Date: December 30 2022

# Introduction to Spreadsheets for Data Analysis

Course Overview:

**Module 1**

In this module, we will learn about the:

* basics of spreadsheets
* spreadsheet terminology
* The interface – Navigating Around a Spreadsheet

**Module 2**

* Selecting, Entering and Editing Data
* Copying and Auto-Filling Data
* Formatting Data
* Using Functions and Formulas

**Module 3**

* Cleaning and Wrangling Data Using a Spreadsheet
* The Fundamentals of Data Quality and Data Privacy
* Removing Duplicated and Inaccurate Data
* Removing Empty Rows
* Removing Data Inconsistencies and Whitespaces
* Using the ‘Flash Fill’ and ‘Text to Columns’ Features

**Module 4**

* Analyzing Data Using Spreadsheets
* Filtering and Sorting Data
* Using Common Data Analysis Function
* Creating and Using Pivot Tables
* Creating and Using Slicers and Timelines

At the end of the course,

* Understand how spreadsheets can be used as a data analysis tool.
* Understand when to use spreadsheets as a data analysis tool and their limitations.
* Create a spreadsheet and explain its basic functionality
* Perform data wrangling and data cleaning tasks using Excel.
* Analyze data using filter, sort, and pivot table features within Excel spreadsheets.
* Perform some intermediate level data wrangling and data analysis tasks to address a business scenario.

## How a Data Analyst Uses Spreadsheets?

As a Data Analyst, we can use spreadsheets as a tool for our data analysis tasks, including

* collecting and harvesting data from one or more distributed and different sources.
* Cleaning data to remove duplicates, inaccuracies, errors and resolve missing values to improve the quality of the data.
* Analyzing data by filtering, sorting and interpreting it to determine what useful information can be gleaned from it.
* And visualizing data, to help us tell a story about our data analysis findings to key business stakeholders and any other interested parties within our organizations.

## Spreadsheet Basics:

* Workbooks are the highest-level component in Excel and are represented as a .XLSX file.
* Workbook consists of one or more worksheets each of which is represented by a tab in Excel.

Getting Started using Spreadsheets:

* Viewing, Entering and Editing Data:

Viewing features in Excel, enter some data and edit data.

* Copying, Fulling and Formatting Cells and Data

Entered and edited some data., how to move, copy and fill data, format cells

The Basics of Formulas:

* Basic calculations (SUM)
* Selecting ranges in formulas
* How to copy formulas

The common functions used by Data Analysts:

* Some statistical functions [Formulas tab at the top bar]

Referencing Data in Formulas:

* The difference between relative, absolute, and mixed references
* How to use relative, absolute, and mixed references in formulas
* Formula errors in Excel

**Practice Quiz**:

1. How can you zoom to a specific area of data in an Excel spreadsheet?

* The Zoom to Selection button will zoom in to a specific area of your data.

1. What do you use the AutoFill feature for?

* AutoFill is used to fill cells automatically for you, when the data is identified to be in a sequential series or pattern.

1. What is one of the key components of a typical formula?

* While you can use division and percentages in your calculations in a formula, of these options, only **Reference** is a key component of a typical formula

1. What happens when you use the median calculation but select an even number of values in a range?

* Returns middle figure between the two middle values in the selected range.
* The median is a midway point between a range of values, so if you select an odd number of values, it uses the middle value, but when you select an even number of values, the median is the midway point between the two middle values.

1. What are Excel cell references by default?

* By default, in Excel, cell references are always relative, and you must manually configure a cell reference to be absolute or mixed.

1. When creating formulas, what is a mixed reference?

* A mixed reference has at least one relative cell reference and at least one absolute cell reference.

**Basics of Data Quality and Privacy**:

**Introduction to Data Quality**:

Data analysis can play a pivotal role in business decisions and processes. In order to use the data to make confident decisions, we must have the right information for the project and the data must be free from errors. In this video we will learn how to profile data to discover inconsistencies. Whether we are working with small sets of data or analyzing a spreadsheet with thousands of rows, one of the most difficult parts of the data analysis is finding and keeping clean data.

To help with this process and qualify the data, look for these five traits: **Accuracy, Completeness, Reliability, Relevance and Timeliness.**

**Accuracy:**

* Accuracy is the first and most significant aspect to data quality.
* A data analyst must clean the data set by removing duplicates, correcting formatting

errors, and removing blank rows.

**Completeness**:

* Another important aspect of data quality is determining if the information required to

complete the data set is readily available. Why does this matter as a trait for quality data?

Let’s say we are given the task to calculate the revenues of all sales per region.

After collecting the data, we discover that no regions were specified.

This data would then be considered incomplete and other sources would have to be considered to obtain the data required.

**Reliability**:

* Reliability is another vital factor in determining the quality of the data.
* For instance, let’s say we are given the task to determine the agent revenue by customer.

When gathering the data, we find the agents keep their own records and do not always update the information in the shared company database.

* With those factors in mind, we would then determine that the data in the shared company

database was unreliable and new processes would need to be established to ensure reliable data.

**Relevance:**

* Relevance is another trait of quality data.
* When collecting information, a data analyst must consider if the data being assembled

is really necessary for the project.

For example, when reviewing the data related to the sales revenue per customer, information

such as customer birthdays and other personal information is also included.

* By making the determination early to exclude the personal information from the data set,

the analyst would save themselves from having to review unnecessary information.

**Timeliness**:

* The last factor in determining the quality of the data is timeliness.
* This trait refers to the **availability and accessibility of the selected data**.

Let’s say our sales report is going to be used for weekly employee reviews, but our

report is only refreshed once a month.

This error in refreshing the data would cause our report to become outdated, and would have

serious consequences for employee reviews.

From this, we learned the important role of a data analyst in qualifying data. By considering the five traits of good quality data, an analyst can save time, avoid serious issues, and have data that is free from errors.

**Importing File Data**:

Importing file from different sources like csv, txt etc. to the spreadsheet workbook.

Importing Text:

We can use the ‘Text Import Wizard’ to import data from other formats, such as plain text, or comma-separated value files.

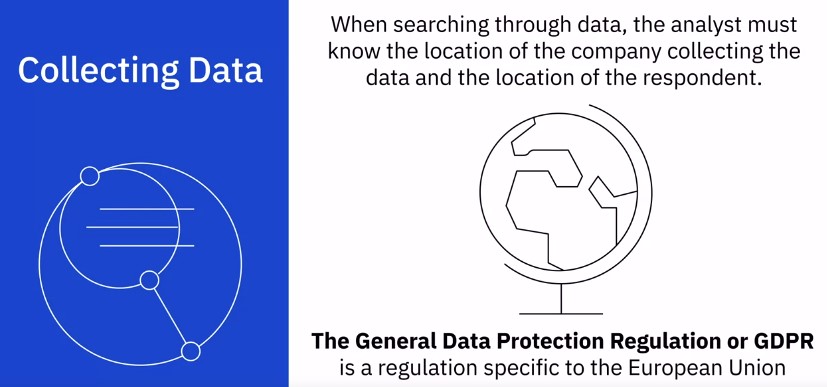
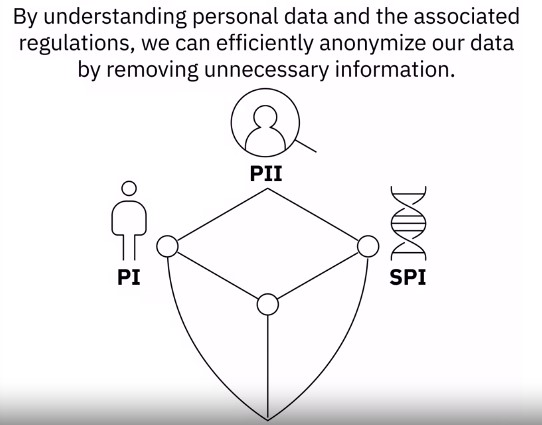
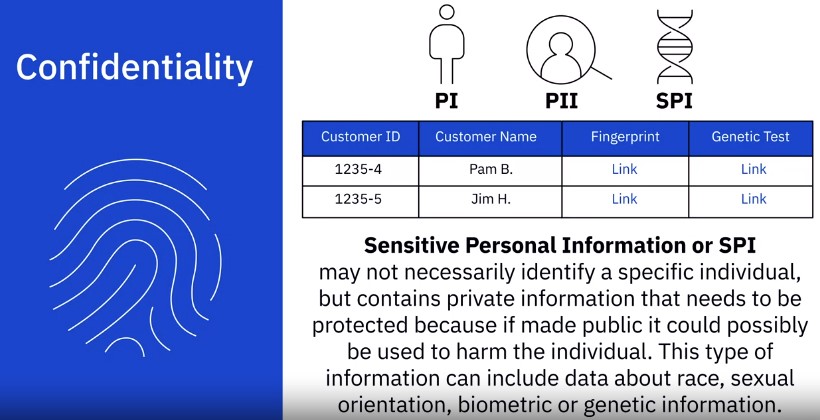
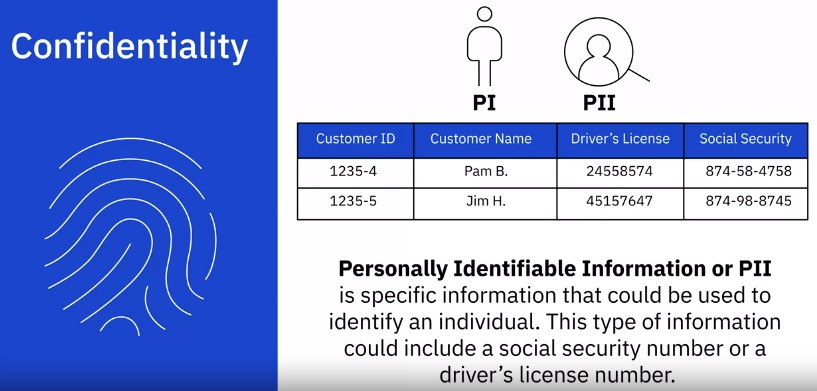
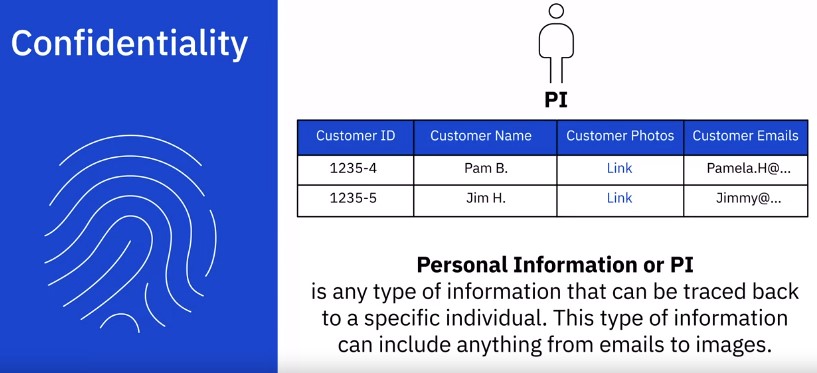
**Basics of Data Privacy**:

The Three Fundamentals of Data Privacy:

* Confidentiality
* Collection and Use
* Compliance

**Delimited file**:

A delimited file is a file that has data fields separated by characters, the most common of which are comma and tab**.**



For example, let’s say a data analyst downloads a spreadsheet of sensitive information. In order to complete the report by Monday morning, the analyst decided to take their work laptop home for the weekend. After driving home, the analyst accidently left the laptop in their car. The next morning, they found their car had been stolen along with the laptop. Because it is the responsibility of the company to keep customer data safe, this was a breach of privacy when the data left company property. This type of action could not only cost the company large amounts of money in fines and penalties, but could also reduce consumer confidence causing a significant impact to revenue. While data privacy applies to most data that is collected, there are some instances where these regulations do not apply. In order for these laws and regulations not to apply, the particular collection of data must be completely anonymous. To make data anonymous means to exclude all data which ties it back to a particular individual. While this approach might not be practical in all circumstances, collecting data with privacy in mind cloud remove privacy limitations and make data collections more accessible.

What is the importance of data quality as it relates to data analysis?

Data quality is of the utmost importance in terms of data and analytics, but the reason behind this is because as soon as what you're presenting does not align with what someone expects, that's the first thing that they tend to go after.

Where did you get the data?

What's happened to the data?

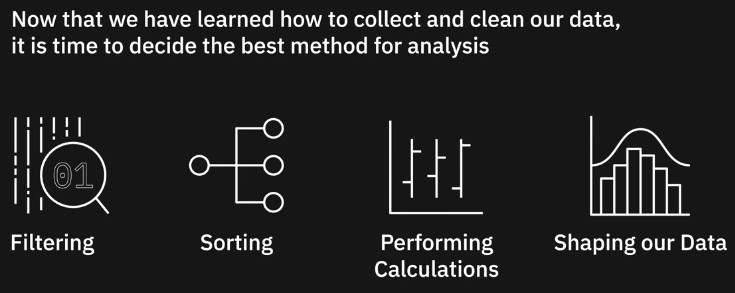
How's it been transformed?

Because people like to think that they know and understand their, their business. And when you start to challenge that if you don't have the ground to stand on of the data that it's quality that it's clean and then it is from a trusted source, that's when you start to get into a lot of discussions. A lot of debate. And ultimately, the plot of what you're trying to present gets lost. The backbone of any successful data analysis project is good quality data. There is a common term in computer science called garbage in garbage out, which is essentially if we read in bad quality data, you can expect to get bad quality results.

What is the Importance of data privacy as it relates to data analysis?

In today's world, data privacy is a huge thing on the tax side, especially of our business we have what we have what we call PII: personal, identifiable information. We have to protect that and so we can't just send things through email. We don't send tax returns or even actually in our business. We don't send things through email. They have sensitive PII data in it. We encrypt it. We make sure the email is encrypted or we use software. Some certain software’s that will allow us to not show the social security numbers or the names or the date of birth and what will happen is it has a certain sequence, and we share that with the client by calling them. We don't put that in an email and we certainly don't put that in the same email with the encrypted information because we want to make sure that you are always safe. So, we have to make sure we're protecting it at all costs.

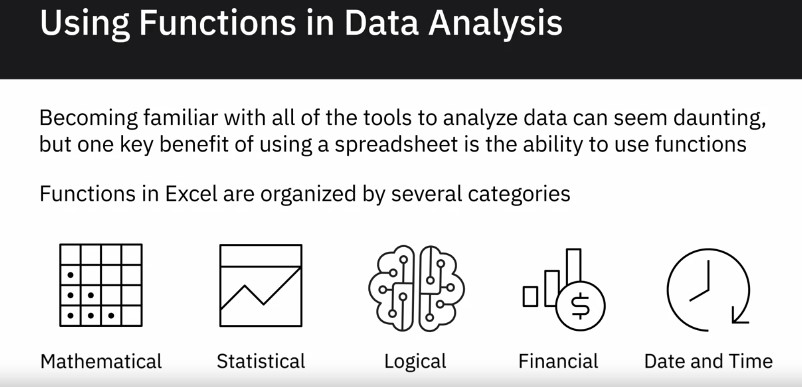
**Data Analysis Basics, Filtering and Sorting Data**:

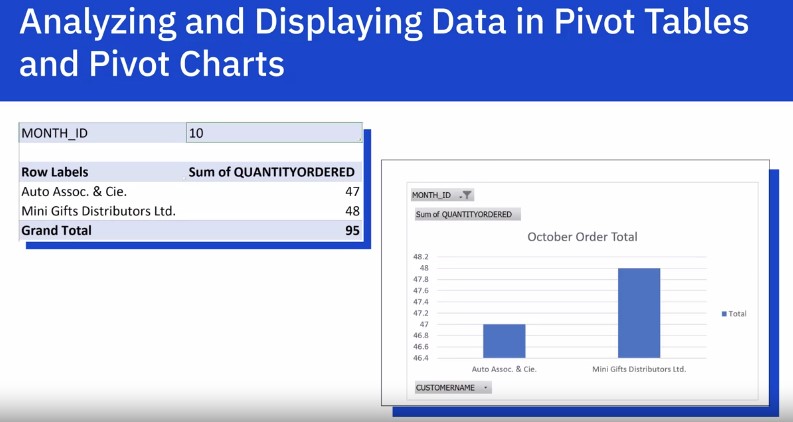
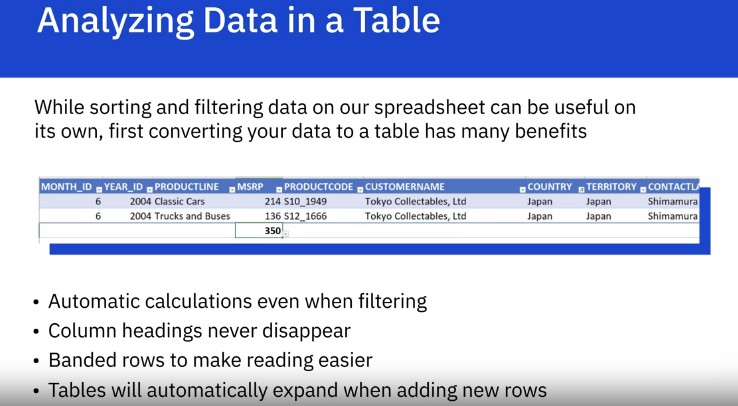


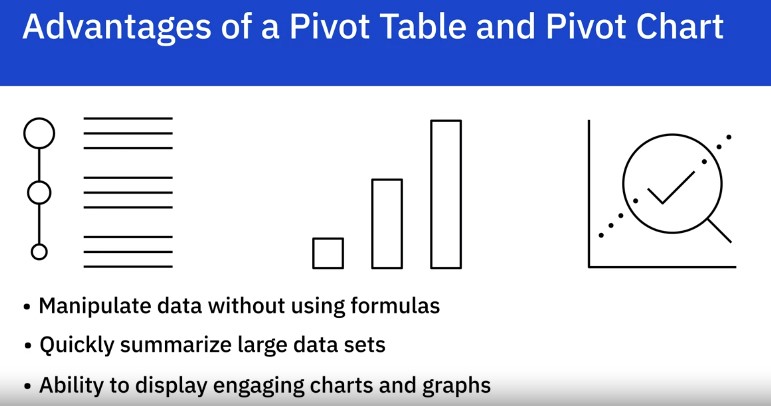
Howe to shape our data?

Shaping our data to provide meaningful information. Deciding how to manipulate our data can sometimes be difficult. Before we make any changes or adjustments, we will need to visualize the final output. Below are some questions to ask before beginning the task. How big is the dataset? What type of filtering is required to find the necessary information? How should the data be sorted? What type of calculations are needed? Now that we have visualized the final output, we must decide the best approach to shape our data.

Filtering and Sorting Data:







Filtering and Sorting Data in Excel:

Why is it important to filter and sort data?

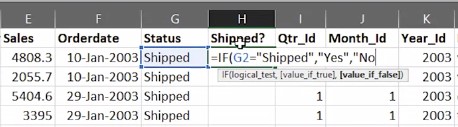
Filtering and sorting are very important as part of analysis. It allows us to create one single view of the data. it provides a function for people to do their own analysis on the data. sorting tends to be highest to lowest or alphabetical.

Filtering means we have a value by which we want to see the data. if we have a bar chart showing product sales over months, filtering allows us to filter down, so we only see one geography or one product line.

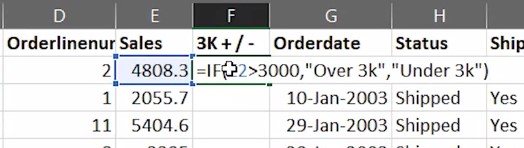
**Some most important functions that Data Analyst used?**

1. **IF Function**:

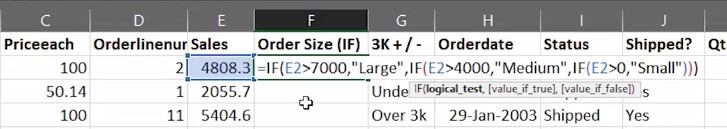
IF Function is one of the most used logical functions in Excel. IF compares a value against criteria, then returns a result based on TRUE or FALSE comparison. It says: “if something is TRUE, return ‘x’, but if it is FALSE, return ‘Y’”.



It displays Yes because G2 has a text Shipped.

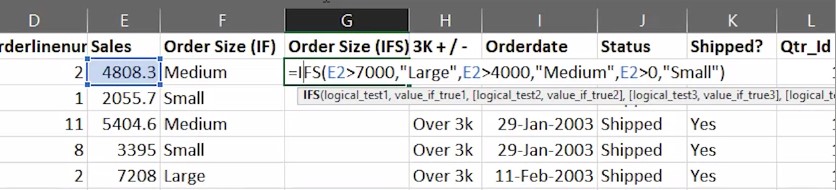


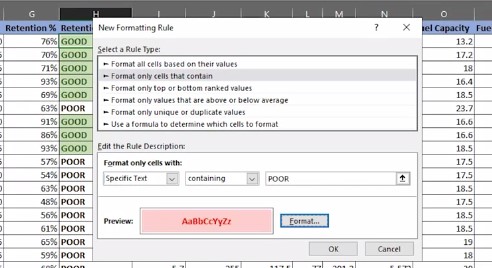
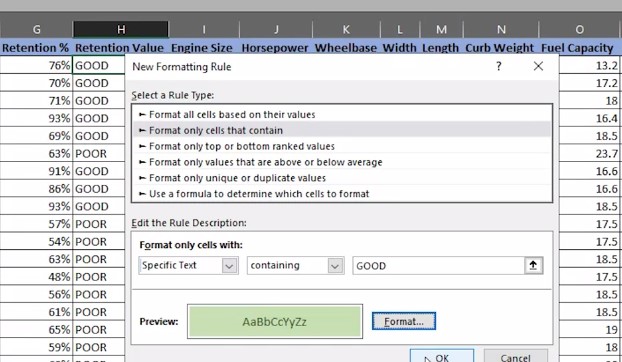
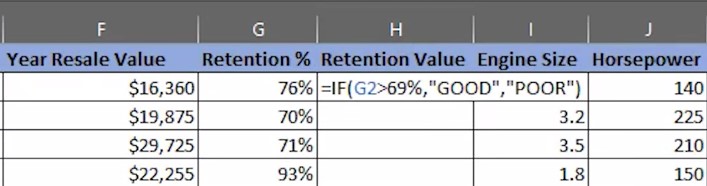
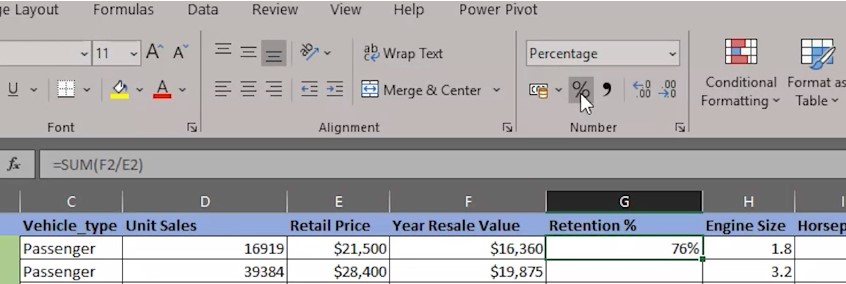
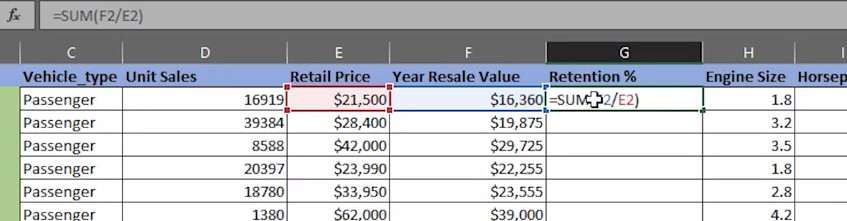
It displays Over 3K.



Excel theoretically supports up to 64 nested ‘IF’s – but it’s not best practice. Having large numbers of ‘IF’s in a formula is hard to manage and understand.

A newer function called ‘**IFS**’ eliminates the need for multiple nested ‘IF’s. IFS only supported in Excel 2019, Excel for M365, and Excel for the web.

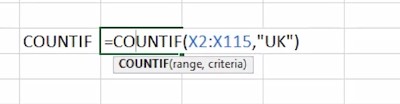




1. **COUNTIF Function:**

COUNTIF is a common statistical Excel function. It counts the number of cell values that meet specified criteria. For instance, how many times does ‘Employee X’s name appear in sales orders? How many times has ‘part number Y’ been ordered this month/year?

Suppose we want to find out how many of the sales orders in the list went to customers based in the UK.

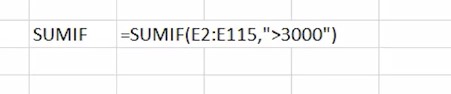


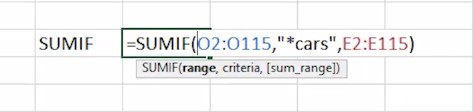
1. **COUNTIFS:**

Newer ‘COUNTIFS’ function eliminates need for multiple nested ‘COUNTIF’s. it only supported on EXCEL 2019, Excel for M365, and Excel for the web.

1. **SUMIF:**

SUMIF function, which is a very commonly used mathematical function in Excel. You use the SUMIF function to sum the values within a specified range that meet specified criteria. For example, you might want to add up only the salaries that are over a specified salary level, or you might want to find the total of all sales of a particular product category.



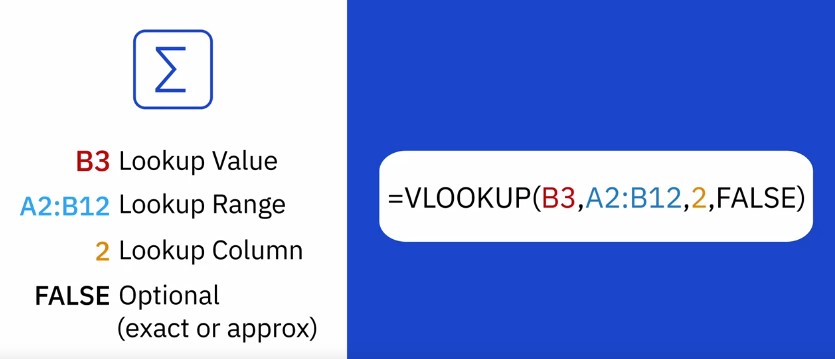


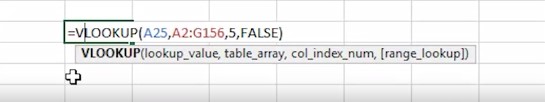
1. SUMIFS:

Newer ‘SUMIFS’ function eliminates need for multiple nested ‘SUMIFS’s. it only supported on EXCEL 2019, Excel for M365, and Excel for the web.

**Using VLOOKUP and HLOOKUP Functions**:

**VLOOKUP** is one of the most commonly used reference-type functions in Excel, and it enables you to find data referenced in a lookup table. It stands for Vertical Lookup and therefore is a useful tool to use when you want to find something in a table or a range by row. Shortly, we will look at HLOOKUP, which stands for Horizontal Lookup, which looks for data by column instead. VLOOKUP works by using a common shared key between the source data and the lookup data in the lookup table. A typical VLOOKUP would look like:





**HLOOKUP** function, which as we mentioned earlier, does the same thing, and works in virtually the same way, as the VLOOKUP function, but it looks for data in columns, rather than rows. So, HLOOKUP looks for a word or value in the top row of a table, and then returns a value in the same column from a row specified in the table array. Therefore, you would use HLOOKUP if your comparison values were situated in a row along the top of a data table. In contrast, you would use VLOOKUP if your comparison values were located in a column to the left of the data you want to find; as they were in the previous task. Of the two functions, VLOOKUP is used far more than frequently than HLOOKUP,

because of the nature of most data tables. The syntax for HLOOKUP is identical to that of VLOOKUP except that you specify a row index number, referenced in a formula by Excel as ‘row\_index\_num’. This indicates the number of the row in the lookup table that contains the value you are looking for.

