CMP020L012A	Data Analytics	2022/23

Week-5

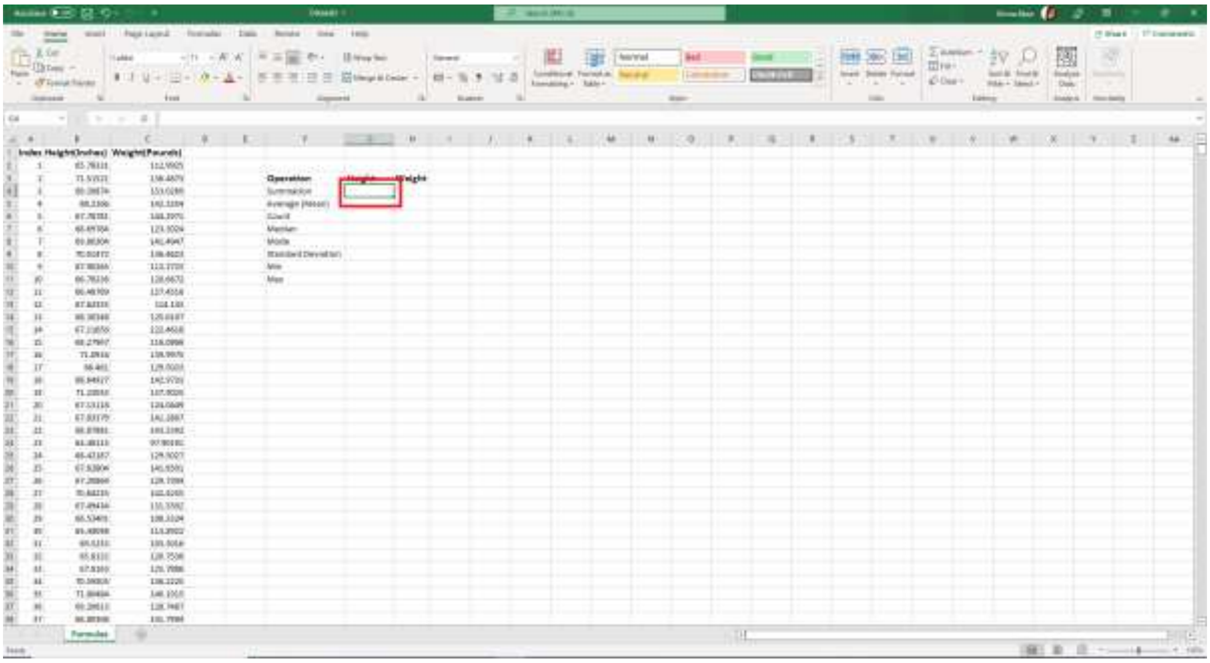
Data Processing and Visualization using Off-the Shelf Tools

Lab Guideline

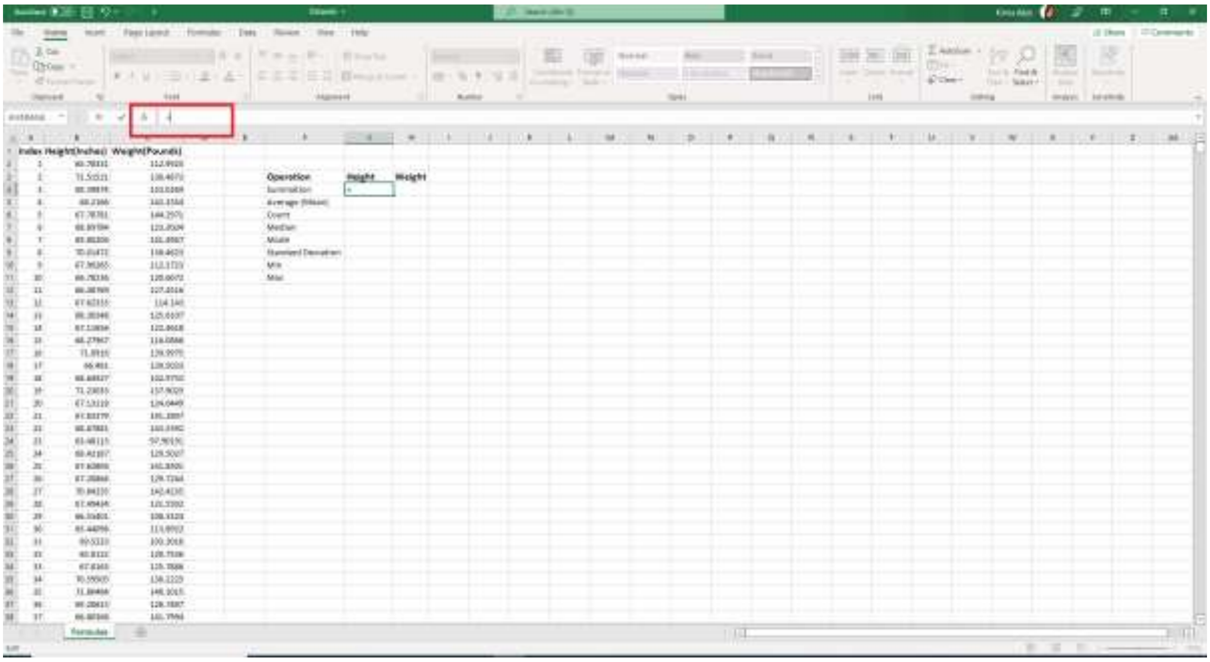
In today’s Lab you will use Microsoft Excel as for performing Data preprocessing operations concerning “Data Cleaning”, “Basic Statistical Analysis” as well as “Visualization”. The dataset(s) you will be using are available in a excel (.csv) file.

Applying Useful Formula and Basic Statistical Analysis:

1. Download the Datasets Excel (.csv) file from Moodle. Open the “Formulae” spread sheet. You will find Height(inches) and weight(pounds) data.
2. At this step you will use readily available formulas available in MS Excel for performing some basic statistical operations.
3. To do so, following are the steps that you would continue:
   a) Assume that, you want to calculate the summation of “Height” data available to you. Before you start the operation, you must select a cell where your desired “Summation” output will be held/stored. After clicking/selecting the cell you type equal “=” in the formula box
   b) Then your formula operation mode is activated.

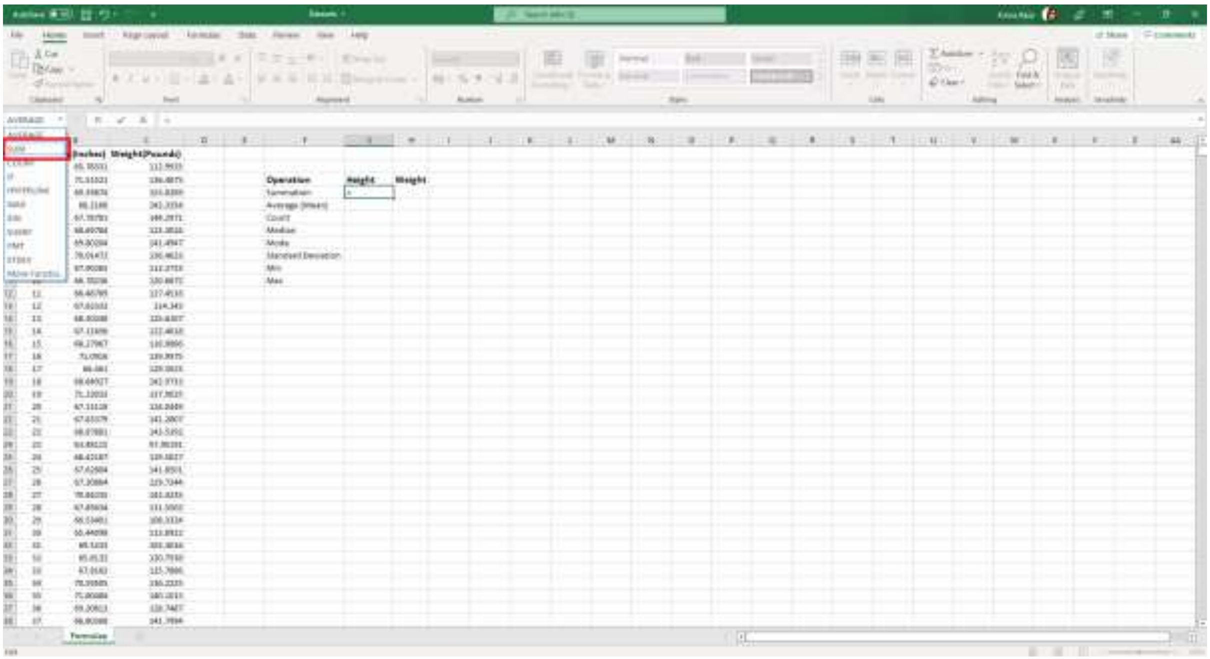


Operation	Height	Weight
Summation		
Average (Mean)		
Count		
Median		
Mode		
Standard Deviation		
Min		
Max		



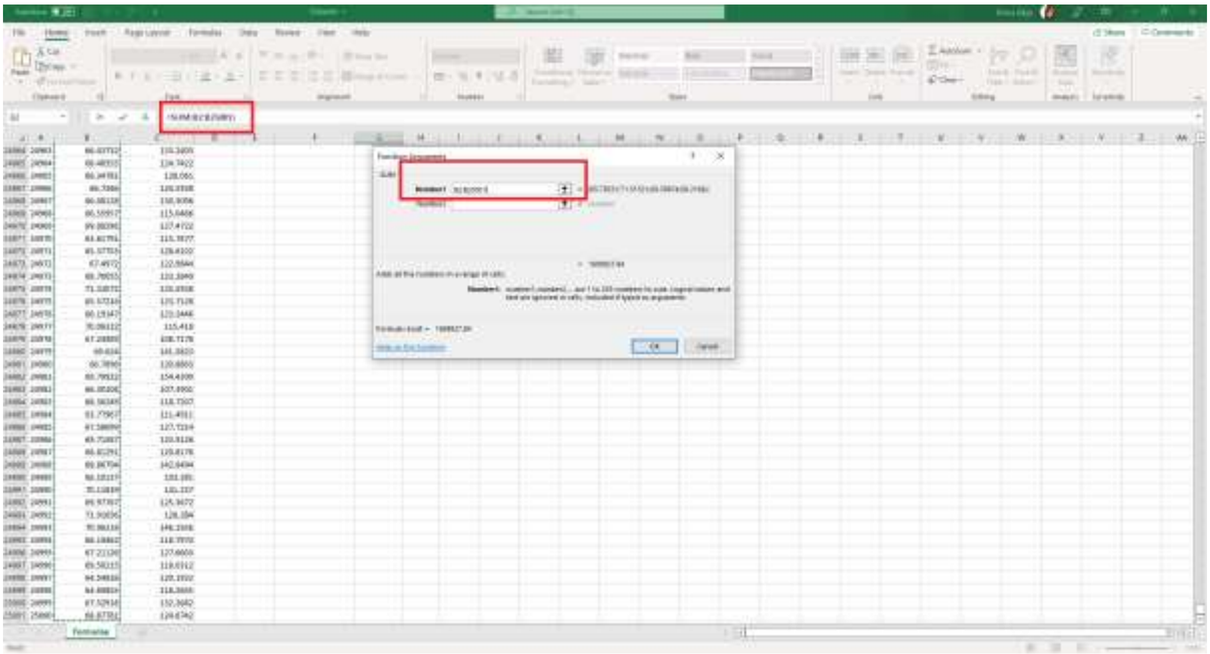
Operation	Height	Weight
Summation	=	
Average (Mean)		
Count		
Median		
Mode		
Standard Deviation		
Min		
Max		

c) Keeping everything consistent till the previous step, click on the top left corner. You will find a formula list available. From there, you will now select "summation" operation.



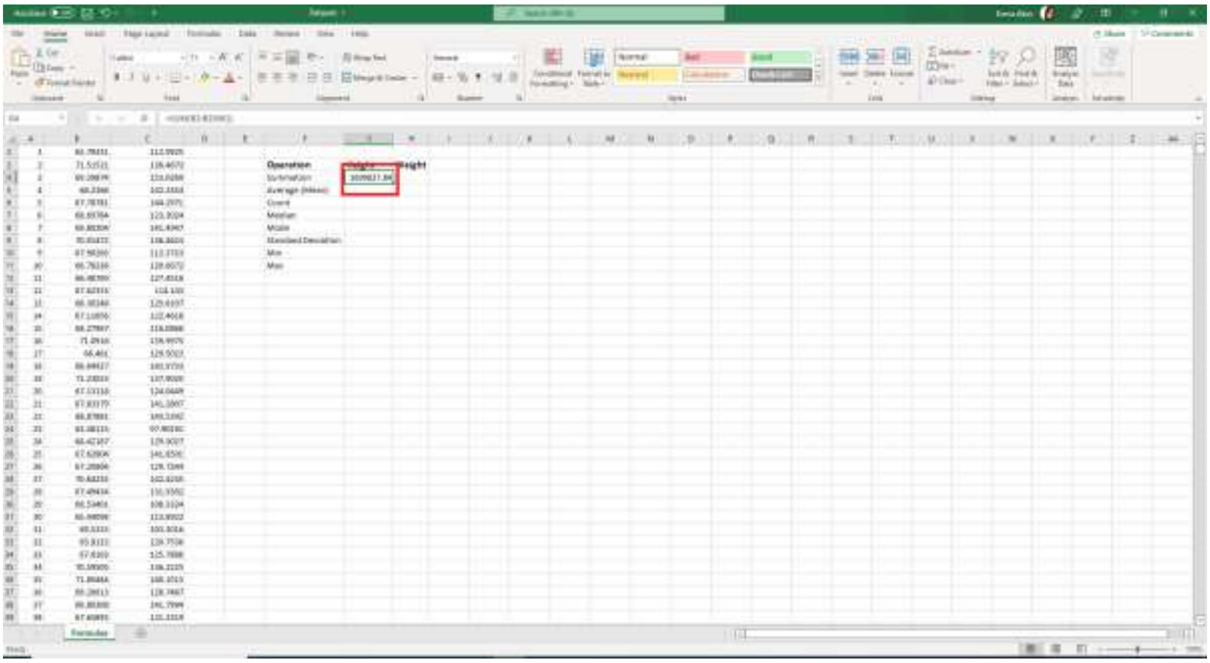
Operation	Height	Weight
Summation	=	
Average (Mean)		
Count		
Median		
Mode		
Standard Deviation		
Min		
Max		

d) Once you select a function from the list a new window will pop up. You have to write "From" and "To" cell number(s) in here. We are taking cell "B2 to cell B25001" As we are up for summing up the total height from the data present here. We write it like,



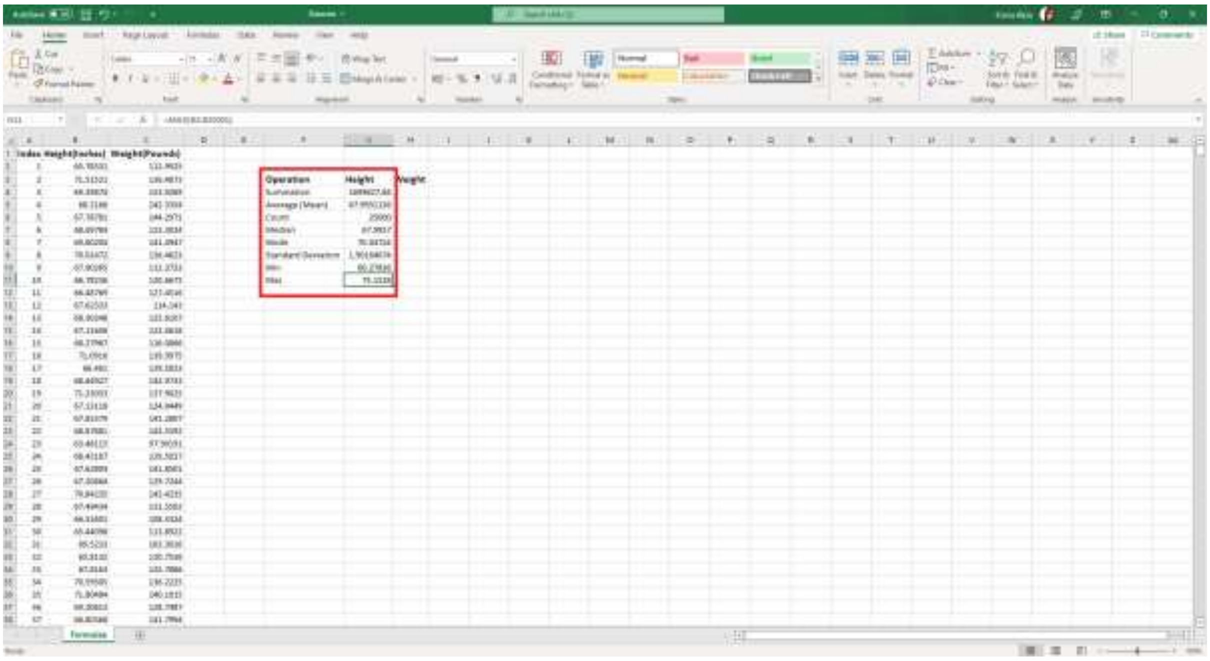
B2:B25001 in the text box.

e) After giving the input and clicking “OK”, you should now get the result of the summation of the “Height” in the cell you selected.



Operation	Height	Weight
Summation	[illegible]	
Average (Mean)		
Count		
Median		
Mode		
Standard Deviation		
Min		
Max		

Following the same steps, you now should be able to perform other statistical operations from the dataset.

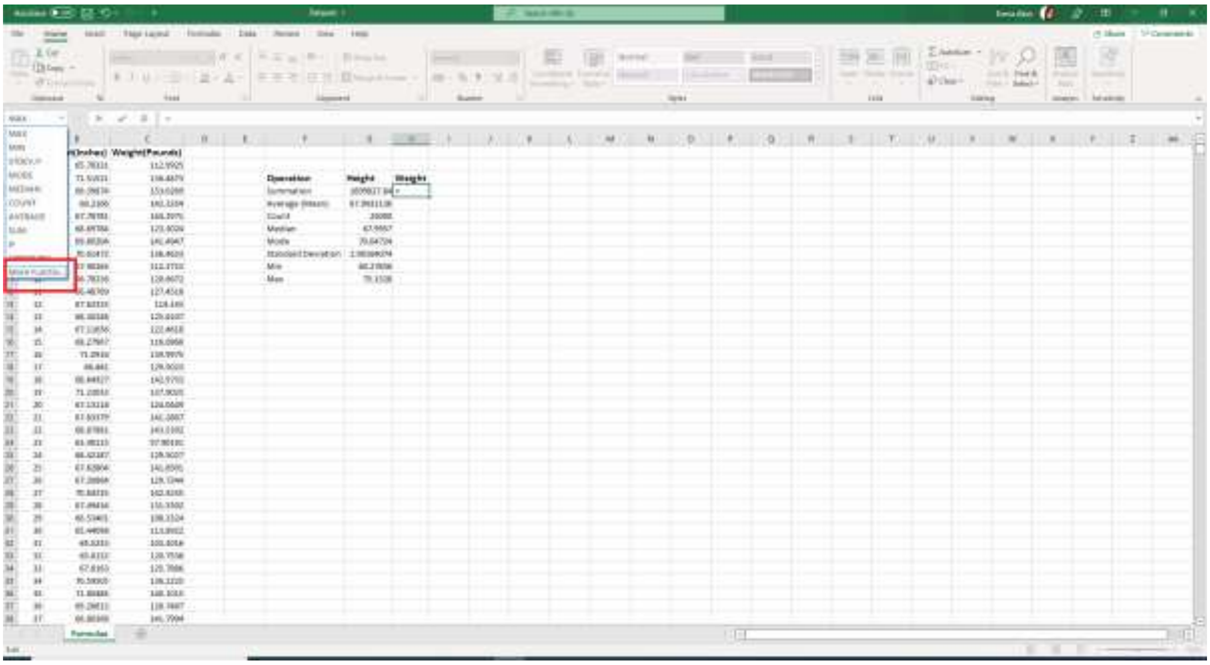


Operation	Height	Weight
Summation	1699827.66	
Average (Mean)	67.9931136	
Count	25000	
Median	67.9957	
Mode	70.04724	
Standard Deviation	1.90168474	
Min	60.27836	
Max	75.1528	

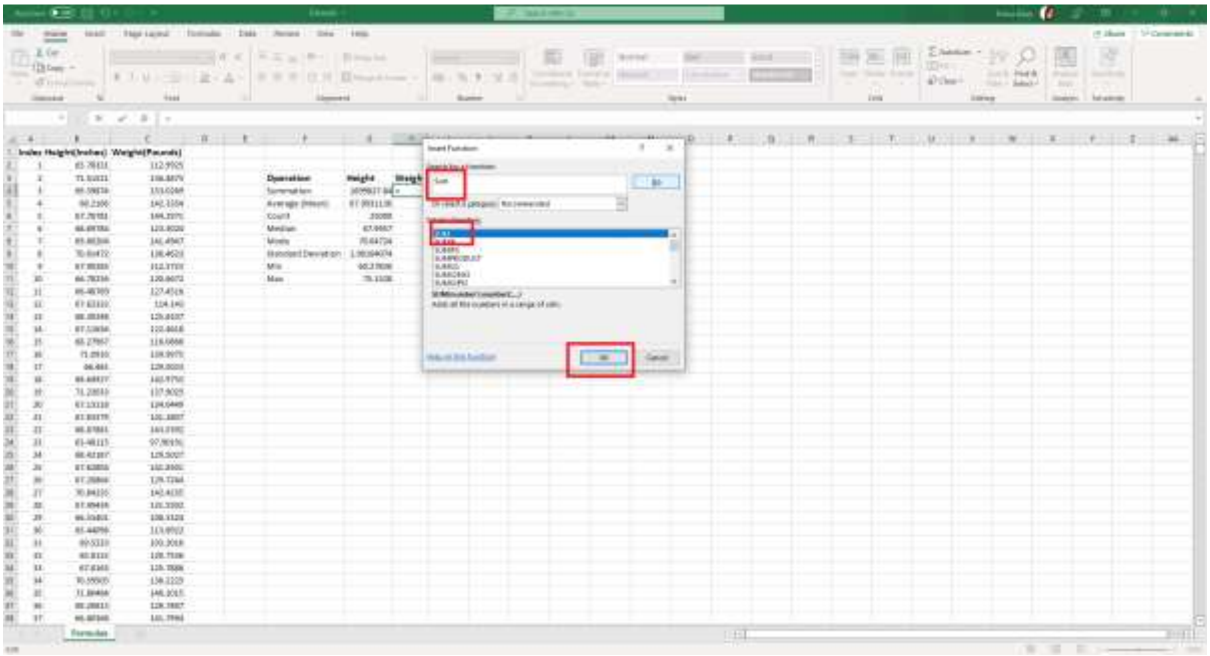
Useful Tips:

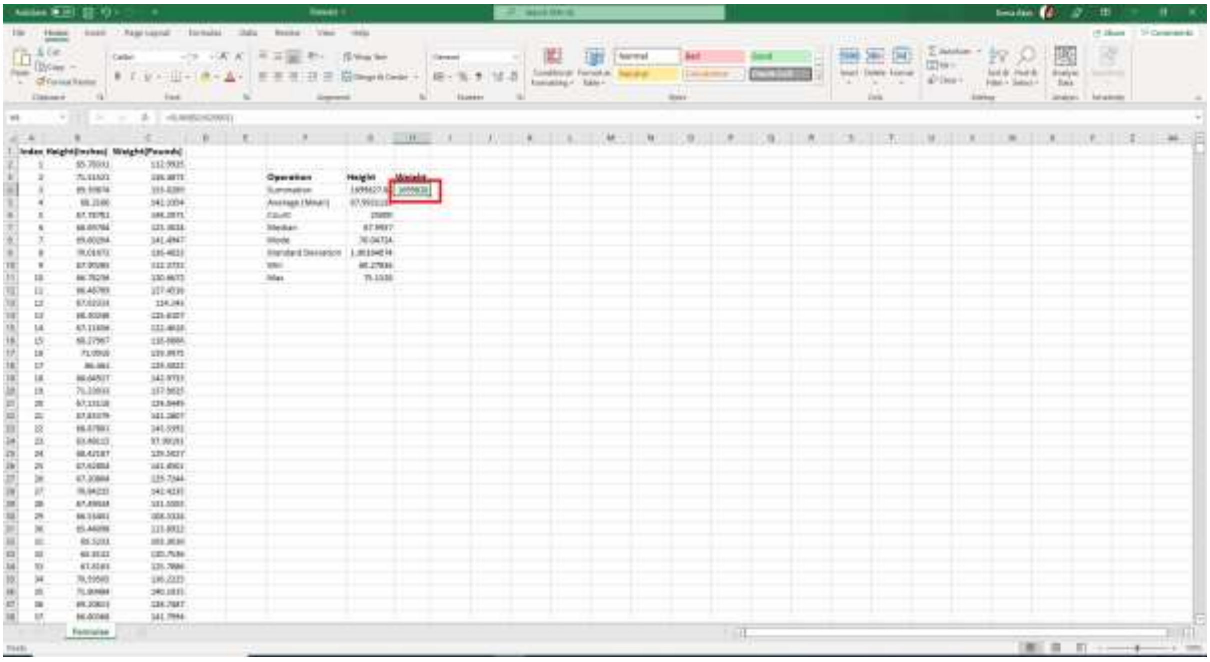
Instead of scrolling down and searching for the operation type. You can manually search for the operation you want to perform by using following steps:

Assume, you want to calculate the summation of the "Weight". So, you select cell. Go to choose function and go for "More Options".



You can manually type/input the operation “Sum”, and it will give you the result in the selected cell.





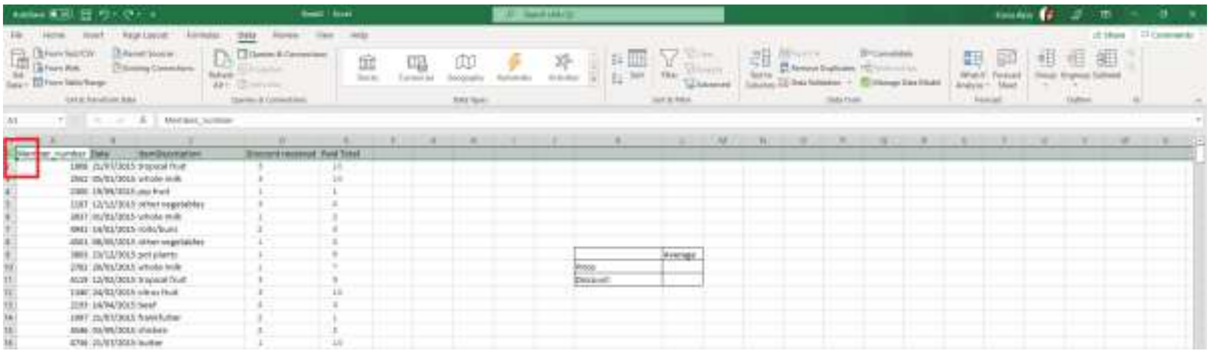
Task1:

Similarly, calculate the Average(Mean), Median, Mode and Standard Deviation of the “Weight” data given in the same dataset.

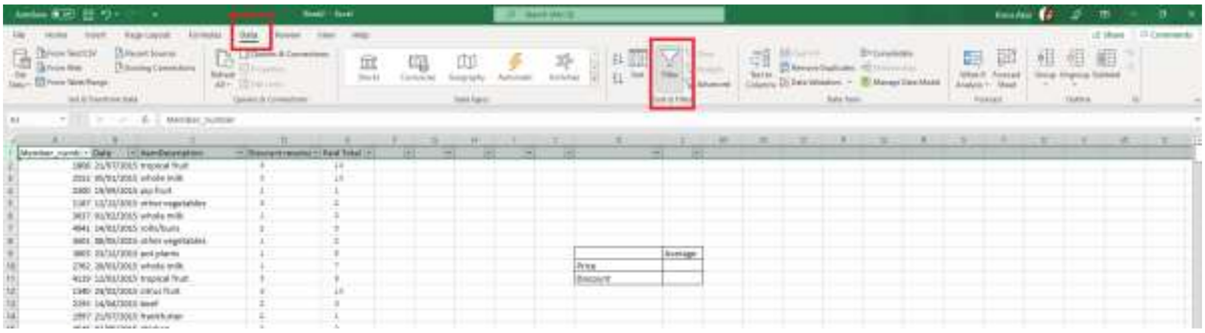
Data Cleaning:

There is another dataset given to you, called “Data Cleaning”. In this Dataset you will find data of a grocery store.

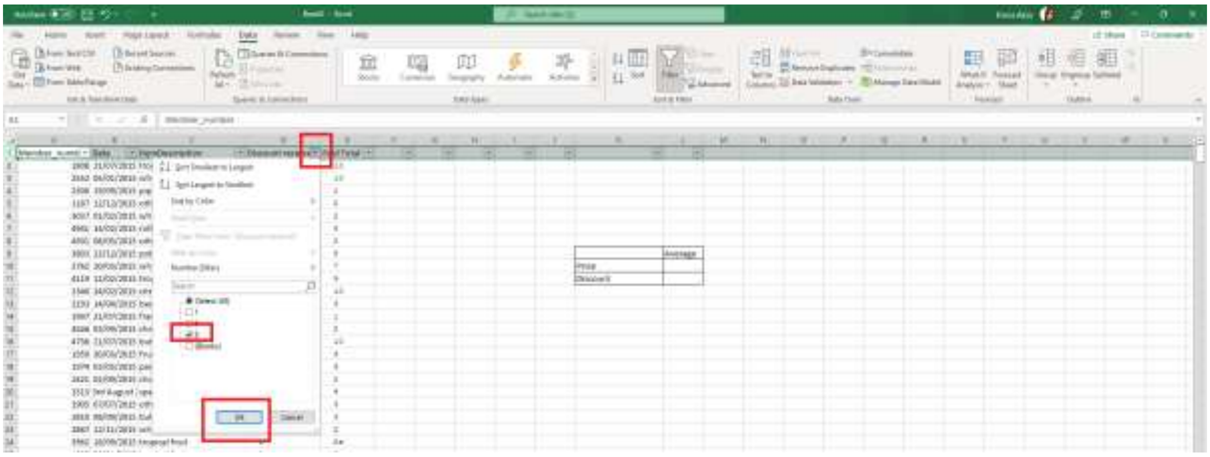
1. First observe if you can see any inconsistent formatting or, missing value problem in the data there.
2. In this step you will see how you can add filters and so, it should be much easier for you to see the inconsistencies in the dataset.
   a) Select the row that contains the Title of your attributes:

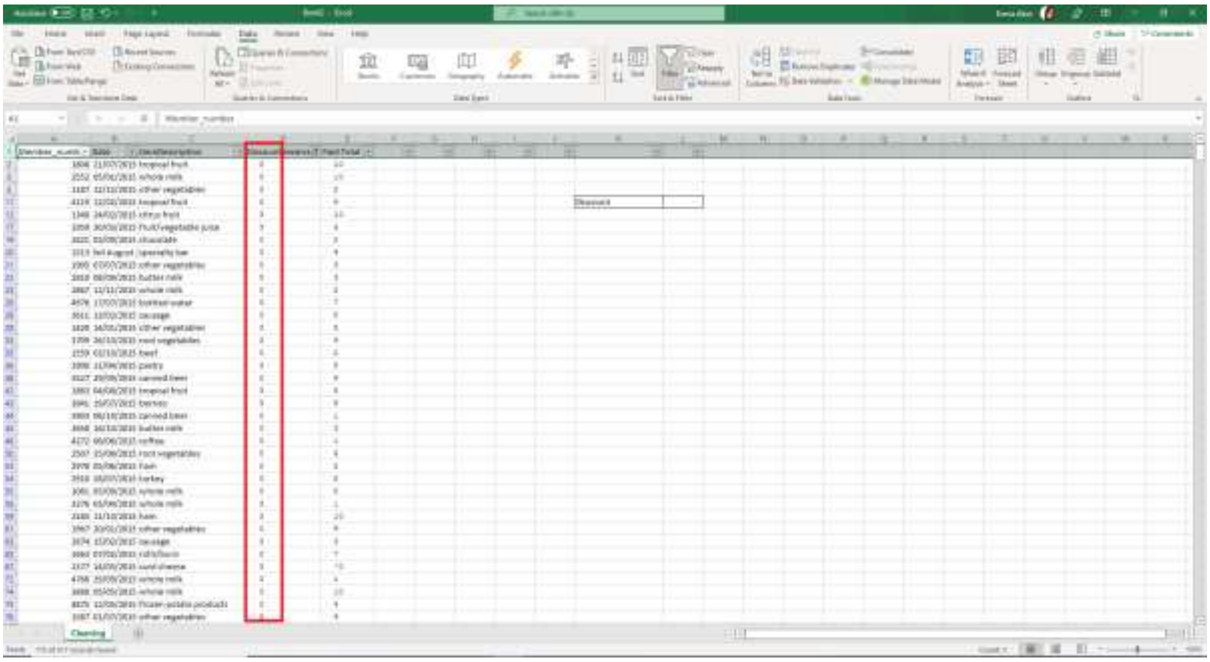


   b) Keeping this row selected, go to option “Data” and select “Filter”. This is how you should be able to add “Filter” on every attribute present in this dataset.

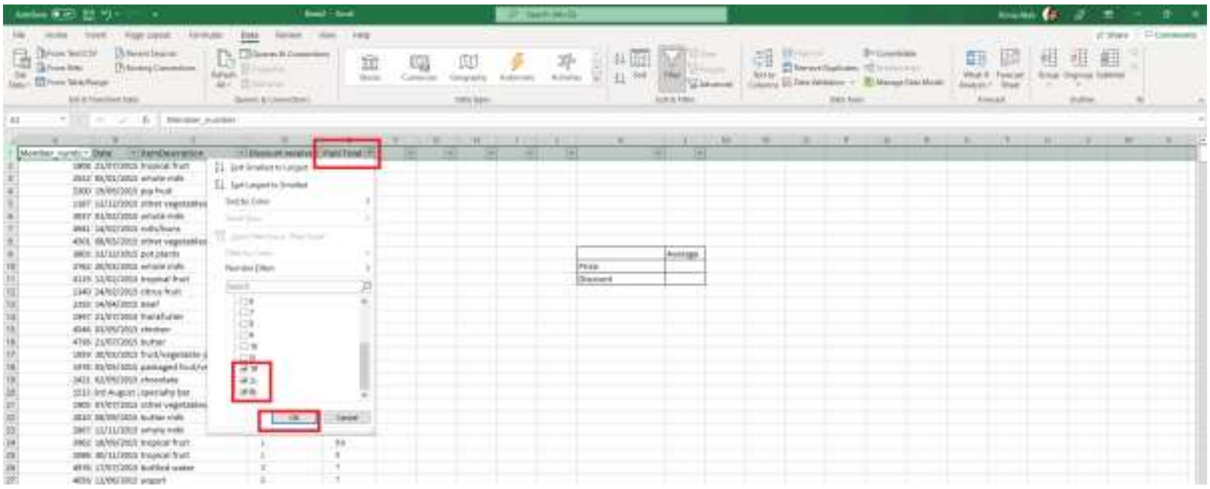


   c) Let’s say, you want to see only the customers got discount of GBP 3. You can see them by applying filter list.

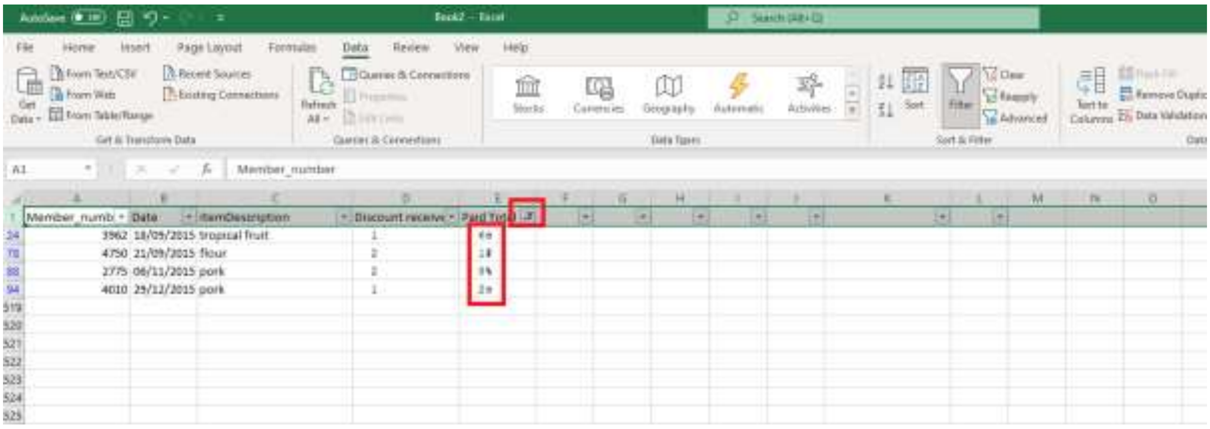




Now, if you use filter on “Paid total” column, you will see all the unique values in the filter list. In the list you will see there are some values which are wrongly recorded (Some special characters and/or, character inputs have come along).



After they are selected. You can now easily see only these values and do the cleaning as they are already spotted from the filter.



A1 Member_number

	Member_numb	Date	itemDescription	Discount receive	Paid Total
24	3962	18/09/2015	tropical fruit	1	[illegible]
78	4750	21/09/2015	flour	2	1#
88	2775	08/11/2015	pork	2	3%
94	4010	29/12/2015	pork	1	[illegible]

You can now correct them and make the data consistent, or, even discard the tuples if there's no hint left to correct them. This attribute contains total paid amount and only removing the characters from the values should work appropriately.

Task:2

Use filter on "Discount" attribute select the records where "Discount" value is missing and discard those tuples.

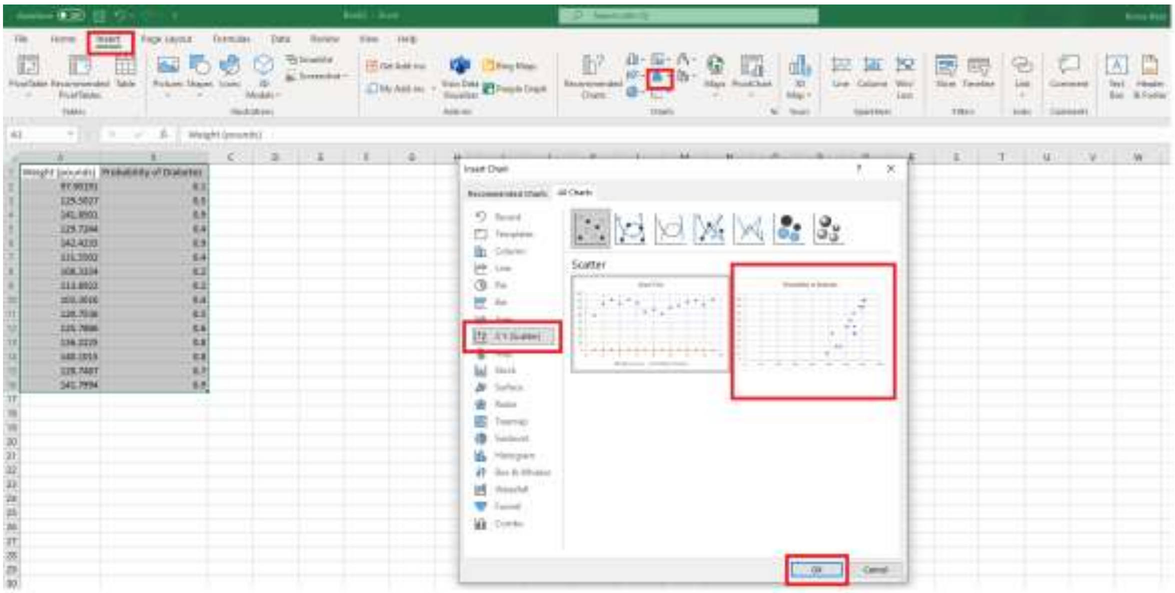
Task:3

Have a look at the date column make the date format consistent, if you observe any inconsistency exists.

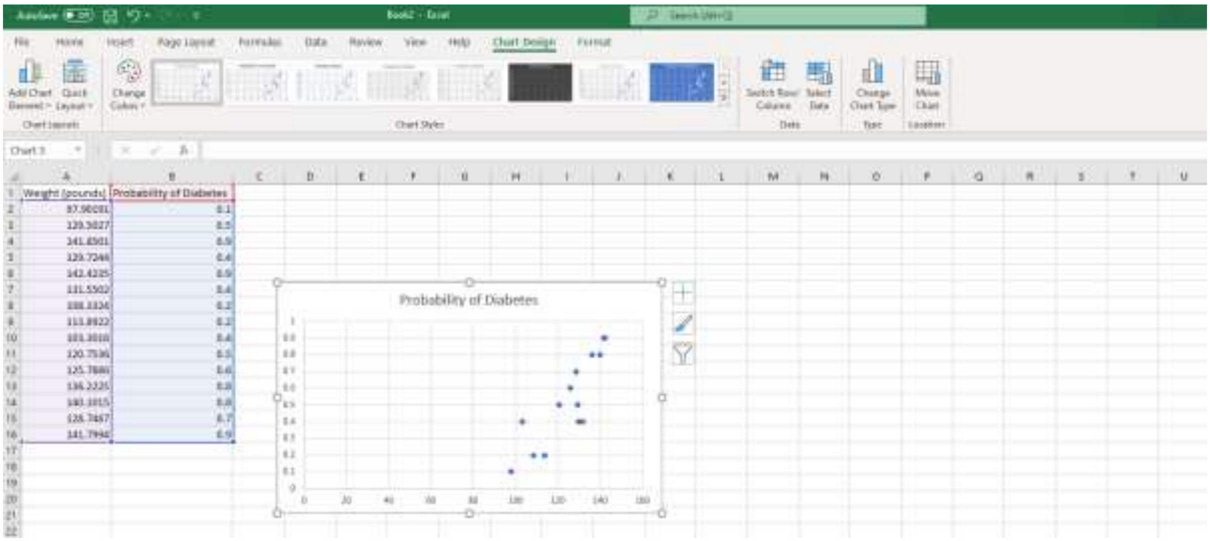
Data Visualization:

As we know already, we can represent useful and meaningful insights from data through Data Visualization. We will now, work on Data representation. For that, you will use “Visualization” spread sheet from the Datasets excel (.csv) file. The Dataset contains data of weight and probability of diabetes.

You first select the data in the Dataset, Go to option “Insert” from above, go to chart and graph option. Pick X. Y. scatter and select “OK”.

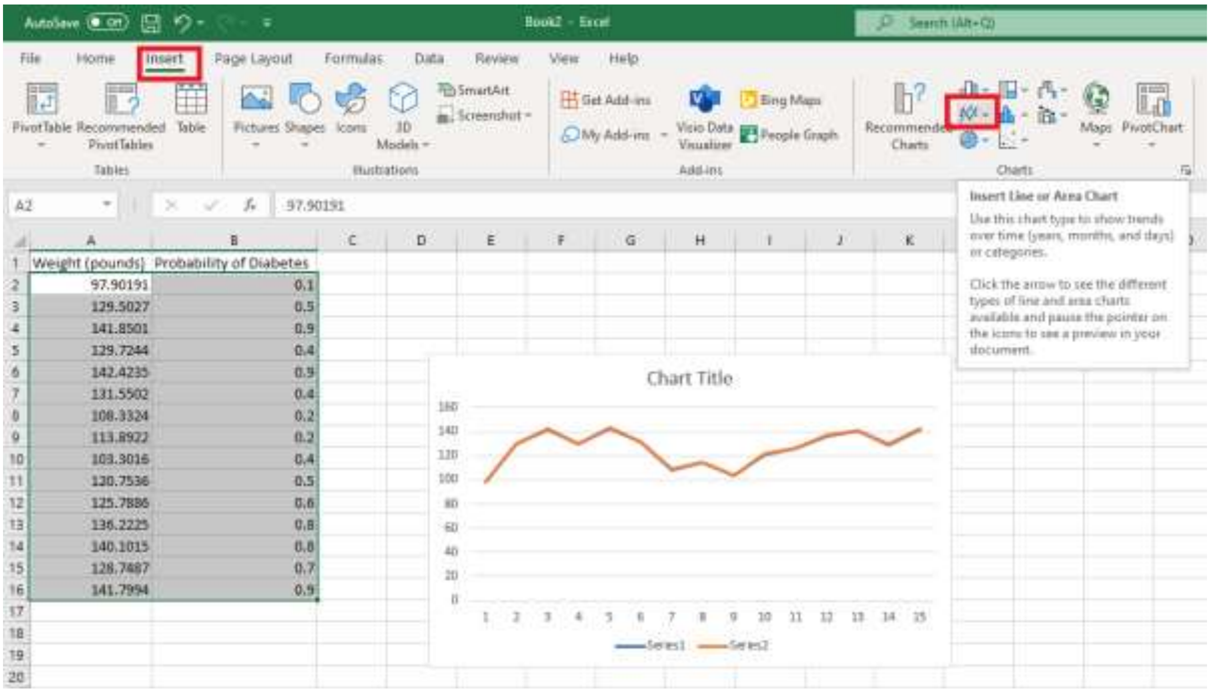


You now get a plotting of probability of diabetes against the weight data.



Weight (pounds)	Probability of Diabetes
97.90081	0.1
129.3027	0.5
141.6501	0.9
129.7244	0.4
142.4235	0.9
131.5502	0.4
108.3324	0.2
113.8922	0.2
103.3016	0.4
120.7536	0.5
125.7886	0.6
136.2225	0.8
140.1015	0.8
128.7467	0.7
141.7994	0.9

You con use the same data and find a different type of visualization output of it.



A2 | 97.90191

	A	B
1	Weight (pounds)	Probability of Diabetes
2	97.90191	0.1
3	129.5027	0.5
4	141.8501	0.9
5	129.7244	0.4
6	142.4235	0.9
7	131.5502	0.4
8	108.3324	0.2
9	113.8922	0.2
10	103.3016	0.4
11	120.7536	0.5
12	125.7886	0.6
13	136.2225	0.8
14	140.1015	0.8
15	128.7487	0.7
16	141.7994	0.9

Insert Line or Area Chart

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

Click the arrow to see the different types of line and area charts available and pause the pointer on the icons to see a preview in your document.

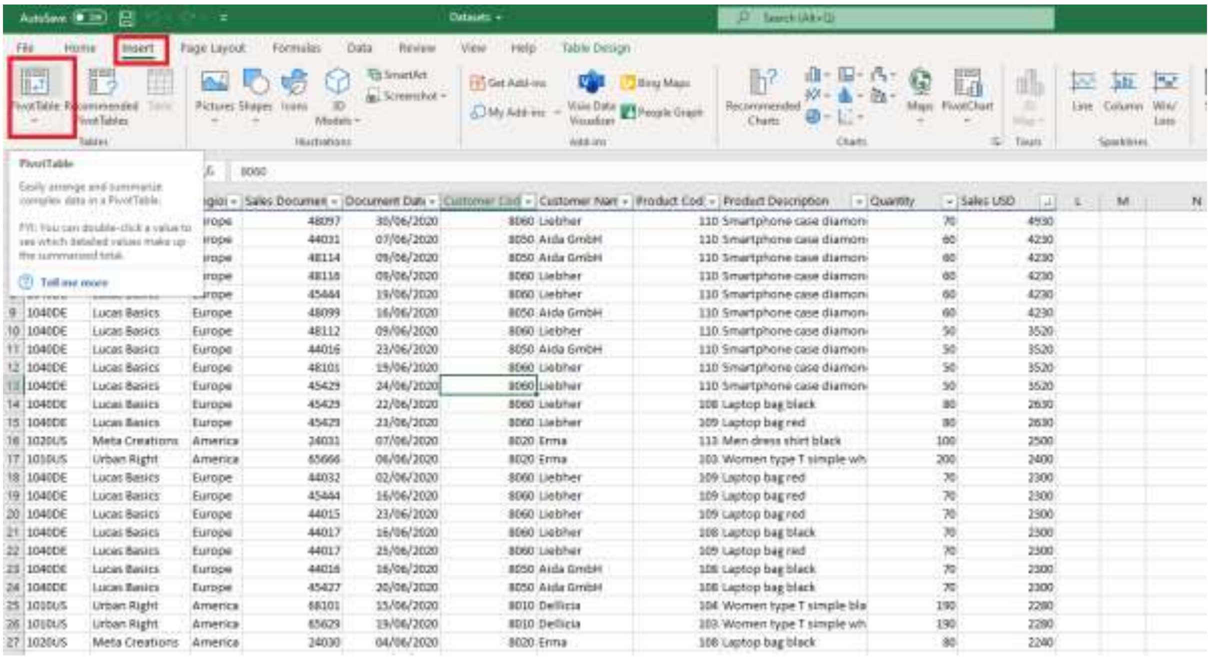
Task:4

Now, represent the sales data, available in “Sales” spreadsheet using a Bar Chart (Column).

Pivot Table:

We will now work on Pivot table. For that, you will use “Company Dataset” spreadsheet from the same dataset file excel file.

Open the spread sheet. Go to “Insert”, Pivot Table.



PivotTable

Easily arrange and summarize complex data in a PivotTable.

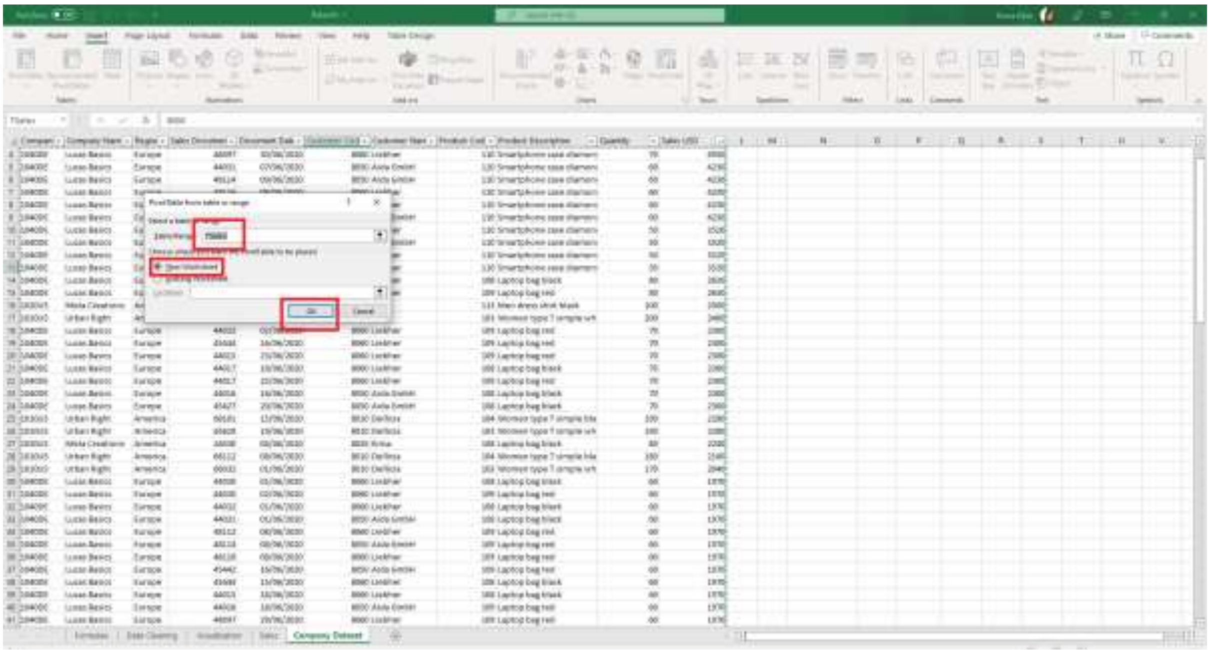
FYI: You can double-click a value to see which detailed values make up the summarized total.

Tell me more

8060

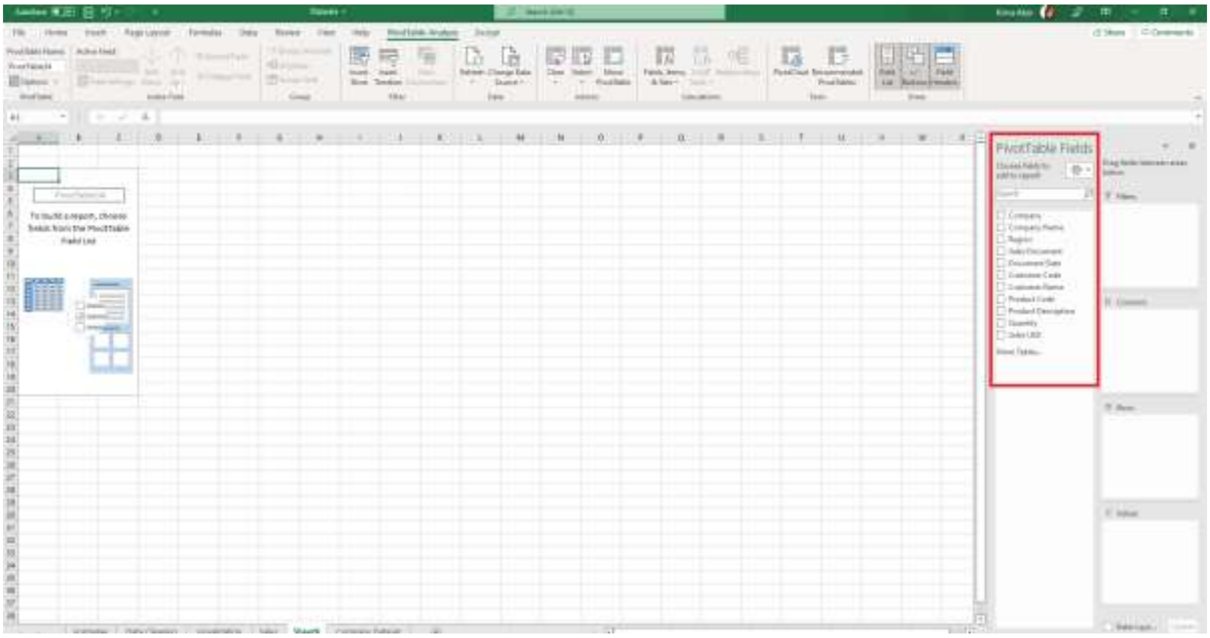
			gion	Sales Documen	Document Dat	Customer Cod	Customer Nam	Product Cod	Product Description	Quantity	Sales USD	L	M	N
			rope	48097	30/06/2020	8060	Liebher	110	Smartphone case diamon	70	4930			
			rope	44031	07/06/2020	8050	Aida GmbH	110	Smartphone case diamon	60	4230			
			rope	48114	09/06/2020	8050	Aida GmbH	110	Smartphone case diamon	60	4230			
			rope	48118	09/06/2020	8060	Liebher	110	Smartphone case diamon	60	4230			
			rope	45444	19/06/2020	8060	Liebher	110	Smartphone case diamon	60	4230			
9	1040DE	Lucas Basics	Europe	48099	16/06/2020	8050	Aida GmbH	110	Smartphone case diamon	60	4230			
10	1040DE	Lucas Basics	Europe	48112	09/06/2020	8060	Liebher	110	Smartphone case diamon	50	3520			
11	1040DE	Lucas Basics	Europe	44016	23/06/2020	8050	Aida GmbH	110	Smartphone case diamon	50	3520			
12	1040DE	Lucas Basics	Europe	48101	19/06/2020	8060	Liebher	110	Smartphone case diamon	50	3520			
13	1040DE	Lucas Basics	Europe	45429	24/06/2020	8060	Liebher	110	Smartphone case diamon	50	3520			
14	1040DE	Lucas Basics	Europe	45429	22/06/2020	8060	Liebher	108	Laptop bag black	80	2630			
15	1040DE	Lucas Basics	Europe	45429	21/06/2020	8060	Liebher	109	Laptop bag red	80	2630			
16	1020US	Meta Creations	America	24031	07/06/2020	8020	Erma	113	Men dress shirt black	100	2500			
17	1010US	Urban Right	America	65666	06/06/2020	8020	Erma	103	Women type T simple wh	200	2400			
18	1040DE	Lucas Basics	Europe	44032	02/06/2020	8060	Liebher	109	Laptop bag red	70	2300			
19	1040DE	Lucas Basics	Europe	45444	16/06/2020	8060	Liebher	109	Laptop bag red	70	2300			
20	1040DE	Lucas Basics	Europe	44015	23/06/2020	8060	Liebher	109	Laptop bag red	70	2300			
21	1040DE	Lucas Basics	Europe	44017	16/06/2020	8060	Liebher	108	Laptop bag black	70	2300			
22	1040DE	Lucas Basics	Europe	44017	25/06/2020	8060	Liebher	109	Laptop bag red	70	2300			
23	1040DE	Lucas Basics	Europe	44016	16/06/2020	8050	Aida GmbH	108	Laptop bag black	70	2300			
24	1040DE	Lucas Basics	Europe	45427	20/06/2020	8050	Aida GmbH	108	Laptop bag black	70	2300			
25	1010US	Urban Right	America	68101	15/06/2020	8010	Dellicia	104	Women type T simple bla	190	2280			
26	1010US	Urban Right	America	65629	19/06/2020	8010	Dellicia	103	Women type T simple wh	190	2280			
27	1020US	Meta Creations	America	24030	04/06/2020	8020	Erma	108	Laptop bag black	80	2240			

You will select ok for Tsale table already created for you.

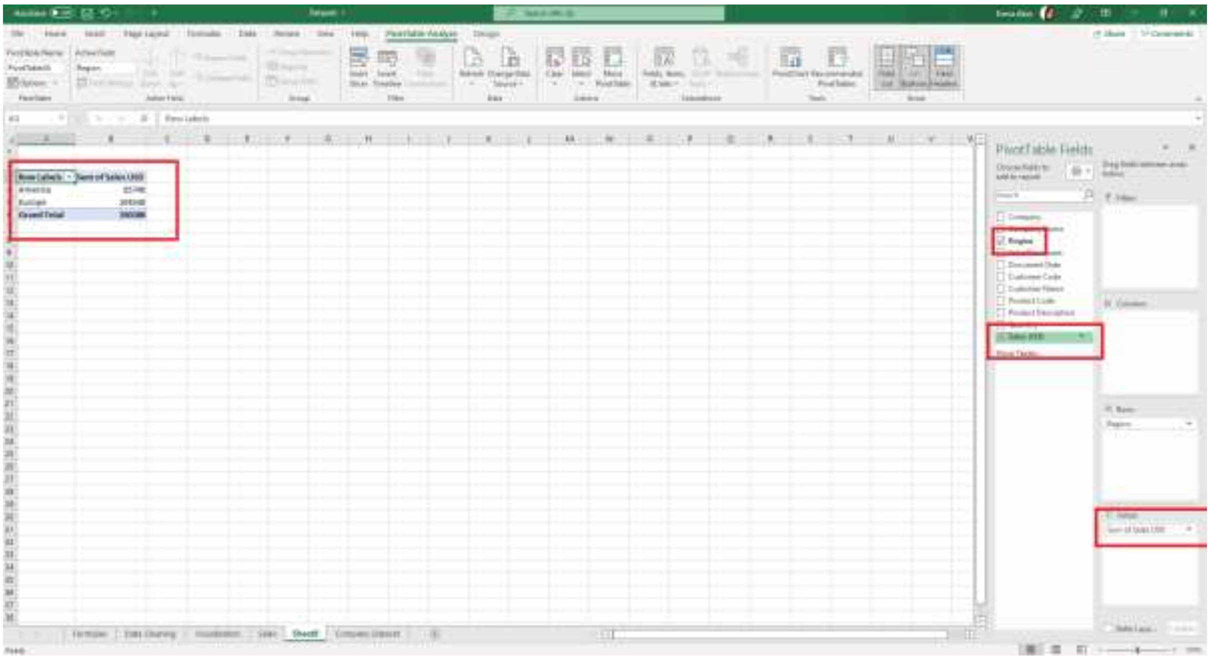


Kimia Aksir

Once the pivot gets opened in a new worksheet. You will see by manually selecting the parameters from the table you can generate any report you want.



Let’s find the total sales (USD) in different regions.



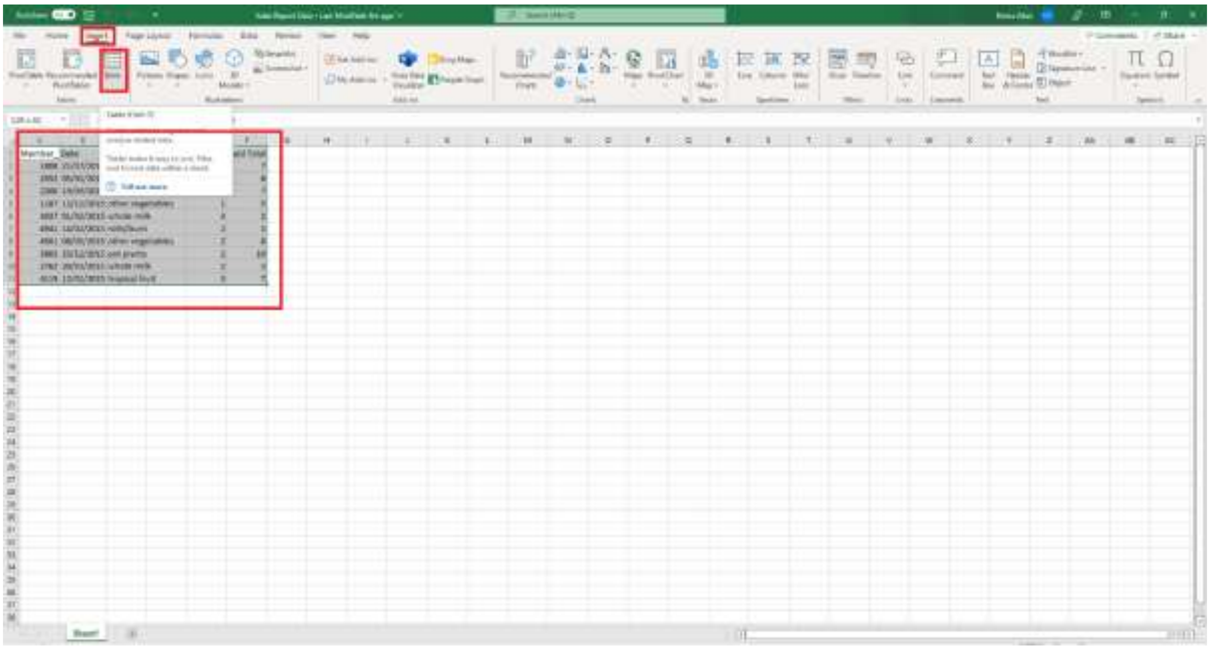
Row Labels	Sum of Sales USD
America	[illegible]
Europe	[illegible]
Grand Total	[illegible]

Task:5

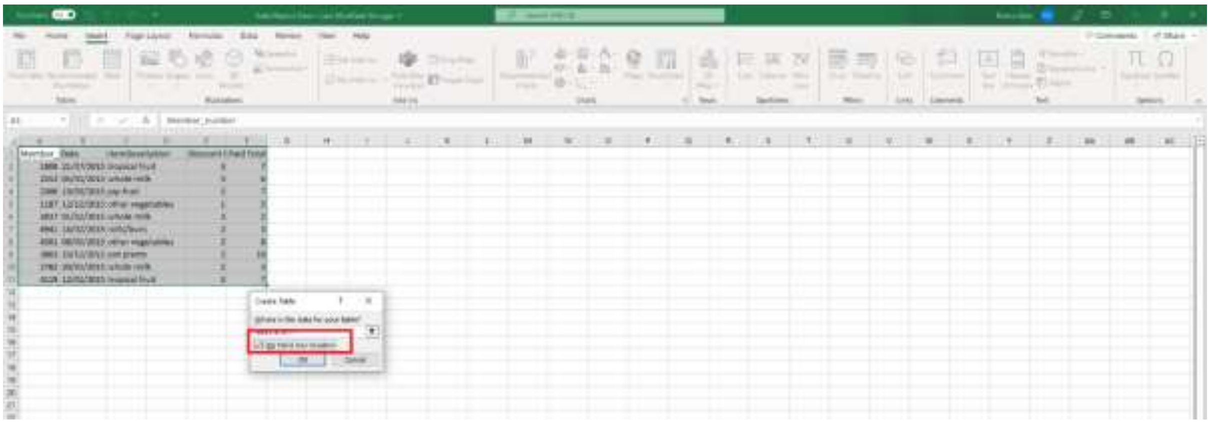
Find the Region wise Average of Sales (USD)

You can now create and manage a Pivot Table on your own. Use “Sales Report Data” dataset and try to follow the steps given below.

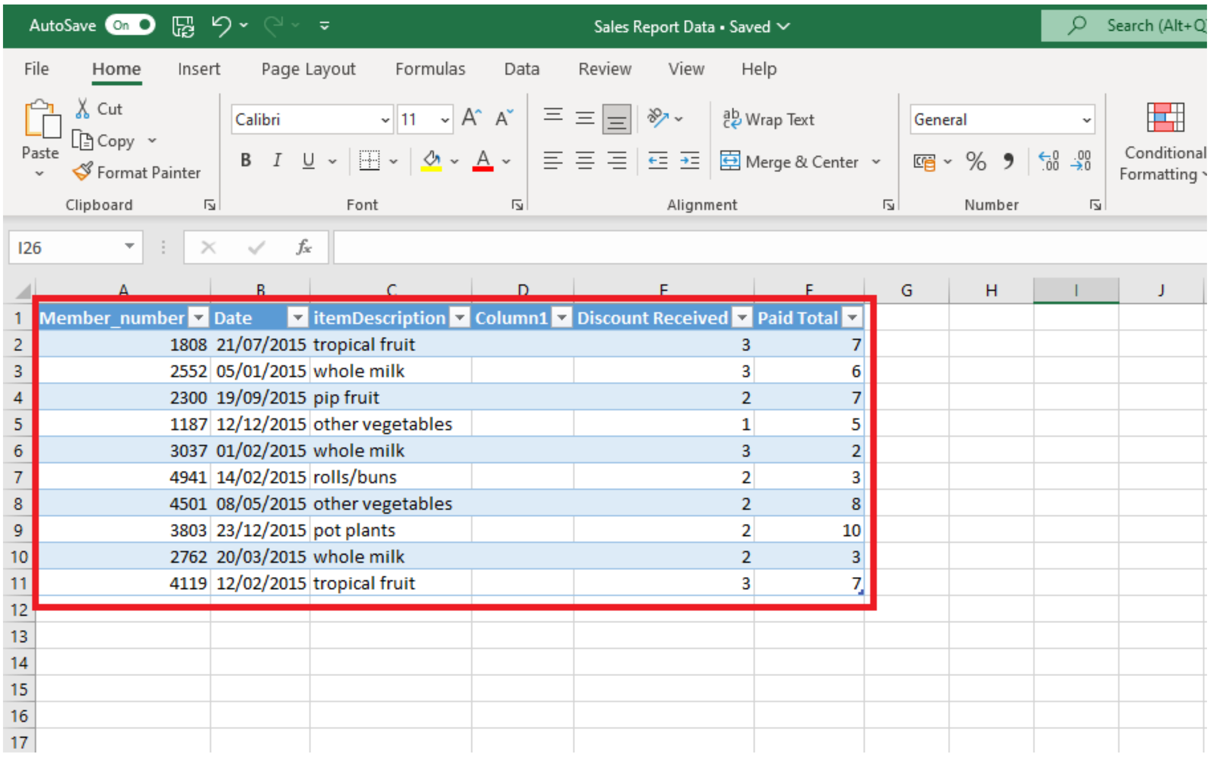
Selecting the Dataset Go to “Insert” menu tab from the top and select “Table”:



Remember to tick “My table has header” manually if it’s not automatically selected. Because without Header Pivot Table won’t work!

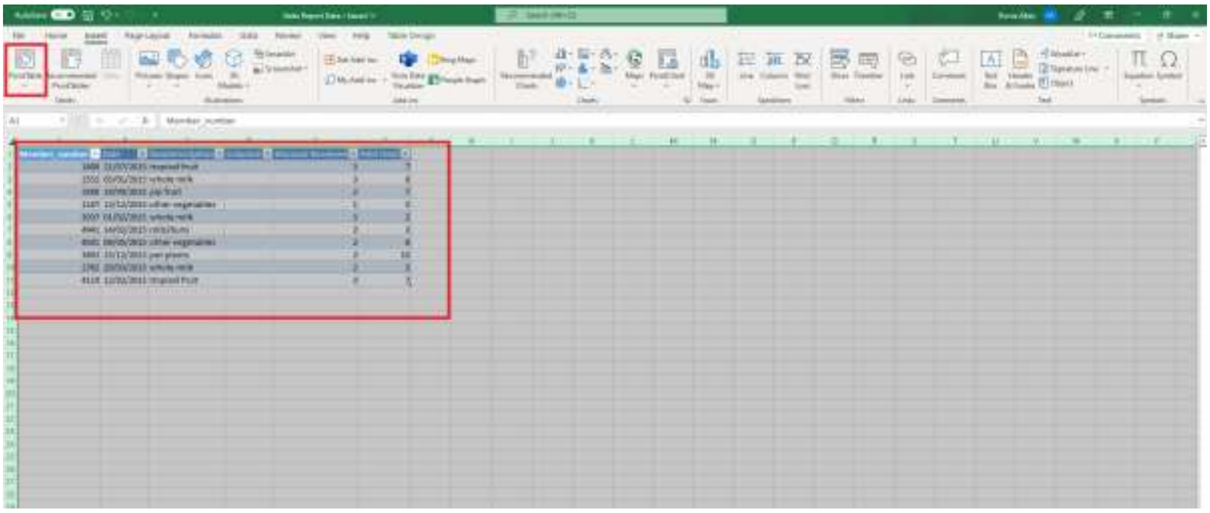


Finally, You get your table as shown below with all the headers present in it:

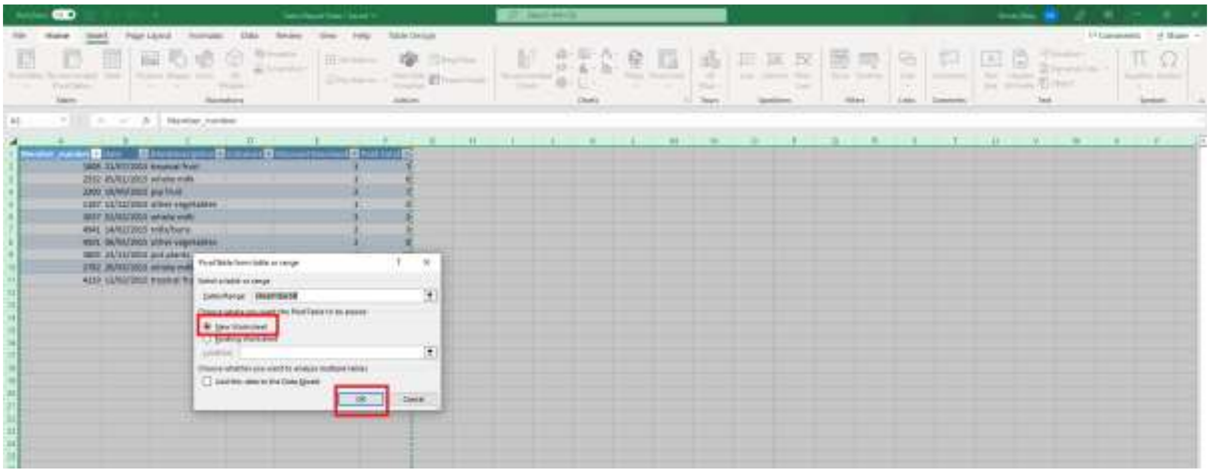


Member_number	Date	itemDescription	Column1	Discount Received	Paid Total
1808	21/07/2015	tropical fruit		3	7
2552	05/01/2015	whole milk		3	6
2300	19/09/2015	pip fruit		2	7
1187	12/12/2015	other vegetables		1	5
3037	01/02/2015	whole milk		3	2
4941	14/02/2015	rolls/buns		2	3
4501	08/05/2015	other vegetables		2	8
3803	23/12/2015	pot plants		2	10
2762	20/03/2015	whole milk		2	3
4119	12/02/2015	tropical fruit		3	7

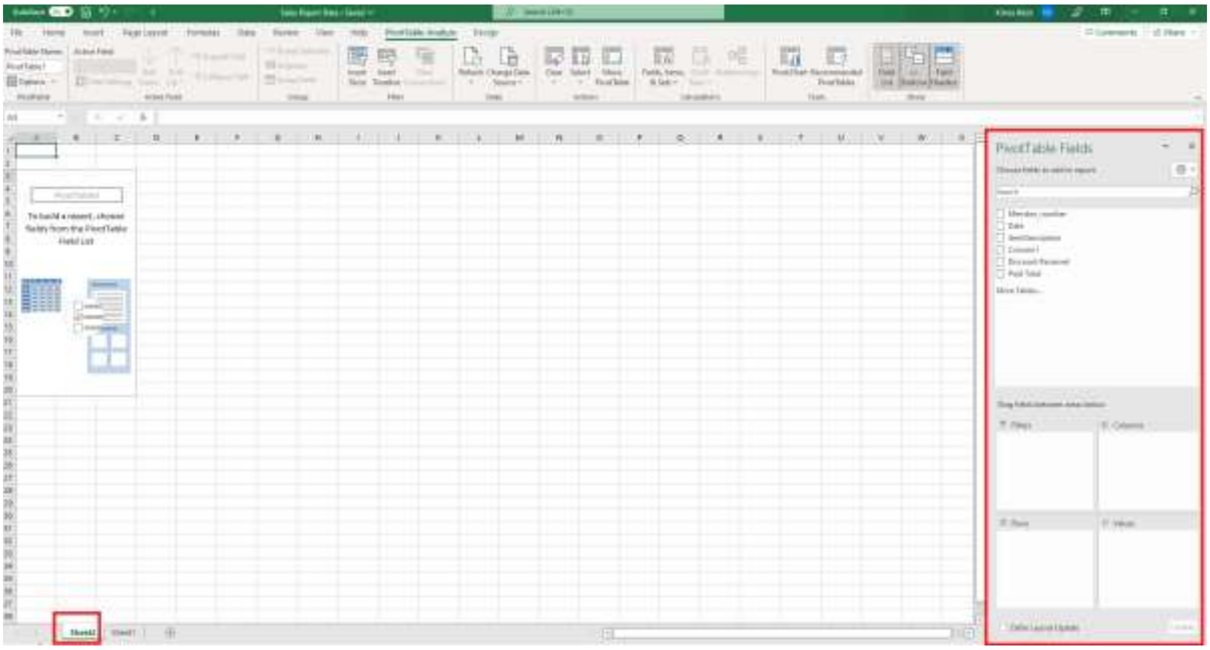
Keeping the table selected, go to the tab “Insert” and then “Pivot table”. We may also want to choose “Recommended Pivot”, however, choosing the former option will always allow us to customize the features on our own.



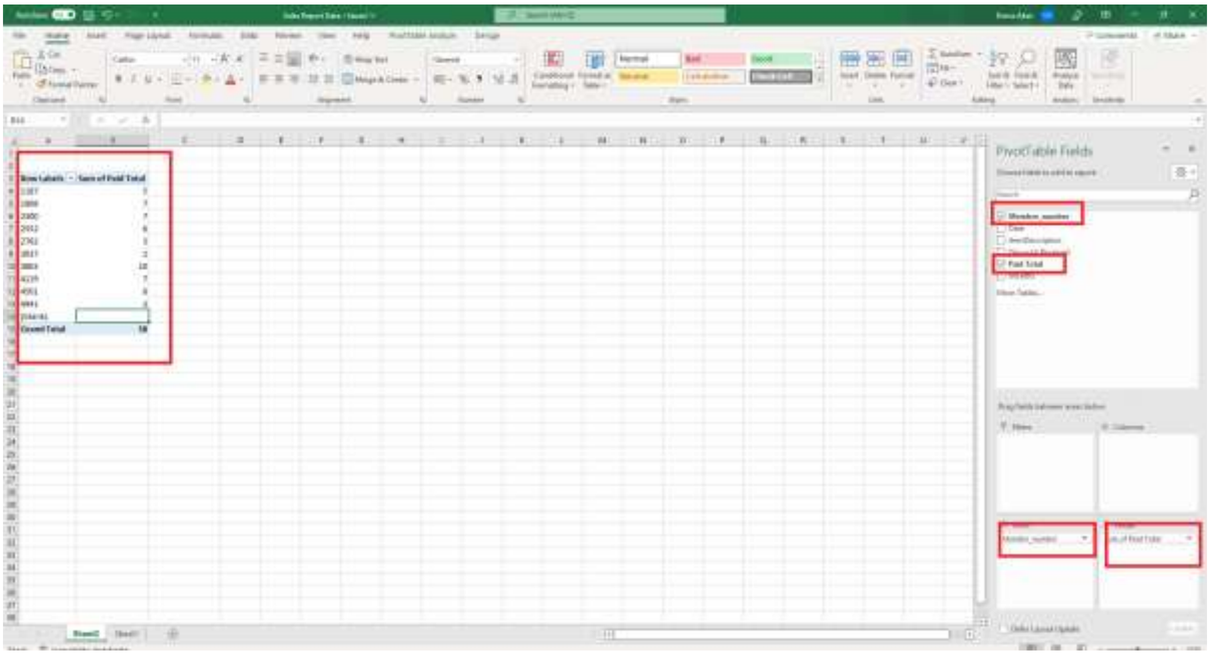
Choosing “New Worksheet” for presenting the Pivot table is recommended here. However, you can also keep the Pivot on the same worksheet along with the table populated with data.



As you can see, in a separate worksheet the Pivot table features appeared. Now, you can generate any report or, check the pattern of the data in the dataset using this pivot table.



Let’s generate a report on each member and they paid total.



Row Labels	Sum of Paid Total
1187	5
1888	7
2300	7
2412	6
2761	1
3517	2
3883	10
4239	7
4551	8
4441	2
(blank)	
Grand Total	58

You can also try to add a few for rows in the table and check how it reflects in the Pivot table.