



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

BABM
Data Handling & Business Intelligence
ESSAY

Module: Data Handling & Business Intelligence

Code: LCBB5000

Lecture: Dr Anupam Mazumdar

Name: Miroslav Penkov

Student ID: 2133625

SUBMITTED ON THE 17/08/2023

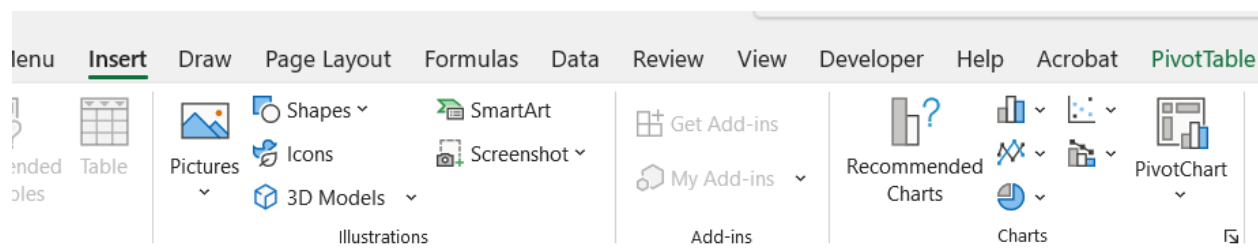
Microsoft excel is a modern tool that have made the life easier and better by providing the opportunity to analyze and present different data set which can be used to identify certain characteristics or make strategic decisions.

Regarding pre processing of data, Excel makes the life easier by providing a user friendly interface where people can sort, filter, or modify the data as they need to make the data more suitable for their use. Even Microsoft Excel provides the benefit to customize data and apply required formula to create customized and calculated columns in the software (Carlberg, 2014). However, the data entry may required manual inputs which can create different types of typing error, thus diverging the result from the original.

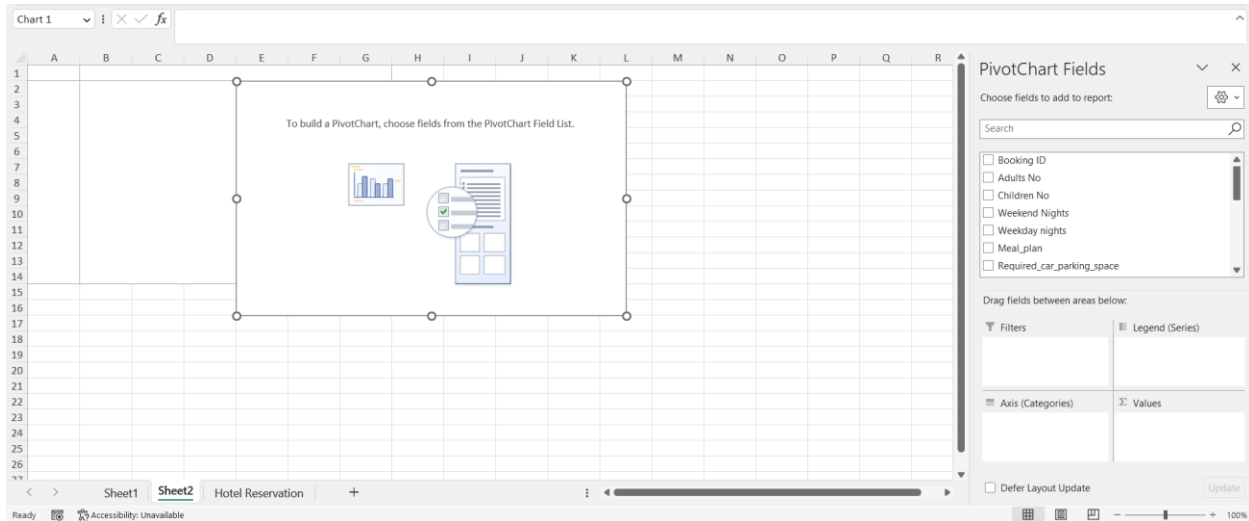
Analyzing the data using Microsoft Excel is very effective and efficient with the help of all the functions available in the software and many of them will be required in this report such as AverageIF. However, the statistical analysis that can be done within the Microsoft Excel can be limited when compared to some other software (Carlberg, 2014).

Visualizing the data through Microsoft Excel is beneficial as several charting options are available in the software and producing these charts are very easy and requires minimal time. Using these charts, the complex data can be easily presented to any audience, even the non technical people can understand the data with a glance (Carlberg, 2014). However, the advanced visualization feature of Microsoft Excel may be limited when compared to other software that are being used for this task.

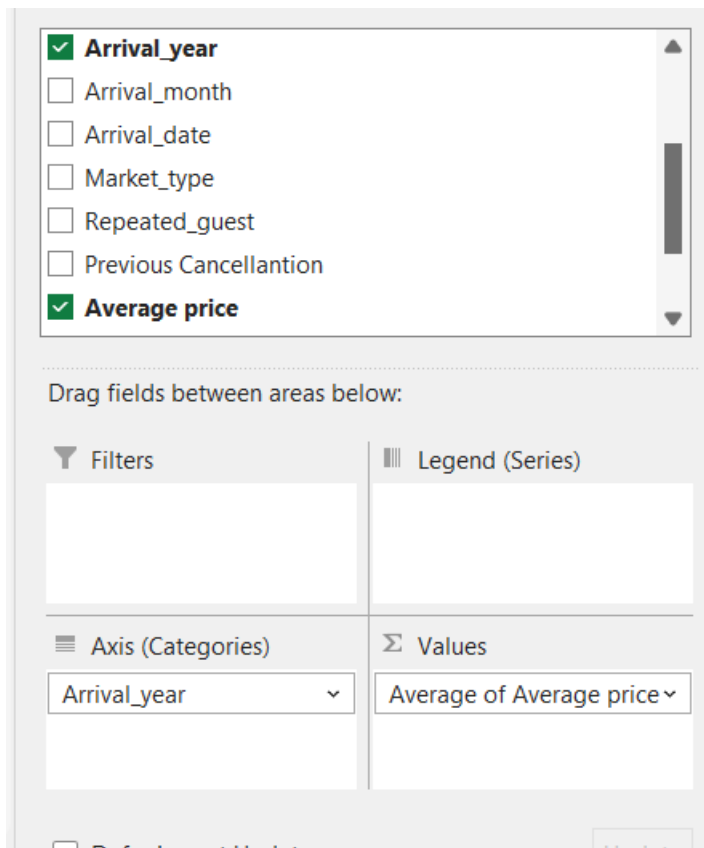
The goal of this report is to analyse the hotel reservation data which can be found from the hotel reservation dataset given. To get an overall idea about the trend of the price fluctuation over the years, the Pivot Table that is available in the Excel can be a valueable feature (Schmuller, 2013).



From Insert>PivotChart Option was selected to create Pivot Table and Chart which makes it easier to find the average based on each year which is the current goal (Schmuller, 2013).



The Pivot Chart and Table was inserted in the Excel as a new Sheet called Sheet2. There are multiple options provided on the right hand side called PivotChart Fields which will be helpful to find the anticipated results.

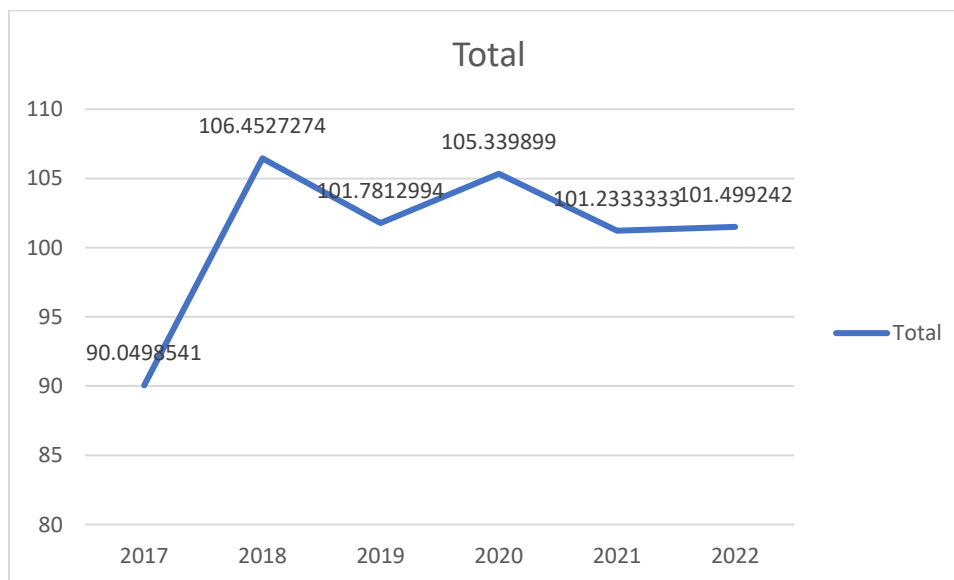


From the PivotChart Fields, the Arrival year and average price was selected which will be used to get the required data. By ticking both of these boxes, all the values of these two columns were

added on the bottom right corner box called “Values” as a Sum of Arrival year and sum of Average Price (Dierenfeld and Merceron, 2012). However, this is not the result that is required which is why the boxes needed some changes. To get the expected result, Arrival Year data was dragged on the other box called Axis (Categories) and on the Values box, the Sum of Average Price was changed to Average of Average Price. By doing so, an average price of each year was found as a table in the sheet as well as a graph was found that represents the average data of each year in the graph.

Arrival_year	Total
2017	90.0498541
2018	106.4527274
2019	101.7812994
2020	105.339899
2021	101.2333333
2022	101.499242

This is the table found in the Sheet2 where it is evident that the lowest average of a year happened during 2017. This is the only year when the average went lower than 100. Among all the years, 2018 had the most average price and the value is equal to 106.4527274.



This is the graph that was generated as a Pivot Chart based on the average price that was found in the Pivot chart (Camões, 2016). From this graph, it is easy to understand that the trend of the average price based on each year, and the year 2018 had the highest among them. The price was lowest back in 2017 and 2018 when it became the prime, but during the 2019, the business was

not as profitable so the average price went down which increased close to the peak time 2018 in the next year during 2020. The average price went down in the next year during 2021 and during 2022, the price is almost similar to the one found in 2021. The lowest average price among all the years in reference happened in 2017 and the highest average price was found in 2018.

The task is to use another excel function that will be useful to find the values of average price of each year. The average value is already calculated in the previous part of this part, so the new method of finding average should be able to provide same values to ensure that both the methods used and values found were right.

The AverageIF function can be used for this task to get the average value even though there is a function available in excel called Average (Schmuller, 2013). However, Average function will not work in this project because the average function is suitable to find an average value of a whole column, consecutive values of a column, or so. In this project, the values that are recorded in the excel file are not sorted based on the year of operation, instead the years that are recorded are shuffled. For example, if a value is found to be a part of 2020, the next value can be found to be part of 2017, and so on. Due to these reasons, the average function will not work.

To find the average, AverageIF function is available in Excel which can be useful for this project. The AverageIF function requires three conditions, the first one is the first value which will be used to check the criterion that will be used to check the argument, then the argument of this function that is required is provided, after that, the final criterion is written which will be basically checked with the criterion (Schmuller, 2013).

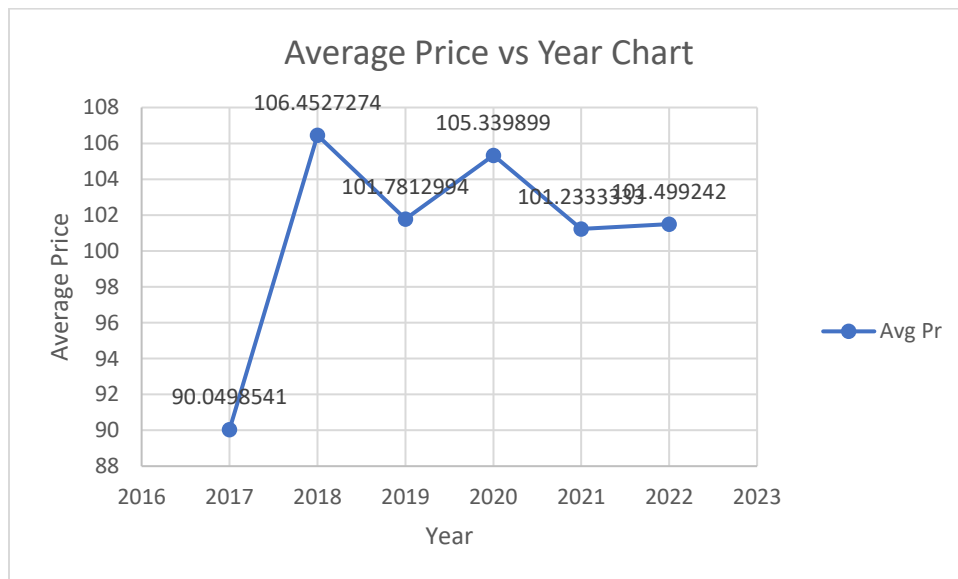
Two new columns were created called Yr and Avg Pr which basically represents Year and Average Price respectively but names were written in short so that the columns can easily be separated from the existing columns. In the Yr column, all the years were written in a chronological order to make it easier to understand. The values were written on the right side cell of each year to get the respective values of that year and the formula that was inputted on the right side column was AverageIF. The formula for 2017 was written as

“=AVERAGEIF(J\$2:J\$36276,2017,Q\$2:Q\$36276)”

By applying this formula for each of the years, the average price for each year were found which are shown below.

Yr	Avg Pr
2017	90.04985
2018	106.4527
2019	101.7813
2020	105.3399
2021	101.2333
2022	101.4992

Here the average values found are same as the average values found from the Pivot Table.



School_s etting	School_type	C a s e	Teaching_meth od	Student_ No	Student_I D	Gende r	Lunch	Pretest	Posttest
Urban	Non-public	6OL	Standard	20.00	2FHT3	Female	Does not qualify	62.00	72.00
Urban	Non-public	6OL	Standard	20.00	3JIVH	Female	Does not qualify	66.00	79.00
Urban	Non-public	6OL	Standard	20.00	3XOWE	Male	Does not qualify	64.00	76.00
Urban	Non-public	6OL	Standard	20.00	556O0	Female	Does not qualify	61.00	77.00
Urban	Non-public	6OL	Standard	20.00	74LOE	Male	Does not qualify	64.00	76.00
Urban	Non-public	6OL	Standard	20.00	7YZO8	Female	Does not qualify	66.00	74.00
Urban	Non-public	6OL	Standard	20.00	9KMZD	Male	Does not qualify	63.00	75.00
Urban	Non-public	6OL	Standard	20.00	9USQK	Female	Does not qualify	63.00	72.00
Urban	Non-public	6OL	Standard	20.00	CS5QP	Male	Does not qualify	64.00	77.00
Urban	Non-public	6OL	Standard	20.00	D6HT8	Female	Does not qualify	61.00	72.00
Urban	Non-public	6OL	Standard	20.00	DZMKU	Male	Does not qualify	61.00	73.00
Urban	Non-public	6OL	Standard	20.00	FH7B9	Male	Does not qualify	64.00	74.00
Urban	Non-public	6OL	Standard	20.00	J19VG	Male	Does not qualify	66.00	78.00
Urban	Non-public	6OL	Standard	20.00	JQM2V	Female	Does not qualify	60.00	71.00
Urban	Non-public	6OL	Standard	20.00	MEUC4	Female	Does not qualify	64.00	77.00
Urban	Non-public	6OL	Standard	20.00	R4U8H	Male	Does not qualify	64.00	73.00
Urban	Non-public	6OL	Standard	20.00	TH7KI	Male	Does not qualify	63.00	70.00
Urban	Non-public	6OL	Standard	20.00	U1FV7	Male	Does not qualify	67.00	73.00
Urban	Non-public	6OL	Standard	20.00	WC5I6	Female	Does not qualify	63.00	71.00
Urban	Non-public	6OL	Standard	20.00	ZBQ4T	Male	Does not qualify	64.00	73.00
Urban	Non-public	ZNS	Standard	21.00	0CRO6	Male	Does not qualify	60.00	68.00
Urban	Non-public	ZNS	Standard	21.00	1QMDI	Male	Does not qualify	60.00	70.00
Urban	Non-public	ZNS	Standard	21.00	3CFUK	Male	Does not qualify	57.00	66.00
Urban	Non-public	ZNS	Standard	21.00	447O0	Male	Does not qualify	57.00	70.00
Urban	Non-public	ZNS	Standard	21.00	4IDEM	Female	Qualifies for reduced/free lunch	56.00	65.00
Urban	Non-public	ZNS	Standard	21.00	5UQND	Male	Does not qualify	58.00	67.00

This is the excel file where all the data inputs are provided to be inputted in the SPSS software for further analysis of data. The column L was not created in the default mode of the excel file, but it will be required to calculate the values through SPSS which is why this column was created. To create the column, simple excel formula known as IF formula was used in order to convert all the “Lunch” column input into a numerical data. The numerical data only needs to have a 0 and 1 input which will be useful in the SPSS software. The formula that was written here is

=IF(I2="Does not qualify",0,1)

Using this formula, the cell of the “Lunch” column that is written as Does not quality got turned into 0, and the other response that was found in the column was assigned 1 (Schmuller, 2013). Once this column was created, the rest of the tasks were done using the SPSS which was helpful for further analysis of the data.

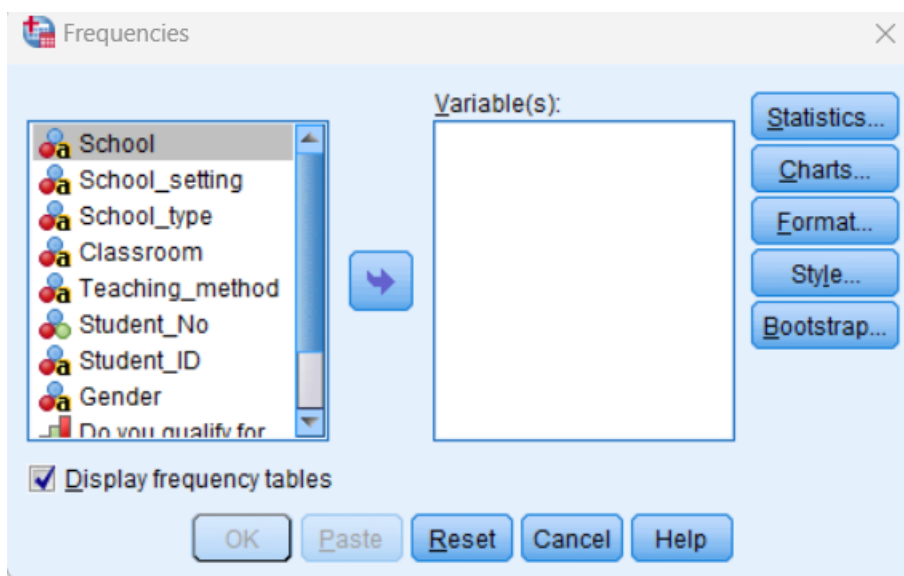
All the data were copied in the SPSS file which provided the view that is provided above to start the analysis in the SPSS. However, there are several tasks that needs to be completed in the SPSS to get the required results.

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	School	String	6	0		None	None	6	Right	Nominal	Input
2	School_sett...	String	14	0		None	None	14	Right	Nominal	Input
3	School_type	String	11	0		None	None	11	Right	Nominal	Input
4	Classroom	String	9	0		None	None	9	Right	Nominal	Input
5	Teaching_m...	String	15	0		None	None	15	Right	Nominal	Input
6	Student_No	Numeric	10	0		None	None	10	Right	Nominal	Input
7	Student_ID	String	10	0		None	None	10	Right	Nominal	Input
8	Gender	String	6	0		{1, Male}...	None	6	Right	Nominal	Input
9	Lunch	Numeric	32	0	Do you qualify f...	{0, No}...	None	32	Right	Ordinal	Input
10	Pretest	Numeric	7	0		None	None	7	Right	Scale	Input
11	Posttest	Numeric	8	0		None	None	8	Right	Scale	Input
12											
13											
14											
15											
16											
17											
18											
19											
20											
21											
22											
23											
24											
25											
26											
27											
28											
29											
Data View Variable View											

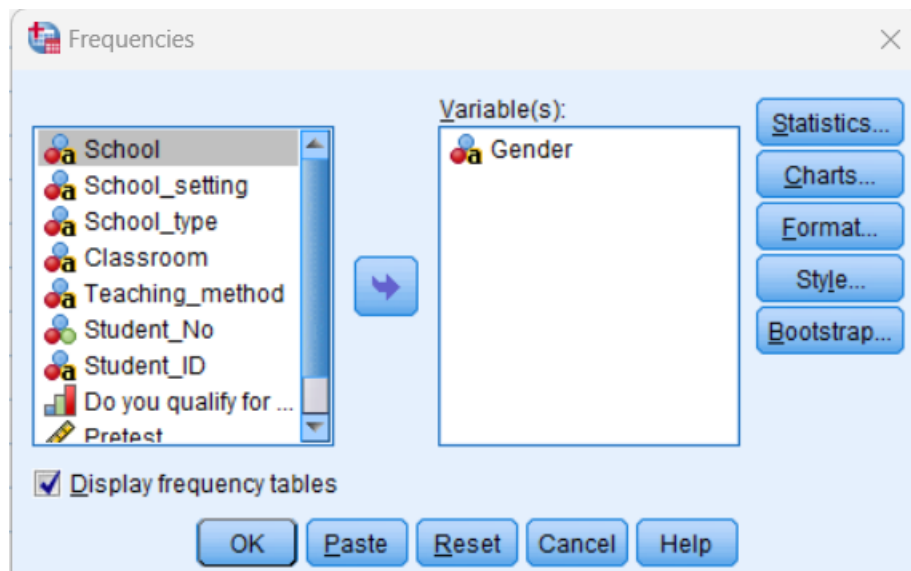
This is the Variable View tab available in the SPSS which will be useful to work in the SPSS and the first task that was done is changing the name of each column according to the name provided in the Excel file (Griffith, 2009). Once done, the type of each of the variable was assigned where the type of the data that is inputted in the relevant columns were considered. For example, the name of the school was words, thus the type of the variable was chosen as string, and all other data which consists words or letters are set similarly (Griffith, 2009). The columns that consists of only numerical value are assigned to be numerical type so that relevant calculations can be done. The values were assigned for Gender and Lunch as they will only consists of either Yes or No answer, thus 0 was assigned to be no, and 1 was assigned to be Yes. In terms of Gender, similar task was performed where Male was assigned as 1 and Female was assigned as 0 (Griffith, 2009). The Measure needed some tweak as well because the Pretest and Posttest were basically marks which can be greater or lower from one another, so these were set as Scale (Griffith, 2009). Having lunch can be a positive and bigger reward than not having it, so it was assigned as Ordinal (Griffith, 2009).

School	School_setting	School_type	Classroom	Teaching_method	Student_No	Student_ID	Gender	Lunch	Pretest	Posttest	VBF	VBF
ANKYI	Urban	Non-public	60L	Standard	20	2FH3	Female	0	62	72		
ANKYI	Urban	Non-public	60L	Standard	20	3JIVH	Female	0	66	79		
ANKYI	Urban	Non-public	60L	Standard	20	3XOWE	Male	0	64	76		
ANKYI	Urban	Non-public	60L	Standard	20	55600	Female	0	61	77		
ANKYI	Urban	Non-public	60L	Standard	20	74LOE	Male	0	64	76		
ANKYI	Urban	Non-public	60L	Standard	20	7YZ08	Female	0	66	74		
ANKYI	Urban	Non-public	60L	Standard	20	9KMZD	Male	0	63	75		
ANKYI	Urban	Non-public	60L	Standard	20	9USQK	Female	0	63	72		
ANKYI	Urban	Non-public	60L	Standard	20	CSSQP	Male	0	64	77		
ANKYI	Urban	Non-public	60L	Standard	20	D6HT8	Female	0	61	72		
ANKYI	Urban	Non-public	60L	Standard	20	DZMKU	Male	0	61	73		
ANKYI	Urban	Non-public	60L	Standard	20	FH7B9	Male	0	64	74		
ANKYI	Urban	Non-public	60L	Standard	20	J19VG	Male	0	66	78		
ANKYI	Urban	Non-public	60L	Standard	20	JQM2W	Female	0	60	71		
ANKYI	Urban	Non-public	60L	Standard	20	MEUC4	Female	0	64	77		
ANKYI	Urban	Non-public	60L	Standard	20	R4U8H	Male	0	64	73		
ANKYI	Urban	Non-public	60L	Standard	20	TH7KI	Male	0	63	70		
ANKYI	Urban	Non-public	60L	Standard	20	U1FV7	Male	0	67	73		
ANKYI	Urban	Non-public	60L	Standard	20	WC5I6	Female	0	63	71		
ANKYI	Urban	Non-public	60L	Standard	20	ZBQ4T	Male	0	64	73		
ANKYI	Urban	Non-public	ZHS	Standard	21	0CR06	Male	0	60	68		
ANKYI	Urban	Non-public	ZHS	Standard	21	1QMDI	Male	0	60	70		
ANKYI	Urban	Non-public	ZHS	Standard	21	3CFUK	Male	0	57	66		
ANKYI	Urban	Non-public	ZHS	Standard	21	44700	Male	0	57	70		
ANKYI	Urban	Non-public	ZHS	Standard	21	4IDEM	Female	1	56	65		
ANKYI	Urban	Non-public	ZHS	Standard	21	5UQNP	Male	0	58	67		

This is the final form of the SPSS data where the last column was copied and pasted in the Lunch column, and this is how it looked at the end with some necessary tweaks (Sujarweni and Utami, 2019). At this point, the SPSS file was ready for the analysis and anticipated results.



By clicking Analyze > Descriptive Statistics> Frequencies, the tab shown above was found which is the tab where related tasks was performed to get the respected results (Sujarweni and Utami, 2019).



First of all, only the Gender was clicked and sent to the right side and clicking OK initiated the analysis which provided the table with the Gender related data (Sujarweni and Utami, 2019).

Gender					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Female	1056	49.5	49.5	49.5
	Male	1077	50.5	50.5	100.0
	Total	2133	100.0	100.0	

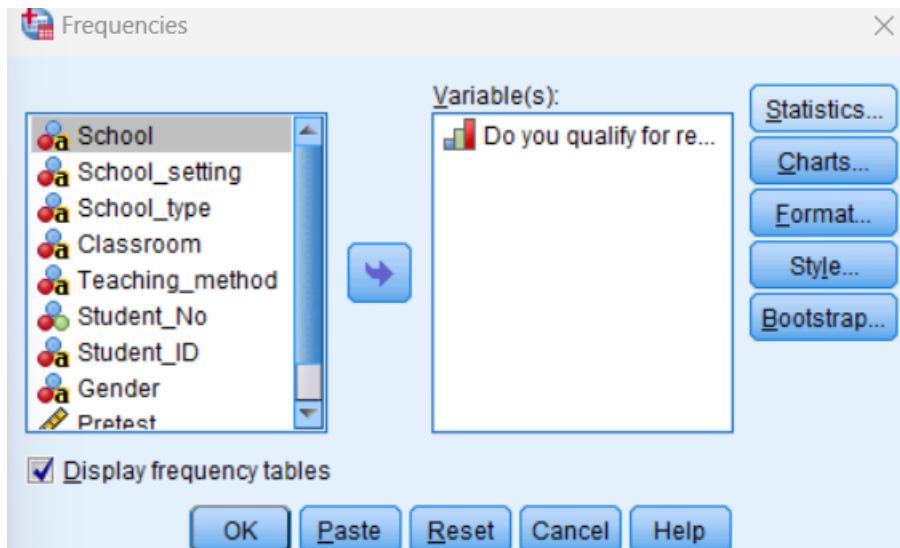
From the table that SPSS created, it is found that the number of Female in the class was 1056 and the number of Male present in the class was 1077.

By clicking Analyze > Descriptive Statistics> Frequencies again, removing Gender from the right side and selecting “Do you qualify for reduced/free Lunch?” on the right side provided with the relevant result (Sujarweni and Utami, 2019).

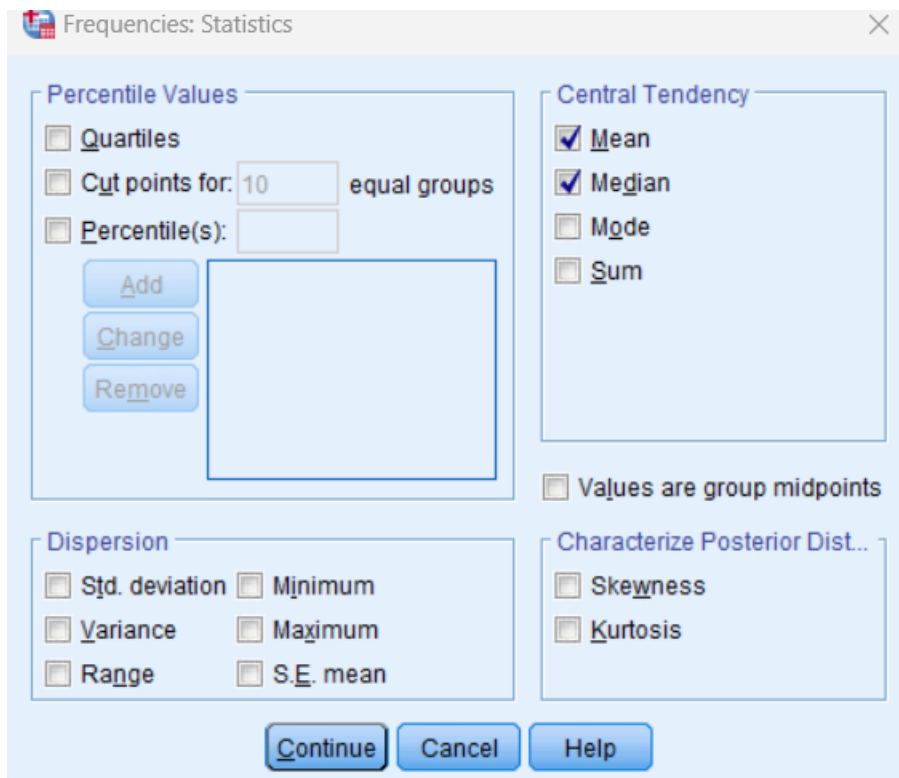
Lunch					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	1215	57.0	57.0	57.0
	1	918	43.0	43.0	100.0
Total		2133	100.0	100.0	

It is found that among 2133 students, 1215 students do not qualify for reduced/free lunch, but 918 students do qualify for reduced/free lunch.

Now the task is to create the Mean and Median of the students who qualify or does not qualify for free lunch, and to get the releavent data, the Analyze > Descriptive Statistics> Frequencies tab was selected again.



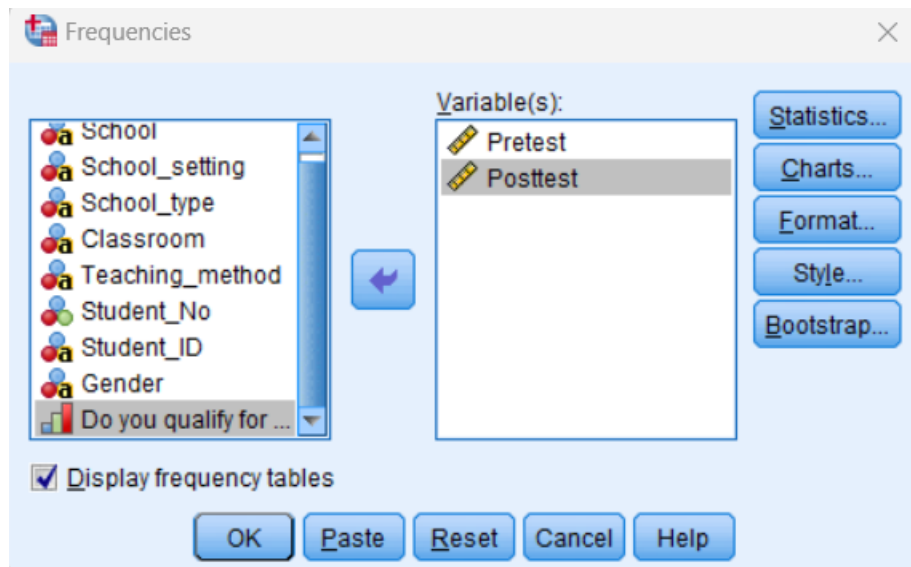
In this tab, the “Do you qualify for reduced/free lunch?” was already on the right side, so the top right option called Statistics was pressed.



By clicking Statistics, this new tab came up where the Mean and Median options were ticked to get these values and Continue was pressed (Rajathi and Chandran, 2019).

Statistics		
Lunch		
N	Valid	2133
	Missing	0
Mean		.43
Median		.00

By doing so, this result was found which shows mean of the students who does or does not qualify for Lunch is 0.43 and there are no median for these data (Rajathi and Chandran, 2019).



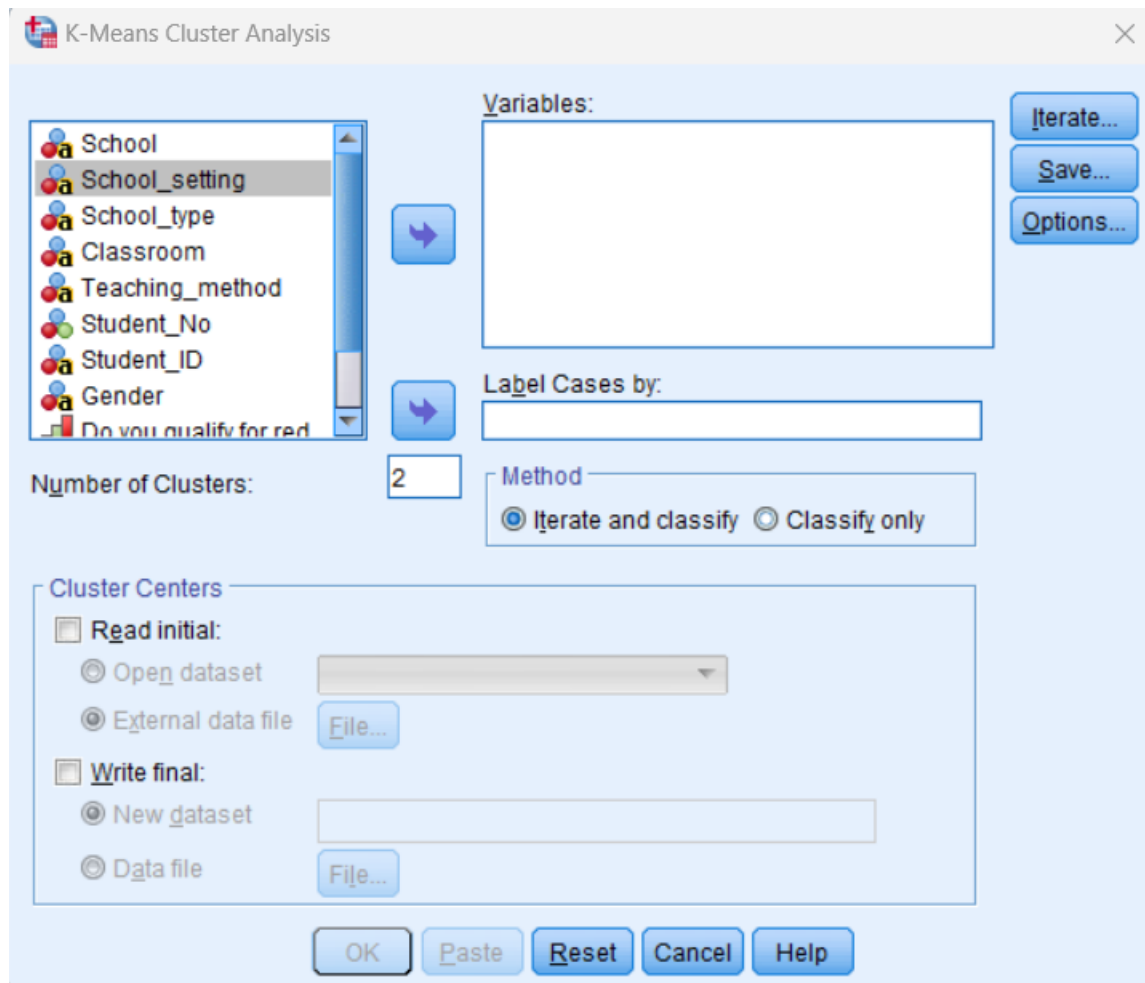
Now the task was to find the Mean and Median of the students based on the Pretest and Posttest marks, so Analyze > Descriptive Statistics> Frequencies tab was opened where the Pretest and Posttest were sent on the right side. From Statistics option, Mean and Median was chosen to get the relevant results (Rajathi and Chandran, 2019).

Statistics			
		Pretest	Posttest
N	Valid	2133	2133
	Missing	0	0
Mean		54.9559	67.1022
Median		56.0000	68.0000

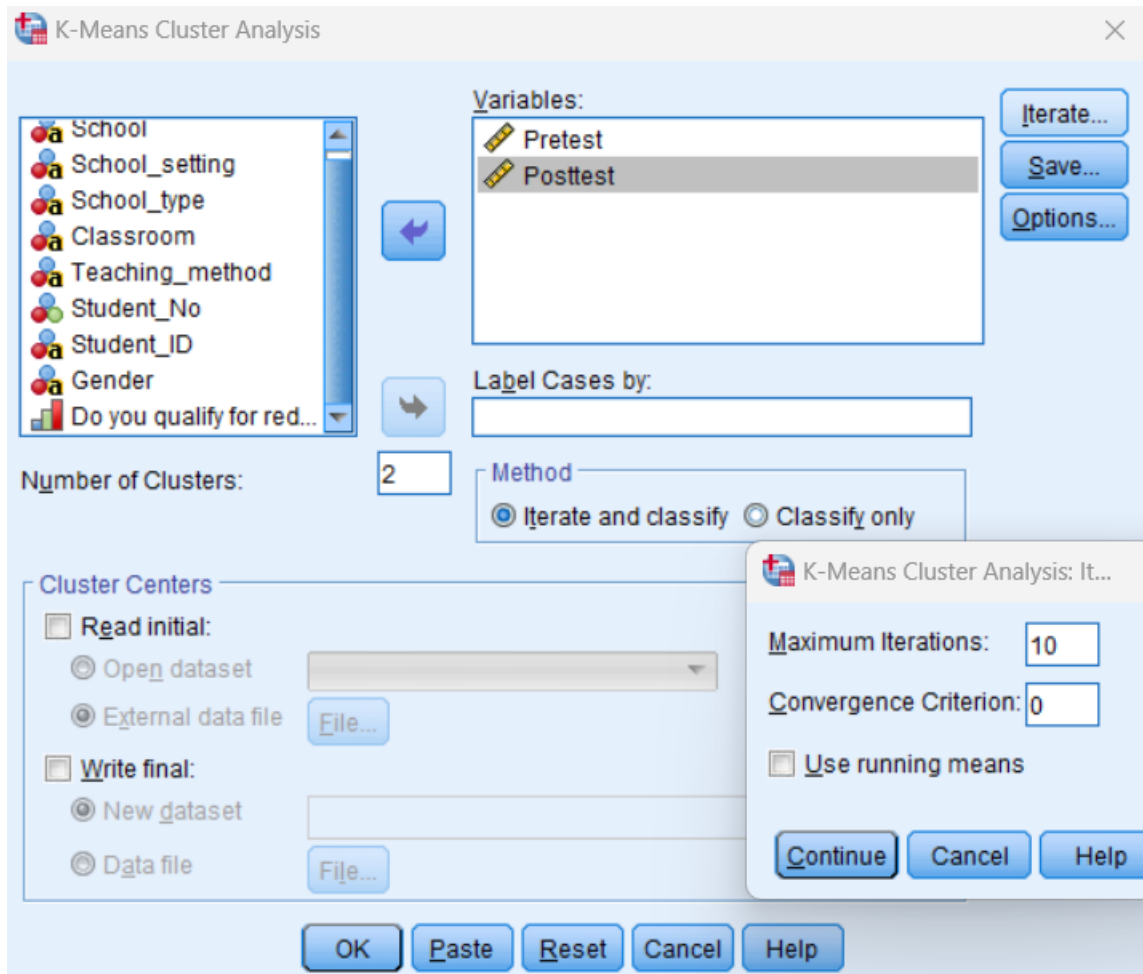
As shown in the above table, the mean and median value for Pretest marks were found to be 54.9559 and 56.0000 respectively, where as the Posttest mean and median was found to be 67.1022 and 68.0000 respectively.

The difference between the Mean values of Pretest and Posttest is found to be = $67.1022 - 54.9559 = 12.1463$

The difference between the Median values of Pretest and Posttest is found to be = $68.00 - 56.00 = 12.00$



At this stage, by clicking Analyze>Classify>K means Cluster, this tab was created which provides the opportunity to run the K Means Cluster Analysis on the student's data (Rajathi and Chandran, 2019).



By selecting Pretest and Posttest on the right tab, and choosing iterate option on the most right corner provided the opportunity to choose the maximum iterations which was chosen to be 10. By clicking continue and OK, SPSS have performed the required iterations and provided with the results (Meyers, Gamst and Guarino, 2013).

Initial Cluster Centers

	Cluster	
	1	2
Pretest	23.00	93.00
Posttest	35.00	100.00

Iteration History^a

Iteration	Change in Cluster Centers	
	1	2
1	30.191	33.606
2	.885	.791
3	.611	.532
4	.212	.181
5	.227	.200
6	.150	.128
7	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 7. The minimum distance between initial centers is 95.525.

Final Cluster Centers

	Cluster	
	1	2
Pretest	43.05	65.32
Posttest	54.70	77.90

Number of Cases in each Cluster

Cluster	1	993.000
	2	1140.000
Valid		2133.000
Missing		.000

Different Data Mining Methods is helping in different sectors to make the best use of the technologies in hand.

1. Associated Rule Mining is a method of data mining that helps to discover relationship that exists within a set of data (Aggarwal, 2015). Supermarkets are using this data mining to increase sales such a customer who is willing to buy baby food may also buy diapers.
2. Classification is another data mining technique where the data are categorized based on some sort of feature or characteristics that these data possess (Aggarwal, 2015). Email inbox is being spam free as the email can identify the spam emails by reading the mail.
3. Clustering is the process that helps to create groups among the data which are similar to each other and the grouping is based on the characteristics that these data possess (Aggarwal, 2015). Ecommerce websites can advertise by identifying the preference of the customers and grouping them to be in the same group for the same type of advertisements.
4. Regression analysis is a well known data mining method that helps to understand and identify relationship between the dependent and independent variables that exists within a dataset (Aggarwal, 2015). Risk assessments at the insurance companies are being done using the regression analysis by analyzing the age, gender, location, etc.
5. Unusual pattern can often be found in a set of data, and anomaly detection data mining method helps to identify these (Aggarwal, 2015). Fraud are being prevented by identifying the unusual transactions using this data mining.

Several Text Mining Methods are available that is helping in different sectors to make the best use of the technologies in hand.

1. Sentiment can be expressed in various ways, and a text mining method known as sentiment analysis helps to identify what is expressed through it (Jo, 2019). Social media is using it to identify whether the review of a product is positive or negative which helps to create a better opportunity to boost the better products.
2. Name of something can needs to be identified for various purposes which is done through the Named Entity Recognition (Jo, 2019). News organizations can use this technique to classify a huge number of name of people or location with this data mining.

3. When a method is used with the sole purpose to identify topics present in a set of data can be called topic modelling (Jo, 2019). The emerging trends of people can easily be identified through the social media using it.
4. Automation is often used in statistics and using it to categorize texts can be called text classification (Jo, 2019). Customer support centers are enjoying the advantages of this text mining method by providing support tickets automatically based on customer's message.
5. Automation of identifying a set of information and extracting it without manual or with minimal manual input is known as information extraction (Jo, 2019). Time and efforts required during a manual document analysis for a legal procedure is being saved through this text mining method.

	Advantages	Disadvantages	Practical Examples
SPSS	Comprehensive statistical analysis capabilities (Milovanović and Perišić, 2020)	Steeper learning curve for complex analyses	Analyzing survey data to identify significant factors influencing customer satisfaction
	Advanced data manipulation and transformation	Relatively expensive compared to other software options	Conducting a cluster analysis to segment customers based on purchasing behaviour
	Robust handling of large datasets	Requires knowledge of statistical concepts and techniques (Milovanović and Perišić, 2020)	Performing regression analysis to predict sales based on advertising expenditure
	Provides graphical visualization tools (Milovanović and Perišić, 2020)	Limited flexibility for customization	Examining survey responses to conduct factor analysis and extract underlying latent variables
Excel	Familiar and widely available (Fraser and Fraser, 2013)	Limited statistical analysis capabilities	Tracking monthly sales and calculating basic summary statistics

	Ease of data entry and basic calculations (Fraser and Fraser, 2013)	Limited capacity to handle large datasets	Creating a budget spreadsheet and performing simple financial calculations
	Flexible for customizing and formatting data	Prone to errors and inconsistencies due to manual entry	Creating a pivot table to analyze sales by product category
	Integration with other Microsoft Office tools	Lack of advanced data manipulation features	Organizing and analyzing survey responses with basic charts and graphs

References

- Aggarwal, C.C., 2015. *Data mining: the textbook* (Vol. 1). New York: springer.
- Camões, J., 2016. Data at work: Best practices for creating effective charts and information graphics in Microsoft Excel. New Riders.
- Carlberg, C., 2014. Statistical analysis: Microsoft excel 2013. Que Publishing.
- Dierenfeld, H. and Merceron, A., 2012. Learning analytics with excel pivot tables.
- Fraser, C. and Fraser, 2013. Business statistics for competitive advantage with Excel 2013. Springer.
- Griffith, A., 2009. *SPSS for Dummies*. John Wiley & Sons.
- Jo, T., 2019. *Text mining* (Vol. 45). Cham, Switzerland: Springer International Publishing.
- Meyers, L.S., Gamst, G.C. and Guarino, A.J., 2013. *Performing data analysis using IBM SPSS*. John Wiley & Sons.
- Milovanović, M. and Perišić, J., 2020. Advantages and limitations of using SPSS in teaching statistics. *MEFkon 2020 INNOVATION AS AN INITIATOR OF THE DEVELOPMENT "INNOVATIONS IN THE FUNCTION OF DEVELOPMENT"*, p.274.
- Rajathi, A. and Chandran, P., 2019. *SPSS for You*. MJP Publisher.
- Schmuller, J., 2013. Statistical analysis with excel for dummies. John wiley & sons.
- Sujarweni, V.W. and Utami, L.R., 2019. *The master book of SPSS*. Anak Hebat Indonesia.