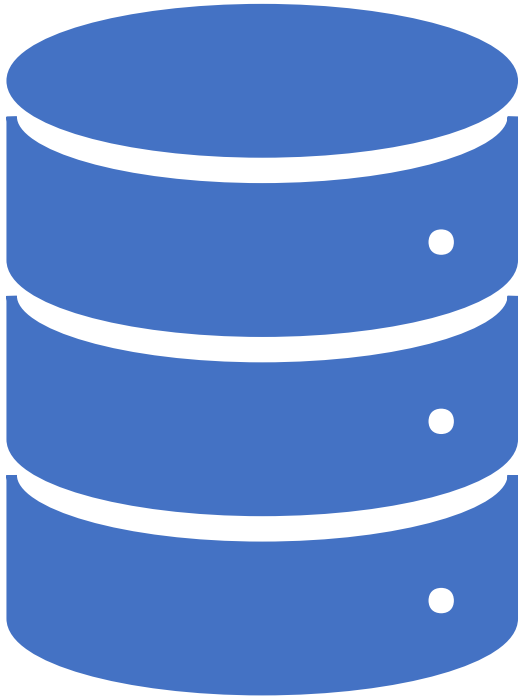


Data Management

01. Course Objectives

- Define Data cleaning
- Steps to validate and clean data
  - Explain concatenated functions, logical values/operators & formulas
  - Introduce AND or NOT functions (AND function, IF with AND function, OR NOT function, complex nested function cells, ignoring numbers in calculations)
  - Explain lookup functions (VLOOKUP and INDEX- MATCH)
- Using Excel Power Query to prepare data
  - Describe the steps to use to power query features
  - Explain the Pivot tables for data preparation
- How to measure data quality
  - Find patterns and trends
  - Find data outliers and how to handle it
  - How statistics can help
- Analyzing data using Excel
  - Create charts & graphs

Data Management

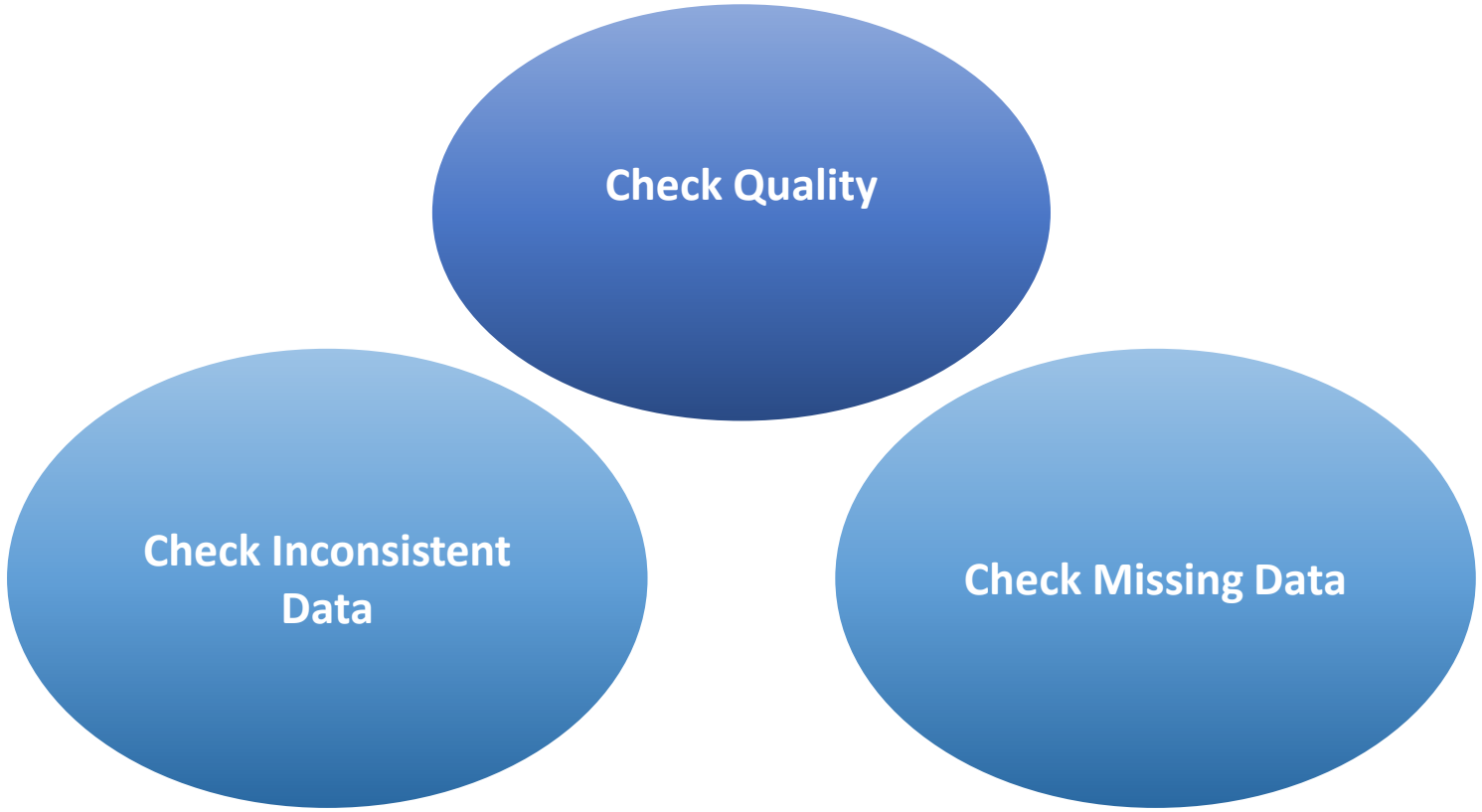
- It is the practice of collecting, keeping, and using data securely, efficiently, and cost-effectively
- The goal is to optimize the use of data so that you help people make decisions and take actions that maximize the benefits to everyone
- Why data management is important ?
  - Achieve goals and objectives
  - Make correct decisions
  - Innovate, Improve and Enhance…

Data Management Cycle

1. Design Log-frame or Matrix ("Collect first, analyze later **X**")
2. Data Collection
3. **Data Cleaning**
4. **Data Preparation**
5. **Data Visualization (Excel)**

Data Cleaning

- It is the process of ensuring data is correct, consistent and usable. You can clean data by identifying errors or corruptions, correcting or deleting them, or manually processing data as needed to prevent the same errors from occurring.
- Most aspects of data cleaning can be done through the use of software tools, but a portion of it must be done manually. Although this can make data cleaning an overwhelming task, it is an essential part of managing company data.



Missing Data

- If you find missing data, check with data provider and ask for compete dataset
- Shall I predict it ?
  - Imputation using the mean or median
  - Imputation using “most frequent” or zero values especially with categorical type of data
  - Eliminate the records that has missing data – sometimes hard ?

Inconsistent Data

Date Format:

- Use Format cell and convert it to the right template and filter issues in wrong dates
- Add a validation rule on the column in the excel template

Number, same thing to add a validate rule on the format or you use format cell

Drop down lists

- Fixes mis-spelling issues

Duplicates

- You can highlight all duplicates using excel by defining your criteria

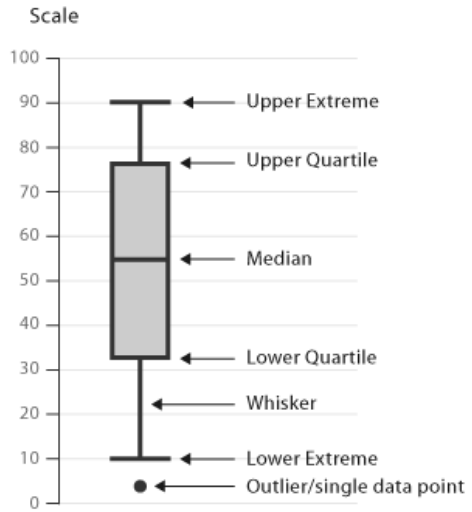
Quality Check

Meta data quality check list:

- Average duration of survey – To figure out if its fake or not
- GPS points – you can use KOBO Mapping if possible

Find data outliers or patterns/trends:

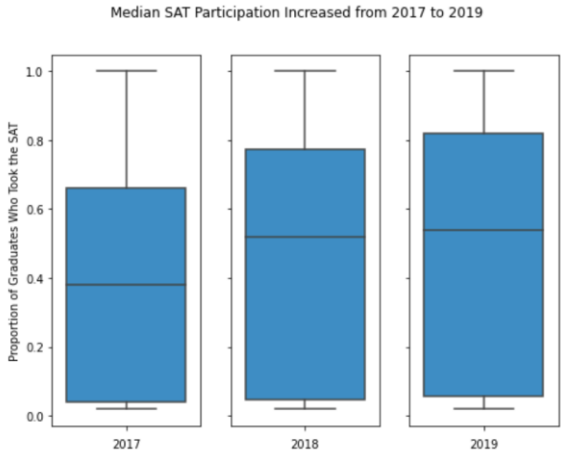
- **Method 1**: Calculate Max and Min values
- **Method 2**: Find Outliers in your data
  - Try to use the box and whiskers chart to check the variance in your data and show outliers in the data:
    - Min
    - Q1
    - Median or Q2
    - Q3
    - Max



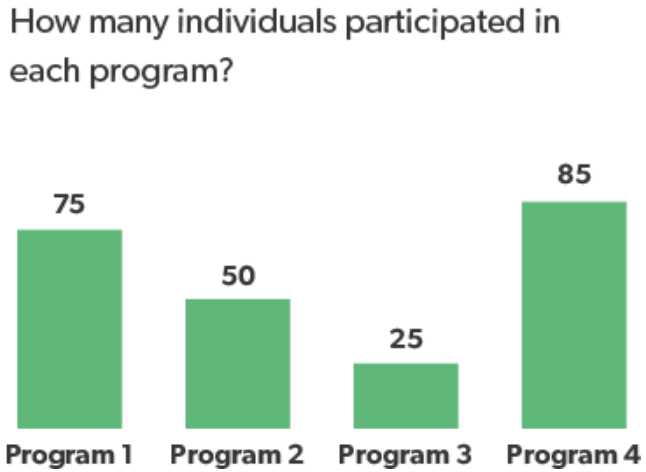
How to use Excel to clean data

Find data outliers or patterns/trends:

- Method 3: Find trends and patterns:
  - Trends: The general tendency of a set of data to change (Up or down).
    - You can use Mean or Median
  - Patterns: It is set of data that repeats itself in a predictable way.
    - You can use frequencies of the answers as shown in the right graph



Increasing trend in the median



Pattern spotted: Most of the individuals participated in program 1 & program 4

Data Cleaning and Verification Log

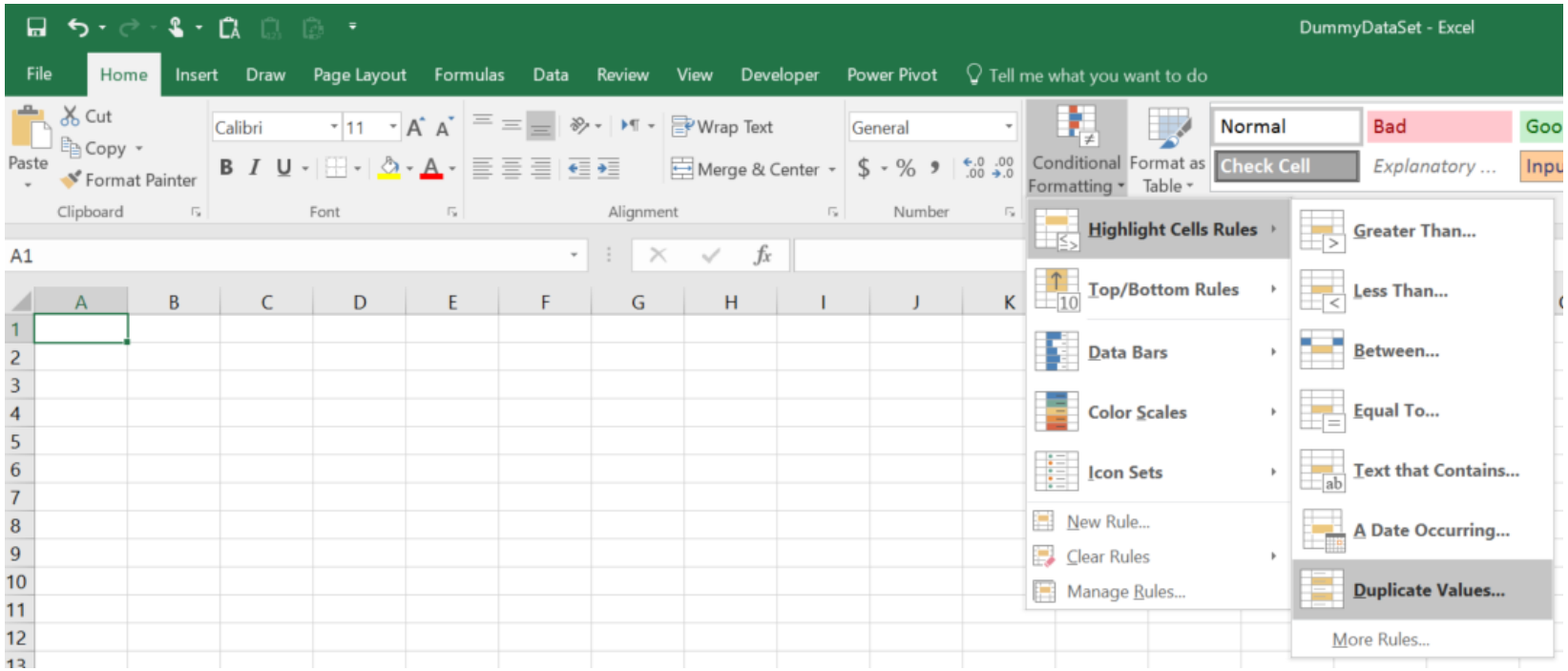
- Create a change log within your workbook, where you will store all information related to modified fields
- This will serve as an audit trail showing any modifications, and will allow a roll back to the original value if required
- Within the change log, store the following fields:
  - Table (if multiple tables are implemented)
  - Column, Row
  - Date changed
  - Changed by
  - Old value
  - New value
  - Comments

Introduction to MS Excel

- Conditional Formatting
- Formulas
- Arithmetic Operators
- Relative Cell Referencing Vs. Absolute Cell Referencing
- Common Functions
- Text Functions
- Advanced Functions
- Pivot table
- Power Query
- Charts & Graphs

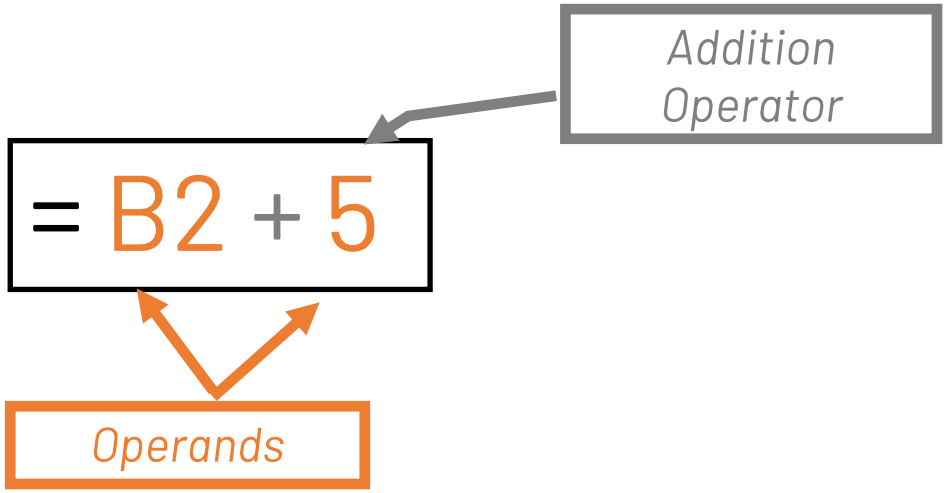
Conditional Formatting

Conditional formatting lets you format cells based on their value using conditions



Formulas

- Formulas are equations that perform calculations on values in your worksheet.
- A formula starts with an equal sign (=).
- Formulas contain two types of components:
  - **Operators**: Operations to be performed
    - Arithmetic operators: * / + - ^
    - Relational operators: >, <, <=, >=,< >,=
  - **Operands**: Values to be operated on



Arithmetic Operators

Operator	Purpose	Example
+	Addition	A1+B1
-	Subtraction	A1-B1
*	Multiplication	A1*B1
/	Division	A1/B1
^	Exponentiation	A1^3
%	Percentage	A1%

Relative Cell Referencing Vs. Absolute Cell Referencing

- By default, all cell references are relative references. When copied across multiple cells, they change based on the relative position of rows and columns. For example, if you copy the formula =A2*B2 from row 2 to row 6, the formula will become =**A6*B6**
- Relative references are especially convenient whenever you need to repeat the same calculation across multiple rows or columns

- Unlike relative references, absolute references do not change when copied or filled
- You can use an absolute reference to keep a row and/or column constant. An absolute reference is designated in a formula by the addition of a dollar sign **($)**. It can precede the column reference, the row reference, or both.

Common Functions

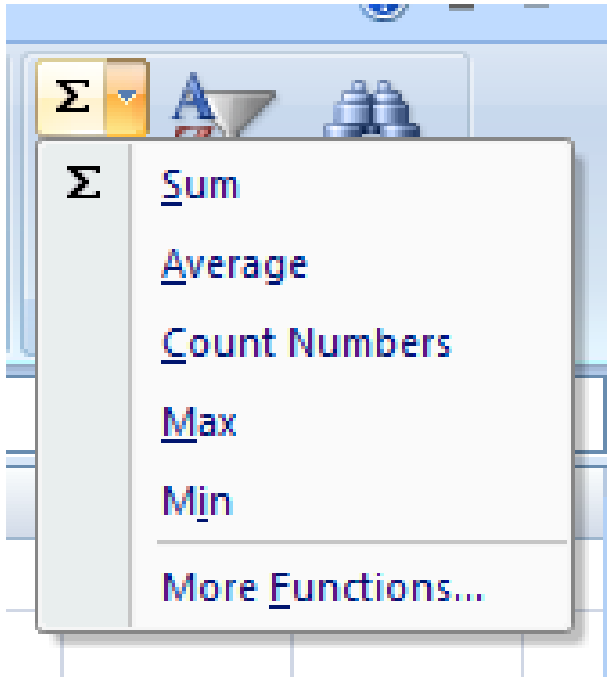
- **SUM**
  Calculates the sum of a range of cells

- **MAX**
  Displays the largest value in a range of cells

- **MIN**
- Displays the smallest value in a range of cells

- **COUNT**
  Calculates the number of values in a range of cells

- **AVERAGE**
  Calculates the average of values in a range of cells



Text Functions

Many functions are used to manipulate text values:

- **Right( )**
- **Left( )**
- **Mid( )**
- **CONCATENATE(A1," ",B1)**
- **Lower( )**
- **Upper( )**
- **Len( )**
- **Proper( )**
- **Trim( )**

Advanced Functions

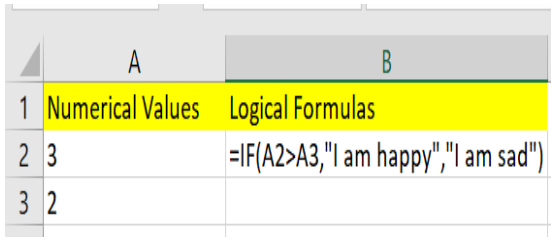
A logical value can be one of only two values:

- An IF statement allows you to make logical comparisons between a value and what you expect. In its simplest form, the IF function says: **IF(Something is True, then do something, otherwise do something else)**

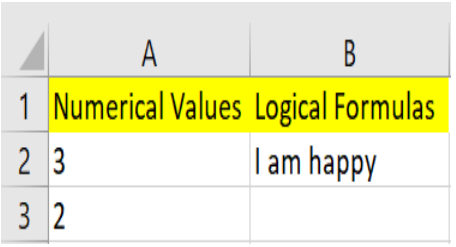
Operator	Meaning
>	Greater than
<	Less than
>=	Greater than or equal to
<=	Less than or equal to
<>	Not equal to

Examples

3 > 2	True
3 < 2	False



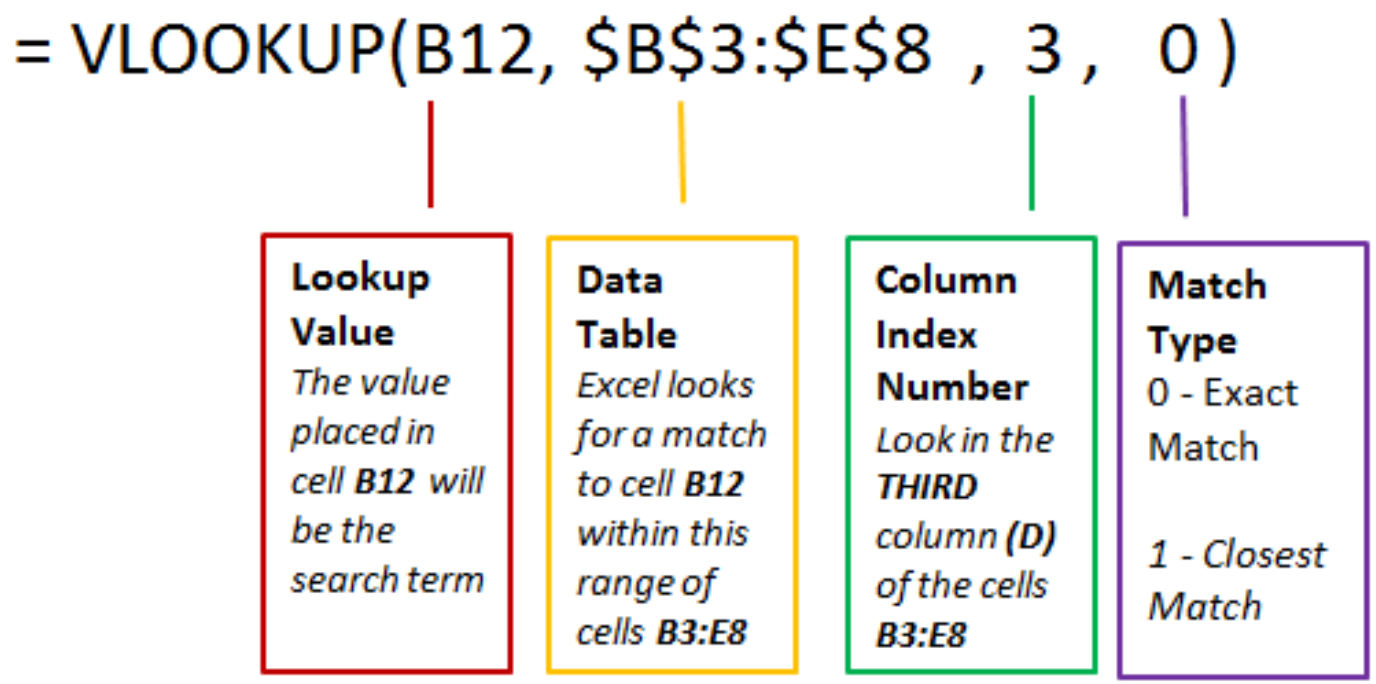
	A	B
1	Numerical Values	Logical Formulas
2	3	=IF(A2>A3,"I am happy","I am sad")
3	2	



	A	B
1	Numerical Values	Logical Formulas
2	3	I am happy
3	2	

Advanced Functions

Use VLOOKUP when you need to find things in a table or a range by row



Advanced Functions

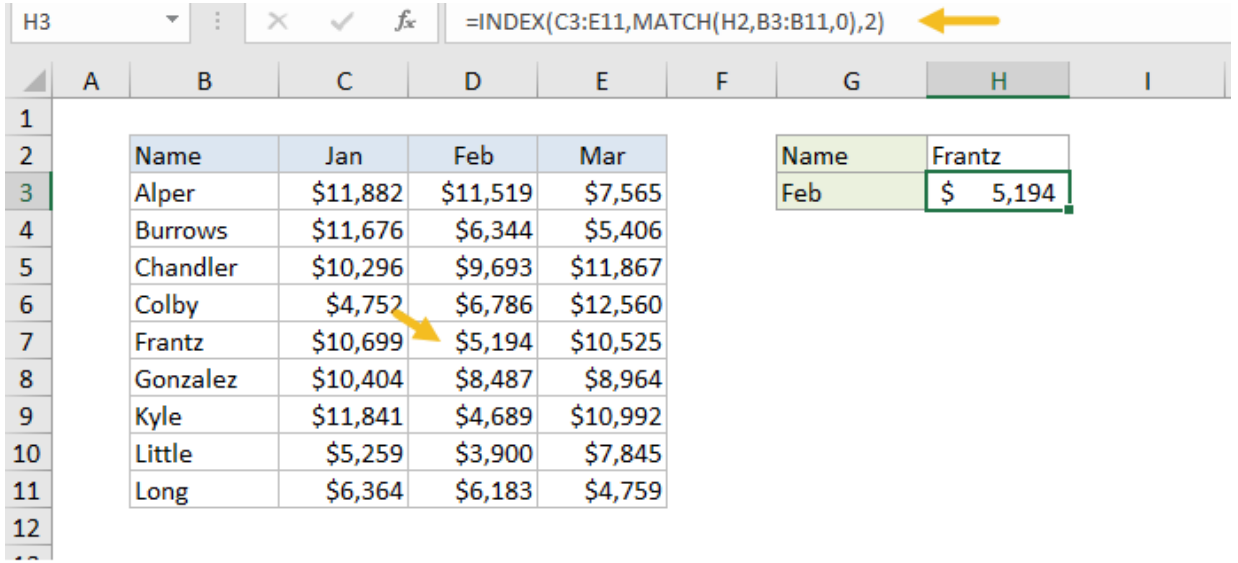
Use INDEX/Match to retrieve the value at a given location in a range.

=INDEX (array, row_num, [col_num])

Array - A range of cells, or an array in which you seek to find the content of cell you search by

Match – Match (Search by cell, column of the matched values in the array, exact match=0, semi match=1)

Col_num - The column position in the reference that you want to retrieve



Advanced Functions

Use offset – Cascading Selects e.g. (State, Townships when you build up a data collection tool using excel sheet)

E.g.
=OFFSET(**Reference**,**Rows**,**Cols**,**Height**, **Width**)

- **Reference**: In our formula, the reference is **state_start of the state list**
- **Rows:** How many rows down from the reference cell should our range start? MATCH(state_cell_num,**state_list**,0)
- **Cols**: How many columns away from the starting range? We want a range that is 1 column to the right of the state_start reference.
- **Height:** How many rows in the selected range? The COUNTIF function counts the number of times that State is entered in the StateColumn**- >** COUNTIF(state_list , state_cell_num)
- **Width:** How many columns in the selected range? E.g. 1

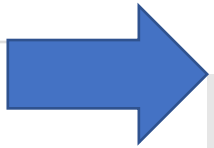
Let us build a simple data collection tool using Excel

- Let us build a 4Ws matrix:
  - Who (Org name)
  - Where (Region, Township)
  - When: Project start and end dates
  - What: Activity ?

Advanced Functions

AND Functions are used to determine if all conditions in a test are TRUE.

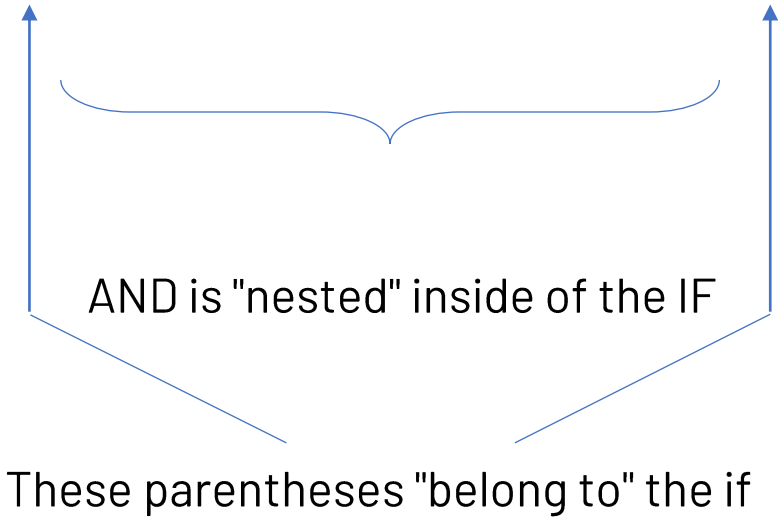
	A	B
1	Numerical Values	Logical Formulas
2	3	=AND(A2<A4,A3>A5)
3	2	
4	100	
5	50	



	A	B
1	Numerica	Logical Formulas
2	3	FALSE
3	2	
4	100	
5	50	

Advanced Functions

- You can use an AND inside of an IF
- This is called a NESTED FUNCTION CALL



Advanced Functions

OR And NOT Functions (Cont.):

OR

- Takes any number of parameters
- Returns TRUE if ANY of the parameters evaluate to TRUE, otherwise returns FALSE

NOT

- Takes ONLY ONE parameter. Returns the "opposite" of the value of the parameter
- Returns FALSE if the parameter value is TRUE
- Returns TRUE if the parameter value is FALSE

Complex Nested Function Cells:

- =IF(AND(A2>A3, OR(B2=B3,C2<C3)), 500, 1000)
- =IF(NOT(AND(A2>A3, OR(B2=B3,C2<C3))), 500, 1000)

Advanced Functions

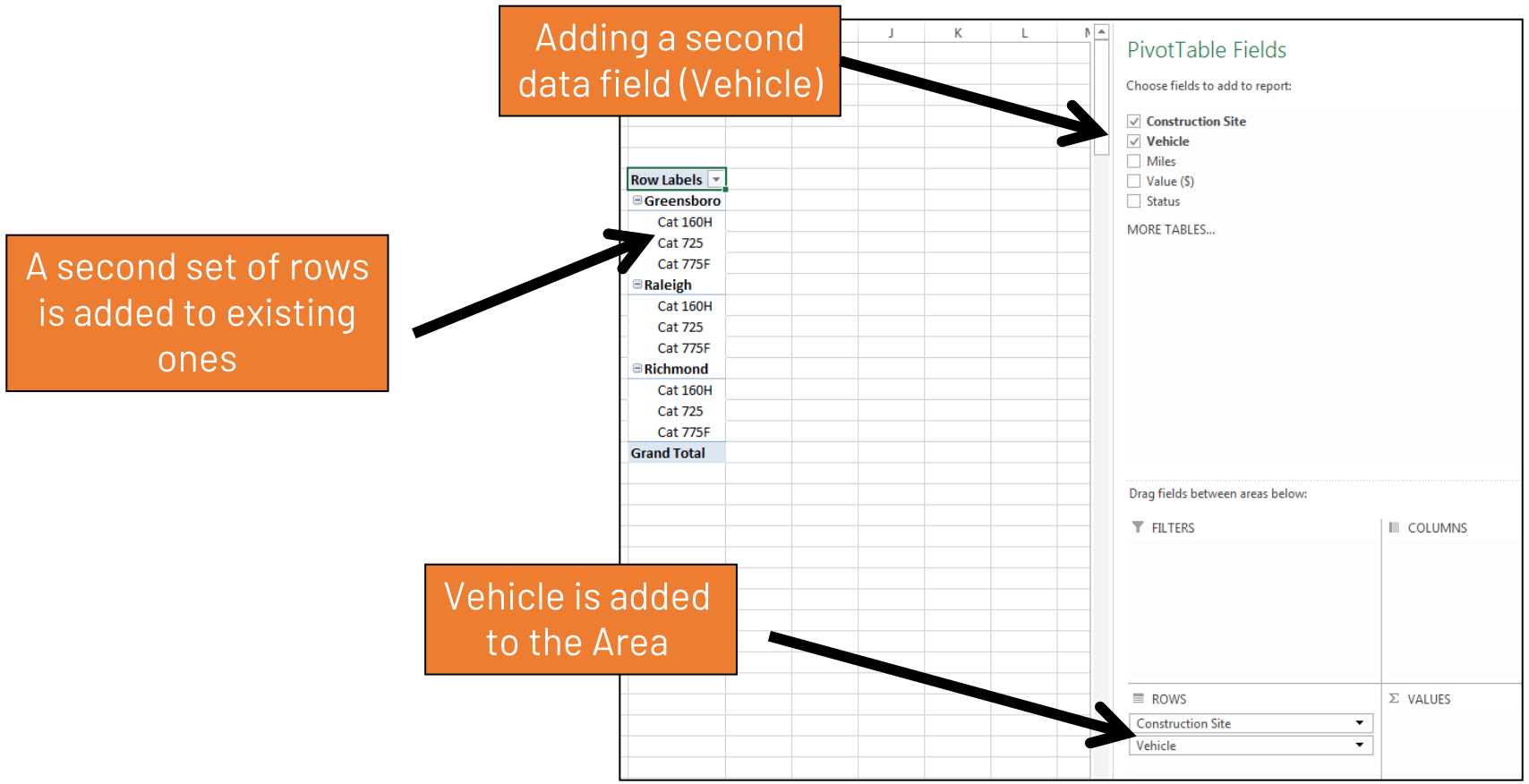
- NOW ( ) : Current datetime
- Today( ): Current date, time is 0:00
- Day( ): It extracts the day out of a date
- Month( ): It extracts the month out of a date
- Year( ): It extracts the year out of a date
- Date( ): it formats the date e.g. Date (year, month, day)
- DATEDIF(A1,A2,"Y"): It extracts difference in years "Y" between A1, and A2
- Text(A1, "dd/mm/yyyy"): if formats the date

Pivot Tables

- A pivot table is a special type of summary table unique to Excel that allows you to analyze your data
- Pivot tables are great for summarizing values in a table because they do their magic without making you create formulas to perform the calculations
- Pivot tables also let you play around with the arrangement of the summarized data
  - Data field
  - Row field
  - Column field
- It's this capability of changing the arrangement of the summarized data on the fly simply by rotating row and column headings that gives the pivot table its name

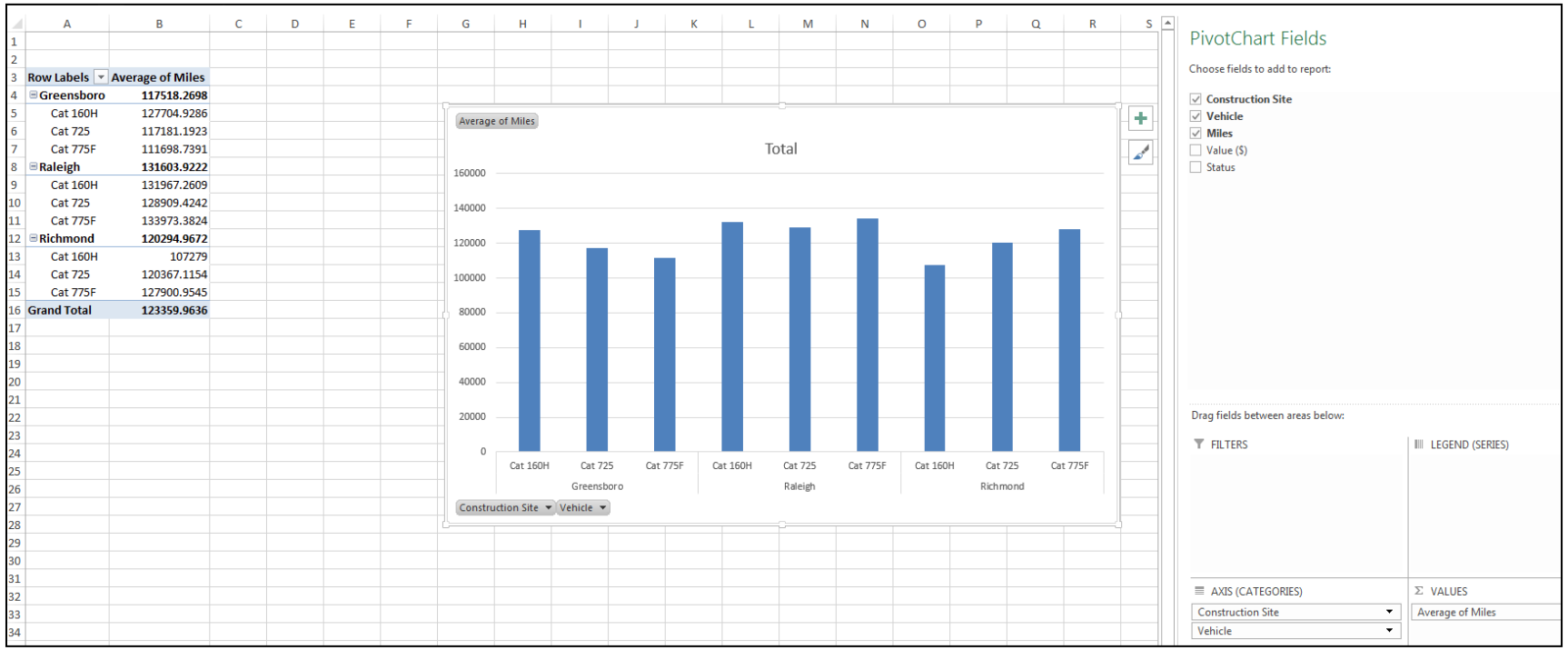
Building The Pivot Table

Adding more variables starts making the pivot table more useful…



Building The Pivot Table

You can also build pivot tables and pivot charts at the same time....



Data Re-Structuring

- It is a process to transform the data into a specialized format to be used for different analysis purposes
- Why do we need that? To analyze our data:
  - Create reports
  - Maps
  - Dashboards
  - Infographics
  - Graphs

Data Re-Structuring Challenge

With Power Query (called Get & Transform Data in previous Excel versions), you can import or connect to external data, and then shape that data, for example remove a column, change a data type, or merge tables, in ways that meet your needs. Then, you can load your query into Excel to create charts and reports. Periodically, you can refresh the data to make it up-to-date.

- **Connect:** Make connections to data in the cloud, on a service, or locally
- **Transform:** Shape data to meet your needs, while the original source remains unchanged
- **Combine:** Integrate data from multiple sources to get a unique view into the data
- **Load** : Complete your query and load it into a worksheet or Data Model and periodically refresh it.

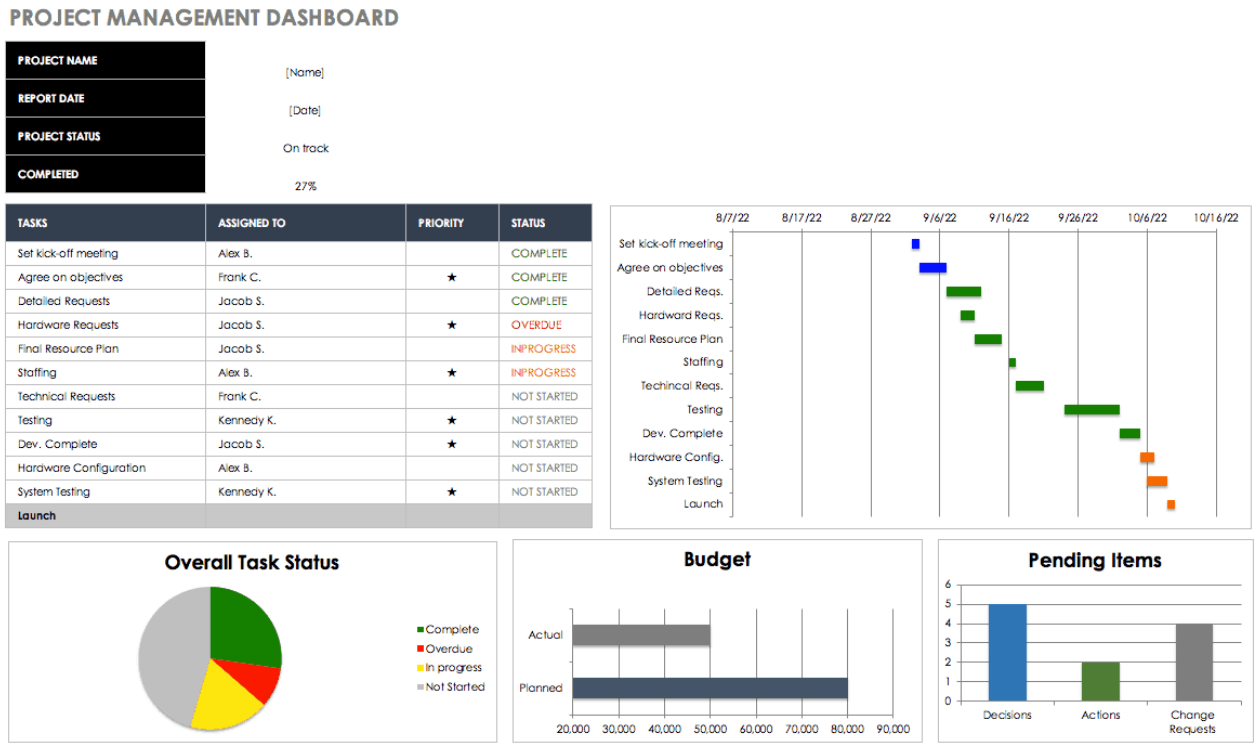
Data Re-Structuring Challenge

How can Power Query ease my life to structure and transform the format of my data ?

- Please open “NA.XLSX” Dataset file

Creating a dashboard using Excel

Let us build Needs Assessment dashboard using Excel, please use NA.XLSX file



Thank You