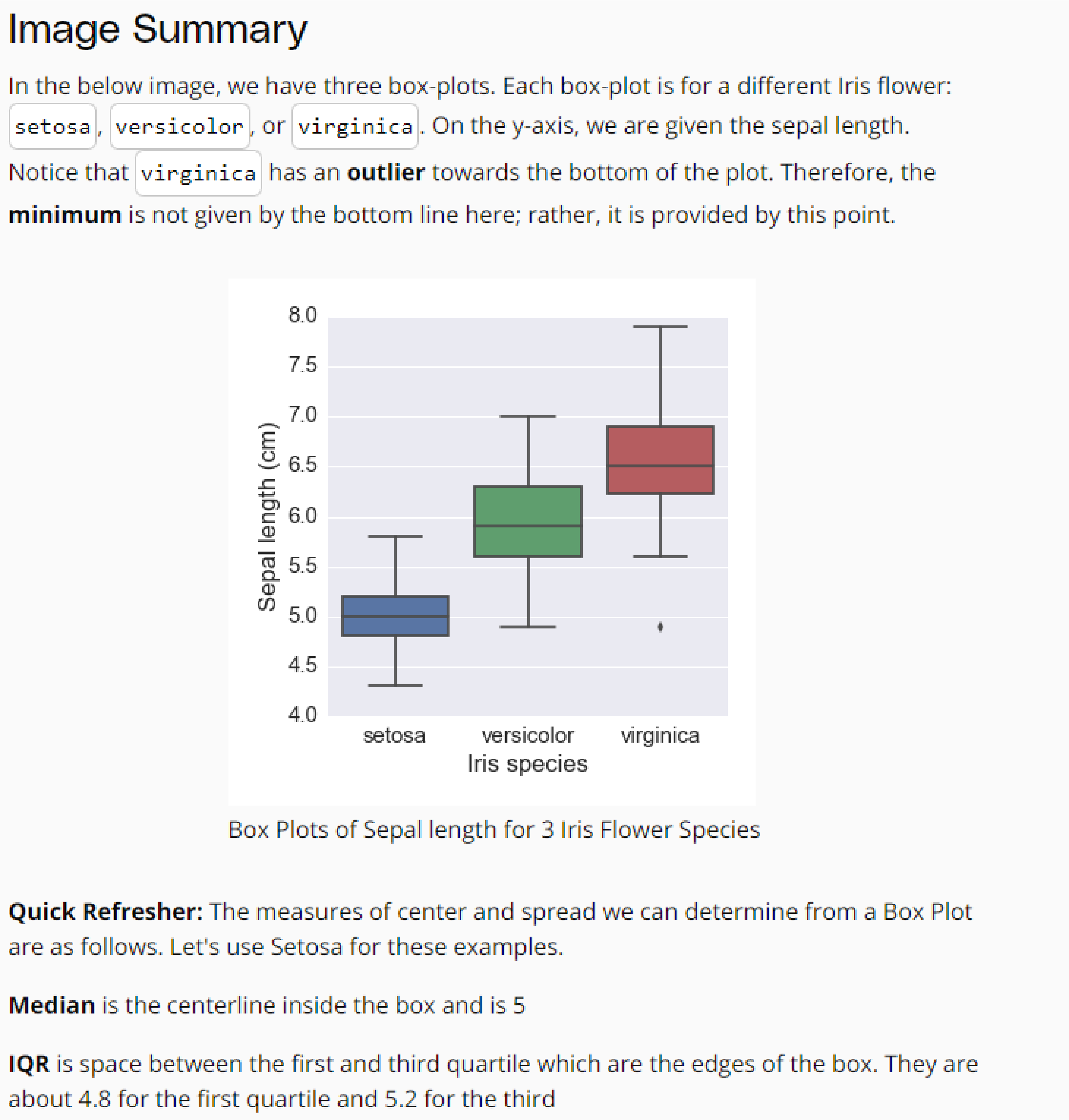


Image Summary

In the below image, we have three box-plots. Each box-plot is for a different Iris flower: `setosa`, `versicolor`, or `virginica`. On the y-axis, we are given the sepal length. Notice that `virginica` has an **outlier** towards the bottom of the plot. Therefore, the **minimum** is not given by the bottom line here; rather, it is provided by this point.

Box Plots of Sepal length for 3 Iris Flower Species

Quick Refresher: The measures of center and spread we can determine from a Box Plot are as follows. Let's use Setosa for these examples.

Median is the centerline inside the box and is 5

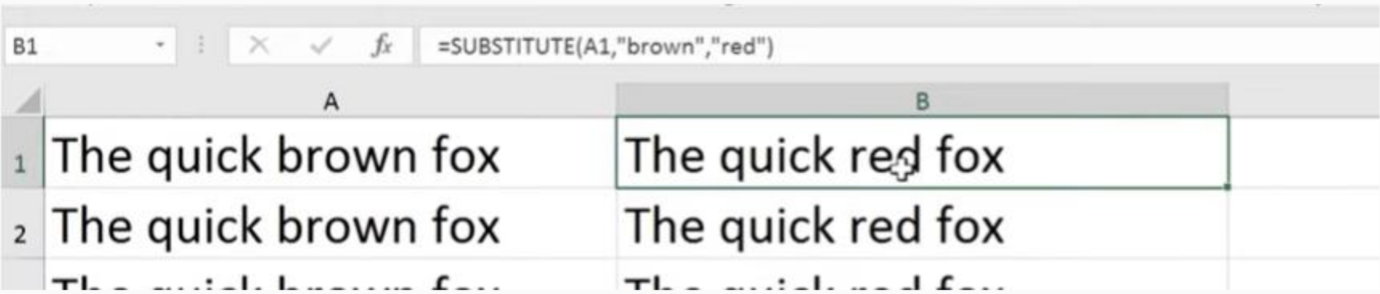
IQR is space between the first and third quartile which are the edges of the box. They are about 4.8 for the first quartile and 5.2 for the third

EXcel

Text String: String of letters, numbers, and punctuation that is not treated numerically.

While the **SUBSTITUTE** function sounds similar to find/replace, it is used for different purposes. Find/replace gets rid of the old data, while `SUBSTITUTE` will not change the original cell, instead showing the transformed data in a new cell.

SUBSTITUTE uses the syntax `SUBSTITUTE({text}, {old_text}, {new_text})`, where `{text}` is the cell to change, `{old_text}` is the string sequence to be replaced, and `{new_text}` is the new string in place of the old one.

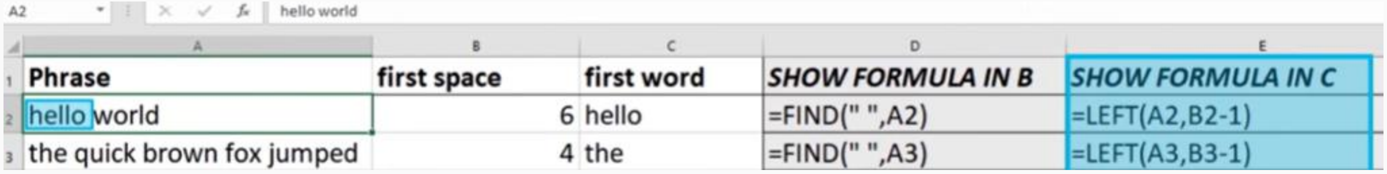


=SUBSTITUTE(A1,"brown","red")

	A	B
1	The quick brown fox	The quick red fox
2	The quick brown fox	The quick red fox

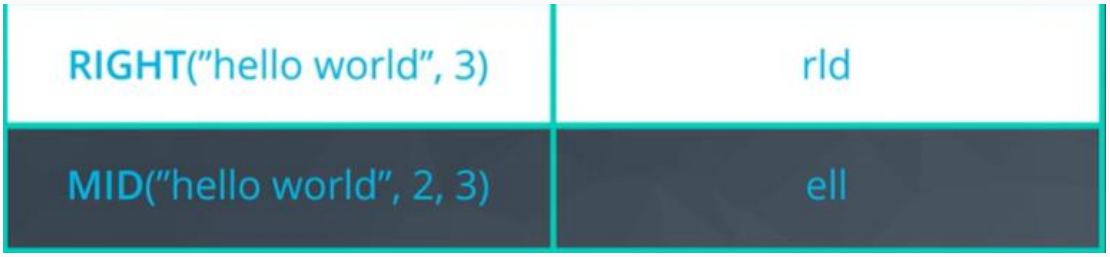
`FIND` and `LEFT` can be used to extract text. `FIND` can be given a substring and a cell to return the position in a string where the substring was found. `LEFT` can then be used to extract a certain number of characters from a cell, starting from the left side.

`RIGHT` therefore extracts from the right side, while `MID` can extract from some starting point in the middle of a cell.

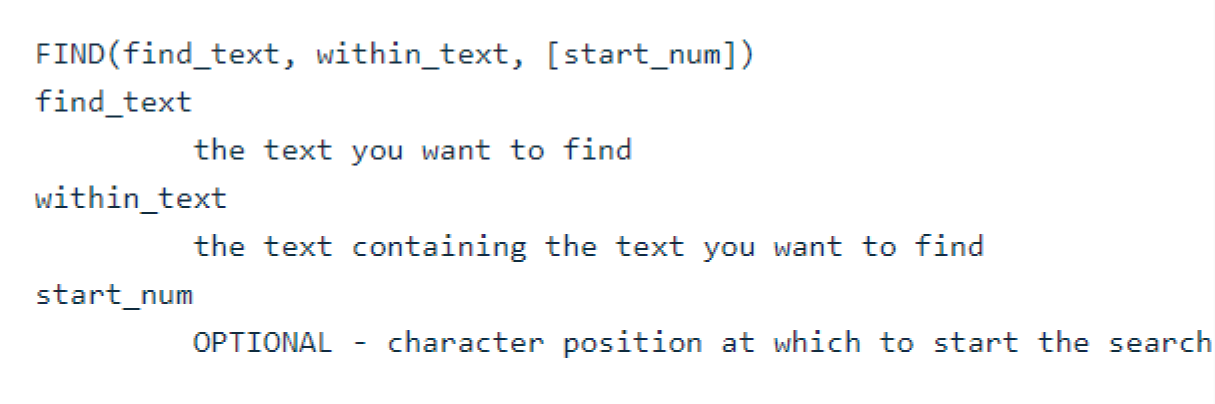


hello world

	A	B	C	D	E
1	Phrase	first space	first word	*SHOW FORMULA IN B*	*SHOW FORMULA IN C*
2	hello world	6	hello	=FIND(" ",A2)	=LEFT(A2,B2-1)
3	the quick brown fox jumped	4	the	=FIND(" ",A3)	=LEFT(A3,B3-1)



RIGHT("hello world", 3)	rld
MID("hello world", 2, 3)	ell



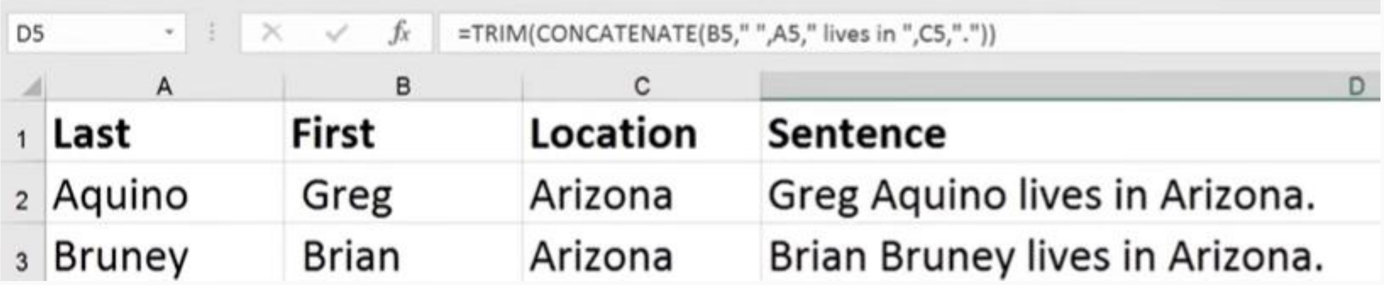
```
FIND(find_text, within_text, [start_num])
find_text
         the text you want to find
within_text
         the text containing the text you want to find
start_num
         OPTIONAL - character position at which to start the search
```

`CONCATENATE` will join together two or more strings. It's important to note that this will not automatically add spaces between them, so make sure to add spaces as formula parameters if you need them.

`TRIM` will help to remove excess whitespace from a string.

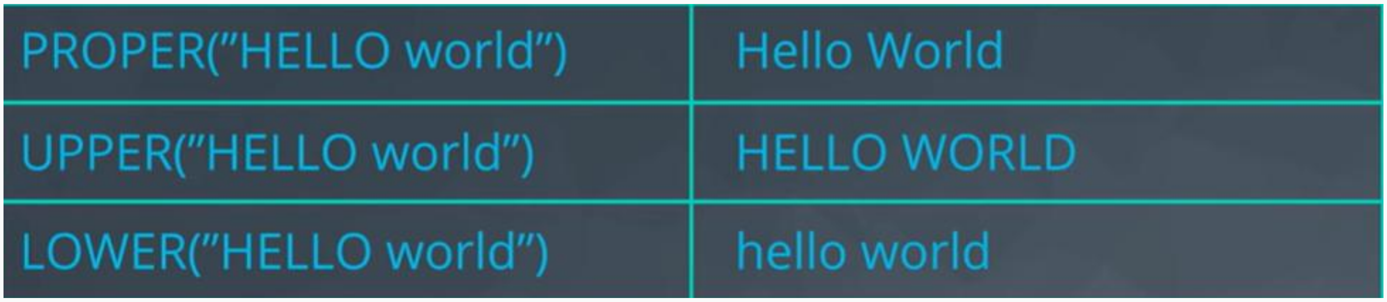
`PROPER` sets the first letter of each word to upper case, with the rest lowercase.

`UPPER` sets all letters to upper case, while `LOWER` sets all letters to lowercase.



=TRIM(CONCATENATE(B5," ",A5," lives in ",C5,"."))

	A	B	C	D
1	**Last**	**First**	**Location**	**Sentence**
2	Aquino	Greg	Arizona	Greg Aquino lives in Arizona.
3	Bruney	Brian	Arizona	Brian Bruney lives in Arizona.



PROPER("HELLO world")	Hello World
UPPER("HELLO world")	HELLO WORLD
LOWER("HELLO world")	hello world

Math operations are one of the most common spreadsheet usages. These are used similarly to what one might expect (with a leading equals sign):

- `+` for addition
- `-` for subtraction
- `*` for multiplication
- `/` for division

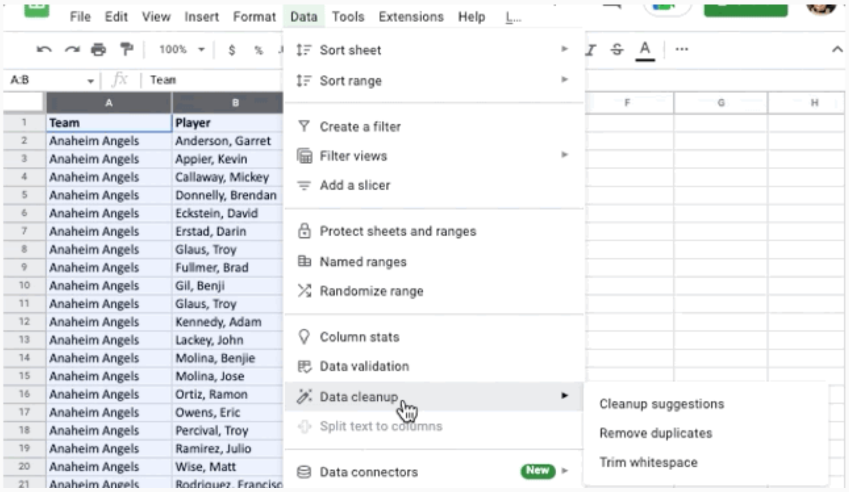
There are also the functions `SUM` and `AVERAGE`, which behave as their names suggest - summing or averaging two or more cells, numbers or a range of cells.

Removing Duplicate Rows with Google Sheets

If you are using Google Sheets you can use **Data Cleanup**

Select ***Data*** > ***Data cleanup*** > ***Remove duplicates***.

Next, select ***Data has header row*** and choose the rows you want analyzed.

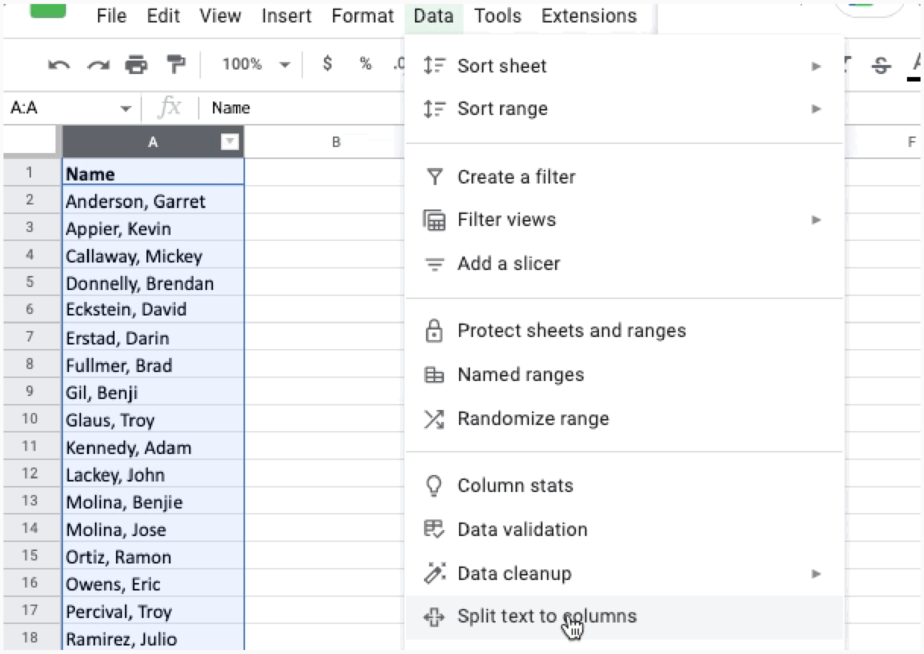


	A	B
1	Team	Player
2	Anaheim Angels	Anderson, Garret
3	Anaheim Angels	Appier, Kevin
4	Anaheim Angels	Callaway, Mickey
5	Anaheim Angels	Donnelly, Brendan
6	Anaheim Angels	Eckstein, David
7	Anaheim Angels	Erstad, Darin
8	Anaheim Angels	Glaus, Troy
9	Anaheim Angels	Fullmer, Brad
10	Anaheim Angels	Gil, Benji
11	Anaheim Angels	Glaus, Troy
12	Anaheim Angels	Kennedy, Adam
13	Anaheim Angels	Lackey, John
14	Anaheim Angels	Molina, Benjie
15	Anaheim Angels	Molina, Jose
16	Anaheim Angels	Ortiz, Ramon
17	Anaheim Angels	Owens, Eric
18	Anaheim Angels	Percival, Troy
19	Anaheim Angels	Ramirez, Julio
20	Anaheim Angels	Wise, Matt
21	Anaheim Angels	Rodriguez, Francisc

Splitting Columns in Google Sheets

If you are working in Google Sheets, you click ***Data*** > ***Split text to columns***.

If you want to override the default separator, click in the ***Separator*** box. You can choose from comma, semicolon, period, or space -- or you can select ***Custom*** to add your own separator.



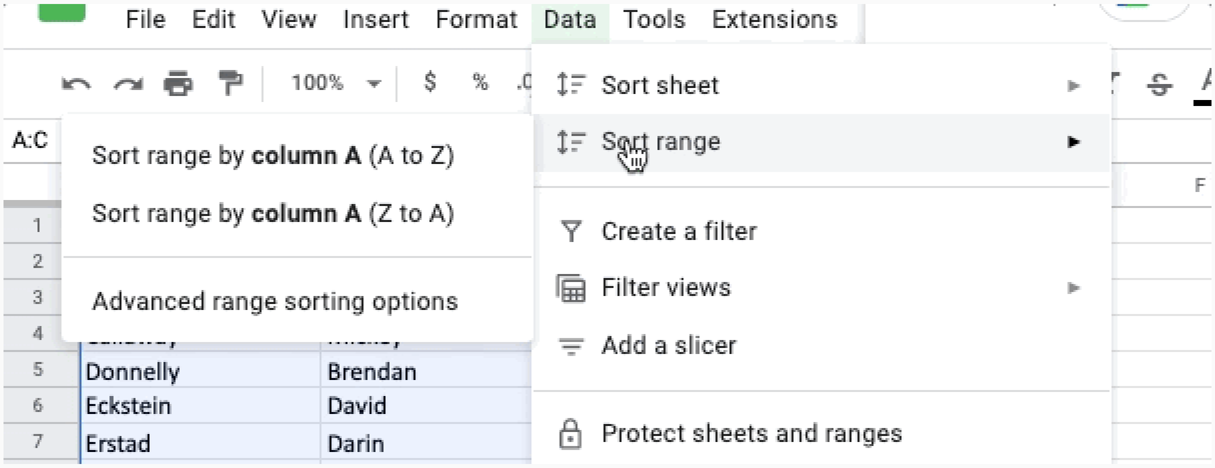
	A	B
1	Name	
2	Anderson, Garret	
3	Appier, Kevin	
4	Callaway, Mickey	
5	Donnelly, Brendan	
6	Eckstein, David	
7	Erstad, Darin	
8	Fullmer, Brad	
9	Gil, Benji	
10	Glaus, Troy	
11	Kennedy, Adam	
12	Lackey, John	
13	Molina, Benjie	
14	Molina, Jose	
15	Ortiz, Ramon	
16	Owens, Eric	
17	Percival, Troy	
18	Ramirez, Julio	

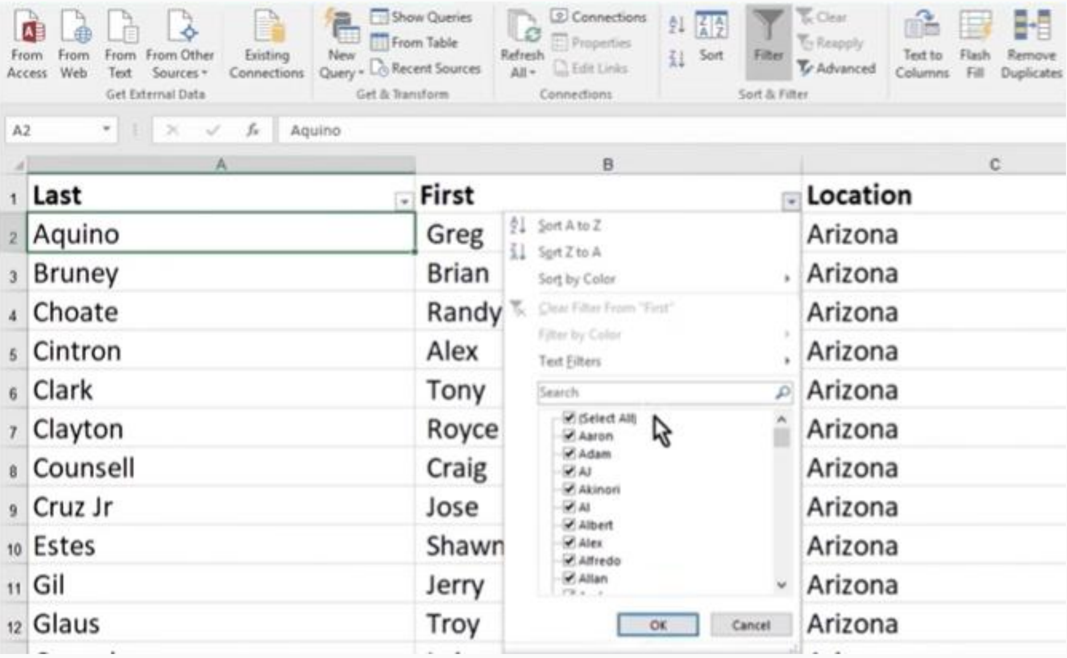
Sorting with Google Sheets

If you are using Google Sheets, the process is similar.

1. Select the range you want to sort.
2. Click ***Data*** > ***Sort range*** > ***Advanced range sorting options***.
3. Select ***Data has header row***.
4. Select sort columns from the drop down.

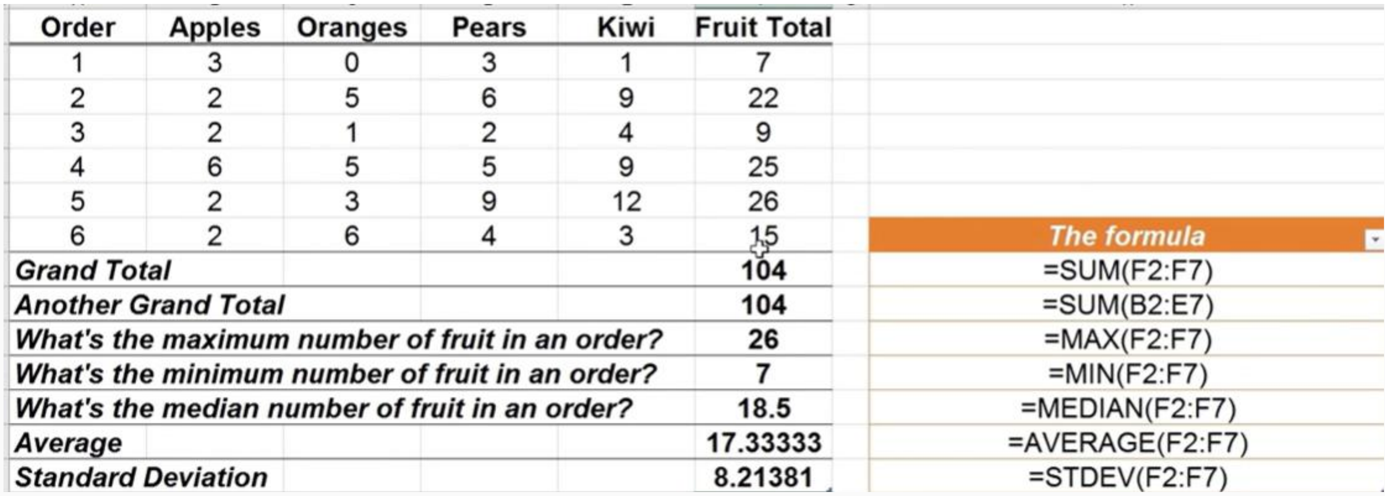
You can select ***Add another sort column*** to add another sort level.





	A	B	C
1	Last	First	Location
2	Aquino	Greg	Arizona
3	Bruney	Brian	Arizona
4	Choate	Randy	Arizona
5	Cintron	Alex	Arizona
6	Clark	Tony	Arizona
7	Clayton	Royce	Arizona
8	Counsell	Craig	Arizona
9	Cruz Jr	Jose	Arizona
10	Estes	Shawn	Arizona
11	Gil	Jerry	Arizona
12	Glaus	Troy	Arizona

Data Filter

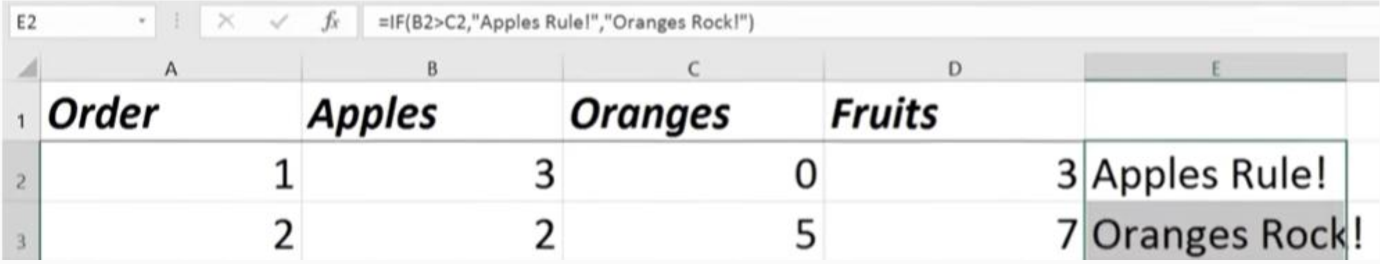


Order	Apples	Oranges	Pears	Kiwi	Fruit Total	
1	3	0	3	1	7	
2	2	5	6	9	22	
3	2	1	2	4	9	
4	6	5	5	9	25	
5	2	3	9	12	26	
6	2	6	4	3	15	*The formula*
Grand Total					104	=SUM(F2:F7)
Another Grand Total					104	=SUM(B2:E7)
What's the maximum number of fruit in an order?					26	=MAX(F2:F7)
What's the minimum number of fruit in an order?					7	=MIN(F2:F7)
What's the median number of fruit in an order?					18.5	=MEDIAN(F2:F7)
Average					17.33333	=AVERAGE(F2:F7)
Standard Deviation					8.21381	=STDEV(F2:F7)

.The IF function can return different values based on whether a condition is true or false. The first parameter is the condition, the second is what the cell value should be if the condition is true, and the optional third parameter is the cell value if the condition is false (skipping the third parameter will otherwise just show "FALSE" in the cell).

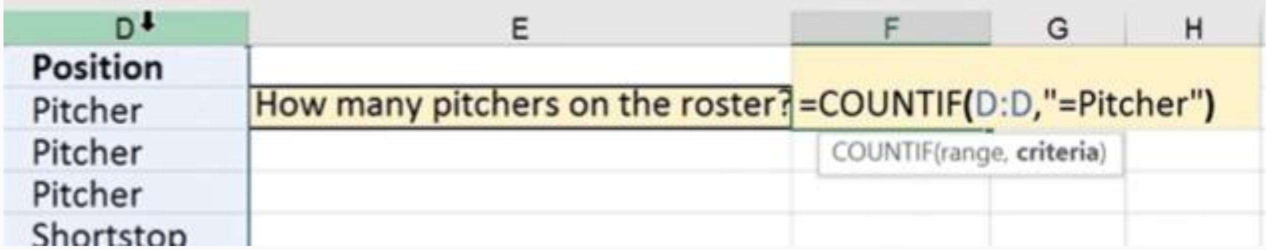
Comparison Operator: Compare the relative size or equality of two values with these operators. The result is a logical value of either true or false. The operators are as follows:

- > for greater than
- < for less than
- = for equal
- >= for greater than or equal to
- <= for less than or equal to
- <> for not equal (note that are no equal signs!)

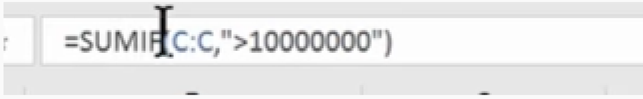


E2 =IF(B2>C2,"Apples Rule!","Oranges Rock!")

	A	B	C	D	E
1	Order	Apples	Oranges	Fruits	
2	1	3	0	3	Apples Rule!
3	2	2	5	7	Oranges Rock!



D	E	F	G	H
Position				
Pitcher	How many pitchers on the roster?	=COUNTIF(D:D,"=Pitcher")		
Pitcher		COUNTIF(range, criteria)		
Pitcher				
Shortstop				

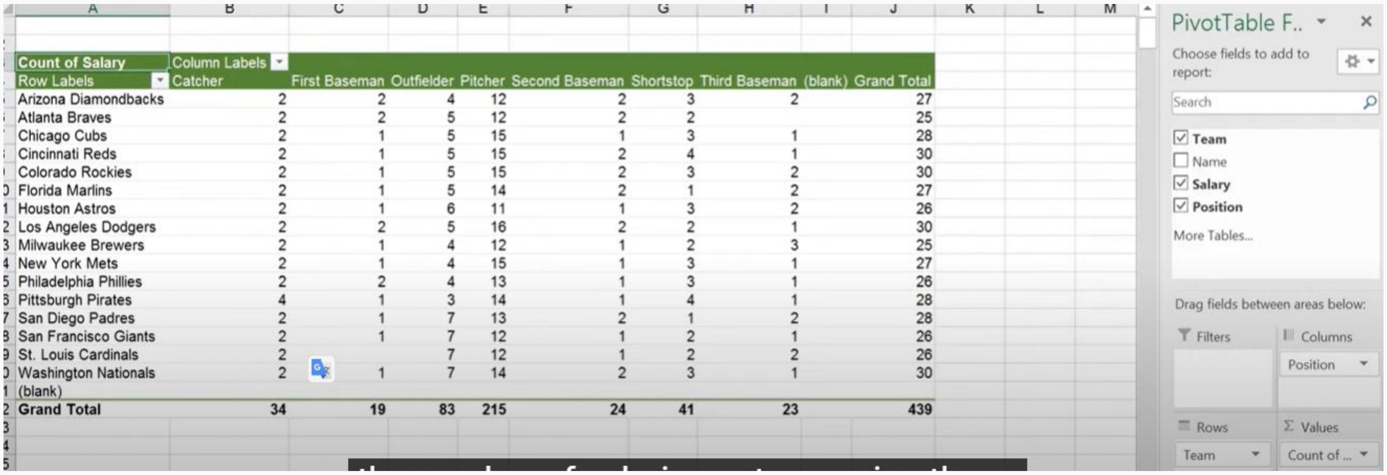


=SUMIF(C:C,">10000000")

to get salaries greater than 10M

In **Google Sheets** pivot tables can be found in the `Insert` menu:

- Select the data
- Click `Insert` and then `Pivot table`
- Choose `New sheet` or `Existing sheet` and click `Create`
- Select values for Rows, Columns, and Values



Count of Salary	Column Labels								
Row Labels	Catcher	First Baseman	Outfielder	Pitcher	Second Baseman	Shortstop	Third Baseman	(blank)	Grand Total
Arizona Diamondbacks	2	2	4	12	2	3	2		27
Atlanta Braves	2	2	5	12	2	2			25
Chicago Cubs	2	1	5	15	1	3	1		28
Cincinnati Reds	2	1	5	15	2	4	1		30
Colorado Rockies	2	1	5	15	2	3	2		30
Florida Marlins	2	1	5	14	2	1	2		27
Houston Astros	2	1	6	11	1	3	2		26
Los Angeles Dodgers	2	2	5	16	2	2	1		30
Milwaukee Brewers	2	1	4	12	1	2	3		25
New York Mets	2	1	4	15	1	3	1		27
Philadelphia Phillies	2	2	4	13	1	3	1		26
Pittsburgh Pirates	4	1	3	14	1	4	1		28
San Diego Padres	2	1	7	13	2	1	2		28
San Francisco Giants	2	1	7	12	1	2	1		26
St. Louis Cardinals	2		7	12	1	2	2		26
Washington Nationals	2	1	7	14	2	3	1		30
(blank)									
Grand Total	**34**	**19**	**83**	**215**	**24**	**41**	**23**		**439**

PivotTable F..

Choose fields to add to report:

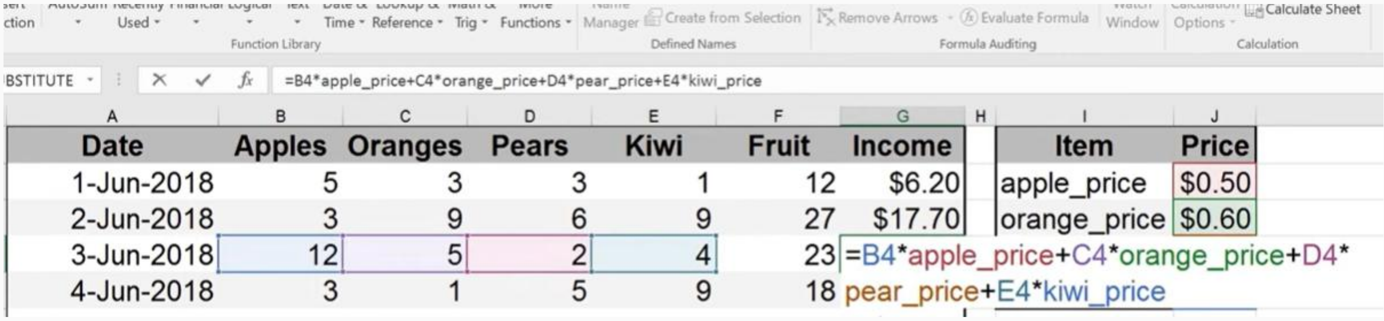
☑ Team
□ Name
☑ Salary
☑ Position

More Tables...

Drag fields between areas below:

Filters | Columns: Position | Rows: Team | Values: Count of ...

Naming Range: select from item and then “create from selection”

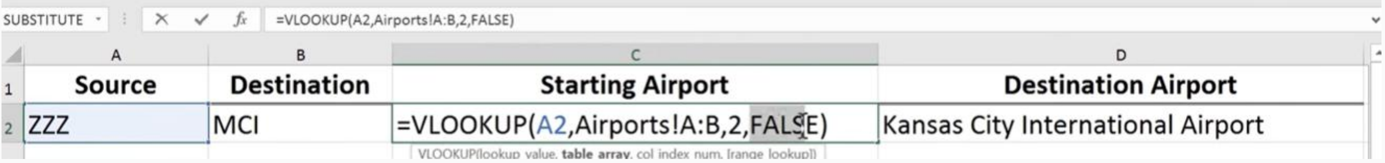


`=B4*apple_price+C4*orange_price+D4*pear_price+E4*kiwi_price`

Date	Apples	Oranges	Pears	Kiwi	Fruit	Income
1-Jun-2018	5	3	3	1	12	$6.20
2-Jun-2018	3	9	6	9	27	$17.70
3-Jun-2018	12	5	2	4	23	=B4*apple_price+C4*orange_price+D4*pear_price+E4*kiwi_price
4-Jun-2018	3	1	5	9	18	

Item	Price
apple_price	$0.50
orange_price	$0.60

Lookup Function: A function that uses a keyword and index to "look up" a value in a table. There are both horizontal and vertical lookup functions, although we will focus on a vertical one called `VLOOKUP`.



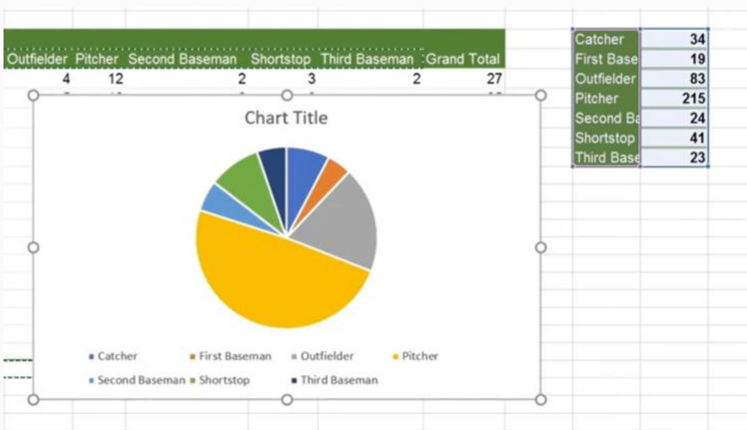
`=VLOOKUP(A2,Airports!A:B,2,FALSE)`

Source	Destination	Starting Airport	Destination Airport
ZZZ	MCI	=VLOOKUP(A2,Airports!A:B,2,FALSE)	Kansas City International Airport

VLOOKUP(lookup_value, table_array, col_index_num, [range_lookup])

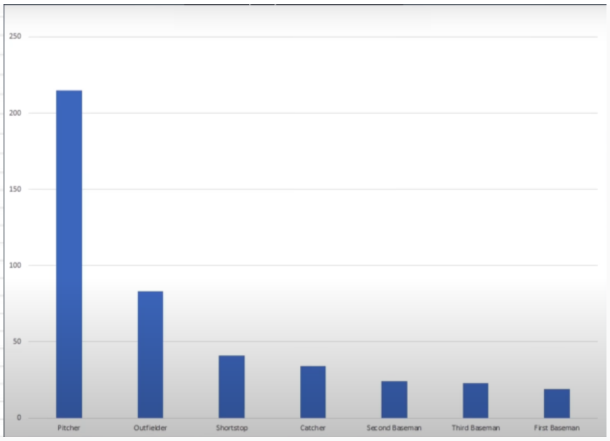
Bar or pie chart?

- Use bar or column charts to compare category values with each other.
- Use a pie chart to show the proportionality of categories.

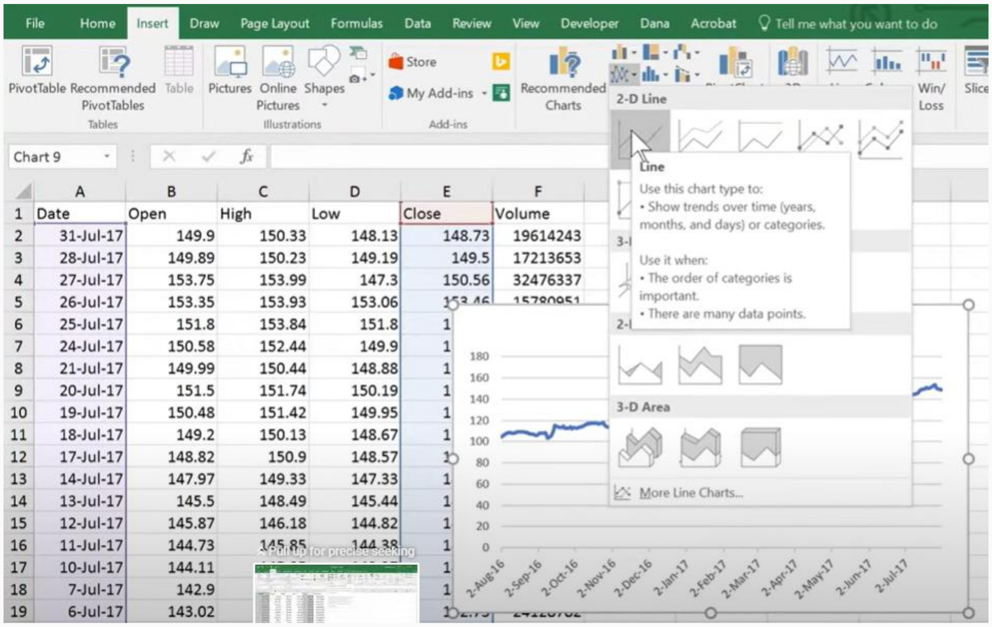


Outfielder	Pitcher	Second Baseman	Shortstop	Third Baseman	Grand Total
4	12	2	3	2	27

Catcher	34
First Base	19
Outfielder	83
Pitcher	215
Second Ba	24
Shortstop	41
Third Base	23



If we have a list of numerical data, such as the list of stock prices over time, a line chart gives us a better picture of the data set.



Line

Use this chart type to:
• Show trends over time (years, months, and days) or categories.

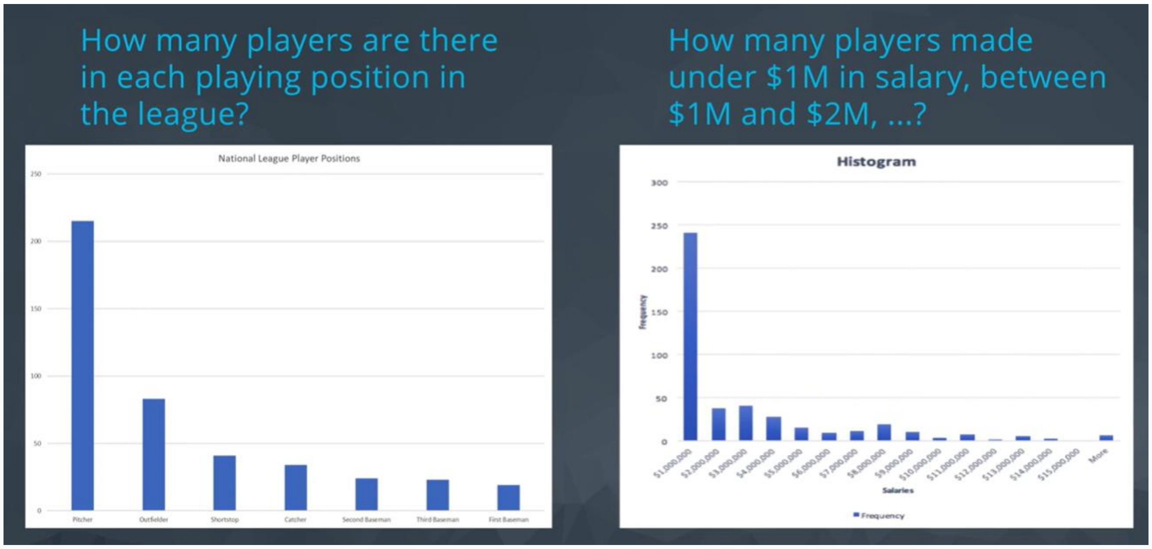
Use it when:
• The order of categories is important.
• There are many data points.

	A	B	C	D	E	F
1	Date	Open	High	Low	Close	Volume
2	31-Jul-17	149.9	150.33	148.13	148.73	19614243
3	28-Jul-17	149.89	150.23	149.19	149.5	17213653
4	27-Jul-17	153.75	153.99	147.3	150.56	32476337
5	26-Jul-17	153.35	153.93	153.06	153.46	15780951
6	25-Jul-17	151.8	153.84	151.8	1	
7	24-Jul-17	150.58	152.44	149.9	1	
8	21-Jul-17	149.99	150.44	148.88	1	
9	20-Jul-17	151.5	151.74	150.19	1	
10	19-Jul-17	150.48	151.42	149.95	1	
11	18-Jul-17	149.2	150.13	148.67	1	
12	17-Jul-17	148.82	150.9	148.57	1	
13	14-Jul-17	147.97	149.33	147.33	1	
14	13-Jul-17	145.5	148.49	145.44	1	
15	12-Jul-17	145.87	146.18	144.82	1	
16	11-Jul-17	144.73	145.85	144.38	1	
17	10-Jul-17	144.11			1	
18	7-Jul-17	142.9			1	
19	6-Jul-17	143.02			1	

A Histogram is a column chart that measures the frequency of data in a data set and specifically groups numerical values into bins we define.

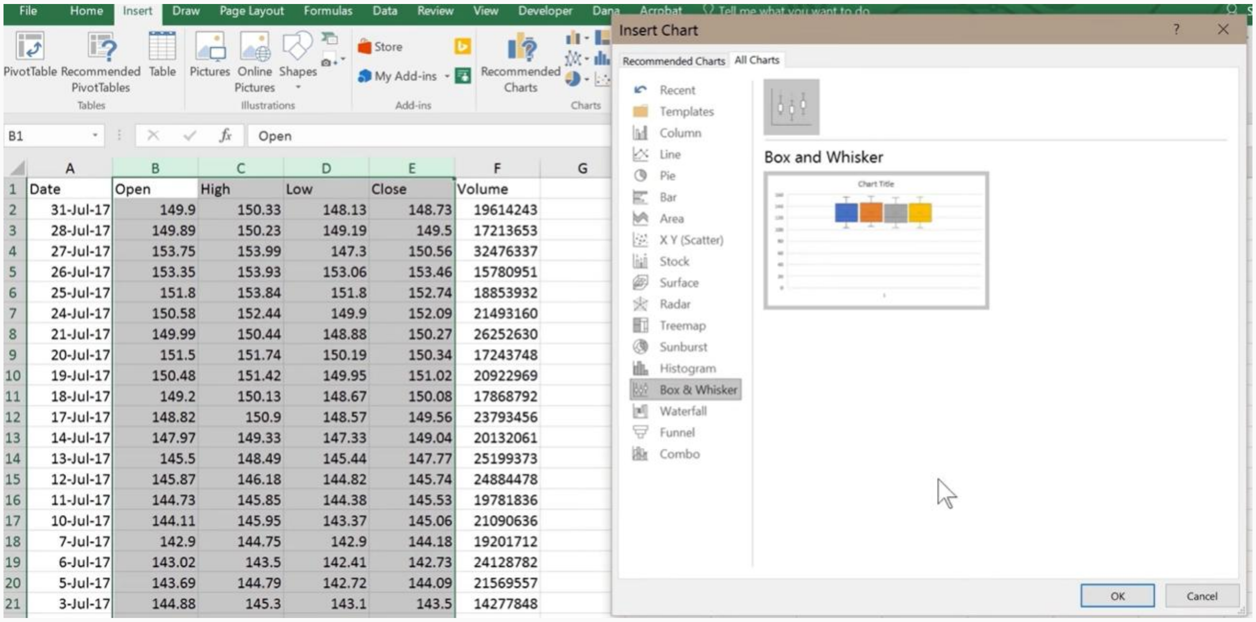
Column Charts vs Histogram

- Recall that we previously created a **column chart** to compare counts of categories within a data set. This kind of chart answers a question like: how many players are there in each playing position in the league?
- But what if we want to ask the question: how many players made under $1 million in salary, and between $1 and $2 million, and between $2 and $3 million in salary? This kind of chart is called a histogram, and the groupings we choose such as, 1) all salaries between $1 and $2 million, and 2) salaries between $2 and $3 million, are the bins.



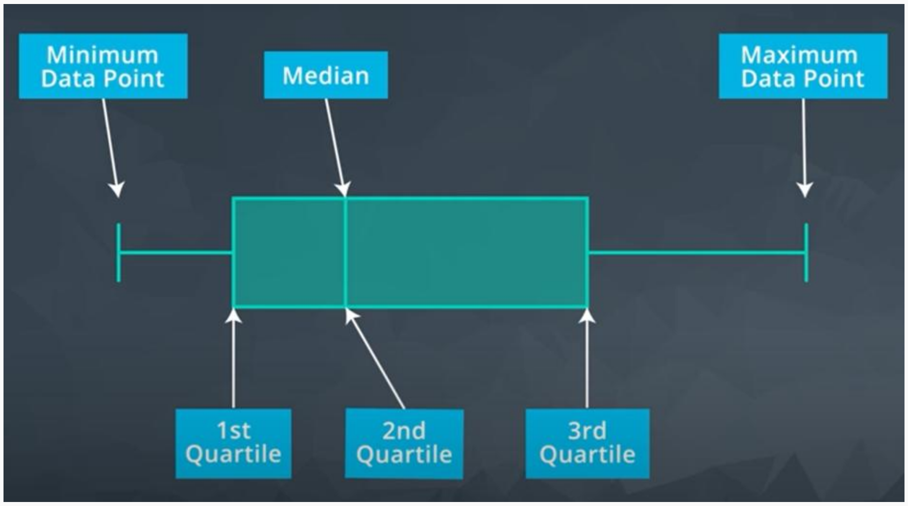
How many players are there in each playing position in the league?

How many players made under $1M in salary, between $1M and $2M, ...?



B1 fx Open

	A	B	C	D	E	F	G
1	Date	Open	High	Low	Close	Volume	
2	31-Jul-17	149.9	150.33	148.13	148.73	19614243	
3	28-Jul-17	149.89	150.23	149.19	149.5	17213653	
4	27-Jul-17	153.75	153.99	147.3	150.56	32476337	
5	26-Jul-17	153.35	153.93	153.06	153.46	15780951	
6	25-Jul-17	151.8	153.84	151.8	152.74	18853932	
7	24-Jul-17	150.58	152.44	149.9	152.09	21493160	
8	21-Jul-17	149.99	150.44	148.88	150.27	26252630	
9	20-Jul-17	151.5	151.74	150.19	150.34	17243748	
10	19-Jul-17	150.48	151.42	149.95	151.02	20922969	
11	18-Jul-17	149.2	150.13	148.67	150.08	17868792	
12	17-Jul-17	148.82	150.9	148.57	149.56	23793456	
13	14-Jul-17	147.97	149.33	147.33	149.04	20132061	
14	13-Jul-17	145.5	148.49	145.44	147.77	25199373	
15	12-Jul-17	145.87	146.18	144.82	145.74	24884478	
16	11-Jul-17	144.73	145.85	144.38	145.53	19781836	
17	10-Jul-17	144.11	145.95	143.37	145.06	21090636	
18	7-Jul-17	142.9	144.75	142.9	144.18	19201712	
19	6-Jul-17	143.02	143.5	142.41	142.73	24128782	
20	5-Jul-17	143.69	144.79	142.72	144.09	21569557	
21	3-Jul-17	144.88	145.3	143.1	143.5	14277848	



Minimum Data Point

Median

Maximum Data Point

1st Quartile

2nd Quartile

3rd Quartile

Which KPI to use?

The decision regarding which KPI a business analyst should use depends on several factors, including which industry or domain they are working in, which business function they are focusing on, and the type of data they have available to them.

Asking Data Questions

The KPIs you use will be determined by the questions you need to ask. As a business analyst, you are tasked with gathering the appropriate data to help solve business problems. To get to that solution, you will need to:

- Identify what needs to change
- Communicate this change to stakeholders in clear manageable chunks of data

How to Ask the Questions

There are several steps needed to determine the questions to ask—

- Identify the business goal and objectives.
- Narrow down the type of data needed to answer questions.
- Identify the KPIs that will be useful to show whether you are making progress on your business goal.
- Conduct the data analysis using the KPIs and use visualizations as part of the analysis.
- Provide recommendations and findings based on the completed data analysis.
- Create succinct and visual presentations for the stakeholders.

Online marketing can take advantage of **cookie tracking**, which allows customer tracking across time and platforms.

Knowing a customer's online roadmap enables companies to pinpoint places for targeted advertising to these customers and other potential customers like them.

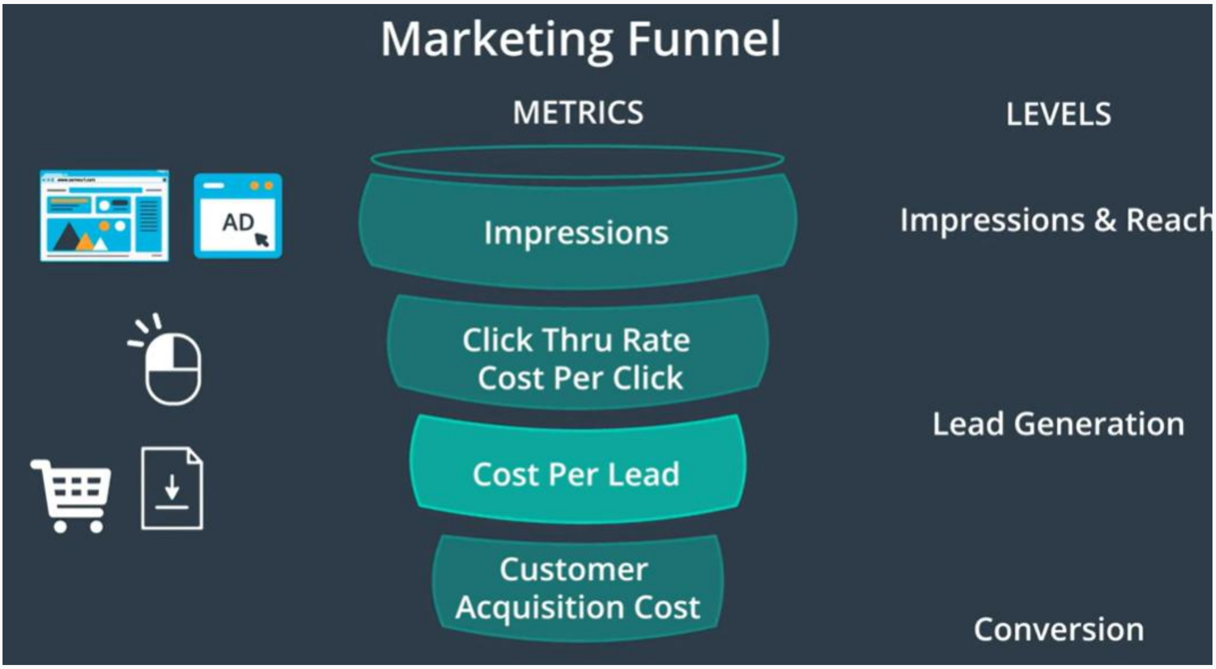
To grow the business, companies need to not only focus on existing customers, but also on new customers. This problem is at the heart of the growth metric. Executive boards, investors, and sales teams are constantly keeping their eye on this critical question about a company's overall health.

The visitor/customer journey described above can be described as 5 stages in which marketing teams used various digital tools to attract potential customers and convert them into actual customers. :

- **Awareness**
- **Interest**
- **Desire**
- **Purchase**
- **Post-purchase**

- **Call To Action (CTA)**: A marketing term that refers to an action a website visitor is supposed to take when given a specific prompt on a website. These can be words or phrases, or icons that prompt and encourage the user to perform the action.
- **Post-Purchase**: Actions customers take after purchasing an item that promotes and increase sales and advocate on behalf of the company. For e.g., coming back and purchasing more items, sharing or liking the company or product on social media, taking pictures of the item, and tagging it on Pinterest.

The **marketing funnel** is the process of tracking and analyzing each step of the customer journey with data.



Marketing Funnel Metrics

Impressions & Reach – building brand and product awareness using ad platforms and search engine optimization (SEO). SEO allows ads to show up for the right mix of search terms as people search online

- **Impressions** – an instance of an advertisement appearing on a website when it is viewed by a visitor.

Lead generation – measures how many visits are made to the website.

- **Click –** every time a website visitor views the ad and clicks it
- **Click Thru Rate** – number of users that clicked an ad or clicked a link sent via email
- **Cost Per Click**
- **Cost Per Lead** – indicates a user has become a potential customer or **lead** because they have expressed interest in the company by downloading a document, creating an account, or providing an email address.

Conversion – when a lead converts to a paid customer

- **Customer Acquisition Cost**

Two Additional Levels

Before we move on, I wanted to share 2 more measures that companies use.

Loyalty

To grow their revenue and company profits, companies don’t just want their customers to buy once from them, but to come back to their website. Especially if the product is not a high-priced product. That customer loyalty allows you to track how many revisits a customer is making after their first purchase, or how many of the customers have continued shopping after their first purchase.

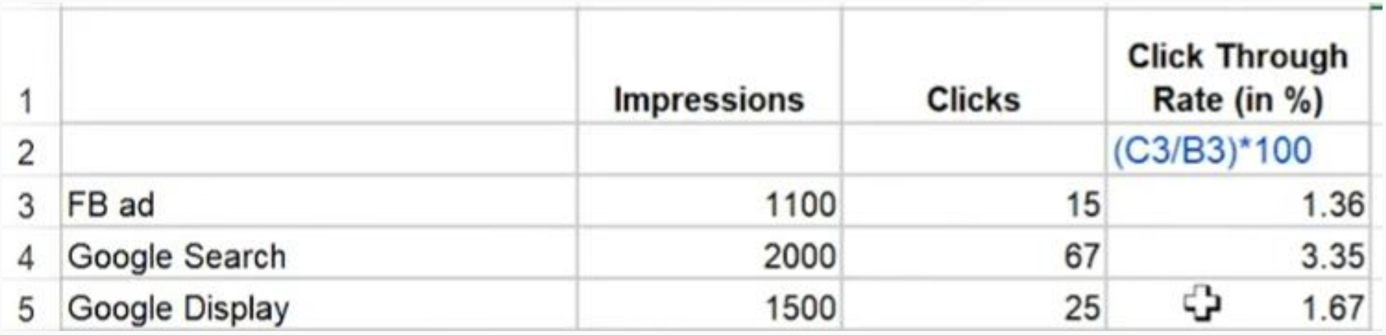
Metrics: Some commonly used metrics include **Repeat Purchase Rate** and **Net Promoter Score**. We will not be going in-depth with these, but please do check out the resources below to learn more about them.

Advocacy

Another level companies sometimes track is whether their customer is advocating for their company. That is, saying good things about the product and services. Leaning on social media provides a great opportunity to do just that.

Metrics: Some commonly used metrics include **Customer Referrals** and **Leads from Social Media**. For example, as the paid customer tweets about the company, likes the product on FB, provides a good rating on Amazon or the company website, analysts can use those metrics, such as ratings and likes to show how many of the customers serve as advocates.

We will not be going in-depth with these last two stage levels, but we have provided some resources below to help you understand these more.



		Impressions	Clicks	Click Through Rate (in %)
1		Impressions	Clicks	Click Through Rate (in %)
2				(C3/B3)*100
3	FB ad	1100	15	1.36
4	Google Search	2000	67	3.35
5	Google Display	1500	25	1.67

The formula used to calculate CTR is: **Click Through Rate (CTR)** = (Clicks/ Impressions) * 100

As potential customers view the ads, some of those potential customers will click the ad and be taken to the website for the company. To be counted at this level, the user needs to click through the ad and the metric we use here is **Click Through Rate**.

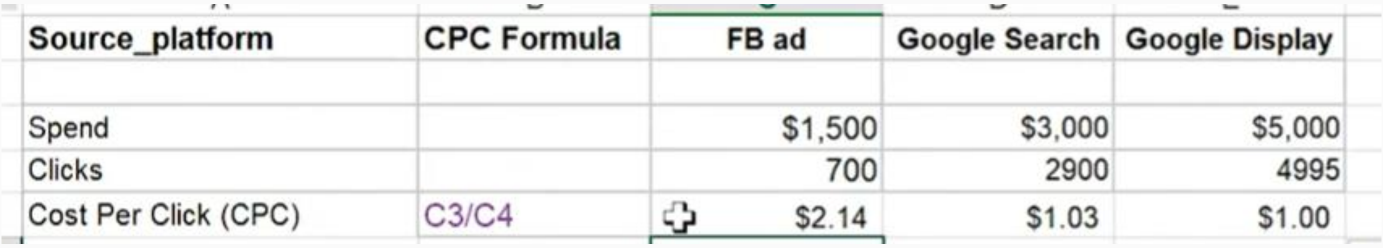
Interpretation of CTR

The Click Through Rate is an informative metric that informs your marketing team whether they should try and increase the number of impressions or when they should reword the ad to increase clicks. Remember, if a person clicks through the ad, it does not mean the customer purchased, but rather they are showing interest in what the ad is about. When your CTR is low, your ad campaign is not generating enough interest. When the CTR increases, it is an indicator of effective and interesting content in your ad campaign, and that maybe you should increase the number of impressions for that ad.

Some points to remember:

- **Click Through Rate (CTR)** is the ratio of users clicking on a link or an ad to the number of total users who received the link or saw the ad.
- CTR measures the success of an advertising or email campaign.
- When the CTR increases, it is an indicator of effective and interesting content in your ad campaign, and that maybe you should increase the number of impressions for that ad.
- In general, a 2% CTR is good, however, the rate will vary by industry.

A related concept called **Unique Click Through Rate** is examined when looking at email campaigns to see how often a link sent through an email was opened by the person receiving the email. If the person receiving the email clicks on the link 5 times, the unique CTR stays one, even though the total CTR is 5. Comparing the unique versus total CTR can help the analyst know if the email campaign reflects the interest among potential customers.



Source_platform	CPC Formula	FB ad	Google Search	Google Display
Spend		$1,500	$3,000	$5,000
Clicks		700	2900	4995
Cost Per Click (CPC)	C3/C4	$2.14	$1.03	$1.00

Cost Per Click (CPC) refers to the cost to get a click on your ad. It helps us gauge the cost of advertising on the specific platform, so we can see which platform is generating more leads.

Since platforms charge you for the number of ads on a page, you can compare the CPC for the different platforms you are advertising on and see which platform is generating more interactions with your website, or generating more traffic to your website.

The formula used to calculate CPC is:

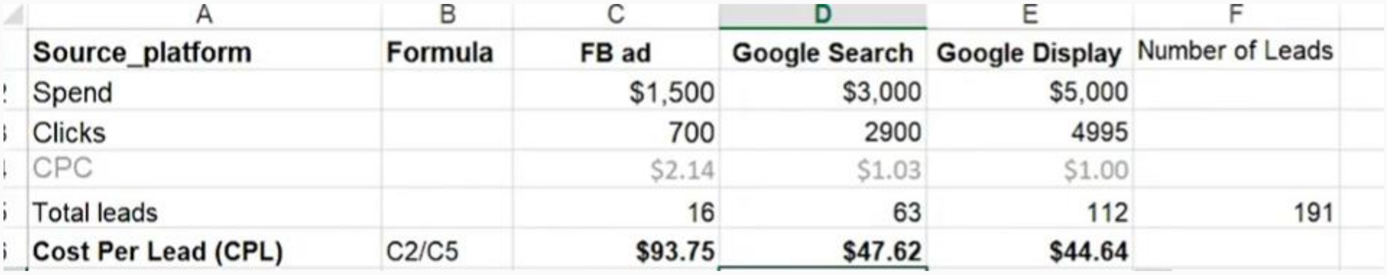
Cost Per Click (CPC)=Cost of Advertising on Source Platform / Number of Viewers who Clicked on the Ad

Interpretation of CPC

CPC is an indicator of the cost-effectiveness of the ad platform and a useful tool to compare and strategize about which marketing platform is yielding a higher impression and reach and resulting in potential leads.

Different ad platforms cost differently and it is important to remember that while one platform might be cheaper it may not necessarily deliver you as many potential customers as another platform. This is an important trade-off that analysts and marketing teams have to consider.

Some marketing channels or platforms convert amazing results but they are small and may not generate as many customers. While you may decide to continue using them, you will also need to identify marketing channels that deliver more potential leads.



A	B	C	D	E	F
Source_platform	**Formula**	**FB ad**	**Google Search**	**Google Display**	Number of Leads
Spend		$1,500	$3,000	$5,000	
Clicks		700	2900	4995	
CPC		$2.14	$1.03	$1.00	
Total leads		16	63	112	191
Cost Per Lead (CPL)	C2/C5	**$93.75**	**$47.62**	**$44.64**	

The formula used to calculate CPL is:

Cost Per Lead (CPL)=Cost of Advertising on Source Platform / Total Number of Leads

Cost Per Lead

Remember, **a lead is when a potential customer visits your website and does something on the website in response to a prompt**, such as share their email , or download a document, create an account. Once the viewer takes that action, we know the viewer is showing some interest for the product or service, and this could possibly lead to a sale. **With Cost Per Lead we are tracking whether the potential customer turned into a lead within a given time period, that could be a 30-day window or 60-day window.**

Interpretation of CPL

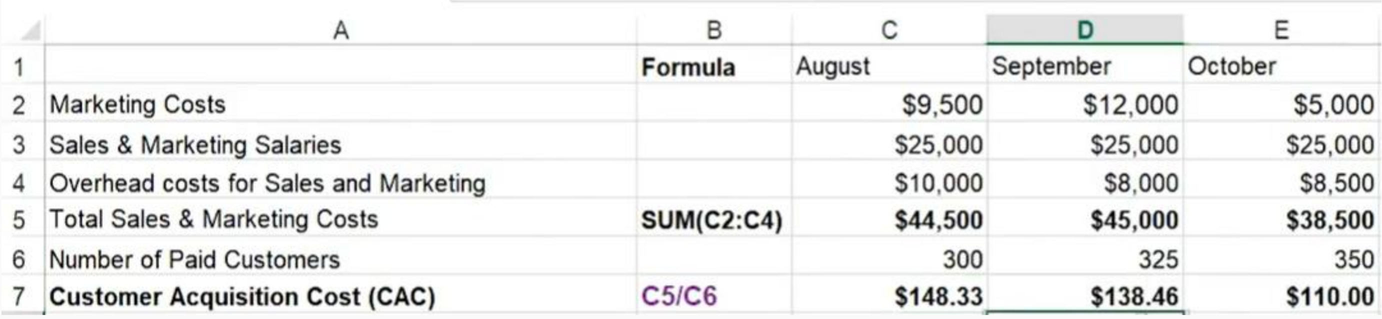
CPL is an indicator of the cost-effectiveness of the ad platform and a useful tool to compare and strategize about which marketing platforms yielded more leads. A low cost per lead means more of this particular type of person is likely to be interested in the product.

Looking at the data above, we can see that Google Display and Google Ads were comparable in terms of the Cost Per Lead. On the other hand, Facebook was costing us more to get to our potential customers.

At the same time, Facebook also generated fewer clicks, so we need to consider if we need to tweak the ad for the Facebook platform or consider other platforms that can generate the same or higher number of clicks for a comparable price.

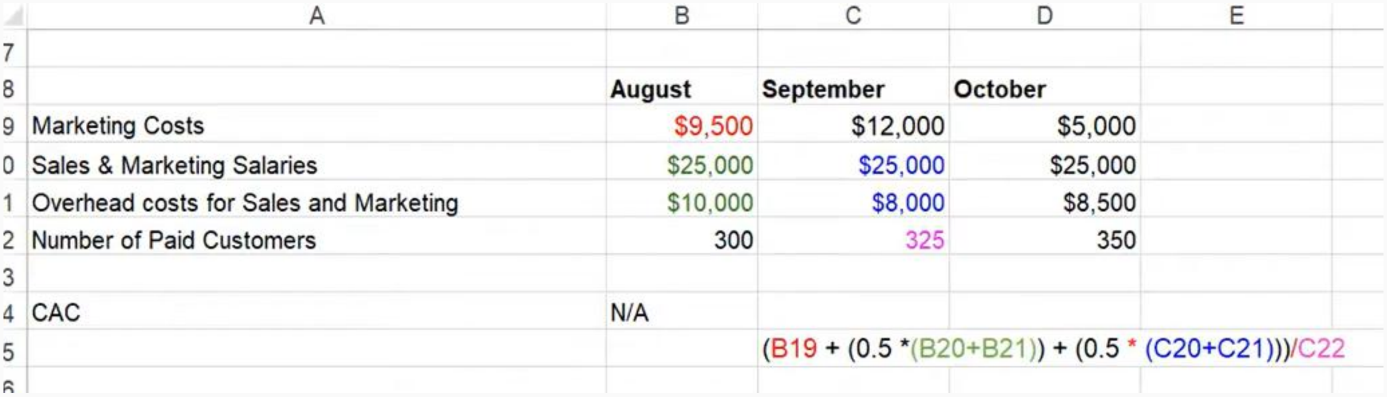
Customer Acquisition Cost (CAC)

Customer Acquisition Cost (CAC) is the metric used in the last step of the marketing funnel and tells us what the cost is to acquire a paying customer.



	A	B	C	D	E
1		**Formula**	August	September	October
2	Marketing Costs		$9,500	$12,000	$5,000
3	Sales & Marketing Salaries		$25,000	$25,000	$25,000
4	Overhead costs for Sales and Marketing		$10,000	$8,000	$8,500
5	Total Sales & Marketing Costs	**SUM(C2:C4)**	**$44,500**	**$45,000**	**$38,500**
6	Number of Paid Customers		300	325	350
7	**Customer Acquisition Cost (CAC)**	C5/C6	**$148.33**	**$138.46**	**$110.00**

Customer Acquisition Cost (CAC)=Total sales & marketing costs / Number of converted customers



	A	B	C	D	E
7					
8		**August**	**September**	**October**	
9	Marketing Costs	$9,500	$12,000	$5,000	
0	Sales & Marketing Salaries	$25,000	$25,000	$25,000	
1	Overhead costs for Sales and Marketing	$10,000	$8,000	$8,500	
2	Number of Paid Customers	300	325	350	
3					
4	CAC	N/A			
5			(B19 + (0.5 *(B20+B21)) + (0.5 * (C20+C21)))/C22		

Sometimes it takes a long time for a lead to convert to a customer. For example:

- A lead may sign up for a free account or download for a few months and then be prompted to become a paying customer then.
- A marketing campaign may intentionally take some time to realize the revenues it is trying to generate.

To account for this 'lag' in revenue, CAC is often calculated based on a company's **average sales cycle. (***averaged across the targeted time period* **)**

Customer Acquisition Cost (CAC) is calculated as:

Prior Month Marketing Costs+Weighted Avg Costs (Overhead+Salaries) / Number of Paid Customers

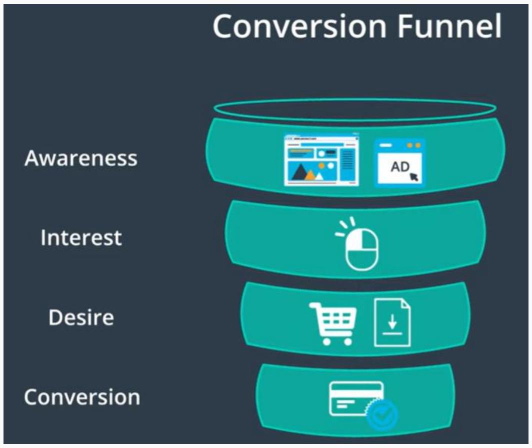
This is the point where a lead, or potential customer, has become a customer by buying something on the website (a product or service). **Most companies try to get that number under 25%.**

The ultimate goal is to increase the lead-to-customer conversions at the bottom of the funnel. Considering the fact that customer shopping cart abandonment is over 60%, each company's goal is to get higher levels of conversions for the minimum cost of sales and marketing. This leads to the concept of **optimizing the marketing funnel**.

Interpreting CAC

The CAC metric is an indicator of how much it costs to acquire a customer. If your customer service team is doing a good job of keeping the paid customers happy, that can lead to future leads and paying customers, and thus keep the cost of acquiring customers low. The company's goal is to keep the CAC low while increasing revenue, as this positively impacts the profit margin and profits.

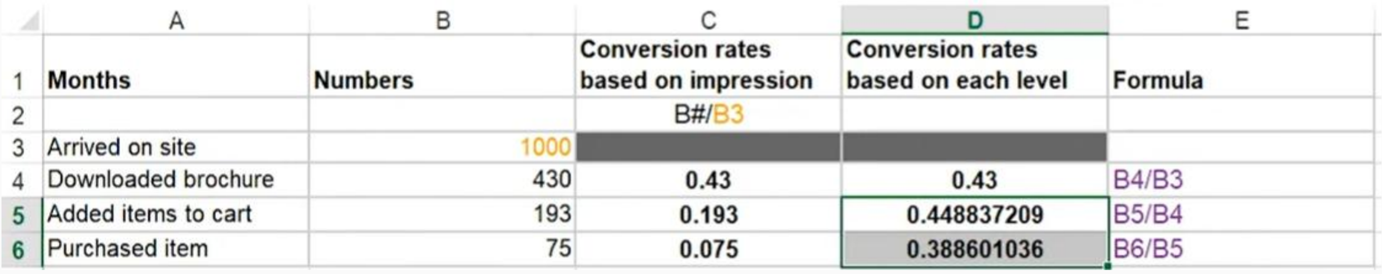
Spending more than 25% of your revenues means you are spending too much to acquire new customers and spending less indicates that you are losing business opportunities.



Optimizing the funnel requires identifying at what level of the funnel your customer loss is the greatest. In other words, are you losing the most customers at the awareness and interest age, or is it when you are converting them into leads?

- If you're losing many of them in the early stages of awareness, you need to focus on the types of ads you're creating, or the ad platforms you're choosing to reach your potential customers.
- If you're losing many of them at the conversion stage, you need to look at your website or online app. It's possible the site is not easy to navigate, and that's why not many customers are converting to paid customers.

So, essentially, you're calculating the success rate at getting a potential customer to do what you want them to do at each level of the marketing funnel, and you compare this number against your impressions.



	A	B	C	D	E
1	Months	Numbers	Conversion rates based on impression	Conversion rates based on each level	Formula
2			B#/B3		
3	Arrived on site	1000			
4	Downloaded brochure	430	0.43	0.43	B4/B3
5	Added items to cart	193	0.193	0.448837209	B5/B4
6	Purchased item	75	0.075	0.388601036	B6/B5

*Conversion rates based on impressions = Numbers/Arrived on site Numbers
*Conversion rates based on each level = Current level Numbers/Prior Level Numbers

Conversion rates based on impressions

Basically, at each touchpoint, we'll divide the number of people at each touchpoint divided by the total number of customers.

- 43 percent roughly of the people who saw the actual ad, downloaded the brochure
- 19 percent of the people who saw the actual ad arrived on the site, added items to the cart
- ~8 percent of the people who saw the actual ad actually purchased the item

Conversion rates based on each touchpoint

Another metric is to calculate the conversion rates based on *the Call to Action* at the previous level in the funnel, as opposed to the initial impression number as we did above.

So now we divide by the *number of people who actually took the call to action of the previous step.*

- 43 percent of that matches what's in the conversion rate based on the impression.
- 45 percent of those people who downloaded the brochure added their items to the cart.
- 39 percent of those people actually purchased the item.

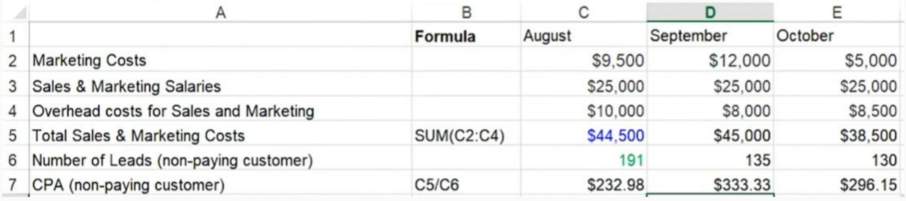
Cost Per Acquisition (CPA)

The impact of marketing campaigns can also be measured in terms of revenue using metrics that capture the financial cost.

- **CPA: cost per acquisition**
- Focuses on sales and marketing costs, including the cost of supplies, labor, marketing, overhead, and sales that it took to convert a non-paying customer(called an acquisition) into a paying customer.

What is the difference between CPA and CAC?

- **CPA –** is focused on the marketing and sales cost, including overhead and salaries, with a focus on sales leads—not actual paying customers
- **CAC –** is focused on actual new customers who have made a purchase



	A	B	C	D	E
1		**Formula**	August	September	October
2	Marketing Costs		$9,500	$12,000	$5,000
3	Sales & Marketing Salaries		$25,000	$25,000	$25,000
4	Overhead costs for Sales and Marketing		$10,000	$8,000	$8,500
5	Total Sales & Marketing Costs	SUM(C2:C4)	$44,500	$45,000	$38,500
6	Number of Leads (non-paying customer)		191	135	130
7	CPA (non-paying customer)	C5/C6	$232.98	$333.33	$296.15

Interpretation of CPA

Cost Per Acquisition provides insight into whether or not the marketing campaigns are successful from a business perspective. For the purposes of calculating the CPA, the cost of the marketing campaigns should not be restricted to the cost of developing the ad, but also other costs of labor and overhead. In other words, CPA allows a business to gauge whether the marketing campaign is generating enough potential leads to cover a broader range of costs other than just direct advertisement costs.

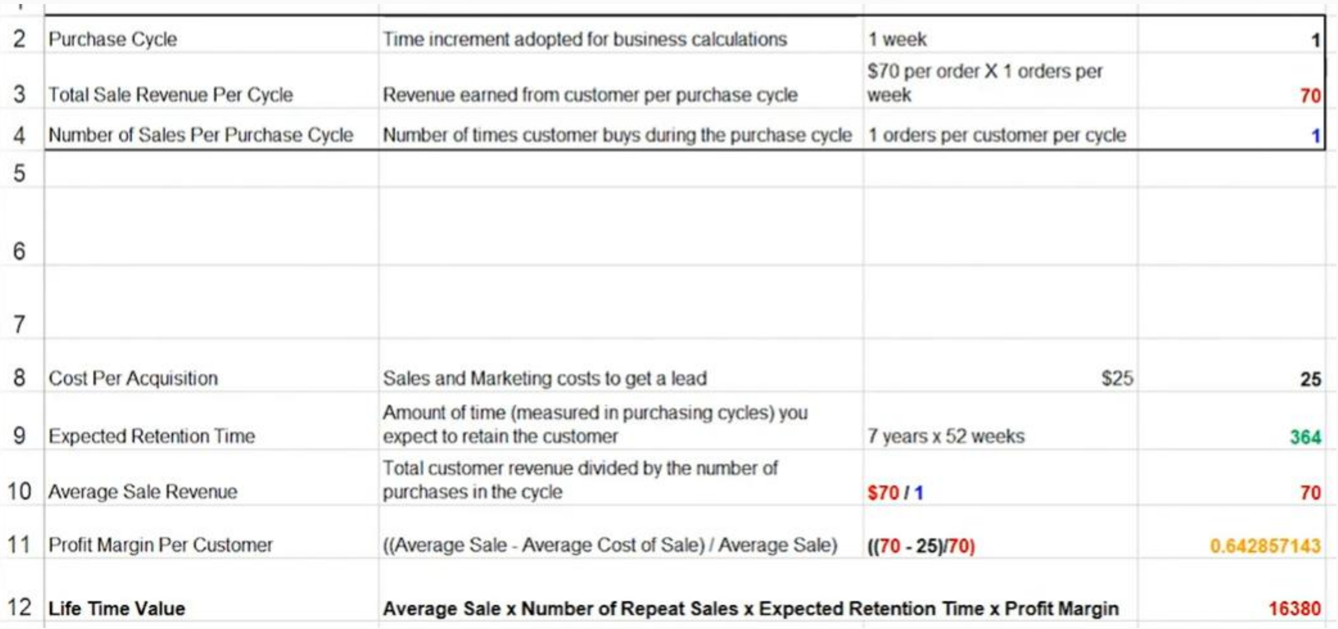
Lifetime Value

When deciding how to spend the marketing budget, you want to focus on some of your best customers – those that will stay for the long term and continue to generate revenue for the company.
These are your **high-value customers** and you want to bring in more of them.
Your goal should be for every dollar spent on marketing efforts. It should provide a higher rate of return and generate revenue multiple times over.

Terms needed to calculate Lifetime Value:
Purchase Cycle: The time increment adopted for business calculations

- **Total Sale Revenue Per Cycle**: Revenue earned from a customer per purchase cycle
- **Number of Sales Per Purchase Cycle**: Number of times customer buys during the purchase cycle
- **Cost Per Acquisition**: (Cost of marketing and sales)/ number of new leads
- **Expected Retention Time**: Amount of time (measured in purchasing cycles) you expect to retain the customer.
- **Average Sale Revenue**: (Total customer revenue/ Number of purchases in the cycle) OR Average revenue received from the customer per transaction during the cycle
- **Profit Margin Per Customer**: ((Average Sale - Average Cost of Sale) / Average Sale)

Lifetime Value (LTV) = Average Sale x Number of Repeat Sales x Expected Retention Time x Profit Margin

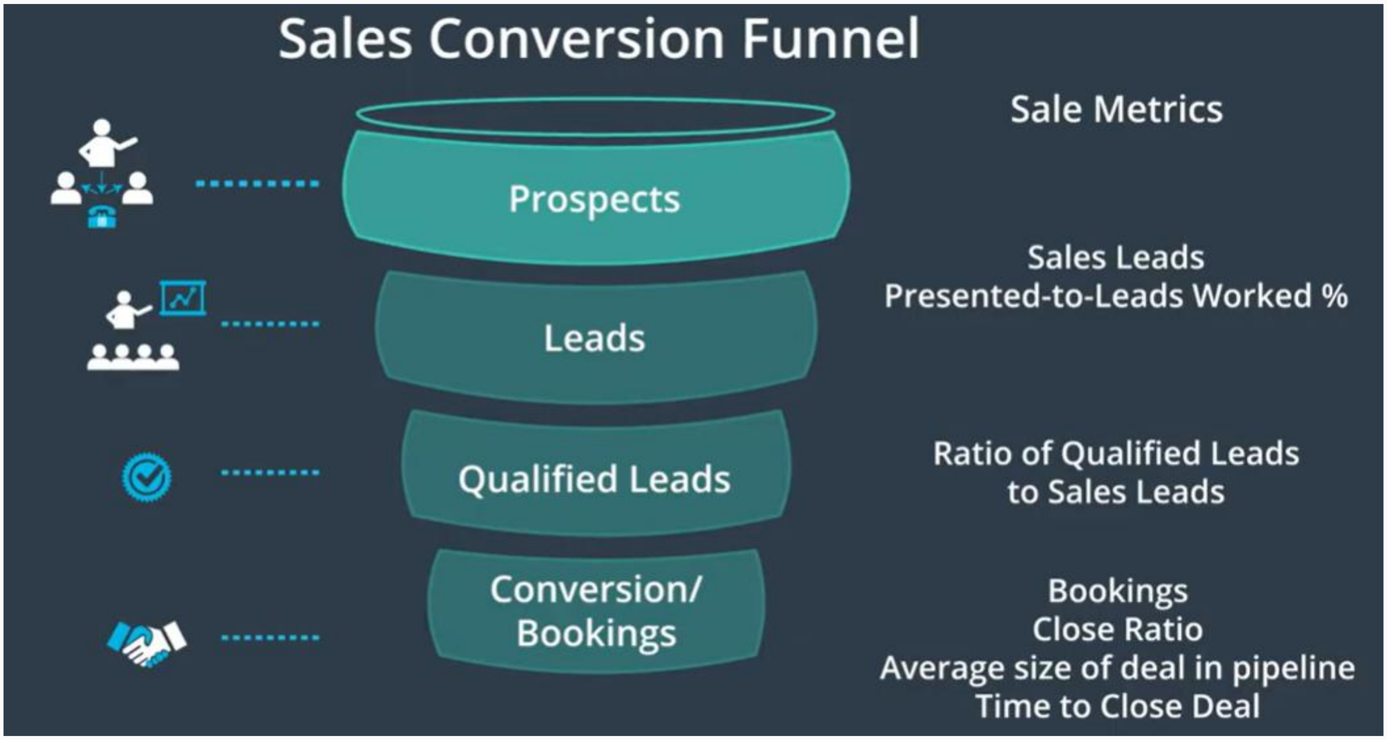


2	Purchase Cycle	Time increment adopted for business calculations	1 week	1
3	Total Sale Revenue Per Cycle	Revenue earned from customer per purchase cycle	$70 per order X 1 orders per week	70
4	Number of Sales Per Purchase Cycle	Number of times customer buys during the purchase cycle	1 orders per customer per cycle	1
5				
6				
7				
8	Cost Per Acquisition	Sales and Marketing costs to get a lead	$25	25
9	Expected Retention Time	Amount of time (measured in purchasing cycles) you expect to retain the customer	7 years x 52 weeks	364
10	Average Sale Revenue	Total customer revenue divided by the number of purchases in the cycle	$70 / 1	70
11	Profit Margin Per Customer	((Average Sale - Average Cost of Sale) / Average Sale)	((70 - 25)/70)	0.642857143
12	**Life Time Value**	**Average Sale x Number of Repeat Sales x Expected Retention Time x Profit Margin**		**16380**

What does this mean to a business? The company must pay far less than this in order to obtain a new paying customer.

Sales metrics borrow some terminology from marketing metrics. Sales can not only focus on the end customer, such as the consumer but also a company that will likely generate customers.

- B2C
- In the case of WeCart from marketing, we have thus far focused on the end consumer and their order as our marketing focus, or *unit of analysis*. This type of model is called a **business-to-consumer** model or *B2C*.
- B2B
- Let's say that we add an additional focus on generating sales by reaching out to grocery stores. This is where We Card partners with the local grocery stores. We can now gain access to their customers and the local grocery stores can access a delivery service. This type of business model is called a **business-to-business** model or *B2B*. So, part of our sales team can focus on the B2B model, where each sales rep is trying to create a sales lead. A sales lead would be a grocery store interested in partnering with We Card. If a grocery store signs a deal, then it becomes a booking or a closed deal.
- **Business to Business (B2B)**: When one business makes a business transaction (goods or services) with another business. Often takes place when one business is providing source materials to the other business to in turn finally sell it to the consumer.
- **Business to Consumer (B2C)**: When a business sells products and services to the final consumer.



Sales pipeline

The sales funnel, also known as the sales pipeline, tracks the number of incoming leads or prospects. These are leads that a sales team member has identified as being potentially interested in the product. Then, the sales team member follows up for an interested lead to ask more about what they're looking for, possibly making presentations.

Here, tracking using a metrics like number of sales leads is effective.

Qualified leads

This is followed by a qualifying process whereas the sales team qualifies the leads, which means they are checking to see if the product offering is within the lead's budget, ultimately in order to identify the ideal buyer and confirm their viable lead.

Here, you track metrics like ratio of qualified leads to sales leads.

Closing deals

Once the sales team has a qualified lead, you end with a closed deal or booking. At this stage, you can also have a lead on hold or last.

Here, you track metrics like bookings, close ratio, and average size of deal in pipeline.

- Bookings is a very important metric for tracking the success of the sales team.
- Close-ratio is the ratio of closed deals to leads from the sales pipeline.

As you may have noticed, sales metrics are measuring the *performance of the sales team* internal to the company. As opposed to tracking the *behavior of individual customers* in the marketing metrics.

For this lesson, we will cover the metrics at the bottom of the funnel. But we have provided links and resources below for you to learn more about the others if you're interested.

New Vocabulary

- **Sales Lead**: A sales lead refers to the number of potential customers who have shown interest or have been identified by the sales team member as being potentially interested in the product.
- **Qualified Lead**: A potential lead who has been vetted by the sales team as meeting key requirements of an ideal buyer. Sales teams check to see if the product offering is within the lead's budget that will make them a viable buyer.
- **Booking**: Booking is a closed deal when the qualified buyer has committed to making the purchase. It is a key metric for tracking the success of the sales team.
- **Sales Pipeline**: Refers to the collection of steps a sales representative takes while navigating incoming leads or prospects through to making the final purchase. It is also used to track how well individual sales representatives are meeting their sales quota.

Bookings are the most important sales metric. **Booking is a won deal that is signed or where the purchaser is committed to buying the product.** Once you have the sales bookings value, you can track it across specific time periods and even product lines.

Calculation

Total Bookings is the sum of all closed deals

Another important metric to keep in mind is the Average Deal Size. This refers to the average deal size in dollars of all of the won deals. Reminder, a won deal is when the account buyer has committed to making the purchase.

Calculation

Average Deal in Size ($) = Total ($) Sale Value of Deals or Orders / (#) of Orders over a Specific Period

The average time to close the deal is the average number of days it takes a member of the sales team to close the deal from the prospect stage to a closed deal.

This metric can be calculated for each sales team member, product, or lead source. **The lead source** refers to whether the prospect inquired through the website or had an **inbound inquiry**. On the other hand, **outbound methods** refer to cold-calling through email lists or phone calls. This means the customer has lower intent to purchase, to begin with, and this lengthens the time to close the deal.

Terminology

- Sum of Total number of days from the first contact to closing the deal for all closed deals
- The average number of days for typical Sales Cycle = Sum(Total number of days to close the deal) for all closed deals / Number of closed deals

Calculation

Step 1. **Sum (Total number of days from the first contact to closing the deal)** for ALL closed deals.
Step 2. **The average number of days for typical Sales Cycle = Sum** (Total number of days from the first contact to closing the deal) for ALL closed deals / **Number of deals**

Measuring your company's growth

Growth for a website or app is counted in the number of users.

- Are we seeing the number of people actually using the site increasing or decreasing?
- If you see your website use as high, are they unique users or the same people coming back?

An important aspect of growth is not just whether you have users but whether they continue to actively use or engage with the website.
In the next few pages, we'll talk about the following metrics:

- Active users
- Stickiness
- Churn rate

Defining 'active'

Defining what constitutes an active user is an important part of calculating your growth metrics.

- Monthly active users tell the number of users who were active in the last month.
- Daily active users are the users who were active in the previous day.

Engagement or actual use of the website will differ depending on the website. For our WeCart business, engagement could take the form of users who are using it to check the prices or those ordering groceries. Once you have your active user accounts, it is easier to see how the visitors are using the website. You can even narrow this down to specific features or market segments.

The ratio of the daily active users compared to the monthly active users is called stickiness.
The stickiness ratio is a very useful tool for several stakeholders, such as investors, to know if the website or app has a potential for growth or either on the trajectory, for growth. Essentially, stickiness tells us if the customers are coming back to the website or app every day, or rather, sticking around to actually engage with it.
A DAU by MAU ratio of 0.5 indicates that the average user is engaged or using the app 15 out of 30 days that month.
Common benchmarks for various industries are as follow:

- Social networking site is 0.5
- Gaming apps, it is 0.1 to 0.2
- ...most other apps strive to achieve a stickiness ratio of 0.2

Why is stickiness important? It is a useful KPI for management and investors.
Investors want to know if this app or website has the potential to make money in the future. For example, if the plan is to introduce advertising into the app, the potential evaluation will depend on whether the app has a large number of users that keep coming back to it. In other words, they need to know the number of active users will likely increase over time.

Interpretation of Stickiness Ratio

- The percentage number means what percent(%) of that months days did the user was online with the site
- A 50% stickiness ratio indicates the average customer used the online app or website 15 of the 30 days in that month. In contrast, a .01 stickiness ratio indicates the average customer used the online app or website < one day that month.
- A higher stickiness ratio indicates the website or online app is engaging the customers.

Churn Rate captures the number of people we retain at the end of a time period.

To calculate the customer churn rate you need 2 simple things:

- **Customers at the beginning of usage interval**
- **Customers at the end of the usage interval**

Just looking at these two numbers will tell you whether you end the interval with the same or fewer customers.
Important note: We only want to calculate the churn rate based on the customers we started the time interval with. When getting the number of customers at the end of the interval, we do NOT add the customers who converted during the interval. The churn rate should only tell you whether the current customers have left or stayed.

Calculation

Churn Rate=(Customers at start of usage interval - Customers at end of usage interval) / Customers at start of usage interval

Terminology

- **Usage Interval: This time period should make sense for the service or product the customers are using. It can range from a day, a week, to a month or quarter. It depends on the service or product the company is providing and how often you would expect a customer to be active on the website.**

Interpretation of Churn Rate

While the churn rate is inevitable, in general, an annual churn rate of 5% is seen as a reasonable benchmark. Keep in mind that the range for churn rates is wider for B2C companies.
As you calculate your annual churn rate, keep in mind a few other "data assumptions" that you need to watch out for.

- **Select a time interval during which you calculate the churn rate that is consistent with the company's subscription or usage model. There is no ideal usage interval - the usage interval depends on the length of time the company expects the user to be active at least once.**
- **Pay attention to different customer segments, especially if they have different churn rates (e.g., by region).**
- **Make sure your data does not include new customers gained during the time interval. Churn rate is focusing on customers who stayed or are active vs. stop being active on the website.**
- **Software as a service (SaaS): SaaS is a software distribution model in which the application is made available on servers hosted by a third-party provider, which in turn provides the software to customers over the Internet.**
- **Subscriber-based service model: Subscriber-based service model is a model where consumers agree to pay a subscription fee to gain access to the service or product.**

Finance Metrics.

When you look at the financial metrics, you are focusing on tracking your performance against your company’s financial goal. You are trying to answer the following questions.

- **How is your revenue comparing to the costs?**
- **How are sales trending against sales goals?**
- **How are sale and marketing lead metrics comparing against acquisitions?**

Items in Profit and Loss Statement

(*aka income statement*)

The following list is a breakdown of the individual items within the Profit and Loss Statement.

- **Revenues: The money a company makes from the sales of its products and services.**
- **Cost of Goods Sold (COGS) or Cost of Sales: These are the direct costs the company incurs to develop the product or service being sold.**
- **Gross Profit: The difference between the revenue earned and the costs summarized in COGS. Gross Profit = Revenue - COGS**
- **Selling, General, and Administrative expenses (SGAs): Includes the following expenses:**
    - **Marketing, sale commissions**
    - **Salaries for office staff**
    - **Supplies and computer hardware**
    - **Note: Some companies list total operating expenses separately from SGAS while others treat them as synonymous with SGAS.**
- **Operating expenses: Expenses incurred outside of direct manufacturing costs:**
    - **Overhead costs**
    - **Legal**
    - **Rent**
    - **Utilities**
    - **Taxes**
    - **Interest**
    - **R&D expenses.**
- **Total Operating Expenses = Sum of SGAs and Operating expenses Total Operating Expenses= SGAs + Operating Expenses**
- **Operating Income: The difference between Gross profit and Total operating expenses Operating Income = Gross Profit - Total Operating Expenses**
    - **Note: Operating Income is also referred to as Earnings Before Interest and Tax (EBIT)**
- **Net Income: Subtracting the Interest and Tax from Operating Income gives the Net Income Net Income = Operating Income - (Interest and Taxes)**

Gross Margin is a statement about the overall profitability of the company.

Calculation

Gross Margin = (Total Sales Revenue – Cost of Goods Sold) / Total Sales Revenue
which is the same as for Gross Profit / Total Sales Revenue
This metric identifies the revenue that remains after accounting for direct costs of production.
Costs to a company can be split into two major groups;

- **Fixed costs**
    - **Fixed costs are expenses that you will incur on a regular, perhaps monthly basis, such as rent, utilities, and employee salaries.**

- **Variable costs**
    - **Variable costs are expenses that move up and down in response to production output. This interpretation is particularly helpful for companies to determine the pricing of the product.**

In other words, it helps them find the breakeven point where the pricing will cover fixed overhead costs for sure.

Contribution margin can also be helpful as a useful tool to dive into the P&L statement! While the typical P&L statement line items tell us the overall profitability of our business, contribution margin can be used to identify which product or product line is contributing the most to our profit margin.

(Looking at the graph in the video)

First of all, we want to calculate the overall contribution margin, the difference between the *sales revenue* and *variable cost*.

Once you get the difference for the total contribution margin, you divide it by the units sold, and that gives you the contribution margin per unit.

You can think of the contribution margin as a percentage of your revenue that'll cover your fixed costs, which you have little control over and have to incur.

- **If the contribution margin per unit > fixed cost, this means you're making a profit on each sale.**
- **If the contribution margin per unit < fixed cost, this means you're making a loss on each sale.**

So in other words, if you want to make a profit, you have to sell your product with at least your contribution margin covered for each unit sale.

Contribution Margin

Contribution Margin tells us the amount of revenue that covers the variable costs and is now available to cover the fixed costs and generate profits. Companies use it to identify which product or product line is contributing the most to the profit margin. It also helps determine the break-even point where the pricing will cover fixed overhead costs and leave enough for profits too.

Fixed costs are also called sunk costs. A good caution to keep in mind is that fixed or sunk costs can increase (for e.g., unexpected rent increases, machinery replacement costs), which is why operational managers prefer the term sunk costs. These sunk costs can prove tricky, because a small increment when taken in bulk, can turn out to be catastrophic for companies, especially start-ups.

Terminology

- **Fixed costs: Expenses incurred on a regular basis, such as monthly rent, utilities, and employee salaries.**
- **Variable costs: Expenses that move up and down in response to production output.**
- **Contribution Margin: The amount of revenue that covers the variable costs and is able to cover the fixed costs.**

Calculation

- **Total Contribution Margin = Total Sales Revenue - Total Variable Cost**
- **Contribution Margin Per Unit: Total Contribution Margin / Number of Units Sold**

As a business analyst, you should make sure to apply these statistical tools as you generate and examine the data using business metrics.

A business metric is one data point that in itself does not tell much about the larger context. Much like other things in life, data makes more sense and has more value if it is looked at within a context.

So, as a business analyst, your data analysis process should include exploratory checks to examine the spread of the data. *You should always be asking and checking to see if the data is spread out equally in each direction and to see if the shape of the distribution resembles a normal distribution or not.*

In the previous lesson, we talked about **skewness** and we come back to it in this lesson. The box plot or histogram as you may remember are useful tools that tell you about skewness. They can alert you to any skewness in the data. Another way to check for skewness is to compare the *mean* and *median* values to see if they are more or less the same.

Creating visualizations to look at data distributions and computing multiple summary statistics like mean and median should be a reflexive habit of a business analyst.

Modeling refers to using certain inputs and using those inputs to predict how a business metric will perform. The time duration of your forecast will depend on the type of business and the business metric:

- Startups – usually forecast 6-12 months out
- Established companies – usually forecast out a few years
- Sales bookings – a few months out

Models are continuously updated on a frequent or periodic basis (monthly, quarterly, yearly), depending on the metric.

Most businesses forecast their sales bookings and financial statements (profit and loss, balance sheet, and cash flow statements). This lesson will focus on the profit and loss statement.

Top-Down Approach

Forecasting requires careful thinking about which approach you want to take to create the models. One approach to modeling is top-down, which takes a macro approach to forecasting.

- You start with the best estimate of the larger size of the market narrowing it down to identify the portion of the market that the company is serving

- Then estimate what it will take to capture that portion of the market.

Top-down is a macro approach, but it is less credible and typically adopted when there's limited historical data.

Bottom-Up Approach

A bottom-up approach takes a micro approach to forecasting.

- This approach starts by looking at historical data. The more data you have, the better. But often, even as little as six months to one year of data is used in this approach.
- The model forecasts are based on this data to make assumptions about how the key metrics will behave, and then we forecast out the revenue based on these assumptions.

Similar to a top-down approach, a bottom-up approach is based on assumptions. However, unlike a top-down approach, it is built on previously attained numbers. These numbers are specific to how the company has performed, and not generalizations that relate to the market as a whole.

Components of a Forecast Model

- **Inputs or Drivers:** These are the inputs that drive the output of the model and can include a combination of **historical data, assumptions,** and scenario analysis.
- **Outputs:** This is the metric being forecasted within the model.

Assumptions

In modeling, assumptions are what you think will come true for the inputs or metrics in the model. It can also be assumptions about circumstances that affect the metrics in the model.

To build your assumptions, you start with information available to you based on historical data, such as continuing to gain market share at the same rate as we have for the last six months or seeing spikes in sales around holidays or on certain days of the week.

As you make assumptions, pay attention to your forecasts about metrics that you have less control over. Be careful about assumptions that you make, and check to see if those assumptions are reasonable and make sense.

Historical Data

Historical data is about what your performance metrics show for the past. For sales, for example, we look at prior sales data from the previous year or months. For financial modeling, we look at prior financial statements, as well as quarterly and monthly results. Above we are using prior year data (e.g., Revenue, COGS) to get our operating income, which we then use to calculate the historical operating margin.

Formulas for Calculating Historical Financial Metrics

Typically, the historical statistics or metrics used to forecast financial metrics in an Income Statement are:

- Revenue Growth
- Gross Margin
- Operating Margin
- Historical Tax Rate
- Historical Interest Expense Rate

The following list provides more information about calculating the historical statistics.

- **Revenue Growth (in %) = (Current Year's Revenue / Previous year's revenue) - 1**
- **Gross Margin = 1 - (Current Year's Cost of COGS / Current Year's Total Revenue)**

Keep in mind the two terms COGS and Cost of Revenue can be used interchangeably.

- **Operating Margin = Current Year's Operating Income / Current Year's Total Revenue**
- **Historical Tax Rate** is the tax rate from the companies previous year's tax rate.
- **Historical Interest Rate** is the interest rate coming from the previous year's Debt Schedule.

Sales Forecasting

You will go through examples of forecasting sales bookings using both a bottom-up and top-down approach. Although there are many ways to implement each of these approaches, the examples will provide a foundation for extending these practices to new examples.

Before you get started, let's revisit the sales funnel. Sales metrics track the conversions from prospects into sales bookings. The layers of the funnel are:

- Prospects
- Leads
- Qualified Leads
- Conversion/Bookings

When using a bottom-up approach, you will use the sales funnel historical data.

Assumptions and KPIs

- **Contract Terms** = Number of months in the contract
- **Price per Unit (by mon)** = Units needed in 1 month X Price per Unit
- **Bookings Forecast** = Price per Unit X Contract Term (month)
- **Closed/Won Probability** = Probability of Closing the deal
- **Weighted Bookings Forecast** = Bookings Forecast * Closed/Won Probability

Top-Down Sales Forecasting

For the top-down sales forecasting model discussed in the video above, the model is focused on bookings per salesperson and is divided into four sections:

Key Seller Assumptions & KPIs

These are the assumptions about the dollar amount needed in bookings and the following questions tell us the average size of opportunity a salesperson would be expected to generate.

- How Productive will the new salesperson be?
- How many opportunities will be generated?
- How much revenue is generated per unit and per opportunity?

Sales Hiring Schedule

This section tells us the time and effort needed to generate the expected bookings and opportunities. In this example, we project out when the salesperson will be hired

Sales Productivity Schedule

Based on historical sales data, we create the assumption that a new sales member would take x number of months to learn the business and "ramp up", in order to generate the bookings and opportunities assumed in the first section.

Model Walk-through

First, we start with the number of opportunities we expect the salesperson to close annually.

- **Average price per unit** – the average price per unit or product for manufacturing.
- **Average units per opportunity** – the average number of units you can expect to sell per opportunity.
- **Average contracts month per opportunity** – the average length of time in months that sales contract can be for.

Next, we get to the **average opportunity size**, the average booking size we expect this salesperson to create on an annual basis

- Take the product of the three numbers above to get this number of bookings.

Next, we determine how to get to these bookings.

- Assume that the **seller ramp** or the length of time we can expect the new seller to reach full productivity after being hired is three months.

The hiring schedule in the model shows that once a person has been hired, a month from then, they will be employed and available to start generating leads.

- Create **dummy coding** for the projected hire date and after that, indicate a 0 for not hired and a 1 for a salesperson having been hired. **Dummy coding** refers to when you use a one and zero as stand-ins or **dummies** for the presence of something happening.

For the center productivity schedule, we want to generate the schedule of when a salesperson will be productive and we want to give that person three months of ramp time.

Finally, we get to our booking projections based on when a seller will be productive.

- Multiply the productivity dummy variable with the projected average booking and the expected monthly opportunities closed to get the monthly bookings generated by a salesperson.

Scenario or Sensitivity Analysis

Scenarios are commonly used for financial forecasting and rely on assumptions to provide some perspective on a company's future. These perspectives typically that the form of:

- **Best Case Scenario**
- **Base Case Scenario**
- **Weak Case Scenario**

The assumptions used in your model will dictate the scenario you are looking at and will directly impact forecasted income for the company.

Transitioning to Spreadsheet Tools

To build out the financial model, let’s first take a look at some Excel functions and tools specifically for Financial modeling. As we introduce you to each of these, we will work towards a financial model, so this will set you up nicely for your final project.

- **Data validation** is a spreadsheet tool that allows you to limit what values are accepted in a cell. You can create drop-down lists of items and restrict cell value to date ranges and numbers.
- **INDEX** is used when you want the cell to have a value chosen from a specified array and row number indicated within the INDEX function.
- **MATCH** is a LOOKUP function that can locate the position of the lookup value within an array only when it meets specific criteria defined in the MATCH function.
- **INDEX AND MATCH** together add a powerful feature for advanced formulas. Together they can give a value from an array (the purpose of the INDEX function) based on a numeric position (which is provided by the Match function).
- And finally... **OFFSET**. Here you can select a start point in the spreadsheet, and tell Excel to return a set of cells that are counted from the starting point.

Excel Steps

The purpose of data validation tools is to confirm that the values within the cell are validated against a criterion. In other words, the values within the cell are confined to specific requirements. There are several criteria, including a provided list of values, date range, range of whole numbers, or decimal values.

To access the Data Validation tool within MS Excel, you **use the Data tab** and choose **Data Validation**.

How to Create a Data Validation Dropdown List

To create a dropdown list using data validation, first **create a pivot table**:

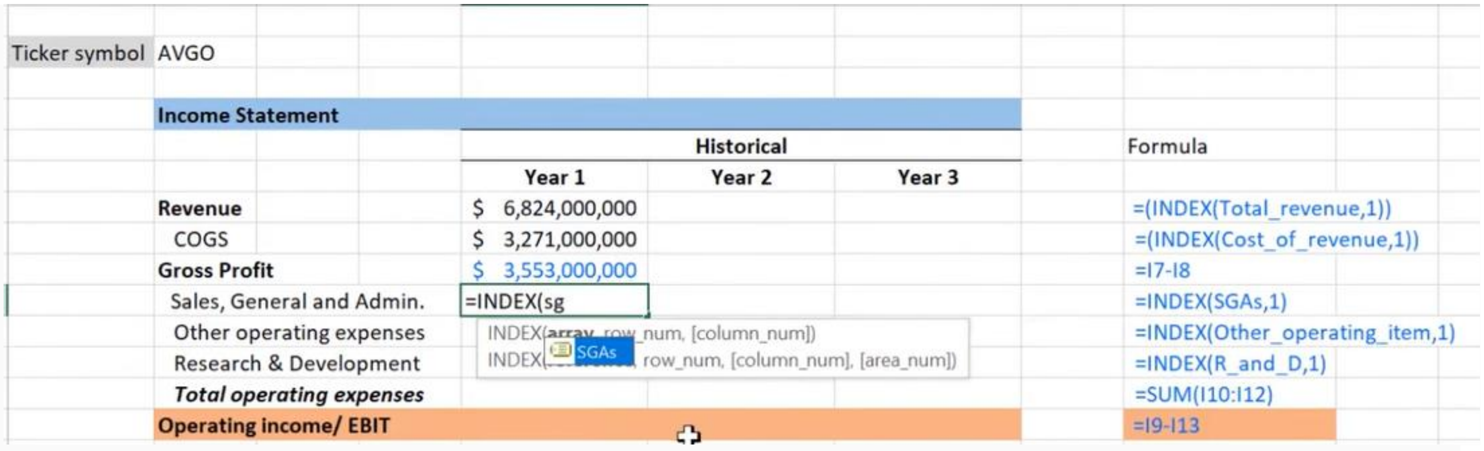
- Highlight the rows of data you want in the dropdown
- Go to Pivot Table under the Insert tab to create a pivot table in a new worksheet (just hit OK)
- From the PivotTable Fields menu, select the name of the field/column you want
- You should now have a list of unique values with all duplication eliminated
- Copy and paste the list of unique values into a new cell and give that cell block a name, for example, company_list.

now that we have a named list of unique values, the second step is to **create a data validation feature**:

- In a new worksheet, go to Data Validation under the Data tab
- Choose your validation criteria. In the video example, we used *List*
- For source, reference back to your named list. In our example, it was *company_list*
- Hit OK and now you should have a dropdown menu with only the unique values from your original data source

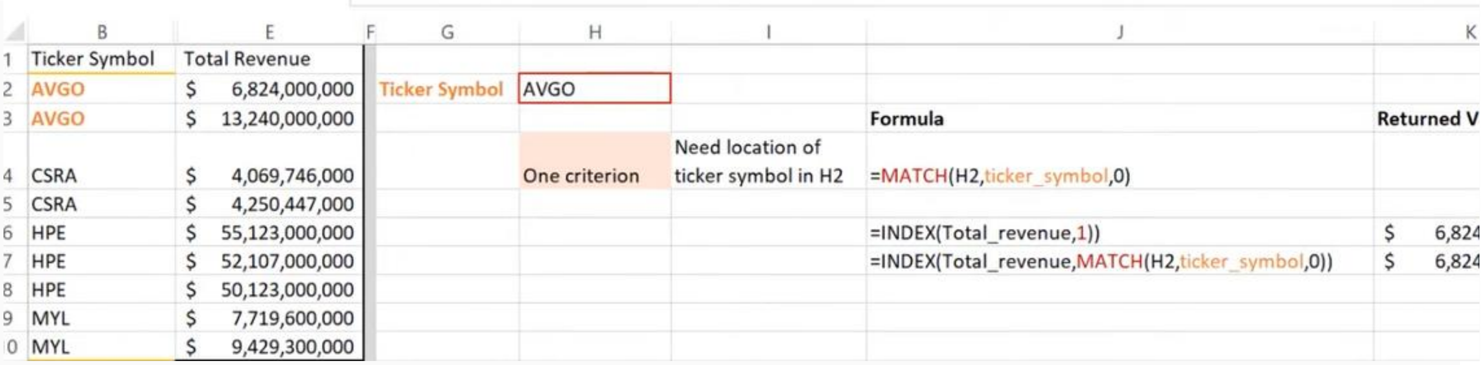
Pro Tip – You can use the Name Manager feature under the Formula tab to see all of the named boxes available in a spreadsheet and you can delete boxes, confirm or edit the source or range the box references.

Index


INDEX(array, row_num, [column_num])
INDEX(
SGAs
row_num, [column_num], [area_num])

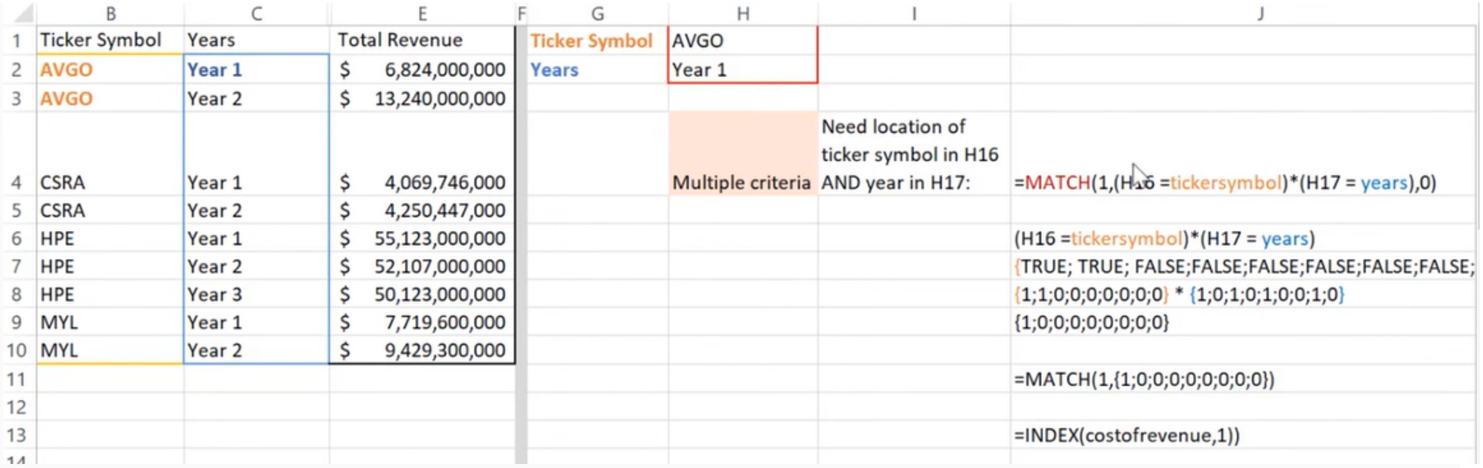
Ticker symbol	AVGO				
	Income Statement				
		Historical			Formula
		Year 1	**Year 2**	**Year 3**	
	Revenue	$ 6,824,000,000			=(INDEX(Total_revenue,1))
	COGS	$ 3,271,000,000			=(INDEX(Cost_of_revenue,1))
	Gross Profit	$ 3,553,000,000			=I7-I8
	Sales, General and Admin.	=INDEX(sg			=INDEX(SGAs,1)
	Other operating expenses				=INDEX(Other_operating_item,1)
	Research & Development				=INDEX(R_and_D,1)
	Total operating expenses				=SUM(I10:I12)
	Operating income/ EBIT				=I9-I13

Match with One Criterion



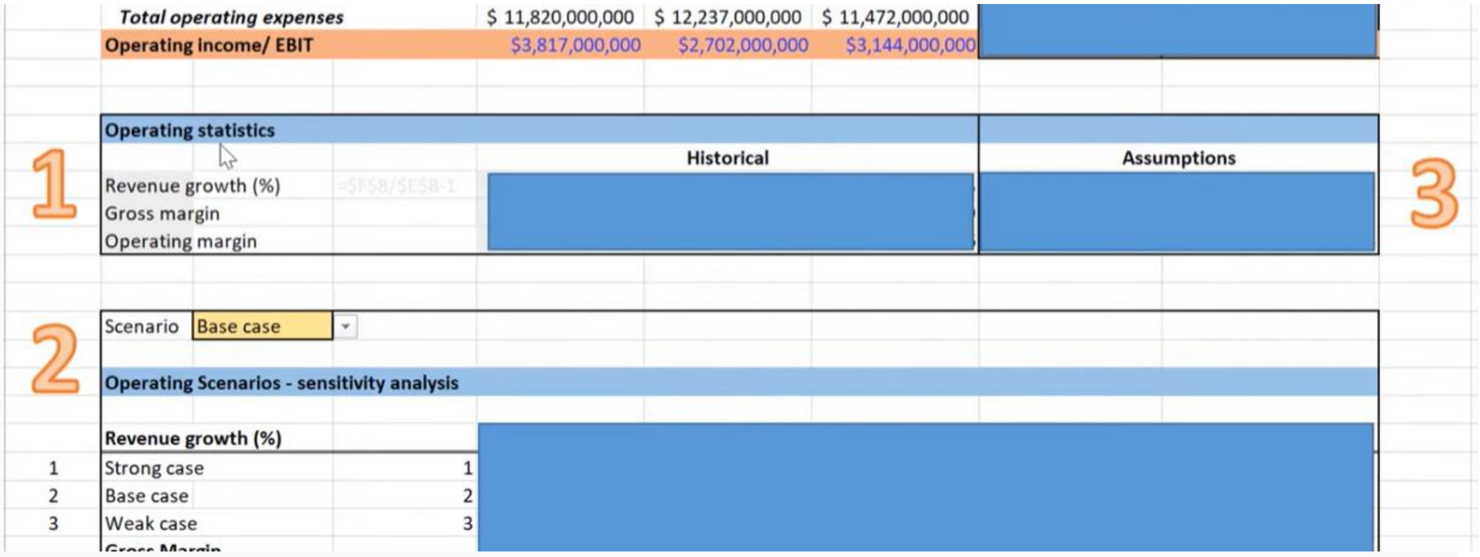
	B	E	F	G	H	I	J	K
1	Ticker Symbol	Total Revenue						
2	AVGO	$ 6,824,000,000		Ticker Symbol	AVGO			
3	AVGO	$ 13,240,000,000					Formula	Returned V
4	CSRA	$ 4,069,746,000			One criterion	Need location of ticker symbol in H2	=MATCH(H2,ticker_symbol,0)	
5	CSRA	$ 4,250,447,000						
6	HPE	$ 55,123,000,000					=INDEX(Total_revenue,1))	$ 6,824
7	HPE	$ 52,107,000,000					=INDEX(Total_revenue,MATCH(H2,ticker_symbol,0))	$ 6,824
8	HPE	$ 50,123,000,000						
9	MYL	$ 7,719,600,000						
0	MYL	$ 9,429,300,000						

Index and Match with Multiple Criteria



	B	C	E	F	G	H	I	J
1	Ticker Symbol	Years	Total Revenue		Ticker Symbol	AVGO		
2	AVGO	Year 1	$ 6,824,000,000		Years	Year 1		
3	AVGO	Year 2	$ 13,240,000,000					
4	CSRA	Year 1	$ 4,069,746,000			Multiple criteria	Need location of ticker symbol in H16 AND year in H17:	=MATCH(1,(H16 =tickersymbol)*(H17 = years),0)
5	CSRA	Year 2	$ 4,250,447,000					
6	HPE	Year 1	$ 55,123,000,000					(H16 =tickersymbol)*(H17 = years)
7	HPE	Year 2	$ 52,107,000,000					{TRUE; TRUE; FALSE;FALSE;FALSE;FALSE;FALSE;FALSE;
8	HPE	Year 3	$ 50,123,000,000					{1;1;0;0;0;0;0;0;0} * {1;0;1;0;1;0;0;1;0}
9	MYL	Year 1	$ 7,719,600,000					{1;0;0;0;0;0;0;0;0}
10	MYL	Year 2	$ 9,429,300,000					
11								=MATCH(1,{1;0;0;0;0;0;0;0;0})
12								
13								=INDEX(costofrevenue,1))

Offset



Total operating expenses	$ 11,820,000,000	$ 12,237,000,000	$ 11,472,000,000
Operating income/ EBIT	$3,817,000,000	$2,702,000,000	$3,144,000,000

Operating statistics

	Historical	Assumptions
Revenue growth (%) =$F$8/$E$8-1		
Gross margin		
Operating margin		

Scenario: Base case

Operating Scenarios - sensitivity analysis

	Revenue growth (%)	
1	Strong case	1
2	Base case	2
3	Weak case	3

Gross Margin

Financial Forecasting Process – Part I

Remember we want to create three forecasting scenarios based on our assumptions:

- Best Case Scenario
- Base Case Scenario
- Weak Case Scenario

To do this, we need to walk through the following four steps to create our financial forecasting model:

- **Calculate operating statistics**
- **Create the scenarios**
- **Create assumptions** – this is where the **OFFSET** function will be used
- **Develop the forecasted scenarios**

Step 1 – Calculate operating statistics

Here we calculate operating statistics based on historical data from the income statement:

- **Gross margin** – 1- (Cost of good sold/Total revenue)
- **Operating margin** – Operating profit/Total revenue
- **Revenue growth** – (Current year revenue/Prior year revenue) -1

Step 2 – Create the scenarios

In this step, we create the three scenarios for each of our operating statistics because they will feed into the forecasting model. The numbers used in this step are based on:

- Historical data
- Business analyst's knowledge of the business
- Research
- Assumptions about the business itself

A key question to ask here would be, "Do you expect revenue growth to stay the same or increase, and if so by how much?"

Step 3 – Create assumptions

When running your models, you always want to avoid hardcoding in numbers. Rather, you want the spreadsheet to change dynamically based on the scenario you choose. This is where the OFFSET function can be used.

Excel Syntax

The purpose of the OFFSET function is to return a range that is a specified number of rows and columns from a reference cell or range.

OFFSET(cell_reference, number of rows to offset by, number of columns to offset by)

In the other words, using OFFSET allows you to move around on a page to pull data. Look at the following table for example;

A	B	C	
1	Fruit	Color	Quantity
2	Apple	Red	25
3	Banana		Yellow10
4	Grapes		Green 7
5	Input	Grapes	

If we are looking for data on the quantity of fruit based on the name input in cell B5 then:

cell_reference – The Quantity column name in the cell C1 would be our

Use MATCH – to find the row of the word in cell B5 based on the words in the range A2:A4

OFFSET Syntax – OFFSET($C$1, MATCH($B$5, $A$2:$A$4,0),0), the 0 at the end tells Excel to stay in column C because we want the quantity value.

Output – 7