

Data Handling and Business Intelligence
LCBB5000

Component 2

Student Name: Ivaylo Bakalov

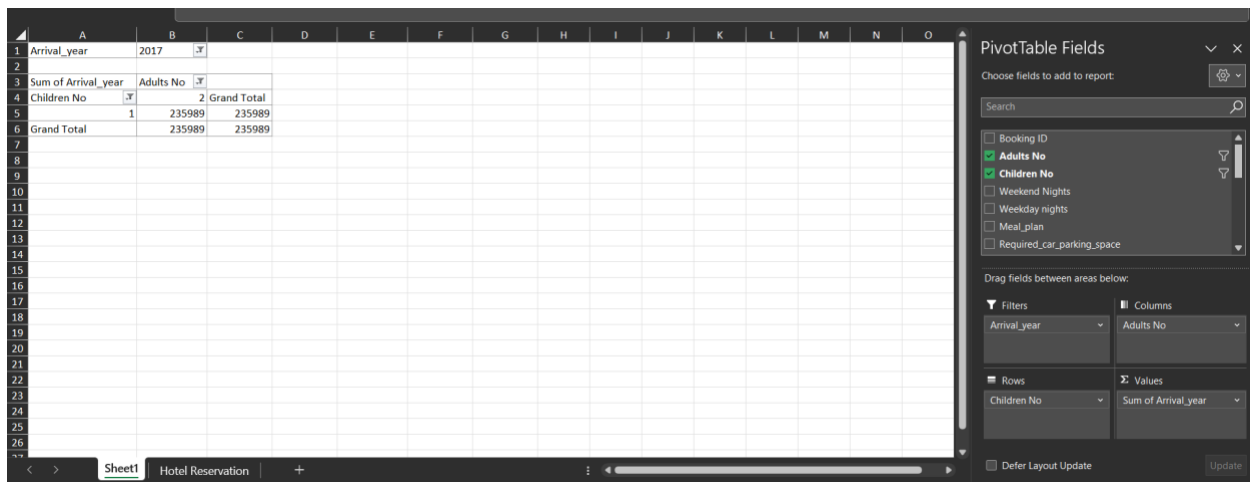
Student ID: 2128950

Lecturer name: Xavier Matieni

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
27541	INN27540	2	0	2	7	Meal Plan 1	0	Room_Type 1	199	2019	10	1	Online	0	0	0	94.5	0	Canc
27542	INN27541	2	0	2	2	Meal Plan 1	0	Room_Type 1	40	2019	3	5	Online	0	0	0	83.9	1	Not_Ca
27543	INN27542	2	0	1	4	Meal Plan 1	0	Room_Type 1	148	2019	12	30	Offline	0	0	0	67	0	Not_Ca
27544	INN27543	2	0	0	1	Meal Plan 1	1	Room_Type 1	1	2019	1	20	Online	0	0	0	111	1	Not_Ca
27545	INN27544	2	0	1	2	Meal Plan 2	0	Room_Type 1	3	2019	8	15	Online	0	0	0	228.33	1	Not_Ca
27546	INN27545	3	0	0	1	Meal Plan 1	0	Room_Type 4	26	2019	5	19	Online	0	0	0	177	2	Not_Ca
27547	INN27546	2	0	2	1	Meal Plan 1	0	Room_Type 1	2	2019	10	4	Online	0	0	0	114	0	Not_Ca
27548	INN27547	2	0	1	2	Not Selected	0	Room_Type 1	28	2019	12	2	Online	0	0	0	88	1	Not_Ca
27549	INN27548	2	0	1	2	Meal Plan 1	0	Room_Type 1	81	2019	4	29	Offline	0	0	0	75	0	Not_Ca
27550	INN27549	2	0	0	3	Meal Plan 1	0	Room_Type 1	93	2019	4	12	Online	0	0	0	96.3	3	Not_Ca
27551	INN27550	1	0	0	4	Meal Plan 1	0	Room_Type 4	8	2019	4	5	Online	0	0	0	92.17	1	Not_Ca
27552	INN27551	1	0	0	4	Meal Plan 1	0	Room_Type 1	5	2019	2	24	Corporate	0	0	0	80	0	Not_Ca
27553	INN27552	2	0	0	1	Meal Plan 1	0	Room_Type 1	105	2019	4	6	Offline	0	0	0	75	0	Canc
27554	INN27553	2	0	0	3	Meal Plan 1	0	Room_Type 1	271	2019	9	21	Offline	0	0	0	101.33	1	Canc
27555	INN27554	1	0	1	1	Meal Plan 1	0	Room_Type 1	5	2019	10	19	Corporate	0	0	0	80	0	Not_Ca
27556	INN27555	3	0	1	0	Meal Plan 1	0	Room_Type 4	89	2019	7	17	Online	0	0	0	137.7	2	Not_Ca
27557	INN27556	2	0	2	1	Meal Plan 1	0	Room_Type 1	11	2019	9	11	Offline	0	0	0	95	0	Not_Ca
27558	INN27557	1	0	2	2	Meal Plan 1	0	Room_Type 1	88	2019	7	18	Online	0	0	0	68	1	Canc
27559	INN27558	2	0	1	3	Meal Plan 2	0	Room_Type 1	210	2019	10	8	Offline	0	0	0	86	0	Canc
27560	INN27559	2	0	0	1	Meal Plan 1	0	Room_Type 1	54	2019	11	24	Complementar	0	0	0	0	1	Not_Ca
27561	INN27560	1	0	1	0	Meal Plan 1	0	Room_Type 1	0	2019	2	29	Corporate	1	0	10	65	1	Not_Ca
27562	INN27561	2	2	2	5	Meal Plan 1	0	Room_Type 6	217	2019	8	7	Online	0	0	0	167.95	0	Canc
27563	INN27562	1	0	2	3	Not Selected	0	Room_Type 1	6	2019	11	24	Online	0	0	0	68.67	0	Not_Ca
27564	INN27563	2	0	2	2	Meal Plan 1	0	Room_Type 1	21	2019	9	5	Online	0	0	0	105	3	Not_Ca
27565	INN27564	2	0	0	3	Meal Plan 1	0	Room_Type 1	19	2019	2	18	Online	0	0	0	82.83	0	Not_Ca
27566	INN27565	2	0	0	2	Meal Plan 2	0	Room_Type 1	102	2019	10	16	Offline	0	0	0	101.5	0	Canc

This file has been obtained which contains the data related to the hotel and the average price and arrival year are the two columns that will be in focus throughout this part. The data that are provided in the datasheet ranges from 2017 to 2022 but the data are not sorted in this datasheet. One way of calculating the average price each year can be shorting based on the years, and finding the data related to each year separately, but the huge amount of data that is provided in this file will make it difficult to do it manually. So, an innovative strategy will be required to achieve the required goal. Microsoft Excel provides several methods that can be utilized in this file in order to obtain the required result, and as the first method, PivotTable and PivotChart can be helpful to obtain the required result (Alexander and Kusleika, 2016).

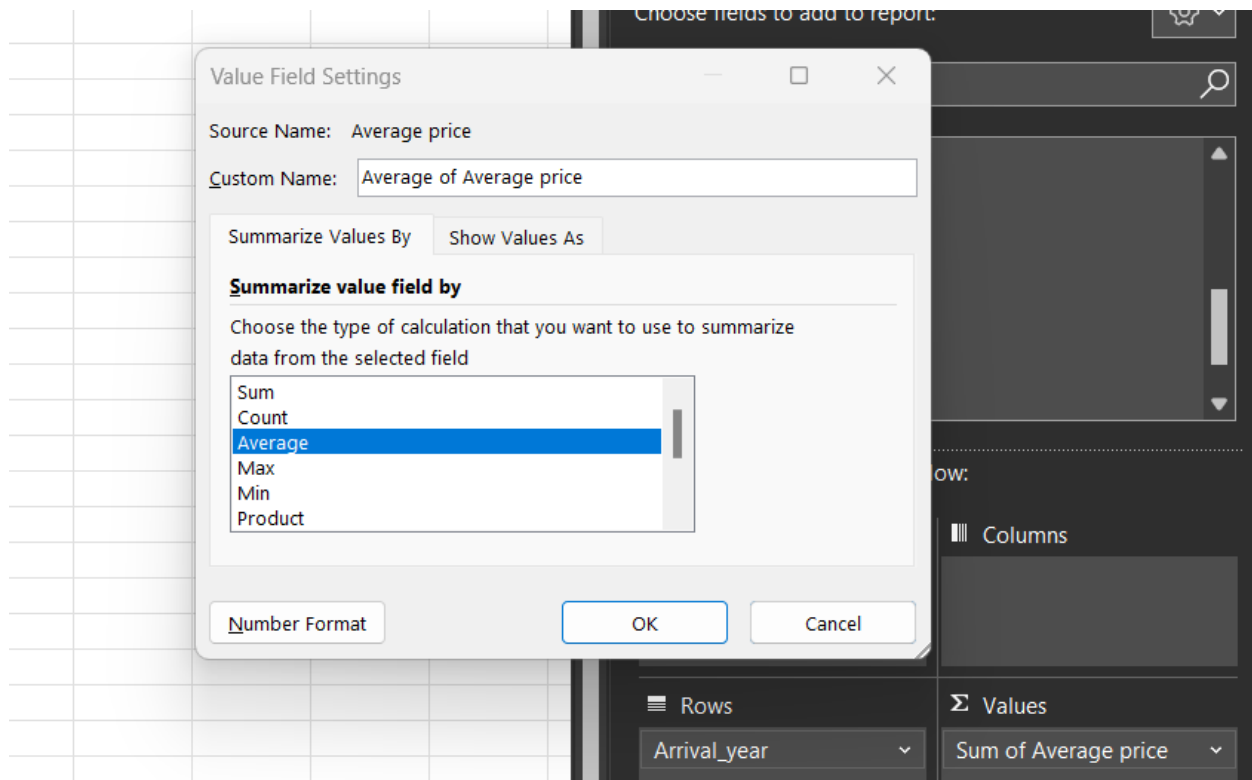
PivotTable and PivotChart is a feature provided by Microsoft Excel where the data analysis can be done easily (Alexander and Kusleika, 2016). Several fields are provided by the software under this feature, and simply selecting the column header that needs to be analyzed is all that is needed to do. Microsoft Excel utilizes its pre recorded functions to obtain the required results based on the fields selected under PivotTable and PivotCharts.



This is the Sheet1 which basically is the PivotTable, and PivotChart can be included in it. Several fields are available in the PivotTable that can help to calculate the required values and obtain the required results (Guerrero et al., 2019).

	A	B	C	D	E	F
1	Arrival_year	2017				
2						
3	Sum of Arrival_year	Adults No				
4	Children No	2	Grand Total			
5	1	235989	235989			
6	Grand Total	235989	235989			
7						
8						
9						

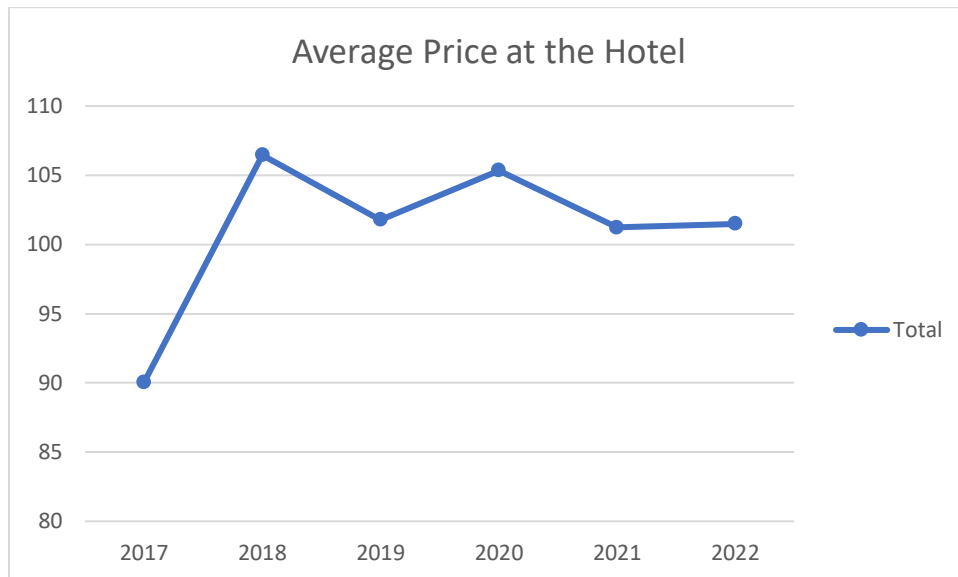
This is the value that is being shown in the PivotTable by default which consists of other values apart from the values that are needed so these needs to change. Also, it is important to ensure that the PivotTable is showing the correct result, and to ensure that, refreshing the sheet can be an effective option (Apte, 2009). Under Data option of Microsoft Excel, it provides the option to Refresh the dataset which is utilized in this case which shows the value has changed.



This is the option which has been chosen in order to obtain the required value from this dataset.

	A	B	C	D	E	F	G
1							
2							
3	Average of Average price						
4	Arrival_year	Total					
5	2017	90.0498541					
6	2018	106.4527274					
7	2019	101.7812994					
8	2020	105.339899					
9	2021	101.2333333					
10	2022	101.499242					
11	Grand Total	103.4235391					
12							
13							
14							
15							
16							
17							
18							
19							
20							
21							
22							

By applying average of the average price values, these are the final values that has been obtained from the PivotTable, and Chart can be utilized at this stage to present the data graphically. Just selecting a value from the PivotTable, and selecting insert option available in the Microsoft Excel, Charts options can be obtained (Pace, 2012). From the charts option, several chart are available, and selecting one of the charts will create the chart automatically based on the data found on the PivotTable.



This is the line chart with markers that has been drawn by Microsoft Excel based on the PivotChart data which better represents the data graphically.

The average value that has been found needs to be checked to verify that the value that has been obtained are correct, and to do so, the method is to evaluate the data using a different method and obtain the results. If the same result is obtained, then it will be ensured that the value that has been obtained is the right one.

To follow a different approach to find the average price from the hotel data, several Microsoft Formula is available. Among them, IF is a formula that can help to obtain the results easily, so IF formula will be utilized (Syrstad and Jelen, 2011).

=IF(J2=2017,Q2,0)

This is the IF formula that has been written under T column which includes values from two different columns to compare the data (Syrstad and Jelen, 2011). The condition provided in the formula is to check the value of J column which is the Arrival year, and if the value found there is

2017, then Excel will go to Q2 cell which basically is the Average Price value and put it as output of the formula. If the value is not equal to 2017, then the formula will output 0 that means the value in the J2 cell is not 2017 (Syrstad and Jelen, 2011).

	J	K	L	M	N	O	P	Q	R	S	T
	Arrival_ year	Arrival_ month	Arrival_ date	Market_type	Repeated _guest	Previous Cancellantion	Previous bookings	Average price	Special requests	Booking status	2017 Price
1											
2	2017	10	2	Offline	0	0	0	65	0	Not_Canceled	65
3	2018	11	6	Online	0	0	0	106.68	1	Not_Canceled	0
4	2018	2	28	Online	0	0	0	60	0	Canceled	0
5	2018	5	20	Online	0	0	0	100	0	Canceled	0
6	2018	4	11	Online	0	0	0	94.5	0	Canceled	0
7	2018	9	13	Online	0	0	0	115	1	Canceled	0
8	2017	10	15	Online	0	0	0	107.55	1	Not_Canceled	107.55
9	2018	12	26	Online	0	0	0	105.61	1	Not_Canceled	0
10	2018	7	6	Offline	0	0	0	96.9	1	Not_Canceled	0
11	2018	10	18	Online	0	0	0	133.44	3	Not_Canceled	0
12	2018	9	11	Online	0	0	0	85.03	0	Not_Canceled	0
13	2018	4	30	Online	0	0	0	140.4	1	Not_Canceled	0
14	2018	11	26	Online	0	0	0	88	0	Canceled	0
15	2018	11	20	Online	0	0	0	90	2	Canceled	0
16	2019	10	20	Online	0	0	0	94.5	2	Not_Canceled	0
17	2018	6	15	Online	0	0	0	115	1	Canceled	0
18	2017	10	5	Offline	0	0	0	96	0	Not_Canceled	96
19	2017	8	10	Online	0	0	0	96	1	Not_Canceled	96
20	2017	10	30	Online	0	0	0	65	0	Canceled	65
21	2019	10	4	Offline	0	0	0	72	0	Not_Canceled	0

After entering the formula on T2 cell, the formula has been applied on each cell of the T column. Instead of manually inputting the formula on each cell, Microsoft Excel provides a shortcut Ctrl+D which instructs Microsoft Excel to copy the formula downward so that the value for the whole column can be obtained within the minimal time. As this step was done, the values were found for each cell on the respective column. However, later during the time when the average price was being calculated, the obtained result was found to be wrong. By investigating the reason behind faulty values, it was found that, the formula was counting “0” as a number in Microsoft Excel. So, this issue needed to be solved before approaching to find the results (Abdulezer, 2003).

By investigating a solution, it was found that Microsoft Excel does not include letters or texts in the average formula and avoids the cell. As a result, the average formula can obtain the result that is required for this task (Abdulezer, 2003). So, the formula used to find the average price of a year, this formula needed some modifications.

=IF(J2=2017,Q2,"Not 2017")

This is the final formula that was written to find the average price of the year 2017. This formula has been able to provide the required result as the average formula was providing the correct answers as needed for this task.

Average of Year 2017												
	N	O	P	Q	R	S	T	U	V	W	X	Y
.type	Repeated _guest	Previous Cancellation	Previous bookings	Average price	Special requests	Booking status	2017 Price	2018 Price	2019 Price	2020 Price	2021 Price	2022 Price
ie	0	0	0	65	0	Not_Canceled	65	Not 2018	Not 2019	Not 2020	Not 2021	Not 2022
ie	0	0	0	106.68	1	Not_Canceled	Not 2017	106.68	Not 2019	Not 2020	Not 2021	Not 2022
ie	0	0	0	60	0	Canceled	Not 2017	60	Not 2019	Not 2020	Not 2021	Not 2022
ie	0	0	0	100	0	Canceled	Not 2017	100	Not 2019	Not 2020	Not 2021	Not 2022
ie	0	0	0	94.5	0	Canceled	Not 2017	94.5	Not 2019	Not 2020	Not 2021	Not 2022
ie	0	0	0	115	1	Canceled	Not 2017	115	Not 2019	Not 2020	Not 2021	Not 2022
ie	0	0	0	107.55	1	Not_Canceled	107.55	Not 2018	Not 2019	Not 2020	Not 2021	Not 2022
ie	0	0	0	105.61	1	Not_Canceled	Not 2017	105.61	Not 2019	Not 2020	Not 2021	Not 2022
ie	0	0	0	96.9	1	Not_Canceled	Not 2017	96.9	Not 2019	Not 2020	Not 2021	Not 2022
ie	0	0	0	133.44	3	Not_Canceled	Not 2017	133.44	Not 2019	Not 2020	Not 2021	Not 2022
ie	0	0	0	85.03	0	Not_Canceled	Not 2017	85.03	Not 2019	Not 2020	Not 2021	Not 2022
ie	0	0	0	140.4	1	Not_Canceled	Not 2017	140.4	Not 2019	Not 2020	Not 2021	Not 2022
ie	0	0	0	88	0	Canceled	Not 2017	88	Not 2019	Not 2020	Not 2021	Not 2022
ie	0	0	0	90	2	Canceled	Not 2017	90	Not 2019	Not 2020	Not 2021	Not 2022
ie	0	0	0	94.5	2	Not_Canceled	Not 2017	Not 2018	94.5	Not 2020	Not 2021	Not 2022
ie	0	0	0	115	1	Canceled	Not 2017	115	Not 2019	Not 2020	Not 2021	Not 2022
ie	0	0	0	96	0	Not_Canceled	96	Not 2018	Not 2019	Not 2020	Not 2021	Not 2022
ie	0	0	0	96	1	Not_Canceled	96	Not 2018	Not 2019	Not 2020	Not 2021	Not 2022
ie	0	0	0	65	0	Canceled	65	Not 2018	Not 2019	Not 2020	Not 2021	Not 2022
ie	0	0	0	72	0	Not_Canceled	Not 2017	Not 2018	72	Not 2020	Not 2021	Not 2022
ie	0	0	0	65	0	Canceled	65	Not 2018	Not 2019	Not 2020	Not 2021	Not 2022
ate	0	0	0	67	0	Not_Canceled	Not 2017	Not 2018	67	Not 2020	Not 2021	Not 2022

The same formula was applied with necessary modification on each cell and the obtained output is shown in the above picture. In addition to that, the same formula was used on the later columns, only change was writing 2018 instead of 2017 and so on. By doing so, the result that was found is shown in the above picture. The benefit of obtaining this result is the cells that are not relevant are containing texts (Bluttman, 2013). For example, the first cell of 2017 price column is related to year 2017, so the value is provided, but the bottom cell is not related to 2017, as a result, the cell is showing a text “Not 2017”. As the cell contains text, the average formula will skip this cell to calculate the average of the whole column. Also, the column heading is strategically named to be a text, so that the average formula can be applied on the whole column directly instead of getting faulty results. By doing so, the average value has provided required results and the task was completed.

=AVERAGE(T:T)

This is the average formula that includes the whole column that means including the empty cells and column headings. However, Microsoft Excel can track these cells and exclude these from the calculation, thus the required result was obtained (Bluttman, 2013).

Z	AA	AB	AC	AD	AE	
Average of Year 2017	Average of Year 2018	Average of Year 2019	Average of Year 2020	Average of Year 2021	Average of Year 2022	
90.04985	106.4527	101.7813	105.3399	101.2333	101.4992	

By applying the formula on each cell with relevant changes, these are the results that has been obtained. These results basically are the average price for each year which was the goal to be obtained through this specific task.



This is the line chart that has been obtained using the results found each year, and the values matches exactly as found in the Pivot Chart which validates the result.

Regarding trend of these data, it can be found that the minimum and maximum average price occurred in two consecutive years which are year 2017, and year 2018. The goal of this kind of analysis is to understand the data as well as forecast the data. It is found that the data raises and

gets reduced in the consecutive year, and this trend is present for each year that has been included in this analysis. In the year 2022, the average price has raised from year 2021, so it can be forecasted that the next year, the average price will drop (Bluttman, 2013). So, it will be wise to invest less in the next year, as the customers will not be booking the hotel as much (Bluttman, 2013). However, it is just a simple forecasting, so there might be other significant factors which are contributing to set the average price each year which needs to be investigated before making major decisions such as investments.

Read Excel File

C:\Users\DELL\Downloads\Students Score.xls (3) Data Handling 2.xlsx

Worksheet: Students Score [A1:K2134]

Range:

☒ Read variable names from first row of data

☒ Percentage of values that determine data type: 95

☒ Ignore hidden rows and columns

☐ Remove leading spaces from string values

☐ Remove trailing spaces from string values

Preview

	School	School...	School...	Classr...
1	ANKYI	Urban	Non-public	6OL
2	ANKYI	Urban	Non-public	6OL
3	ANKYI	Urban	Non-public	6OL
4	ANKYI	Urban	Non-public	6OL
5	ANKYI	Urban	Non-public	6OL
6	ANKYI	Urban	Non-public	6OL
7	ANKYI	Urban	Non-public	6OL

i Final data type is based on all data and can be different from the preview, which is based on the first 200 data rows. The preview displays only the first 500 columns.

OK Paste Reset Cancel Help

The SPSS provides a blank file when opened, but data can be imported at the SPSS so that the required calculation can be made. The student data file was downloaded which will be utilized in the SPSS to obtain the required results. So, selecting File on top left corner of SPSS, Import Data > Excel was chosen as the student file was provided in Excel format (Sujarweni and Utami, 2019). The folder where the data was downloaded were navigated and the excel file was selected to be imported at the SPSS. After selecting the file, the above shown tab came up where the location of the file as well as a sample of the output after importing the file can be seen. This way, the all the variables and data got inputted in the SPSS, but the variables play a big role at the SPSS. So, the variables need to be properly checked if those are set correctly.

	School	School_setting	School_type	Class	Teaching_method	Student_No	Student_ID	Gender	Lunch	Pretest	Posttest	var	var	var
1	ANKYI	Urban	Non-public	6OL	Standard	20	2FHT3	Female	Does not qualify	62	72			
2	ANKYI	Urban	Non-public	6OL	Standard	20	3JRVH	Female	Does not qualify	66	79			
3	ANKYI	Urban	Non-public	6OL	Standard	20	3XOWE	Male	Does not qualify	64	76			
4	ANKYI	Urban	Non-public	6OL	Standard	20	556O0	Female	Does not qualify	61	77			
5	ANKYI	Urban	Non-public	6OL	Standard	20	74LOE	Male	Does not qualify	64	76			
6	ANKYI	Urban	Non-public	6OL	Standard	20	7YZO8	Female	Does not qualify	66	74			
7	ANKYI	Urban	Non-public	6OL	Standard	20	9KMZD	Male	Does not qualify	63	75			
8	ANKYI	Urban	Non-public	6OL	Standard	20	9USQK	Female	Does not qualify	63	72			
9	ANKYI	Urban	Non-public	6OL	Standard	20	C55QP	Male	Does not qualify	64	77			
10	ANKYI	Urban	Non-public	6OL	Standard	20	D6HT8	Female	Does not qualify	61	72			
11	ANKYI	Urban	Non-public	6OL	Standard	20	DZMKU	Male	Does not qualify	61	73			
12	ANKYI	Urban	Non-public	6OL	Standard	20	FH7B9	Male	Does not qualify	64	74			
13	ANKYI	Urban	Non-public	6OL	Standard	20	J9VG	Male	Does not qualify	66	78			
14	ANKYI	Urban	Non-public	6OL	Standard	20	JQM2W	Female	Does not qualify	60	71			
15	ANKYI	Urban	Non-public	6OL	Standard	20	MEUC4	Female	Does not qualify	64	77			
16	ANKYI	Urban	Non-public	6OL	Standard	20	R4U8H	Male	Does not qualify	64	73			
17	ANKYI	Urban	Non-public	6OL	Standard	20	TH7K0	Male	Does not qualify	63	70			
18	ANKYI	Urban	Non-public	6OL	Standard	20	U1FV7	Male	Does not qualify	67	73			
19	ANKYI	Urban	Non-public	6OL	Standard	20	WC596	Female	Does not qualify	63	71			
20	ANKYI	Urban	Non-public	6OL	Standard	20	ZBQ4T	Male	Does not qualify	64	73			
21	ANKYI	Urban	Non-public	ZHS	Standard	21	0CRO6	Male	Does not qualify	60	68			
22	ANKYI	Urban	Non-public	ZHS	Standard	21	1QMDI	Male	Does not qualify	60	70			
23	ANKYI	Urban	Non-public	ZHS	Standard	21	3CFUK	Male	Does not qualify	57	66			
24	ANKYI	Urban	Non-public	ZHS	Standard	21	44T00	Male	Does not qualify	57	70			
25	ANKYI	Urban	Non-public	ZHS	Standard	21	4IDEM	Female	Qualifies for reduced/free lunch	56	65			

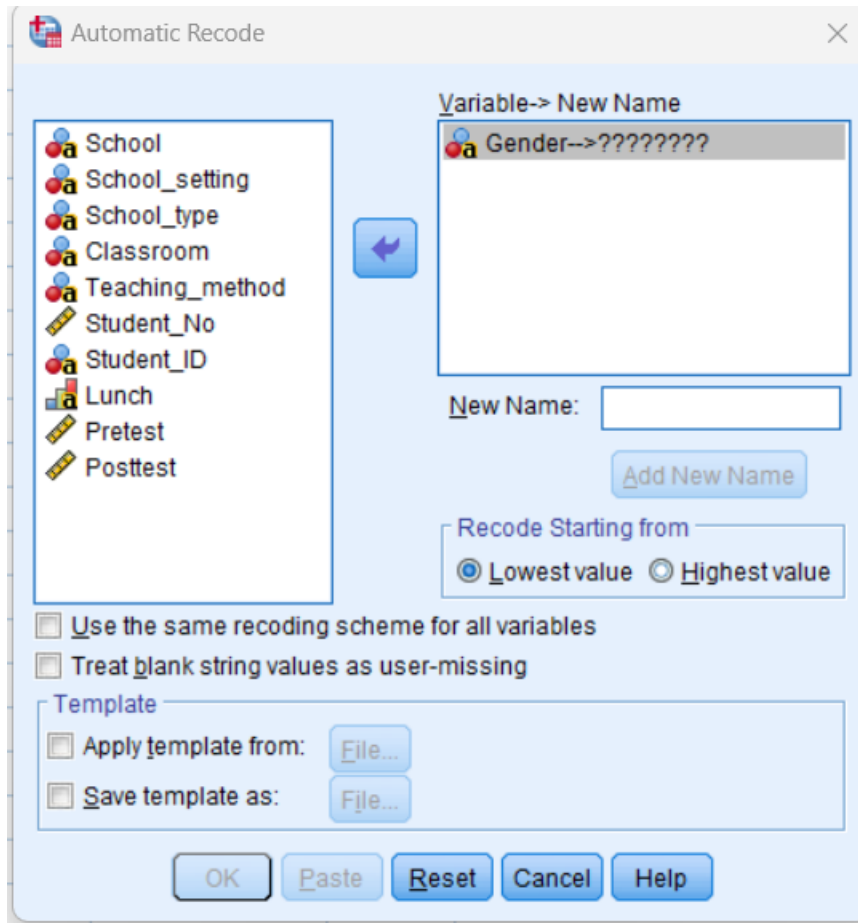
This is the output that was obtained by importing the Excel file at the SPSS software where all the columns has been named properly as set at the Microsoft Excel, and all the values has been inputted from Excel to SPSS.

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	School	String	5	0		None	None	5	Left	Nominal	Input
2	School_sett...	String	8	0		None	None	8	Left	Nominal	Input
3	School_type	String	10	0		None	None	10	Left	Nominal	Input
4	Classroom	String	3	0		None	None	3	Left	Nominal	Input
5	Teaching_m...	String	12	0		None	None	12	Left	Nominal	Input
6	Student_No	Numeric	2	0		None	None	12	Right	Scale	Input
7	Student_ID	String	12	0		None	None	12	Left	Nominal	Input
8	Gender	String	6	0		None	None	6	Left	Nominal	Input
9	Lunch	String	32	0		None	None	32	Left	Nominal	Input
10	Pretest	Numeric	2	0		None	None	12	Right	Scale	Input
11	Posttest	Numeric	3	0		None	None	12	Right	Scale	Input
12											
13											
14											

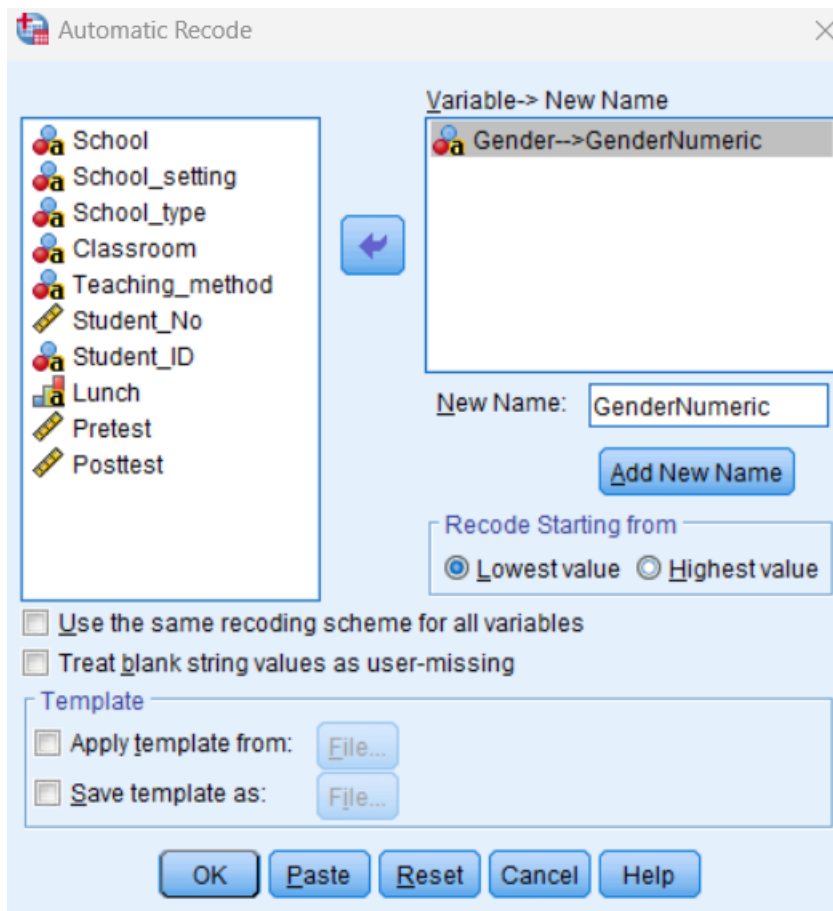
A separate sheet is available at the SPSS called Variable View where all the variable and there characteristics are shown. In this tab, the variables can be modified as needed sot that the anticipated results can be obtained without disruption.

After inspecting this tab, the only change that was required to be made was for Lunch row, and the change was made at the Measure column. The reason behind this change is that the pupil who are getting free lunch are being considered to be better than the pupils who are not rewarded with it. The same logic can not be applied on Gender row as Male and Female has the same weight (Sujarweni and Utami, 2019).

An additional change is required because the Gender or lunch variable contains text which will disrupt the calculation because based on the requirement of the tasks, these data needs to be analysed (Sujarweni and Utami, 2019). So, these data needs to be transformed into numerical values, and SPSS provides the option to do that.



Transform > Automatic Recode was chosen from the ribbon that is available at the top of SPSS, and the shown tab came up where the names of all the variables are listed (Sujarweni and Utami, 2019). From here, the Lunch and Gender variable needs to be transformed into numerical value.



So, the Gender variable was chosen which can be found on the right tab, and a few more information are needed in this tab. First of all, New Name box was filled with GenderNumeric to suggest SPSS that this will be the new name of this set of variable. By pressing Add new name, The question mark was removed and the new name was inputted. Pressing OK brings up a new tab where the changed result can be found (Sujarweni and Utami, 2019).

Your temporary usage period for IBM SPSS Statistics will expire in 4412 days.

NEW FILE.

DATASET NAME DataSet1 WINDOW=FRONT.

GET DATA

/TYPE=XLSX

/FILE='C:\Users\DELL\Downloads\Students Score.xls (3) Data Handling 2.xlsx'

/SHEET=name 'Students Score'

/CELLRANGE=FULL

/READNAMES=ON

/DATATYPEMIN PERCENTAGE=95.0

/HIDDEN IGNORE=YES.

EXECUTE.

DATASET NAME DataSet2 WINDOW=FRONT.

AUTORECODE VARIABLES=Gender

/INTO GenderNumeric

/PRINT.

Gender into GenderNumeric

Old Value New Value Value Label

Female 1 Female

Male 2 Male

This is the tab that came up after pressing ok where it is shown that Female has been changed to 1, and Male has been changed to 2 automatically by SPSS.

AUTORECODE VARIABLES=Lunch

/INTO LunchNumeric

/PRINT.

Lunch into LunchNumeric

Old Value New Value Value Label

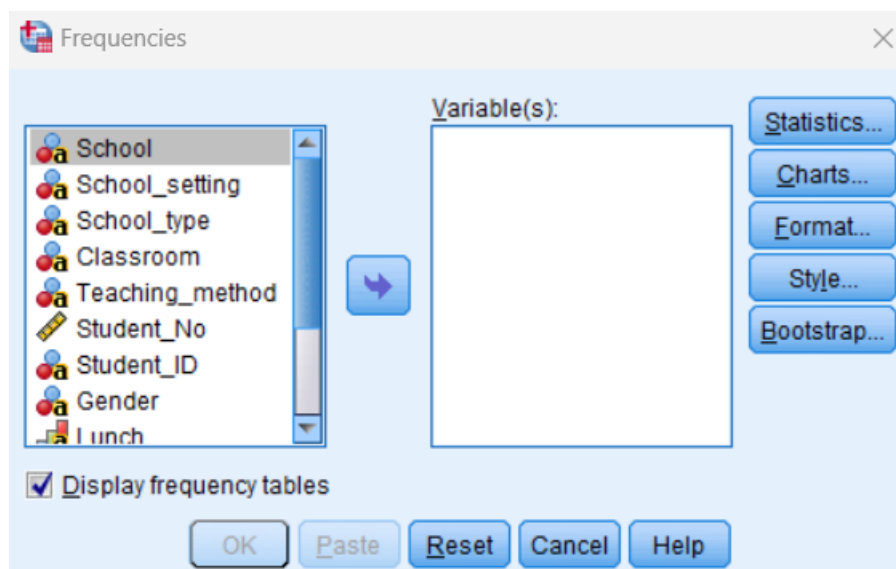
Does not qualify 1 Does not qualify

Qualifies for reduced/free lunch 2 Qualifies for reduced/free lunch

The same process has been followed for lunch column as well which has set 1 and 2 for two different set of texts for this variable and created a new column (Sujarweni and Utami, 2019).

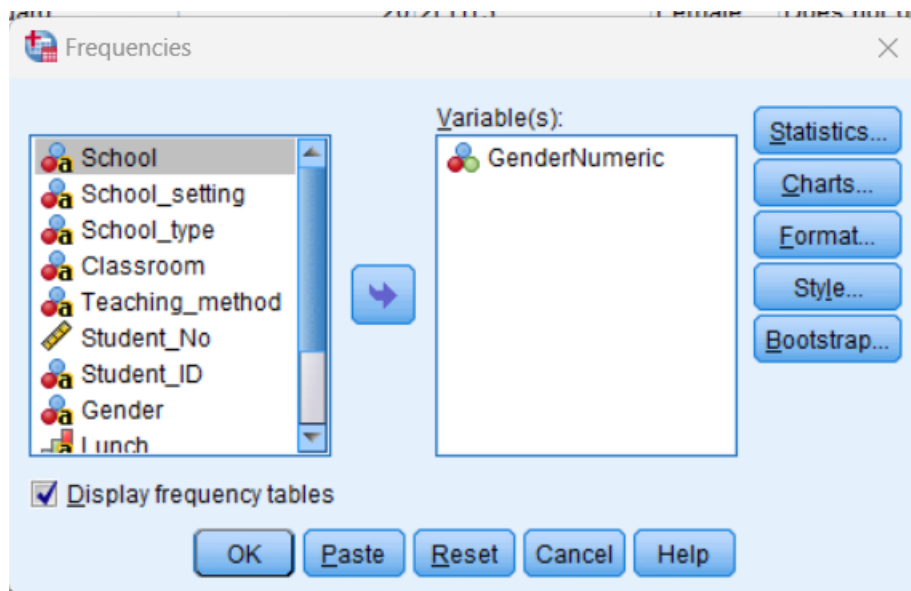
ID	Gender	Lunch	Pretest	Posttest	GenderNumeric	LunchNumeric	var
	Female	Does not qualify	62	72	1	1	
	Female	Does not qualify	66	79	1	1	
	Male	Does not qualify	64	76	2	1	
	Female	Does not qualify	61	77	1	1	
	Male	Does not qualify	64	76	2	1	
	Female	Does not qualify	66	74	1	1	
	Male	Does not qualify	63	75	2	1	
	Female	Does not qualify	63	72	1	1	
	Male	Does not qualify	64	77	2	1	
	Female	Does not qualify	61	72	1	1	
	Male	Does not qualify	61	73	2	1	
	Male	Does not qualify	64	74	2	1	
	Male	Does not qualify	66	78	2	1	
	Female	Does not qualify	60	71	1	1	
	Female	Does not qualify	64	77	1	1	
	Male	Does not qualify	64	73	2	1	
	Male	Does not qualify	63	70	2	1	
	Male	Does not qualify	67	73	2	1	
	Female	Does not qualify	63	71	1	1	
	Male	Does not qualify	64	73	2	1	
	Male	Does not qualify	60	68	2	1	
	Male	Does not qualify	60	70	2	1	
	Male	Does not qualify	57	66	2	1	
	Male	Does not qualify	57	70	2	1	
	Female	Qualifies for reduced/free lunch	66	66	1	2	

The two new columns that has been added by SPSS can be seen in the above picture where Gender and Lunch data has been transformed into numeric values.

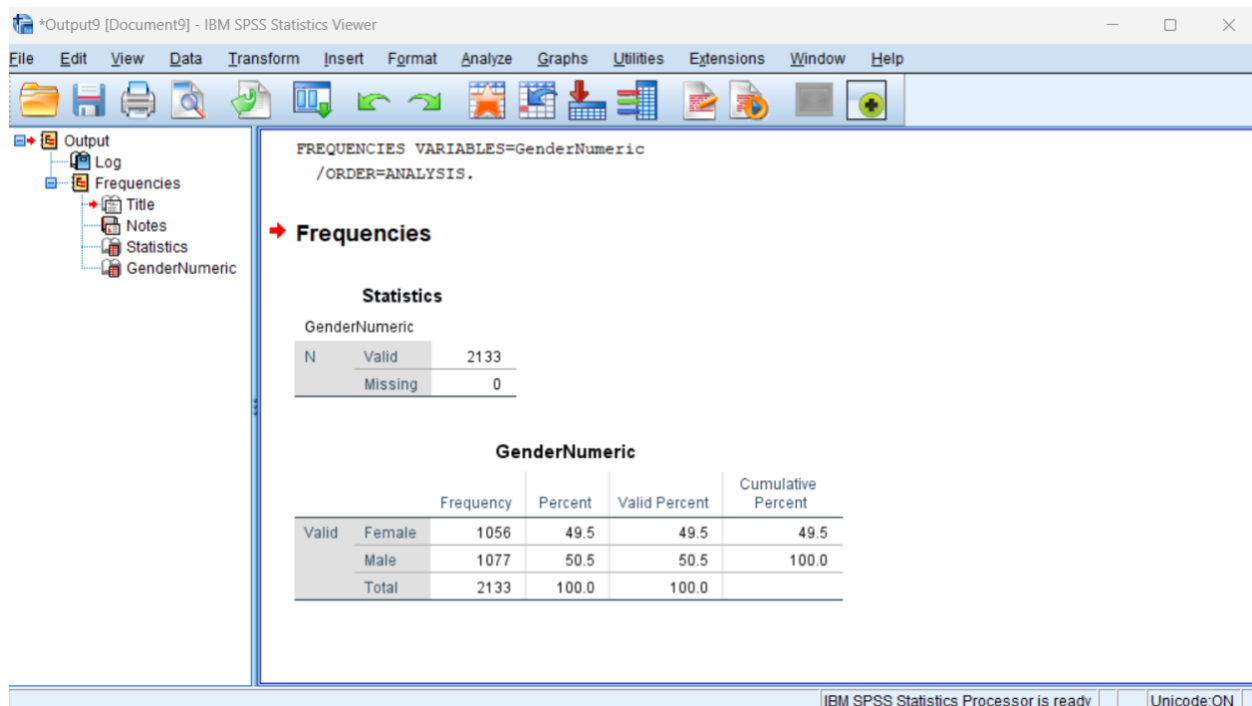


To analyze through SPSS, Analyze option is available at the top ribbon, from there, selecting descriptive statistics > Frequencies provides this tab that is shown in the above picture. In this tab,

name of all variables are provided, as each variables can be analyzed from this tab (Meyers et al., 2013).



The variable GenderNumeric has been chosen to be analyzed to find the number of Male and Female in the class.



This is the result presented through the analysis where it is found that number of Female is 1056, and 1077 males are present in the class.

➔ Frequencies

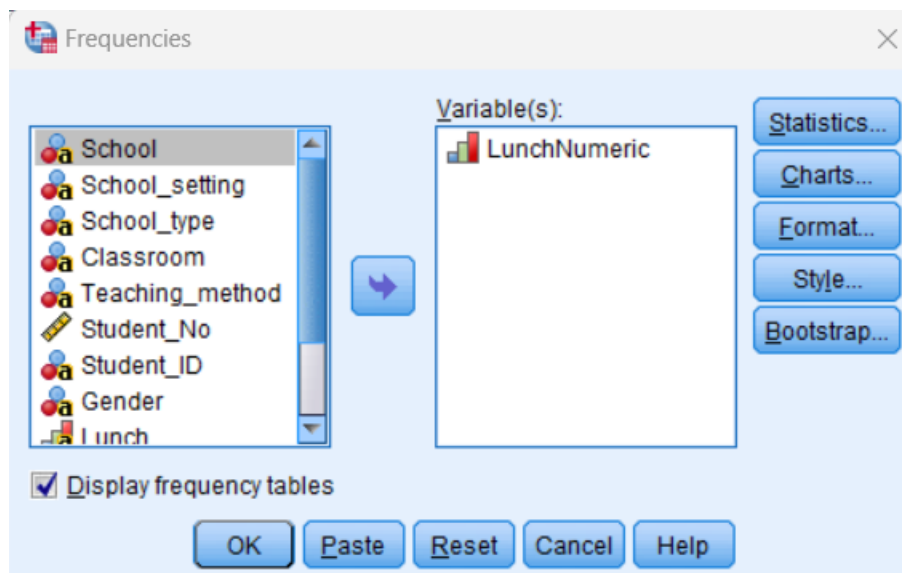
Statistics

LunchNumeric		
N	Valid	2133
	Missing	0

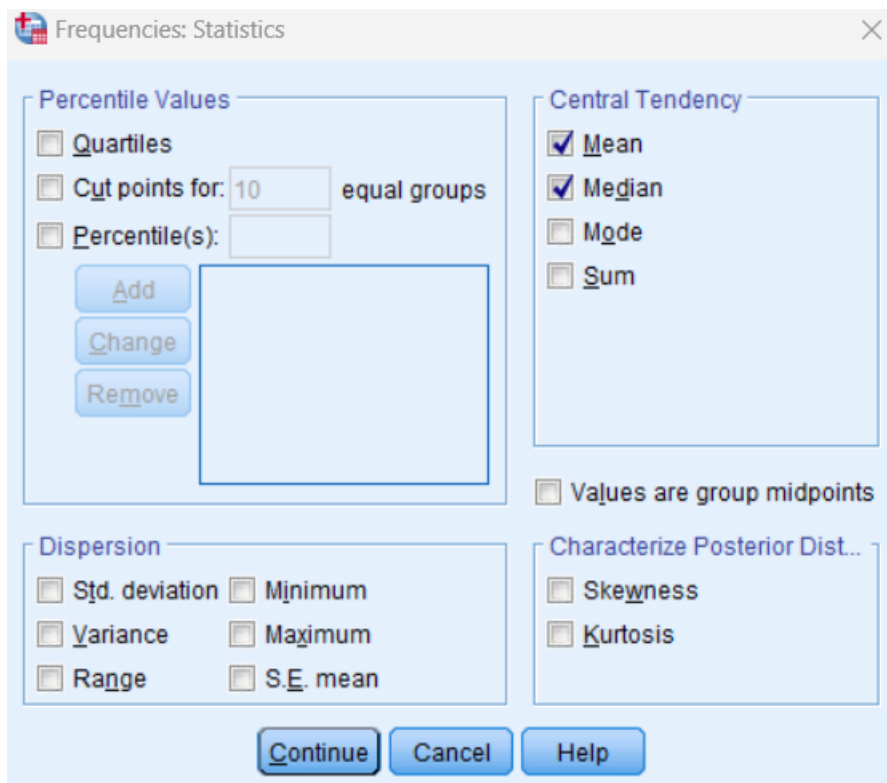
LunchNumeric

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Does not qualify	1215	57.0	57.0	57.0
	Qualifies for reduced/free lunch	918	43.0	43.0	100.0
Total		2133	100.0	100.0	

LunchNumeric Variable was chosen to be analyzed in the same process, and the result is shown above. SPSS has found that there are 1215 pupils who does not qualify, and the rest 918 qualifies for free lunch.



Using the same variable, statistical data needs to be found so statistics shown on the right of the tab has been selected which provided a new tab to obtain inputs (Meyers et al., 2013).



The image shows the 'Frequencies: Statistics' dialog box in SPSS. It is divided into four main sections: 'Percentile Values', 'Central Tendency', 'Dispersion', and 'Characterize Posterior Dist...'. In the 'Percentile Values' section, 'Quartiles' is selected, and 'Cut points for: 10 equal groups' is set. In the 'Central Tendency' section, 'Mean' and 'Median' are selected. In the 'Dispersion' section, 'Std. deviation', 'Variance', 'Range', 'Minimum', 'Maximum', and 'S.E. mean' are selected. In the 'Characterize Posterior Dist...' section, 'Skewness' and 'Kurtosis' are selected. At the bottom, there are 'Continue', 'Cancel', and 'Help' buttons.

Section	Option	Selected
Percentile Values	Quartiles	☒
	Cut points for: 10 equal groups	☐
	Percentile(s):	☐
	Buttons: Add, Change, Remove	
Central Tendency	Mean	☒
	Median	☒
	Mode	☐
	Sum	☐
Dispersion	Std. deviation	☒
	Variance	☒
	Range	☒
	Minimum	☒
	Maximum	☒
	S.E. mean	☒
Characterize Posterior Dist...	Skewness	☒
	Kurtosis	☒

Mean and Median are the two central tendency that has been selected from this tab so that SPSS calculates these data directly (De Sá, 2007).

➔ Frequencies

Statistics		
LunchNumeric		
N	Valid	2133
	Missing	0
Mean		1.43
Median		1.00

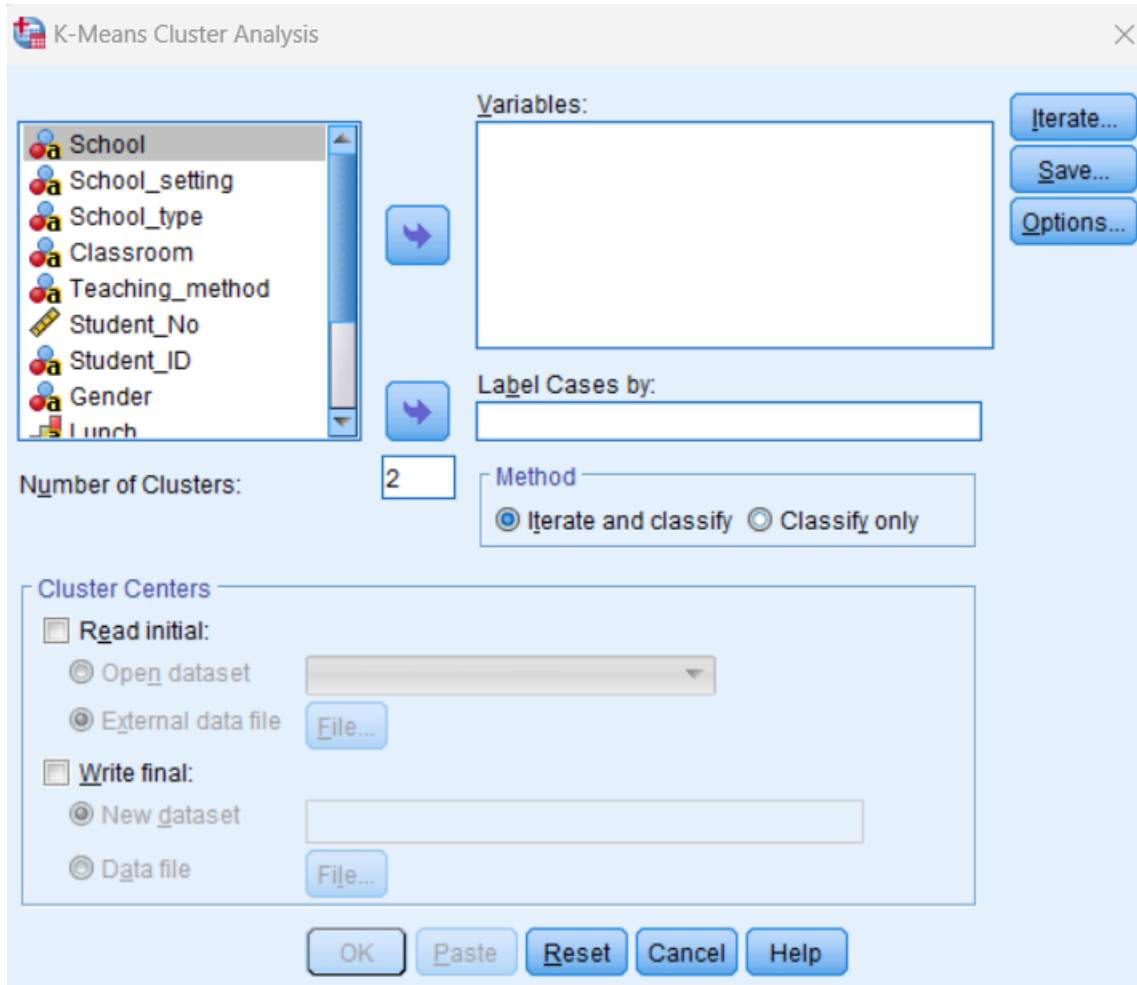
LunchNumeric					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Does not qualify	1215	57.0	57.0	57.0
	Qualifies for reduced/free lunch	918	43.0	43.0	100.0
	Total	2133	100.0	100.0	

This time, the result consists of the mean and media data regarding the lunch variable which is shown in the above picture.

SPSS found that Mean = 1.43

Median = 1.00

That means adding all the values under the variable and dividing with the total number of inputs under the variable gives 1.43 (De Sá, 2007). Also, sorting the data under the Lunch variable, and identifying the exact middle value of the column is 1.00 (De Sá, 2007).



Analyze > Classify provides several clustering options, and from there K means cluster was chosen for this dataset. Here, all the variables are shown as any of the variables can be chosen for clustering (Wendler and Gröttrup, 2016).

Initial Cluster Centers

	Cluster	
	1	2
Pretest	23	93
Posttest	35	100

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	65
Posttest	55	78

Number of Cases in each Cluster

Cluster	1	993.000
	2	1140.000
Valid		2133.000
Missing		.000

Choosing pretest and posttest data, and running the clustering, these are the values that were obtained through the process.

Data Mining Method	Description	Examples
Clustering	Grouping similar data points based on features	- Customer segmentation for targeted marketing
Classification	Assigning predefined labels to data instances (Kantardzic, 2011)	- Spam email detection
Association Rule Mining	Discovering patterns and relationships in data	- Market basket analysis for retail cross-selling
Regression Analysis	Predicting numerical values based on relationships (Kantardzic, 2011)	- Sales forecasting based on historical data
Outlier Detection	Identifying rare or unusual patterns in the data	- Fraud detection in financial transactions

Text Mining Method	Description	Examples
Text Classification	Assigning predefined categories to text documents	- Topic categorization for news articles
Named Entity Recognition (NER)	Identifying and classifying entities in text (Berry and Kogan, 2010)	- Extracting names, organizations, and locations from news articles
Sentimental Analysis	Determining the sentiment expressed in text (Berry and Kogan, 2010)	- Analyzing customer reviews for product sentiment
Text Summarization	Generating concise summaries of text documents	- Summarizing long articles for quick comprehension
Information Extraction	Extracting structured information from	- Extracting key details from legal documents

	unstructured text (Berry and Kogan, 2010)	
--	---	--

Aspect	SPSS	Excel
Advantages	- Statistical Analysis: SPSS is designed specifically for statistical analysis, making it powerful for complex data analysis tasks (Lee and Peters, 2015).	- Familiar Interface: Excel is widely used and familiar, making it accessible for basic data analysis and visualization (Lee and Peters, 2015).
	- Advanced Analytics: Offers a wide range of advanced statistical techniques and machine learning algorithms.	- Quick Data Manipulation: Excel is excellent for quick data manipulation, sorting, and filtering.
	- Data Visualization: Provides robust tools for graphical representation of data, aiding in interpretation.	- PivotTables: Excel's PivotTables are efficient for summarizing and analyzing large datasets.
	- Automation: SPSS allows for the automation of repetitive analysis tasks through syntax and scripting (Lee and Peters, 2015).	- Custom Formulas: Excel's flexibility allows users to create custom formulas for specific analysis needs (Lee and Peters, 2015).
Disadvantages	- Learning Curve: SPSS may have a steeper learning curve for beginners due to its specialized nature.	- Limited Statistical Tools: Excel has limited built-in statistical tools compared to dedicated statistical software like SPSS.
	- Cost: SPSS is a commercial software, and acquiring a license can be expensive for individual users or	- Limited Data Size: Excel may struggle with very large datasets, causing performance issues. (Lee and Peters, 2015)

	small businesses (Lee and Peters, 2015).	
	- Not Ideal for General Data Manipulation: SPSS is not as versatile as Excel when it comes to general data manipulation tasks.	- Limited Version Control: Excel lacks robust version control features, making collaboration challenging.
Practical Examples	Suppose a researcher wants to conduct a complex regression analysis with multiple variables and interactions. SPSS offers a comprehensive set of tools and outputs for such an analysis.	For a small business tracking monthly sales, Excel can be used to quickly create tables, charts, and graphs for visualization without the need for advanced statistical techniques.
	In a psychology research project, SPSS can be used to analyze survey data and perform factor analysis to identify underlying constructs.	A project manager may use Excel to create a Gantt chart for project scheduling and task management.
	An epidemiologist studying disease trends may utilize SPSS for cluster analysis to identify geographical patterns.	Excel can be used to create simple histograms or bar charts to visualize frequency distributions of various data points.

References

- Abdulezer, L., 2003. *Excel Best Practices for Business: Covers Excel 2003, 2002, and 2000*. John Wiley & Sons.
- Alexander, M. and Kusleika, R., 2016. *Excel 2016 formulas*. John Wiley & Sons.
- Apte, D.P., 2009. *Statistical tools for managers: Using MS Excel*. Excel Books.
- Berry, M.W. and Kogan, J. eds., 2010. *Text mining: applications and theory*. John Wiley & Sons.
- Bluttman, K., 2013. *Excel formulas and functions for dummies*. John Wiley & Sons.
- De Sá, J.P.M., 2007. *Applied statistics using SPSS, Statistica, MatLab and R*. Springer Science & Business Media.
- Guerrero, H., Guerrero, R. and Rauscher, 2019. *Excel data analysis*. Springer International Publishing.
- Kantardzic, M., 2011. *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons.
- Lee, N. and Peters, M., 2015. *Business statistics using EXCEL and SPSS*. Sage.
- Meyers, L.S., Gamst, G.C. and Guarino, A.J., 2013. *Performing data analysis using IBM SPSS*. John Wiley & Sons.
- Pace, L., 2012. *The Excel data and statistics cookbook*. Lulu. com.
- Sujarweni, V.W. and Utami, L.R., 2019. *The master book of SPSS*. Anak Hebat Indonesia.
- Syrstad, T. and Jelen, B., 2011. *Using Microsoft Excel 2010*. Pearson Education.
- Wendler, T. and Gröttrup, S., 2016. *Data mining with SPSS modeler: theory, exercises and solutions*. Springer.

