Module 2: Exploratory Data Analysis on Assessment Records using Excel

By
Associate Prof. Ts. Dr. Nurfadhlina Mohd Sharef
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
nurfadhlina@upm.edu.my

Learning Outcomes and Outline

Learning Outcomes

At the end of this module, participants will be able to:

1. Explain functions in Microsoft Excel for data analysis purpose
2. Use data exploration and visualization functions in Microsoft Excel functions to identify patterns and relationship in data

Outline

The outline of this presentation is as follows:

1. Background
2. Data Analysis Toolpak
3. Questioning techniques
4. Descriptive summary
5. Conditional formatting
6. Charting
7. Recommended charts
8. Analyze Data function
9. Quick Analysis function
10. Histogram
11. COUNTIF
12. Deducing Information

Background

Exploratory Data Analysis (EDA) refers to the critical process of performing initial investigations on data so as to discover patterns,to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

EDA is an iterative process comprising the following activities:

1. Generate questions about your data.
2. Identify patterns on the data to deduce information.
3. Search for answers by visualising, transforming, and modelling your data.
4. Use what you learn to refine your questions and/or generate new questions.

Various techniques can be used for EDA such as data preprocessing, descriptive statistics and visualization.

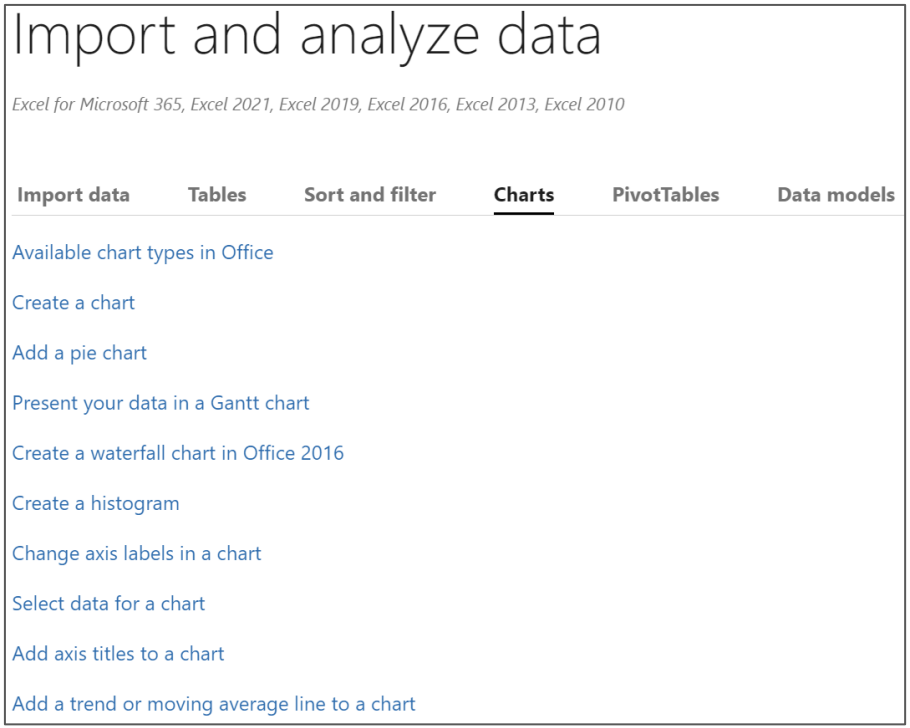
Tools such as Excel and PowerBI can be used to create dashboard visualizations. Programming languages such as Python and Java can be used to manipulate the data and spot trends.

Using Microsoft Excel for Data Analysis

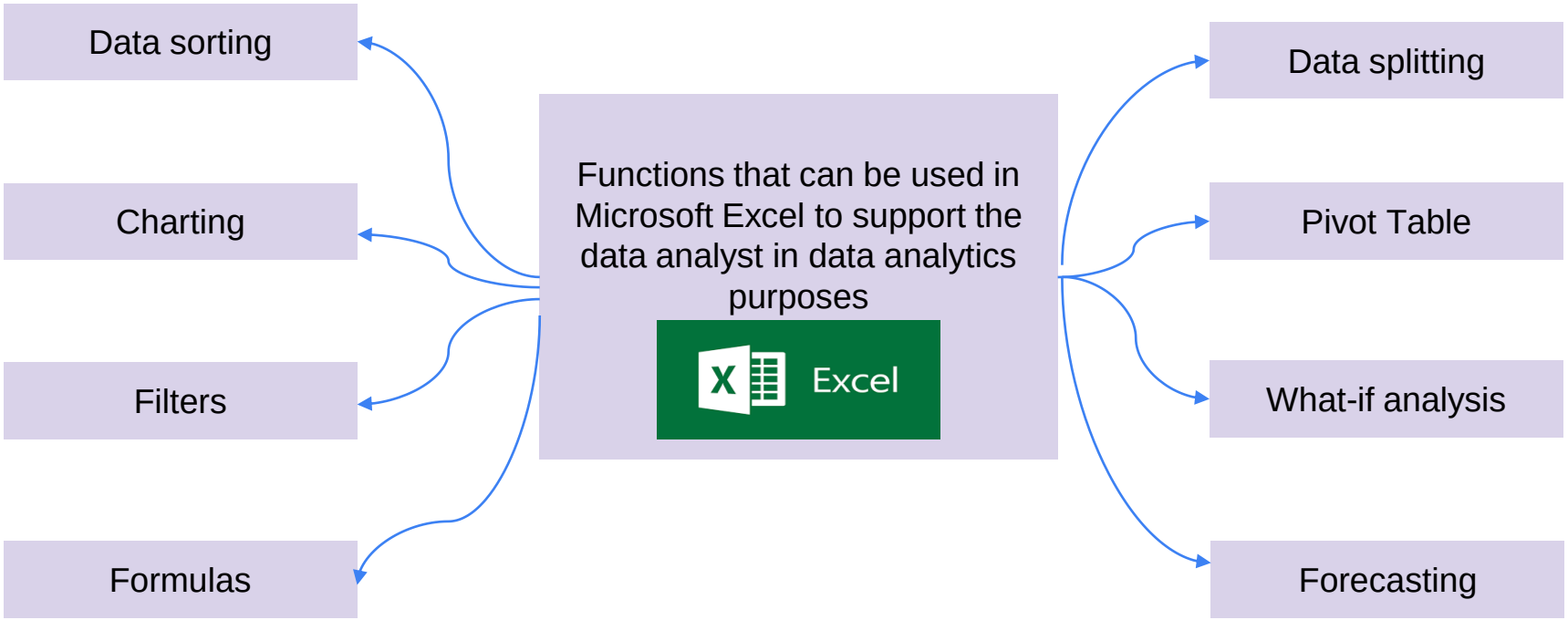
Microsoft Excel is one of, if not the number 1 data analysis software in the world in terms of popularity mainly because its ease of use. It provides the ability to tabulate, analyze and visualize data to assist in data interpretation.

Chart, Formula, Analyze Data function, PivotTable, PivotCharts and Data Analysis Toolpak could support you to identify trends, patterns, and outliers in a data set, facilitating data analysis in seconds and empowering you to understand data through high-level summaries via Excel's Artificial Intelligence ability.

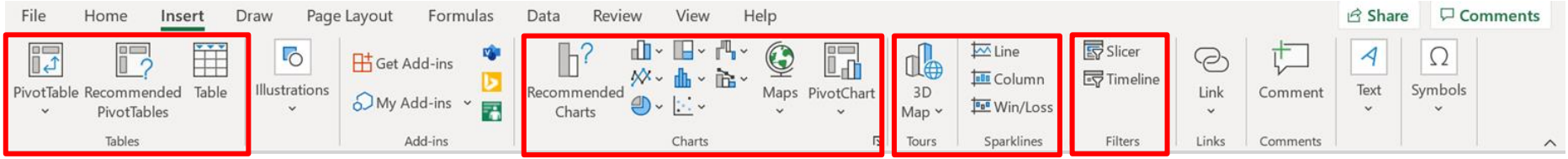
To visit tutorials in Excel, click the image below.



Using Microsoft Excel for Data Analysis (1)



Using Microsoft Excel for Data Analysis (2)



A pivot table is a **table of grouped values that aggregates the individual items** of a more extensive table within one or more discrete categories.

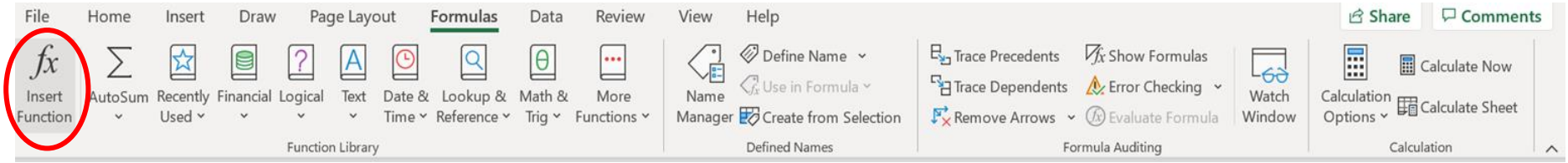
This summary might include sums, averages, or other statistics, which the pivot table groups together using a chosen aggregation function applied to the grouped values.

Various charts can be created to **plot the trends** in the data.

A sparkline is **a tiny chart in a worksheet cell that provides a visual representation of data**. Use sparklines to show trends in a series of values, such as seasonal increases or decreases, economic cycles, or to highlight maximum and minimum values.

Creating filters in the data allows analyst to **focus on specific characteristics**.

Using Microsoft Excel for Data Analysis (3)



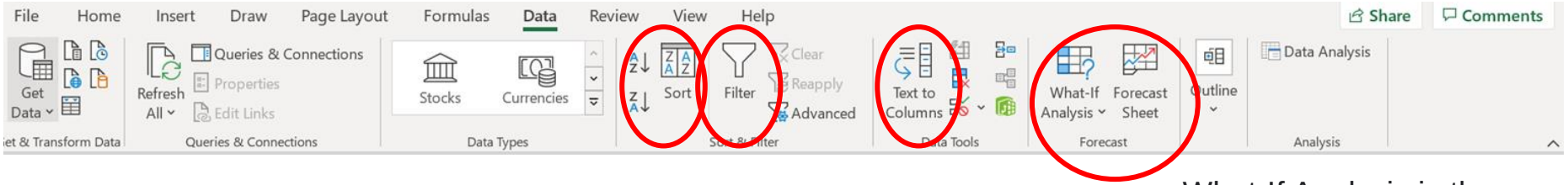
Formulas are **equations that can perform calculations, return information, manipulate the contents of other cells, test conditions, and more.**

These **formulas** return a result, even when it is an error.

Among the formulas that are typically used by data analysts are:

1. SUM: to calculate total
2. COUNT: to identify frequency
3. AVERAGE: to get average value
4. MIN: to identify minimum value
5. MAX: to identify maximum value
6. LEN: to count the number of characters
7. IF: to check a condition
8. VLOOKUP: to refer a value in a vertical list
9. INDEX-MATCH: to lookup value dynamically
10. COUNTIF: to count how many matches the condition

Using Microsoft Excel for Data Analysis (4)



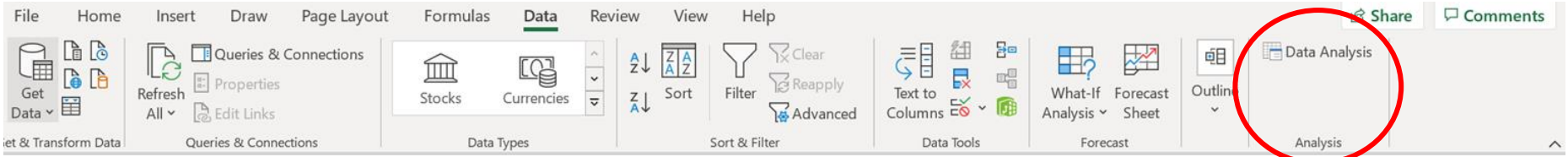
The sorting function allows **data to be arranged according to increasing or decreasing order**

The filter function is to **select some data** based on certain characteristics

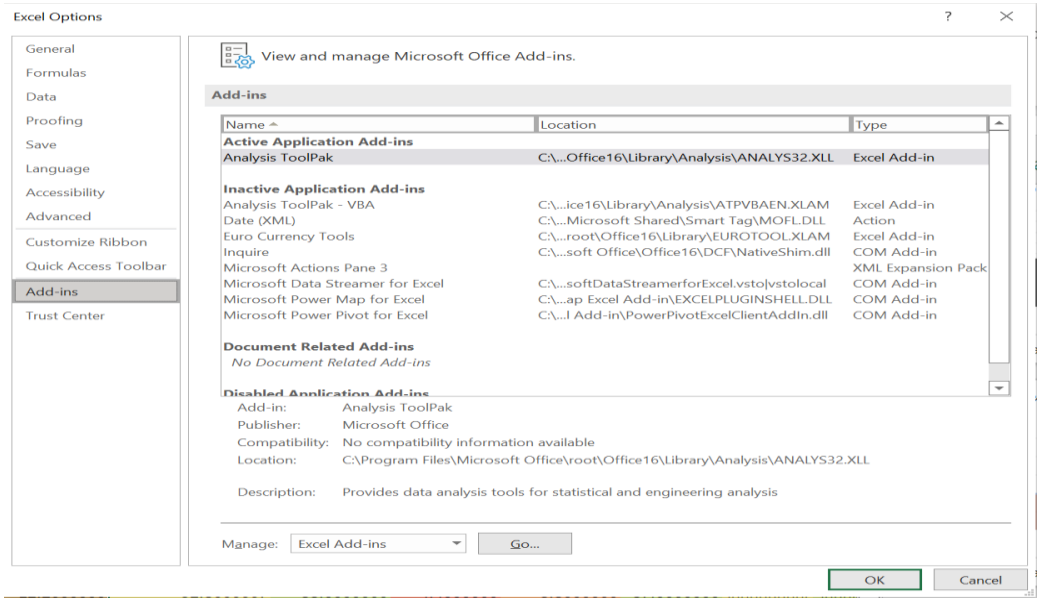
The Text to Columns function **splits data according to occurrence** of certain characters

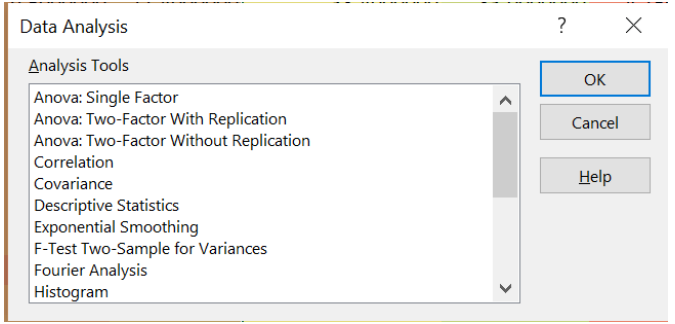
What-If Analysis is the process of **changing the values in cells to see how those changes will affect the outcome of formulas** on the worksheet. Three kinds of What-If Analysis tools come with Excel: Scenarios, Goal Seek, and Data Tables. Scenarios and Data tables take sets of input values and determine possible results.

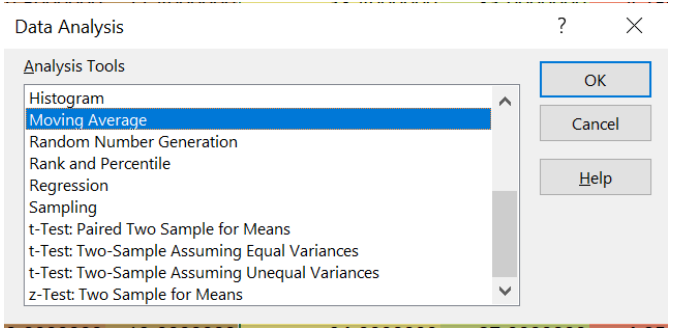
Using Microsoft Excel for Data Analysis (5)



The Data Analysis Toolpak may be added to support for deeper data analysis exercises. To activate this, go to Home>Options>Add-ins.







Activity Time!

Now that you are familiar with the functions that can be used in Microsoft Excel for data analysis, let's do some activities using Microsoft Excel!

12 activities have been prepared for you, as basic learning analytics task. The remaining of this slide presentation contains step-by-step activities based on an example assessment records.

Example data 1: Assessment Record

	A	B	C	D
1	StudentID	Test1-C3 (20%)	Asgmt1(10%)	Asgmt2(15%)
2	S1	10.00	6.17	13.80
3	S10	14.50	8.00	13.80
4	S11	14.50	9.50	13.20
5	S12	18.00	9.50	13.80
6	S13	10.00	9.50	15.00
7	S14	12.00	9.50	13.60
8	S15	17.50	9.50	14.20
9	S16	8.50	9.50	15.00
10	S17	13.50	9.50	11.40
11	S18	13.00	9.50	13.80
12	S19	13.50	9.50	12.00
13	S2	11.00	6.17	0.00
14	S20	4.00	9.50	13.20
15	S21	14.50	9.50	12.00
16	S22	4.50	9.50	13.80
17	S23	5.00	9.50	13.80
18	S24	9.00	9.50	14.00

Assessment record is one of the most common forms of data used for learning analytics because academic institutions usually keep the scores information. The scores are also used as the key performance indicator, not just for the learner, but also the instructor and the institution.

Therefore, typically some analytics mechanism is adopted such as using Excel template, developing a dedicated system or subscribing to a service.

The figure on the left shows a snippet of recorded scores by a batch of anonymised students in the PutraMOOC data analytics course. The columns are the assessment types and weight, while the rows are the scores for each student.

Download the **raw and processed assessment records using the activities in this file at** http://putraoer.upm.edu.my/id/eprint/37

Activity 1: Questioning techniques

Create questions as a driver of your exploration. Identify possible visualization and insights to be gained. Examples are as shown below.

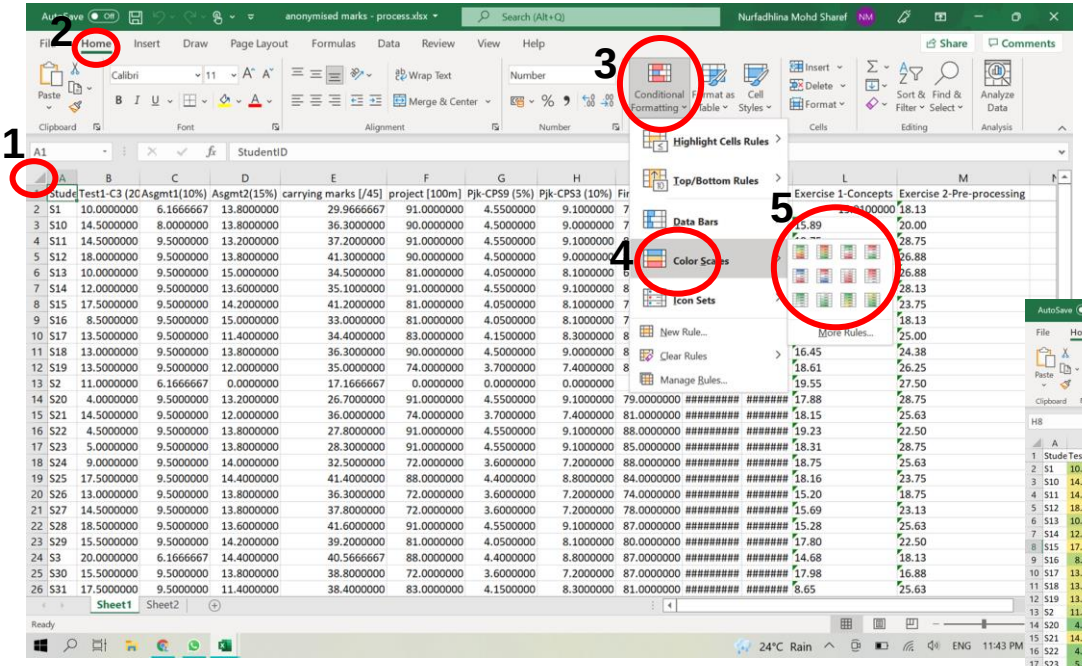
Questions	Metrics	Visualization
What is the average, maximum, minimum, variance and stdev marks in each assessment?	Average, Max, Min, var, Stdev	Table
How many students have scored more than the average score in each assessment?	Frequency	Bar chart
How many students obtained more than 65% of each assessment's maximum score?	Frequency, Percentage	Bar chart

Questions	Metrics	Visualization
Is the performance of each student across assessment consistent?	Percentage	Table
Which assessment has been easy or hard to be scored?	Percentage	Pie chart
What is the student's strength in terms of cognitive, psychomotor and affective skills?	Average	Heatmap

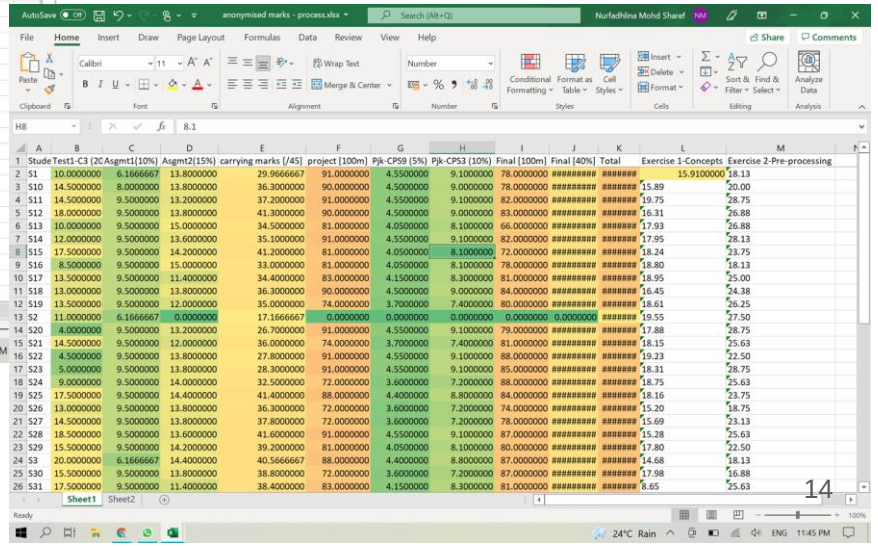
Activity 2: Descriptive summary

1. Using the marks file, provide the descriptive summary: average, maximum (max), minimum (min), standard deviation (stdev) and variance (var). These values represent the distribution of the data.
    a. Open the file
    b. Go to Sheet 1
    c. To get the average of values in column B, go to column B, row 45 and enter =AVERAGE(B2:B44)
    d. To get the max among the values in column B, go to column B, row 45 and enter =MAX(B2:B44)
    e. To get the min among the values in column B, go to column B, row 45 and enter =MIN(B2:B44)
    f. To get the var among the values in column B, go to column B, row 45 and enter =VAR(B2:B44)
    g. To get the stdev among the values in column B, go to column B, row 45 and enter =STDEV(B2:B44)

Activity 3: Conditional formatting

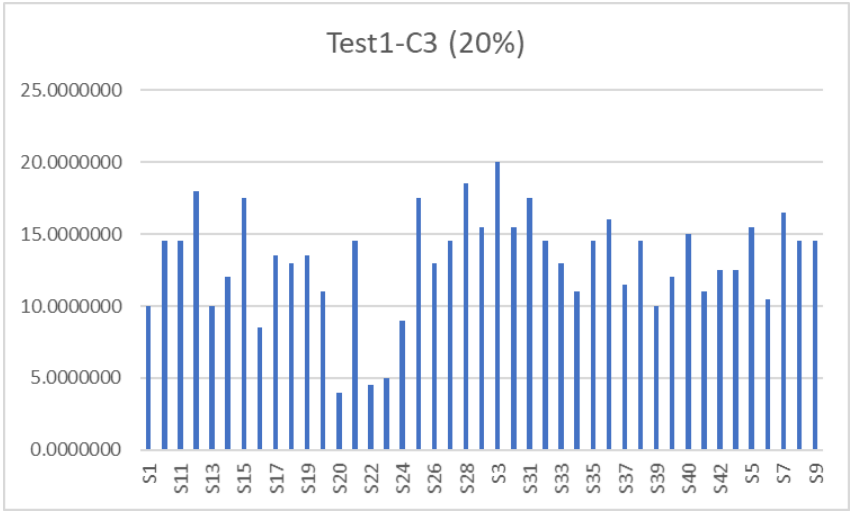


Color scales allows us to see data representation easier visually, using traffic-light color coding



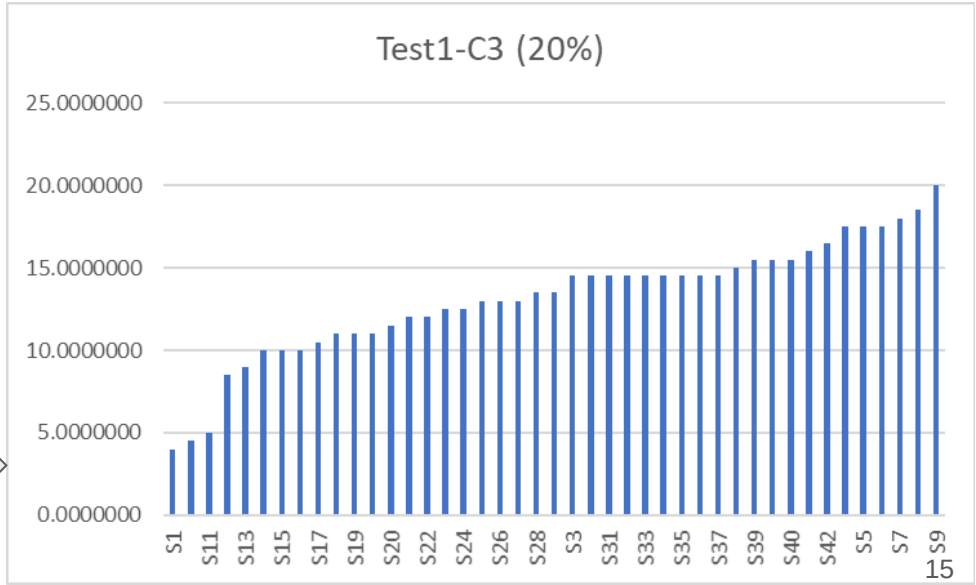
Follow step 1 until 5 in the above figure. Click the corner between column A and row 1. And then, use the Color Scales option in the Conditional Formatting function.

Activity 4: Charting non-sorted and sorted values

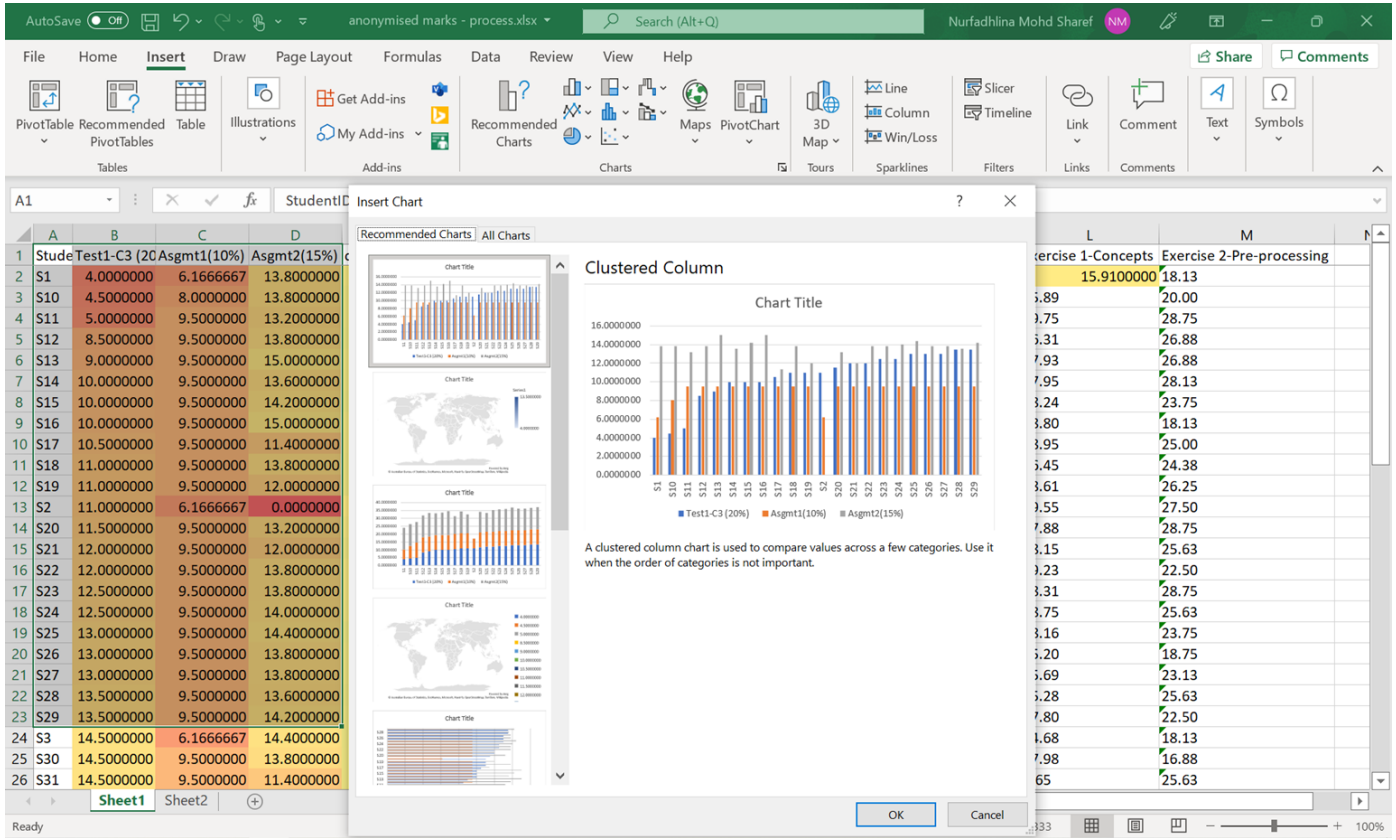


We can plot the chart but to read it requires some time, e.g, to find the student with highest mark

Easier reading because the data is sorted. We can also easily see the gap between the minimum and maximum, and identify the median value.

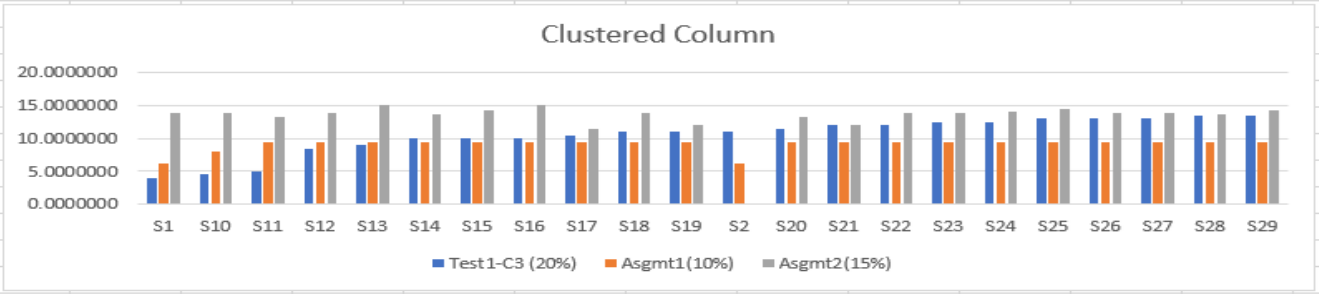


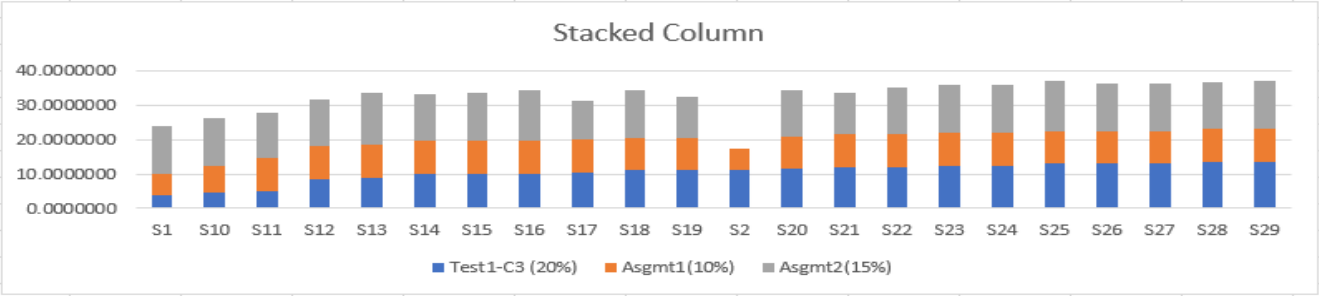
Activity 5: Using Recommended Charts function

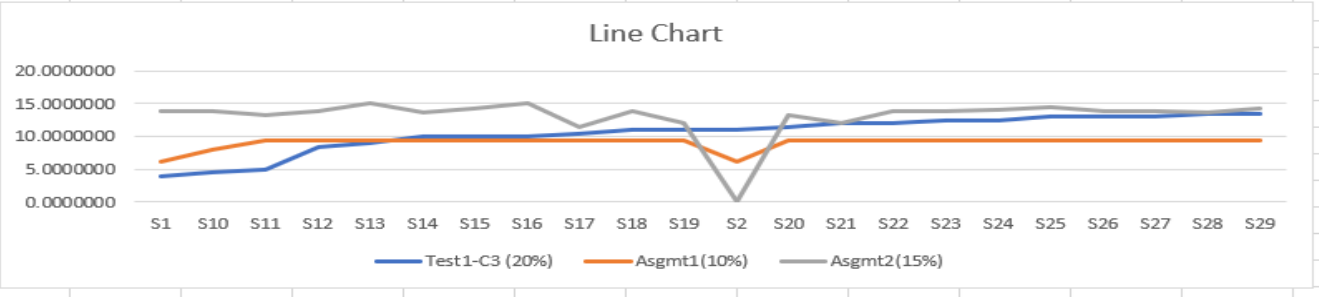


Use the Recommended Charts function for some automatically generatable data plotting.

Activity 6: Using Recommended Charts function







Visualization choice matters!

Explore possible visualizations to identify the one that can best suit your need.

The clustered column shows the score in each assessment clearly, but to get information about the sum of some marks, stacked column is the best.

Meanwhile, the Line Chart makes it easy to compare each student's ranking compared to others in each assessment type.

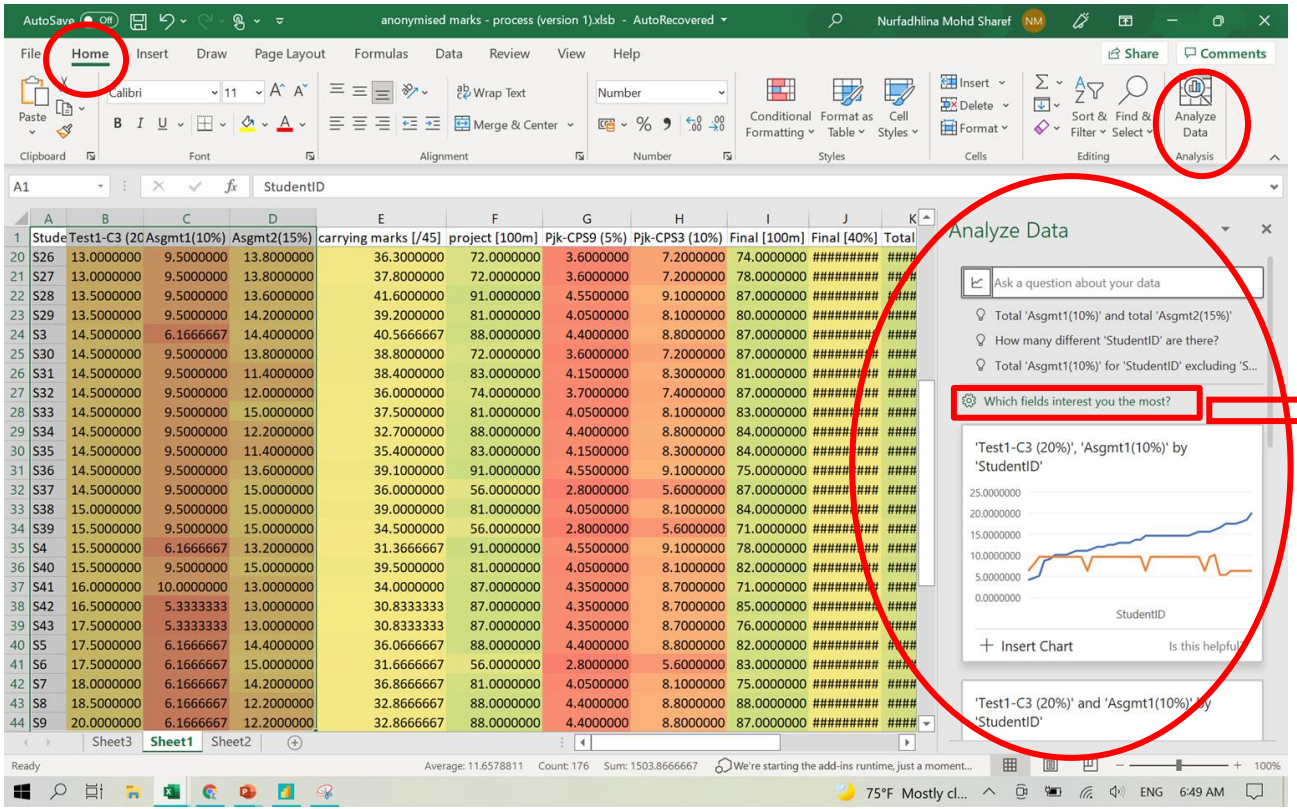
We can also easily see which student has the highest, and the lowest mark.

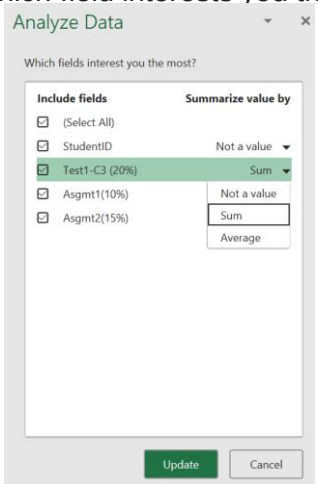
Activity 7: Using Analyze Data function

Simply select a cell in a data range, then select the Analyze Data button on the Home tab.

Analyze Data in Excel will analyze your data, and return interesting visuals about it in a task pane.

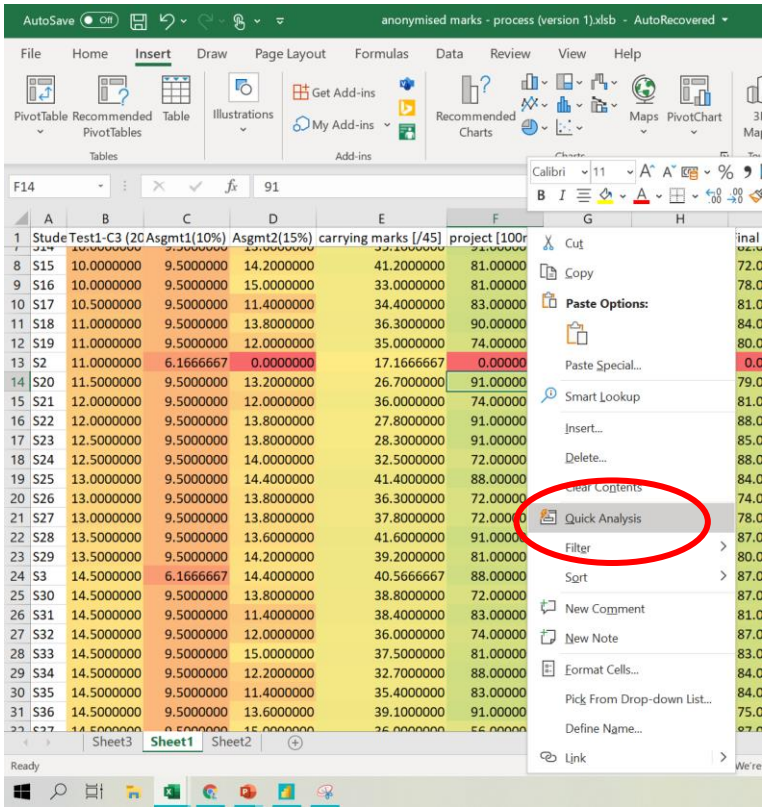
Choose any suggested ideas for data analysis listed or insert any readily made chart. You may also customize it using which field interests you the most.

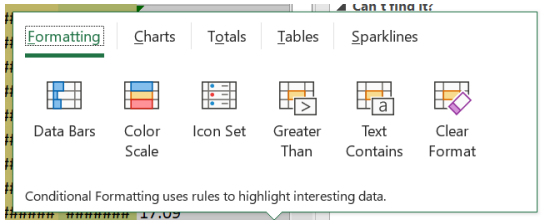


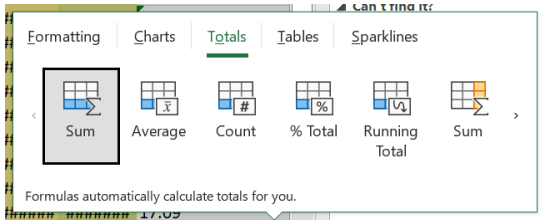


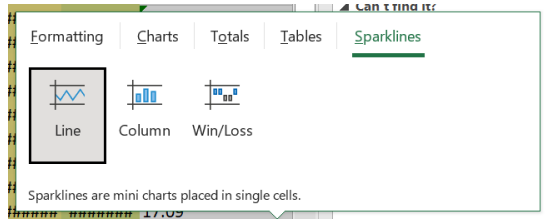
Activity 8: Using Quick Analysis function

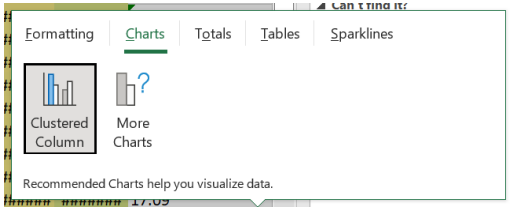
Quick Analysis function takes a range of data and helps you pick the perfect chart with just a few commands. Select a range of cells. Select the Quick Analysis button that appears at the bottom right corner of the selected data. Or, press **Ctrl + Q**.

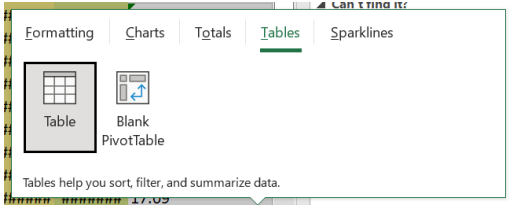












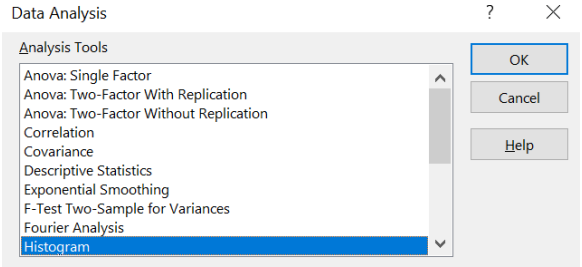
Activity 9: Using Histogram to count frequency of marks of Test1 by a uniformed scale

Histogram is a visualization technique for supporting frequency analysis.

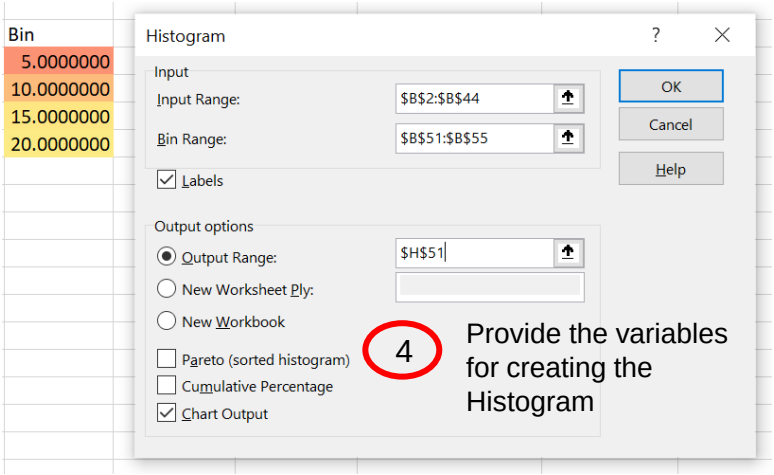
Add the Data Analysis Toolpak to perform Histogram analysis.

2 Go to Data>Data Analysis

3 Choose Histogram



1 Prepare the scale for the bin



4 Provide the variables for creating the Histogram

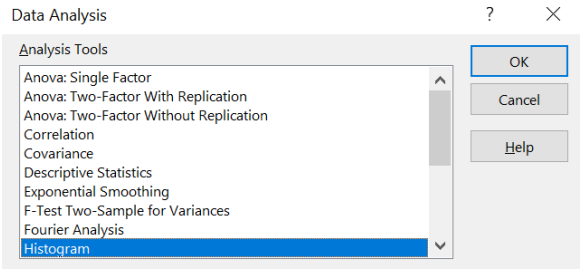
5 The table of answers will be displayed

Bin	Frequency
5.0000000	3
10.0000000	4
15.0000000	24
20.0000000	11
More	0

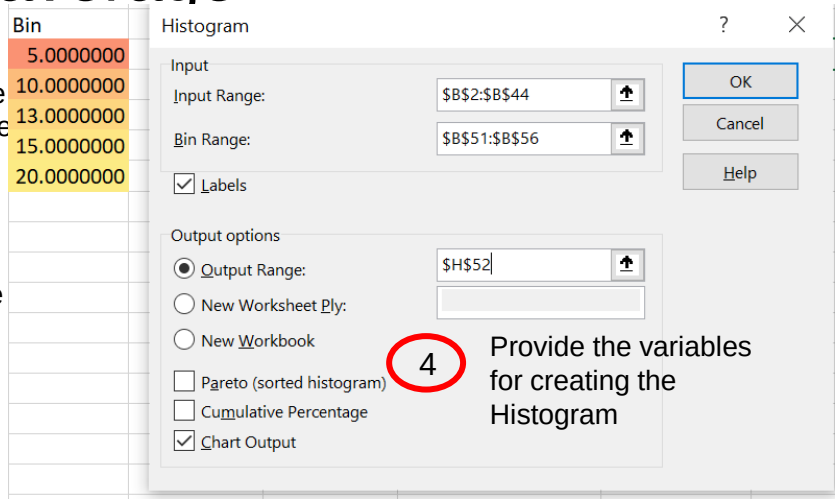
Activity 10: Using Histogram to find frequency of Test1 marks above average

2 Go to Data>Data Analysis

3 Choose Histogram



1 Prepare the scale for the bin by placing the average as one of the bin variable



4 Provide the variables for creating the Histogram

5 The table of answers will be displayed. We can identify the number of students who obtained marks above average (average=13)

Bin	*Frequency*
5.0000000	3
10.0000000	4
13.0000000	12
15.0000000	12
20.0000000	11
More	0

Activity 11: Using COUNTIF to find frequency of Final marks above average

1. The average value of final marks is **79.19, which indicates that many student almost got grade=A (marks for grade=A is 80).**
   We can use this information to identify which students have gotten final marks below the average.
1. Enter this formula in cell N2:
   =IF(K2>(AVERAGE(K2:K46)), TRUE, FALSE)
1. Then, copy this formula and paste into all cells in column K
2. Next, create a cell for counting the frequency of final marks that is larger than the average value of final marks. Enter this formula:
   =COUNTIF(N2:N44,TRUE)
   The result is 35, which indicates that 35 out of 42 students have gotten final marks above 35.
   The instructor can use this **insights** to identify students who have gotten the high and low marks (can use average marks as a benchmark). The instructor can reflect the student's participation, and check whether there has been any comments by the students earlier on their learning problems. The instructor can also perform several additional analytics (using data analysis technique and visuals) to identify the gap among the students, and the trend of score by each student. These information can be used to strategize the next lessons and improve the curricula.

Activity 12: Deducing information

The figures on the right show the formula and the result of the descriptive summary. Based on this data, a few information may be deduced:

a) In terms of average marks, Asgmt2>Test1>Asgmt1
b) In terms of variance and Stdev, Test1>Asgmt2>Asgmt1

However, note that, THIS IS WRONG and we CANNOT compare across the assessments because their weight is different. Test1's weight is 20%, Asgmt1 is 10% and Asgmt2 is 15%.

This is why we need to STANDARDIZE the data so a fair comparison can be made.

In its current state, we can only report separately for each assessment. This limits a deep insight into understanding the student's performance.

Data standardization and normalization need to also be performed when we compare any two data sources. The next module will explain further on how to do this.

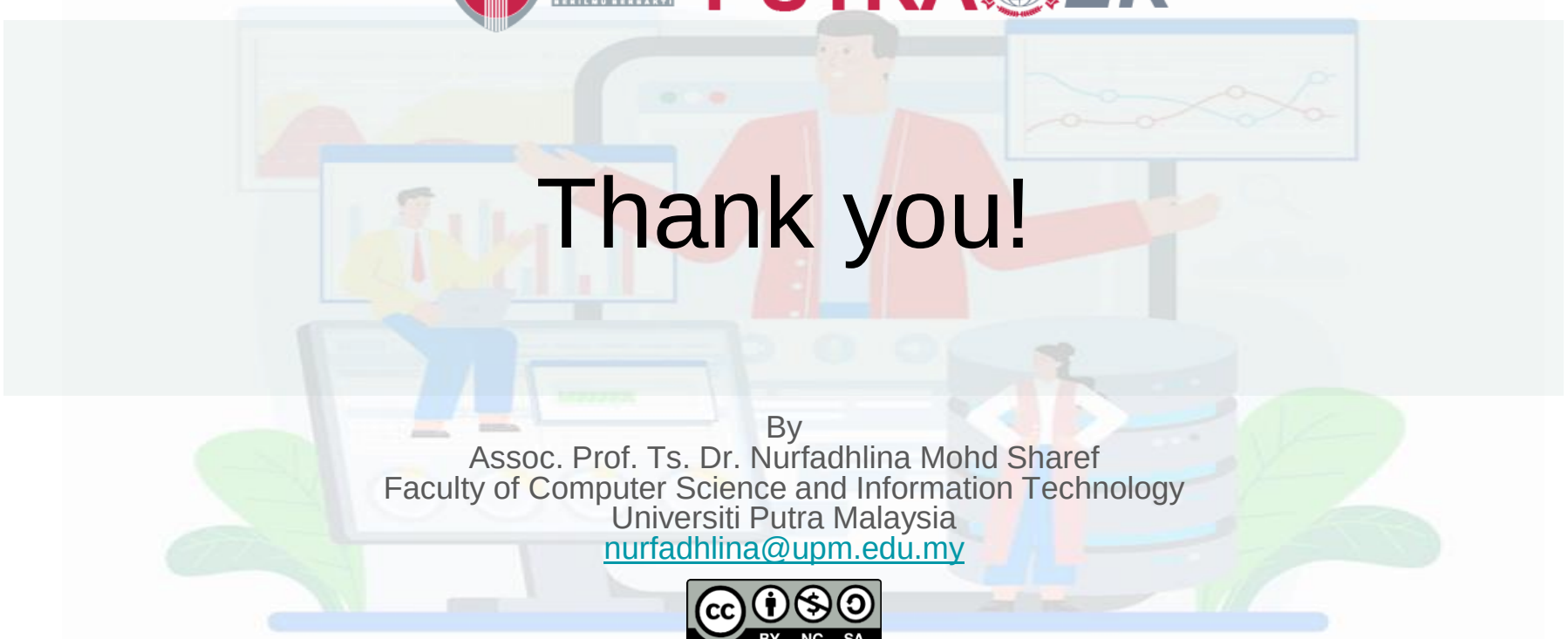
	A	B	C	D	E
1	StudentID	Test1-C3 (20%)	Asgmt1(10%)	Asgmt2(15%)	carrying marks [/45]
35	S4	12	6.16666666666667	13.2	31.3666666666667
36	S40	15	9.5	15	39.5
37	S41	11	10	13	34
38	S42	12.5	5.33333333333333	13	30.8333333333333
39	S43	12.5	5.33333333333333	13	30.8333333333333
40	S5	15.5	6.16666666666667	14.4	36.0666666666667
41	S6	10.5	6.16666666666667	15	31.6666666666667
42	S7	16.5	6.16666666666667	14.2	36.8666666666667
43	S8	14.5	6.16666666666667	12.2	32.8666666666667
44	S9	14.5	6.16666666666667	12.2	32.8666666666667
45	Avg	=AVERAGE(B2:B44)	=AVERAGE(C2:C44)	=AVERAGE(D2:D44)	=AVERAGE(E2:E44)
46	Max	=MAX(B2:B44)	=MAX(C2:C44)	=MAX(D2:D44)	=MAX(E2:E44)
47	Min	=MIN(B2:B44)	=MIN(C2:C44)	=MIN(D2:D44)	=MIN(E2:E44)
48	Var	=VAR(B2:B44)	=VAR(C2:C44)	=VAR(D2:D44)	=VAR(E2:E44)
49	Stdev	=STDEV(B2:B44)	=STDEV(C2:C44)	=STDEV(D2:D44)	=STDEV(E2:E44)

	A	B	C	D	E
1	Stude	Test1-C3 (20	Asgmt1(10	Asgmt2(15%	carrying marks [/45]
35	S4	12.00	6.17	13.20	31.37
36	S40	15.00	9.50	15.00	39.50
37	S41	11.00	10.00	13.00	34.00
38	S42	12.50	5.33	13.00	30.83
39	S43	12.50	5.33	13.00	30.83
40	S5	15.50	6.17	14.40	36.07
41	S6	10.50	6.17	15.00	31.67
42	S7	16.50	6.17	14.20	36.87
43	S8	14.50	6.17	12.20	32.87
44	S9	14.50	6.17	12.20	32.87
45	Avg	13.13	8.59	13.26	34.97
46	Max	20.00	10.00	15.00	41.60
47	Min	4.00	5.33	0.00	17.17
48	Var	12.68	2.41	5.45	21.82
49	Stdev	3.56	1.55	2.33	4.67

Summary

Exploratory data analysis involves

- Understanding your variables
- Using visualization technique to dig into the data
- Cleaning your dataset
- Analyzing relationships between variables



Thank you!

By
Assoc. Prof. Ts. Dr. Nurfadhlina Mohd Sharef
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
nurfadhlina@upm.edu.my