Basics on Data analysis

A step-by-step guide to the data analysis:

Data analysis follows a rigorous step-by-step process where each step requires different skills.

Following are the main steps involved:

A. Defining the question/Objectives
B. Collecting the required data
C. Cleaning the data
D. Plotting/Visualising the data
E. Analysing and sharing the results

In this document, the main steps involved in the data analysis processes are explained: Defining the question/objective, Data collection, Cleaning the data, visualising the data, and analysing and sharing/presenting the results obtained.

A. Defining/framing the questions/Objectives

This is the first step in any data analysis process, sometimes called as the problem statement. Defining your objective or question means – what problem are you going to solve?

To make it simpler for you all, the questions/objectives are already set/framed by your tutors and are available in module 2 tutorial activities.

B. Collecting the required data

Once the questions/objectives are framed, next step is to create a strategy to collect the appropriate data required: determining the type of data required – quantitative data or qualitative data or both. Depending on the objectives set, we have collected the data required for you all to work on the next steps of data analysis.

C. Cleaning the data in Excel

To perform data cleaning, first you need to understand your data, what information/details are collected, the parameters used to collect those details, and meaning of numerical data collected. Once you understand your data, then some basic data cleaning methods can be applied if needed. Few steps of data cleaning are explained in the following sections.

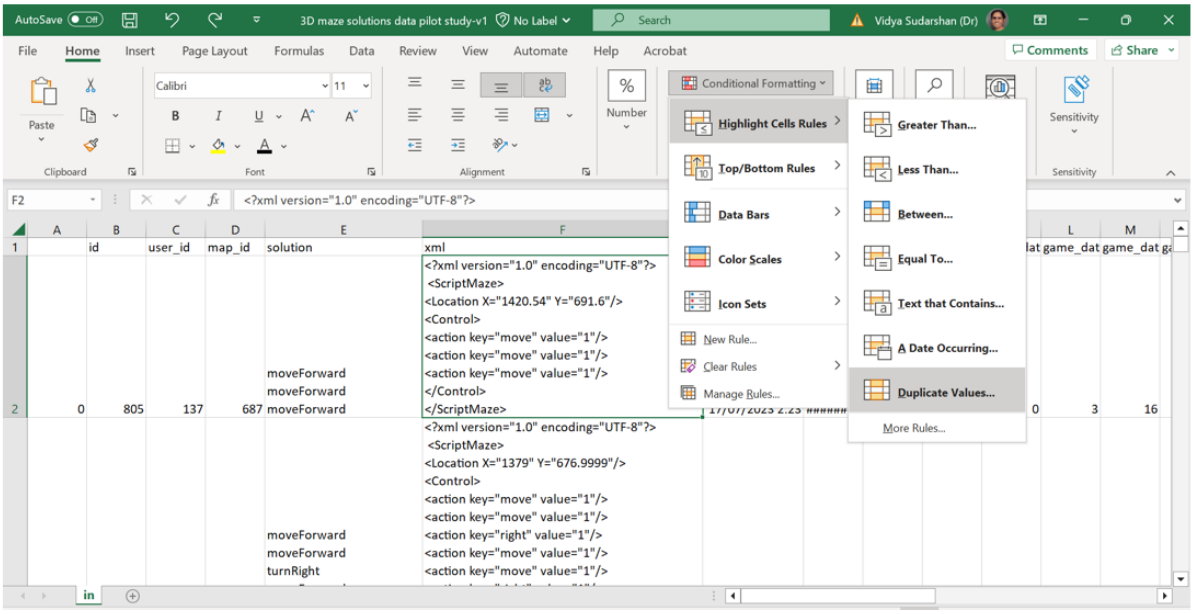
1. Remove duplicates:

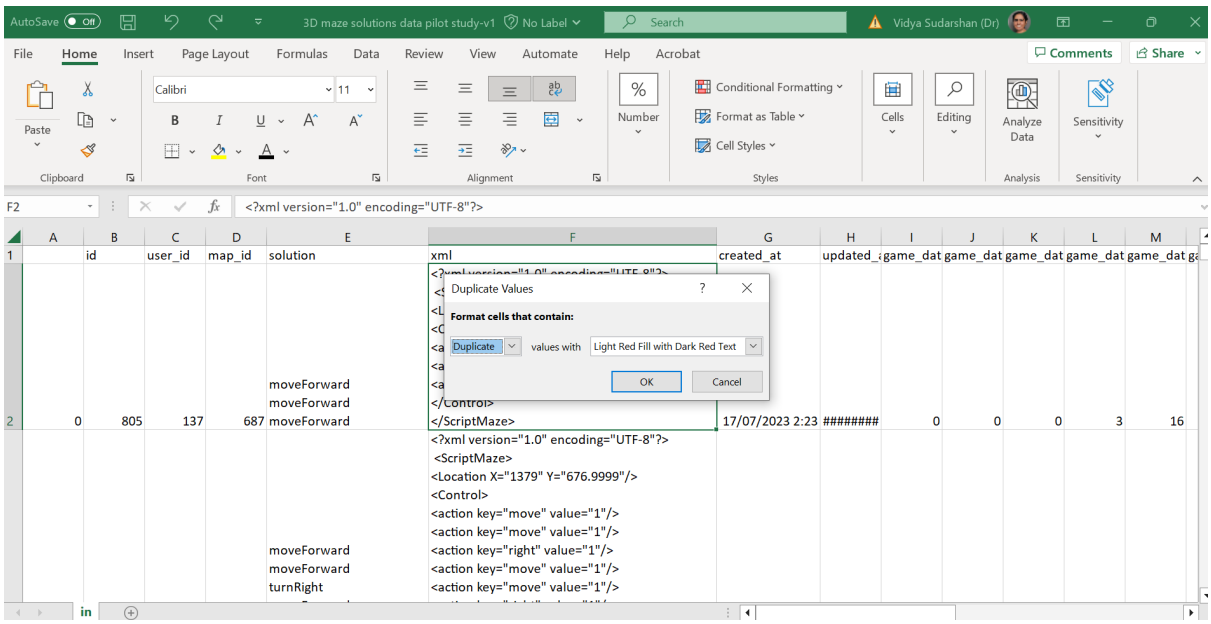
Two things you can do with duplicate data – highlight it or delete it.

<u>Highlight duplicate data:</u>

Select the data and Go to Home → Conditional Formatting → Highlight Cells Rules → Duplicate Values.

Specify the formatting and all the duplicate values get highlighted.



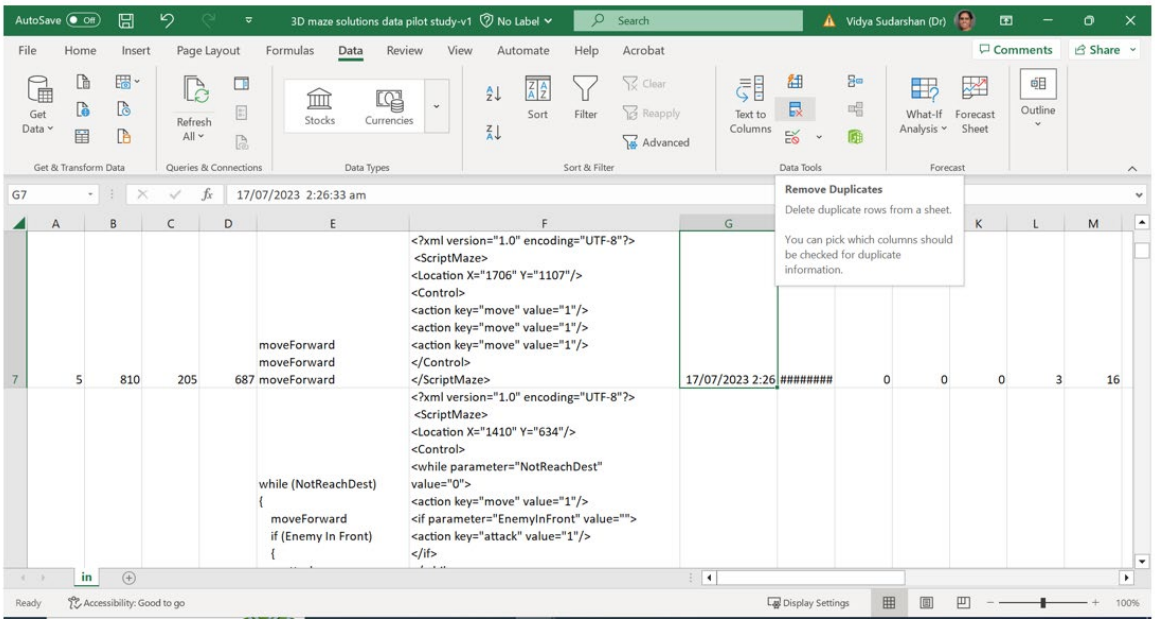


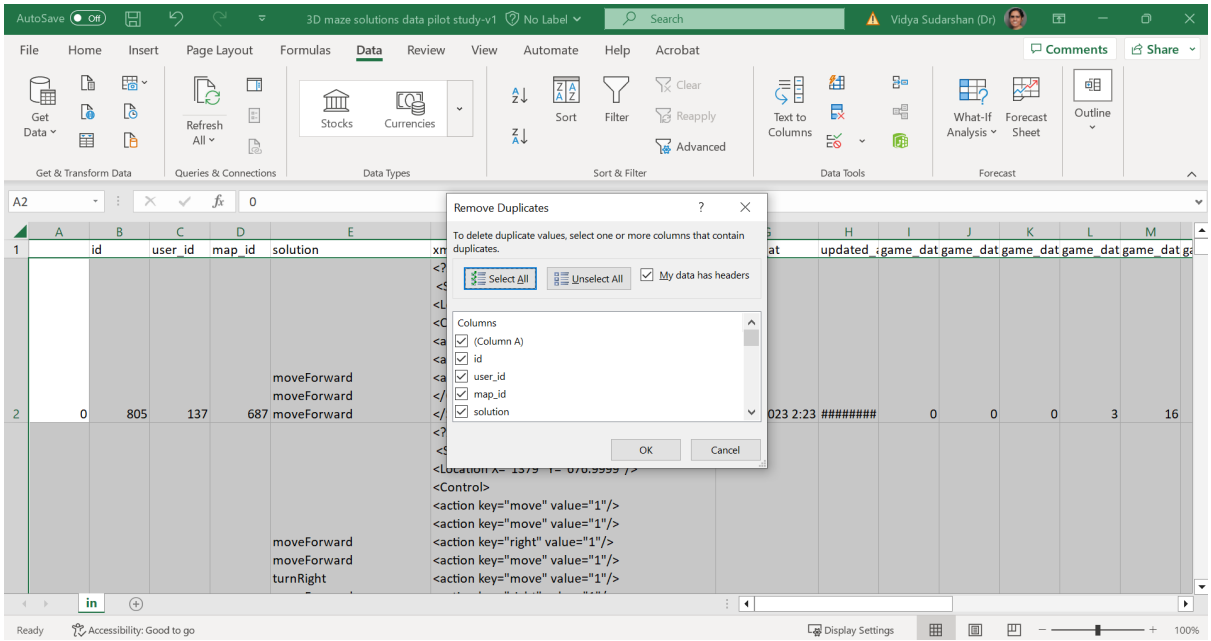
Delete duplicates in data:

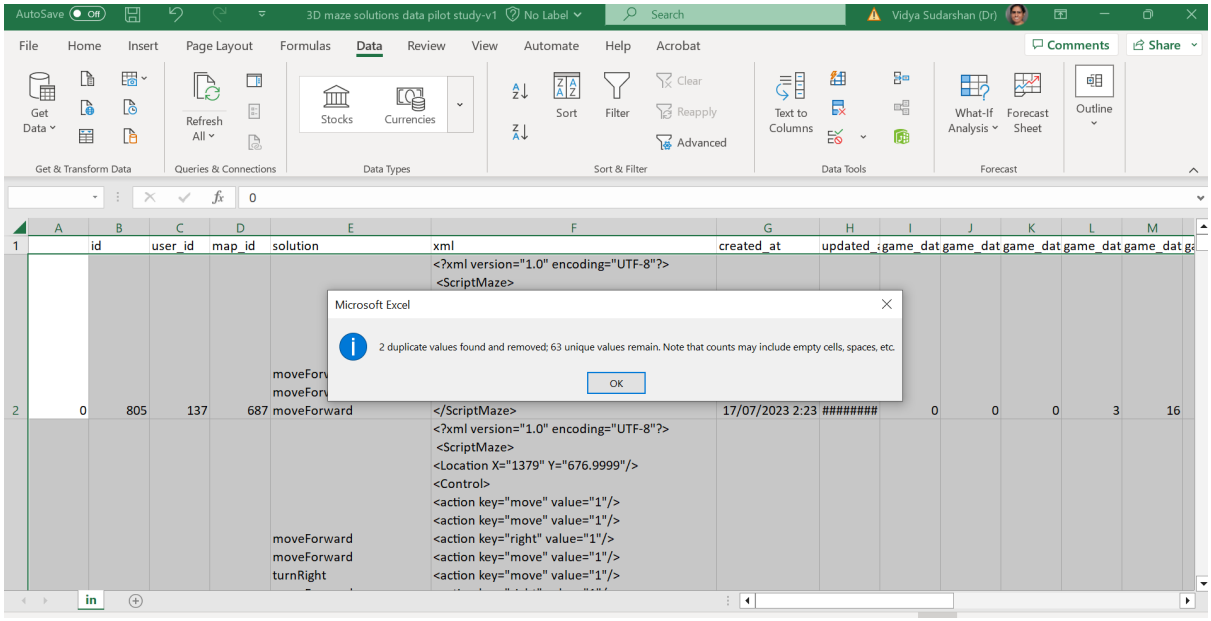
Select the data and Go to Data → Remove Duplicates.

If your data has headers, ensure that the checkbox at the top right is checked.

Select the Column(s) from which you want to remove duplicates and click OK. Usually, it is recommended to select all the Columns – to delete duplicates from entire Rows.



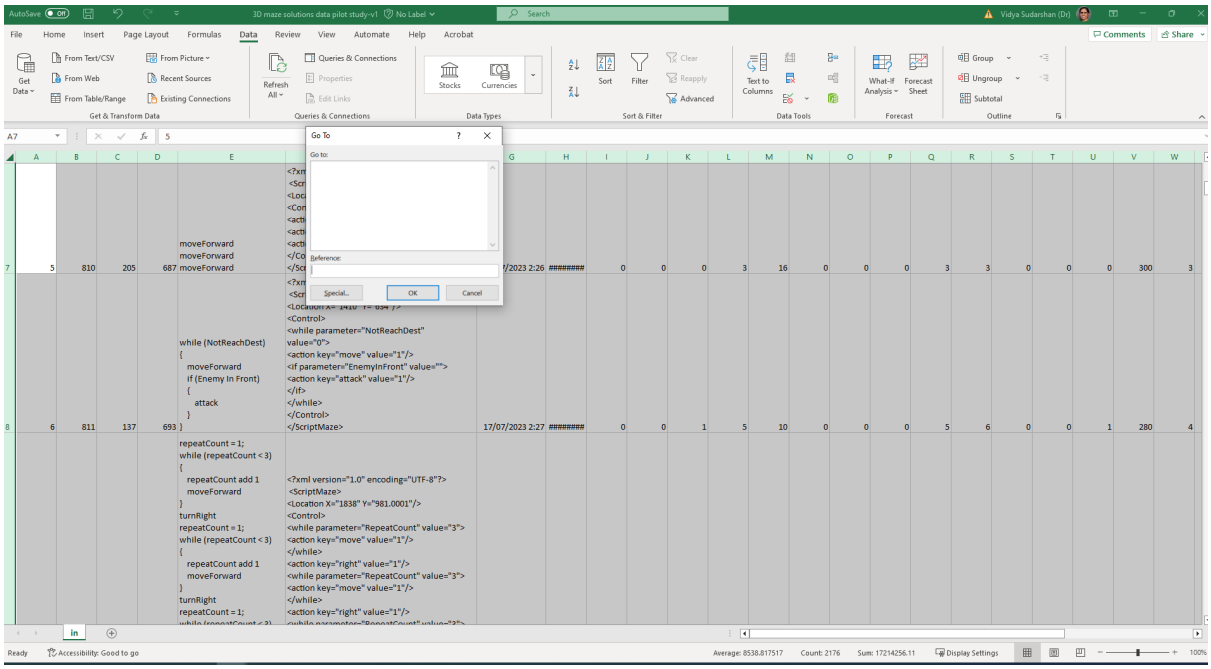




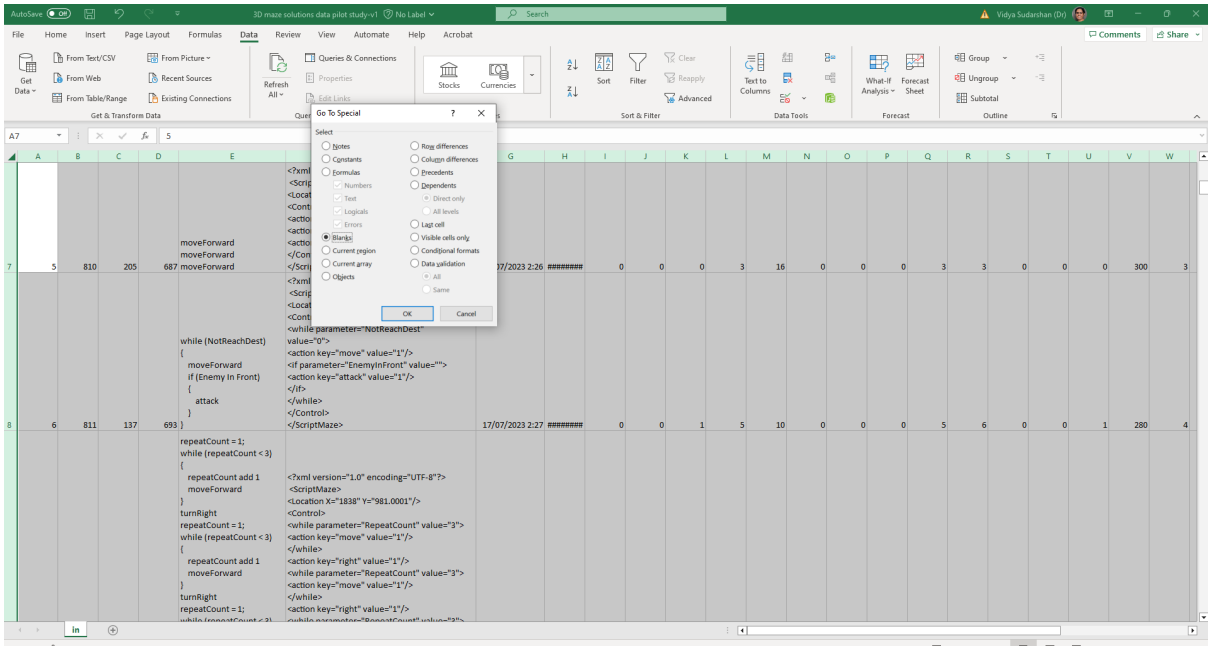
2. Select and Treat all Blank cells:

To fill all blank cells with '0' or 'Not Available', or simply want to highlight it.

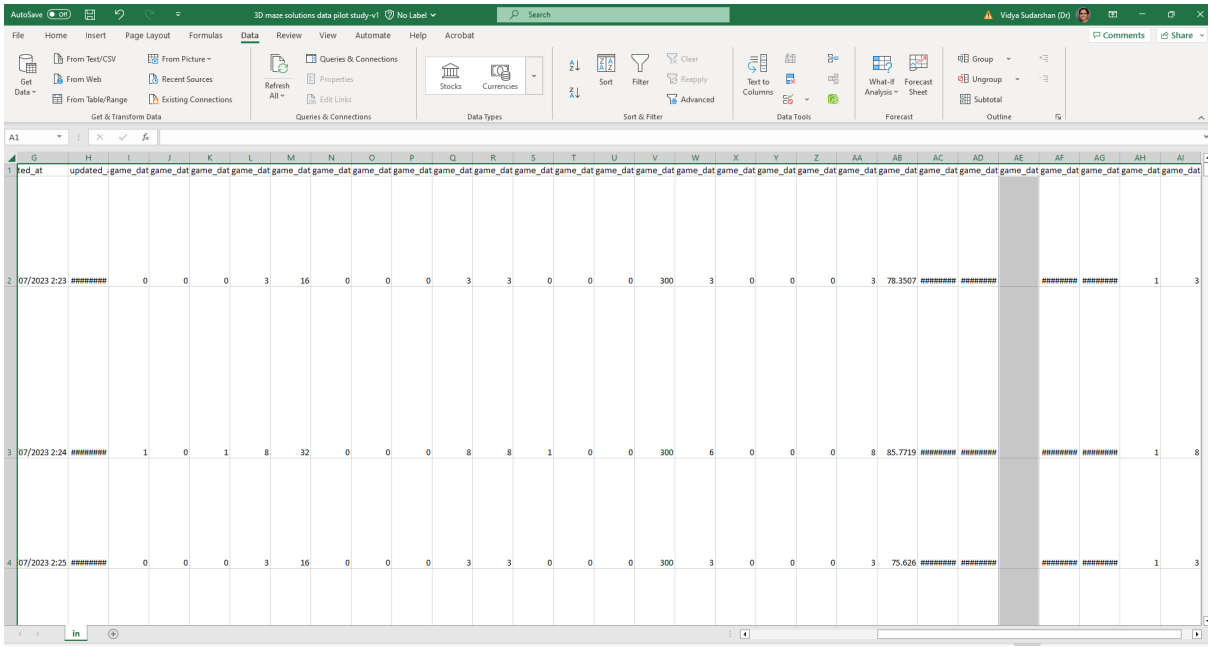
a. Select the entire data set.
b. Press F5 (this opens the Go To dialogue box).
c. Click on the Special….button (at the bottom left). This opens the Go To Special dialog box.



d. Select Blank and Click OK



This selects all the blank cells in your data set.



If you want to enter 0 or Not Available in all these cells, just type it and press Control + Enter.

Or if you want to delete every row in which a blank cell was selected. To do so, hold down your Ctrl key and press the "-" minus key on your keyboard. Excel displays the

Delete dialog. Select the Entire Row option and then choose OK. After you do so, Excel deletes the empty rows.

3. Hide irrelevant data:

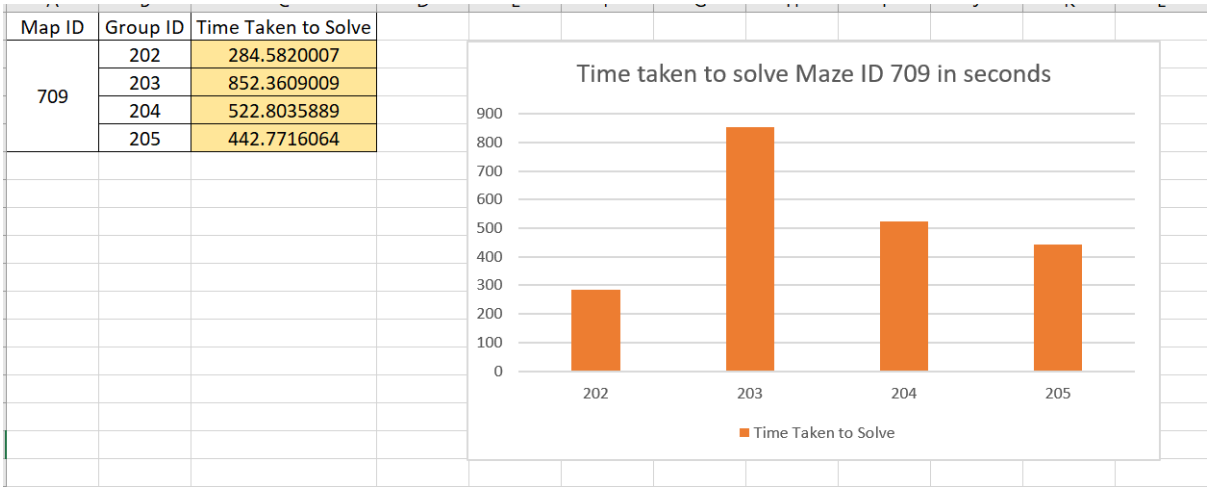
   To hide unwanted columns,

   a. Select the column or any cell within the column (s) you want to hide.
   b. On the Home tab, in the Cells group, click Format → Hide & Unhide → Hide Columns

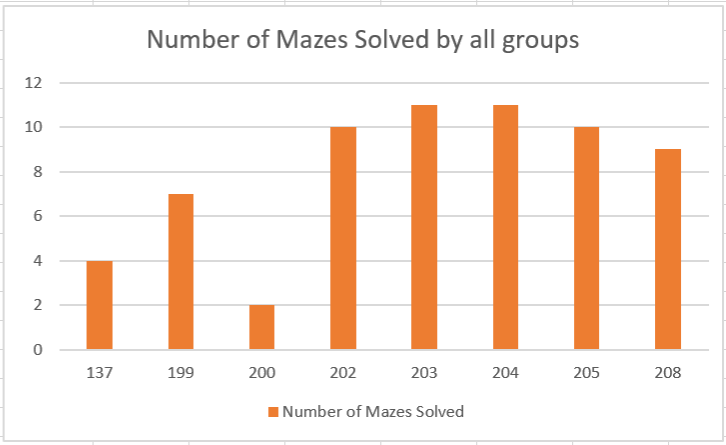
D. Plotting/visualising and presenting the data

Once you have cleaned the data, you can use some data visualization methods to summarize the information to ensure it is meaningful. Visualization helps to understand the data and present it to others in an engaging manner. Charts and graphs can be used to visualize the data you have.

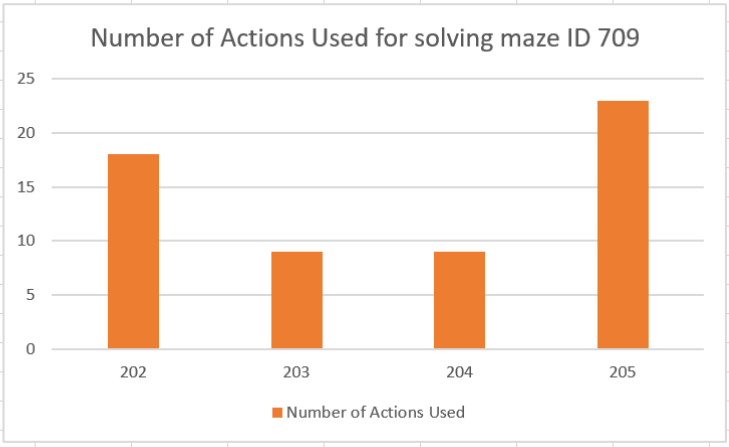
Please refer to the basics of excel document to understand about the plotting of charts or graphs. Sample of bar charts obtained (after data cleaning is done) for our 3D maze data (week 2 activities data) are shown below.



Group ID	Number of Mazes Solved
137	4
199	7
200	2
202	10
203	11
204	11
205	10
208	9



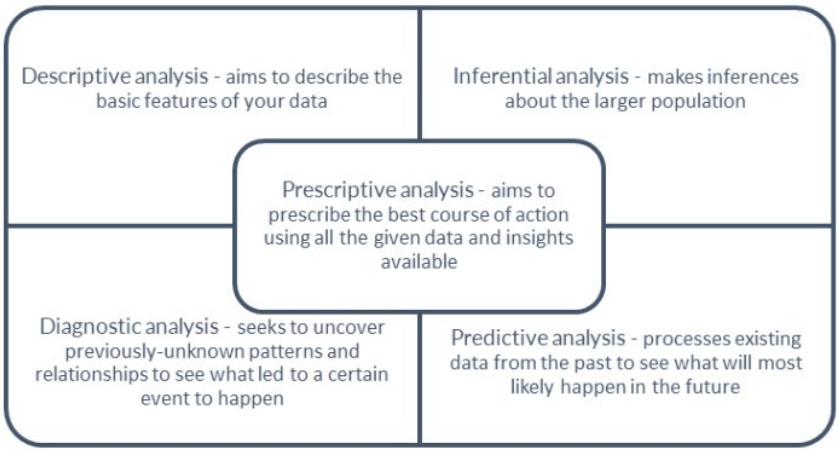
Map ID	Group ID	Number of Actions Used
	202	18
	203	9
	204	9
709	205	23



Please explore other types of graphs in Excel (during your tutorial activity) understand more on this concept.

E. Analysing and sharing/presenting the results

To analyse the data, selection and applying relevant data analysis method is important. What analysis method to use – depends on the problem at hand.



The figure above shows the different data analysis methods available; depending on the problem you are solving you might end up using all the five methods mentioned.

The basic analysis method required for you all at this stage is – Descriptive analysis – which helps to convert quantitative data into information that are easy to understand. In simple words, descriptive analysis describes the basic features of your data by generating simple summaries that show what happened within your study.

Some of the approaches or methods used to describe our data includes: (i) measures of central tendency such as mean/average, median, and mode, (ii) measures of variability such as range, variance, and standard deviation. It also involves graphical representation of data to aid visualization and understanding. These approaches provide a useful starting point for further analysing the data. Even at this stage, visualization can be used (charts and graphs) to present the results of data analysis.

Please complete this reading to understand the step-by-step basics of data analysis before working on your module 2 tutorial activities.