

DATA ANALYSIS **& VISUALISATION**

LCBB5000

MARINELA ANDREICA

ID:2010777

Table of Contents

PART 1	3
Evaluating the use of excel for pre-processing, analysing and visualising the sales and profits of Superstore from the year 2009 to 2012	3
PART 2	13
2.1 Result presenting number of students like vanilla flavor of ice crea.....	13
2.2 Presenting the result of number of students who are male and female.....	15
2.3 Presenting the result of mean and median of students like chocolate and strawberry flavor of ice cream.....	17
2.4 Presenting the result of K-mean clustering along with the step required to follow in the SPSS software	21
2.5 Presenting most common data mining and text mining methods used by business with real world examples	28
2.6 Presenting advantage and disadvantage of using SPSS and Excel using practical relevant examples	29
REFERENCES	30

PART 1

Evaluating the use of excel for pre-processing, analysing and visualising the sales and profits of Superstore from the year 2009 to 2012

Superstore is a UK based furniture store which sales different kinds of products such office supplies, furniture and technology products. to their customers. Data analysis is a process of cleaning, transforming, modelling of the dataset in order to generate useful information and make decision for future benefits (Pallant, 2020). In order to analyse the decline in the sales and profit of the super store over the year on total along with product wise basis, excel function has been used. The excel is one of the best software and provides the user with the different excel function with the help of which the data can be pre-processed, analysed and visualized. This further help the user such as managers of superstore to make the proper decision regarding which products are more profitable and which not. The evaluation of superstore sales and profit data along with its analysis using the excel software is done in the following three steps:

Step 1 Data Pre Processing:

This is the first step of data analysis and visualization where the raw data is pre-processed using the different functions of excel. The data pre-processing is a data mining technique which is basically used for the purpose of converting or transforming the raw data in a useful and efficient format. There are three steps of data pre-processing such as data cleaning, data transforming and data reduction. In the first step, the missing data and noisy raw data has been removed (Betty Jane and Ganesh, 2020). Further, the data is transformed using the normalization, attribute selection, discretization and concept hierarchy generation. In the last stage of the data pre-processing, the users usually reduce by the dataset by removing the irrelevant data from the whole dataset.

For instance, in the case of pre-processing the raw data of Superstore, the users have used the Filter and Sort function of Excel. Steps of Data pre-processing of superstore are as follows:

- 1.

1/4/2009	566.53	-209.6
1/5/2009	789.01	-539.59
1/5/2009	528.6	52.59
1/5/2009	107.41	20.68
1/5/2009	149	-119.48
1/5/2009	334.29	-157.28
1/5/2009	596.21	211.59
1/5/2009	78.81	-93.25
1/6/2009	5178.17	531.69
1/6/2009	535.24	216.43
1/6/2009	124.06	15.74

SuperstoreSales1672486953 **Filter Function**

In the first step, the data related to order date, sales and profit has been gathered and transfer to the new worksheet of excel (name Filter Function) in order to process and transform the sales and profit into year wise (Mihias, 2019).

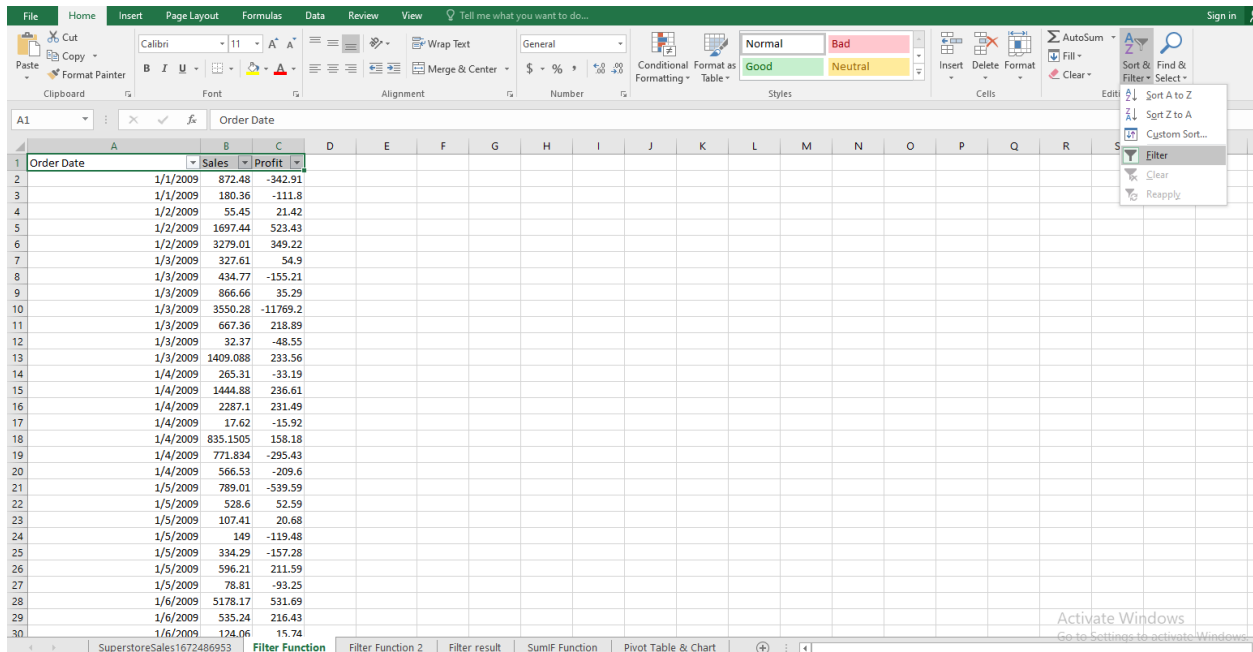
2.

Order Date	Sales	Profit
1/1/2009	872.48	-342.91
1/1/2009	180.36	-111.8
1/2/2009	55.45	21.42
1/2/2009	1697.44	523.43
1/2/2009	3279.01	349.22
1/3/2009	327.61	54.9
1/3/2009	434.77	-155.21
1/3/2009	866.66	35.29
1/3/2009	3550.28	-11769.2
1/3/2009	667.36	218.89
1/3/2009	32.37	-48.55
1/3/2009	1409.088	233.56
1/4/2009	265.31	-33.19
1/4/2009	1444.88	236.61
1/4/2009	2287.1	231.49
1/4/2009	17.62	-15.92
1/4/2009	835.1505	158.18
1/4/2009	771.834	-295.43
1/4/2009	566.53	-209.6
1/5/2009	789.01	-539.59
1/5/2009	528.6	52.59
1/5/2009	107.41	20.68
1/5/2009	149	-119.48
1/5/2009	334.29	-157.28
1/5/2009	596.21	211.59
1/5/2009	78.81	-93.25
1/6/2009	5178.17	531.69
1/6/2009	535.24	216.43
1/6/2009	124.06	15.74

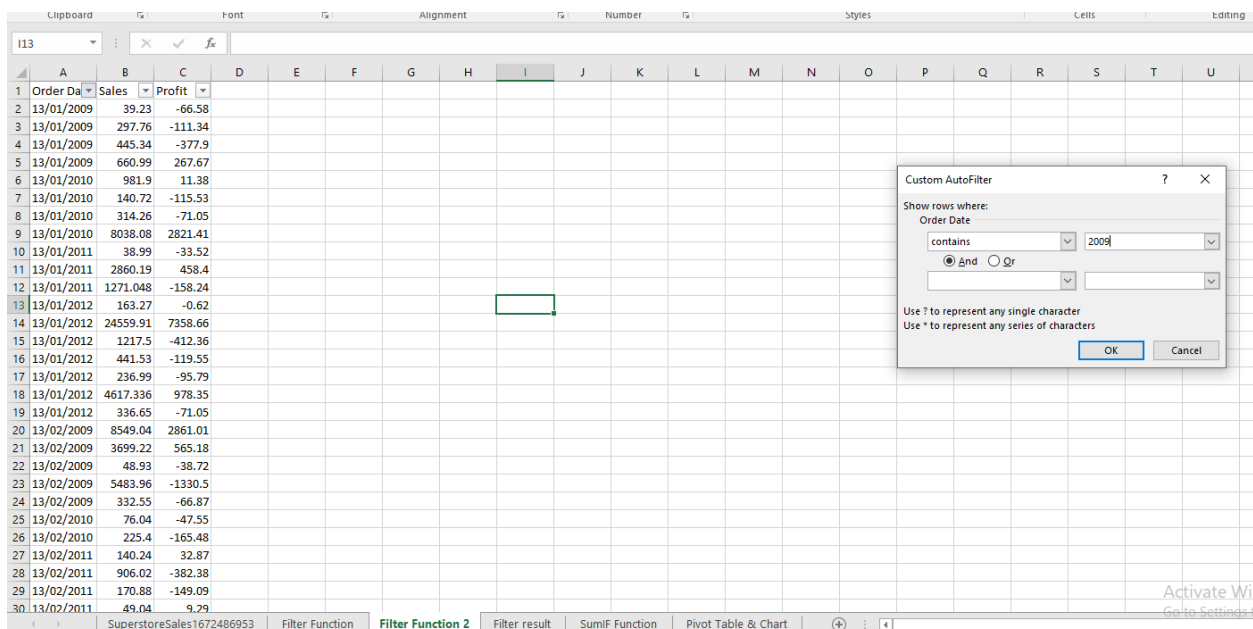
Further, in the second step, with the use of the Sort Function, the order date is Sort from A to Z. This has helped in transforming the sales and profit of 2009, 2010, 2011 and 2012 together

(Aryadoust and Raquel, 2019). However, due to the error in some of the dataset, the sorting is not possible yearly. Hence, to solve this issue, the dataset regarding order date, sales and profit unable to sort is further copy and paste in the Filter Function 2 excel worksheet.

3.



	Order Date	Sales	Profit
1	1/1/2009	872.48	-342.91
2	1/1/2009	180.36	-111.8
3	1/2/2009	55.45	21.42
4	1/2/2009	1697.44	523.43
5	1/2/2009	3279.01	349.22
6	1/3/2009	327.61	54.9
7	1/3/2009	434.77	-155.21
8	1/3/2009	866.66	35.29
9	1/3/2009	3550.28	-11769.2
10	1/3/2009	667.36	218.89
11	1/3/2009	32.37	-48.55
12	1/3/2009	1409.088	233.56
13	1/4/2009	265.31	-33.19
14	1/4/2009	1444.88	236.61
15	1/4/2009	2287.1	231.49
16	1/4/2009	17.62	-15.92
17	1/4/2009	835.1505	158.18
18	1/4/2009	771.834	-295.43
19	1/4/2009	566.53	-209.6
20	1/5/2009	789.01	-539.59
21	1/5/2009	528.6	52.59
22	1/5/2009	107.41	20.68
23	1/5/2009	149	-119.48
24	1/5/2009	334.29	-157.28
25	1/5/2009	596.21	211.59
26	1/5/2009	78.81	-93.25
27	1/6/2009	5178.17	531.69
28	1/6/2009	535.24	216.43
29	1/6/2009	124.06	15.74
30	1/6/2009	124.06	15.74



	Order Date	Sales	Profit
1	13/01/2009	39.23	-66.58
2	13/01/2009	297.76	-111.34
3	13/01/2009	445.34	-377.9
4	13/01/2009	660.99	267.67
5	13/01/2010	981.9	11.38
6	13/01/2010	140.72	-115.53
7	13/01/2010	314.26	-71.05
8	13/01/2010	8038.08	2821.41
9	13/01/2011	38.99	-33.52
10	13/01/2011	2860.19	458.4
11	13/01/2011	1271.048	-158.24
12	13/01/2012	163.27	-0.62
13	13/01/2012	24559.91	7358.66
14	13/01/2012	1217.5	-412.36
15	13/01/2012	441.53	-119.55
16	13/01/2012	236.99	-95.79
17	13/01/2012	4617.336	978.35
18	13/01/2012	336.65	-71.05
19	13/02/2009	8549.04	2861.01
20	13/02/2009	3699.22	565.18
21	13/02/2009	48.93	-38.72
22	13/02/2009	5483.96	-1330.5
23	13/02/2009	332.55	-66.87
24	13/02/2010	76.04	-47.55
25	13/02/2010	225.4	-165.48
26	13/02/2011	140.24	32.87
27	13/02/2011	906.02	-382.38
28	13/02/2011	170.88	-149.09
29	13/02/2011	49.04	9.29
30	13/02/2011	49.04	9.29

In the third step, after transferring the remaining dataset to Filter Function 2 sheet, the Filter Function is used to arrange the order date, sales and profit of one year linewise for example 2009. For this, the step, filter ~~Text~~ Filter ~~Contains~~ ~~Selecting order date 2009~~ ~~OK~~ has been followed.

In this way, the data regarding sales and profit is transformed from date wise to year wise in order to assess the decline in sales and profit of Superstore over the years. The data pre-processing became simple and time saving with the use of “Sort and Filter” function of excel.

Step 2 Data Analysing

This is the second step of the data processing and analysis, where the raw data is analysed in a tabular format so that proper decision can be made by the users (Fernandes and et.al., 2019). For instance, in order to analyse the decline in sales and profit of Superstore over the years, the manager of company has used Pivot Table and SumIf function of Excel. The data analysis and interpretation of Superstore are as follows:

SUMIF:

Table 1: Total Sales and Total Profit of Superstore from year 2009 to 2012

Year	Sales	Profit
2009	4387485	424128.6
2010	3072762	292178.1
2011	1599009	130672.2
2012	1500412	129680.2

Interpretation: The above table has been generated using the “SUMIF” function in an excel (see attached excel). Formula of SUMIF = SUMIF (range, criteria, sum range). On the basis of the above table, it is analysed that the total sales of superstore are 4387485 in year 2009 which reduced to 1500412 in the year 2012. Further, it is also interpreted that the profit of superstore has reduced to 12680.2 in the year 2012 from 424128.6 in the year 2009. This indicate decline in the profit and sales of superstore over the years (Chireshe, 2020).

Table 2: % Decline in sales and profit of Superstore from the year 2009 to 2012

Year	Sales	Profit	% Decline in sales	% Decline in profit
2009	4387485	424128.6	0	0
2010	3072762	292178.1	-30%	-31%
2011	1599009	130672.2	-48%	-55%
2012	1500412	129680.2	-6%	-1%

Interpretation: Further, to assess the decline % of sales and profit of superstore, the Growth Percentage Formula of excel has been used. Formula used to compute the decline % in excel is $\text{current year sales} - \text{previous year sales} / \text{previous year sales} * 100$. On the basis of above table, it is analysed that in the year 2010, the sales and profit of superstore has been decline by 30% and 31% respectively (Winston, 2018). Further, in the year 2011, the sales and profit was declined by 48% and 55% respectively as compared to the 2010-year base sales and profit. However, the decline percentage of sales and profit of superstore was low in the year 2012 as compared to all four years because of lowest decline percentage of 6% and 1% respectively.

PIVOT TABLE:

The PIVOT TABLE is also a function of excel which is used to determine and analyse the total sales and profit of superstore in the year 2009, 2010, 2011 and 2012 by product category (Held, Moriarty and Richardson, 2019). The below mentioned tables stating the sum of sales and profit by product category of all four years as well as individual year using the Pivot table function of excel. In this way, it can be said that PIVOT table function of excel is also one of the way through which data is evaluated and analysed in most appropriate and simplified manner.

Table 3: Sum of sales and sum of profit over the years by product category

Row Labels	Sum of Sales	Sum of Profit
Furniture	3151463.07	308848.58
Office Supplies	8249700.59	896172.38
Technology	3514437.165	316747.02
Grand Total	14915600.82	1521767.98

Interpretation: On the basis of the above table, it is analysed that the Office supplies product of Superstore is more sales and profit generating over the four years because of highest sales and profit among furniture, office supplies and technology.

Table 4: Sum of sales and profit of Superstore in the year 2009 by product category

Row Labels	Sum of Sales	Sum of Profit
Furniture	863122.8	114100.66
Office Supplies	2329861.282	231602.04
Technology	1016912.764	88393.32
Grand Total	4209896.846	434096.02

Interpretation: On the basis of the above table, it is analysed that in the year 2009, office supplies is the highest sales and profit generating product while furniture is lowest sales and profit generating product of superstore.

Table 5: Sum of sales and profit of Superstore in the year 2010 by product category

Row Labels	Sum of Sales	Sum of Profit
Furniture	758832.997	75332.61
Office Supplies	2016495.137	220138.55
Technology	784758.911	69446.17
Grand Total	3560087.045	364917.33

Interpretation: In the year 2010, the highest sales and profit generating product according to above pivot table is office supplies.

Table 6: Sum of sales and profit of Superstore in the year 2011 by product category

Row Labels	Sum of Sales	Sum of Profit
Furniture	742797.159	53560.57
Office Supplies	1883981.193	232509.57
Technology	803166.629	94240.36

Grand Total	3429944.981	380310.5
--------------------	--------------------	-----------------

Interpretation: In the year 2011, the highest sales and profit generating product according to above pivot table is office supplies.

Table 7: Sum of sales and profit of Superstore in the year 2012 by product category

Row Labels	Sum of Sales	Sum of Profit
Furniture	786710.1135	65854.74
Office Supplies	2019362.979	211922.22
Technology	909598.861	64667.17
Grand Total	3715671.953	342444.13

Interpretation: From above table, it is analysed that the highest sales and profit generating product of superstore is office supplies and lowest is technology.

Data Analysis using Regression Function of Excel:

Regression is also a function of excel which is used to analyse the statistical relationship between the sales and profit of superstore. This is done in following ways:

Hypothesis development:

H0 = There is no statistical relationship exist between the sales and profit of superstore of UK.

H1 = There is statistical relationship exist between the sales and profit of superstore of UK

Result:

SUMMARY OUTPUT									
<i>Regression Statistics</i>									
Multiple R	0.581960294								
R Square	0.338677784								
Adjusted R Square	0.338599027								
Standard Error	973.1970493								
Observations	8399								
ANOVA									
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>				
Regression	1	4.07E+09	4.07E+09	4300.29	0.000				
Residual	8397	7.95E+09	947112.5						
Total	8398	1.2E+10							
	<i>Coefficients</i>	<i>Standard Err</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>	
Intercept	-163.7842133	11.85066	-13.8207	5.64E-43	-187.0144371	-140.554	-187.014	-140.554	
Sales	0.194252422	0.002962	65.5766	0	0.188445737	0.200059	0.188446	0.200059	

Interpretation: From the above ANNOVA table, it is identified that the significance F value is 0.00 which is lower than the standard criteria of 0.05. This means the alternative hypothesis is accepted and null hypothesis is rejected. This signifies that there is statistical relationship lies between the sales and profit of superstore (McCullough and Wilson, 2022). As per the r square value (0.33) signifies that there is positive and moderate correlation between sales and profit. It means with the rise in the profit of business, it can be said that the sales of the business are also increased.

Step 3 Visualization of Data

The last stage of data processing and analysis is visualization of data using the charts, diagrams of the excel. There are various types of charts such as line chart, column chart, bar chart, scattered plot etc. with the help of which data regarding the total sales & profit, decline in sales & profit along with the association between the sales and profit of superstore can be analysed and visualized. The visualization of data are as follows:



Figure 1: Graphical presentation of sales and profit of Superstore from year 2009 to 2012

The above column chart is representing the total sales and profit of superstore from year 2009 to 2012 in simplified and appropriate manner.



Figure 2: Graphical presentation of % decline in sales and profit of Superstore from year 2009 to 2012

The above column chart is representing the decline % in the sales and profit of superstore UK from the year 2009 to 2012.

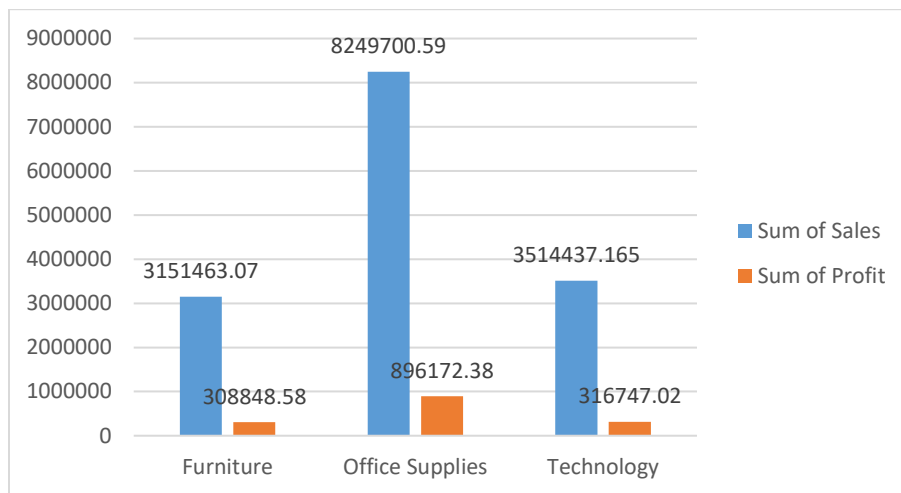


Figure 3: Graphical presentation of total sales and profit of superstore over four years by product category

The above chart or diagram is representing the sum of sales and profit of superstore UK by product category. It is visualized that the sales and profit of office supplies of superstore over the four year is highest. Hence, the company should continue with the product and produce office supplies in large amount.

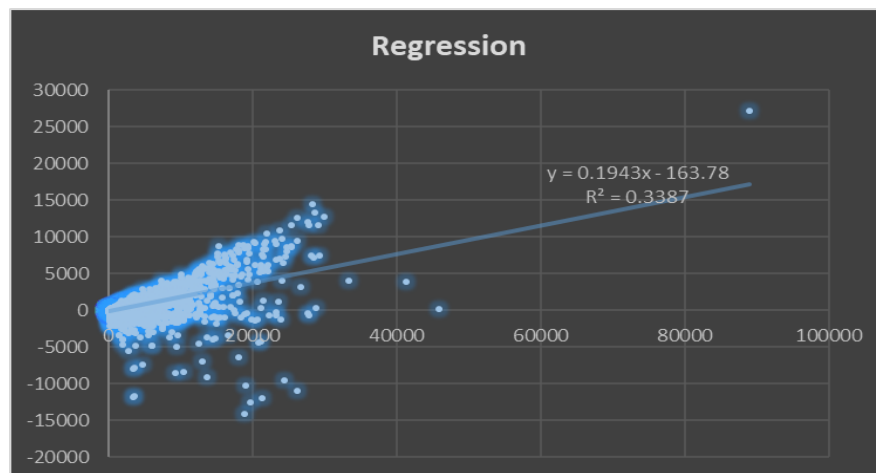


Figure 4: Regression

The above scattered plot diagram is representing straight and positive relation between the sales and profit of superstore UK. On the basis of above diagram, the regression equation of sales and profit is as follows: $Y = 0.1943x - 163.78$. This is helpful in predicting the sales of the superstore based on profit they make.

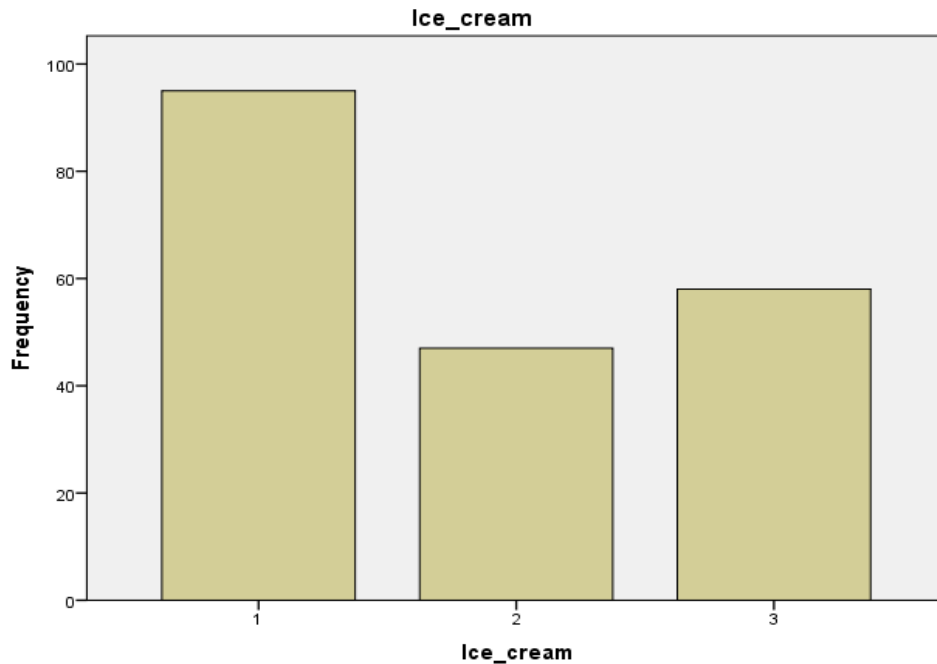
PART 2

2.1 Result presenting number of students like vanilla flavor of ice crème

Frequencies

Statistics		
Ice_cream		
N	Valid	200
	Missing	0

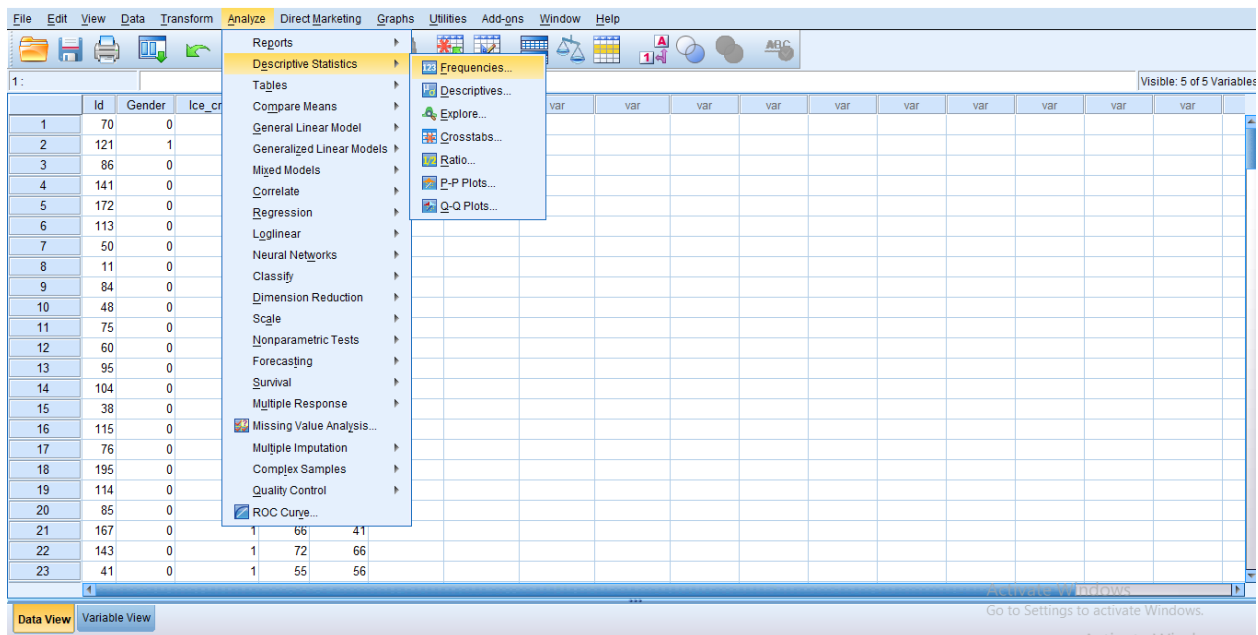
Ice_cream					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	95	47.5	47.5	47.5
	2	47	23.5	23.5	71.0
	3	58	29.0	29.0	100.0
	Total	200	100.0	100.0	



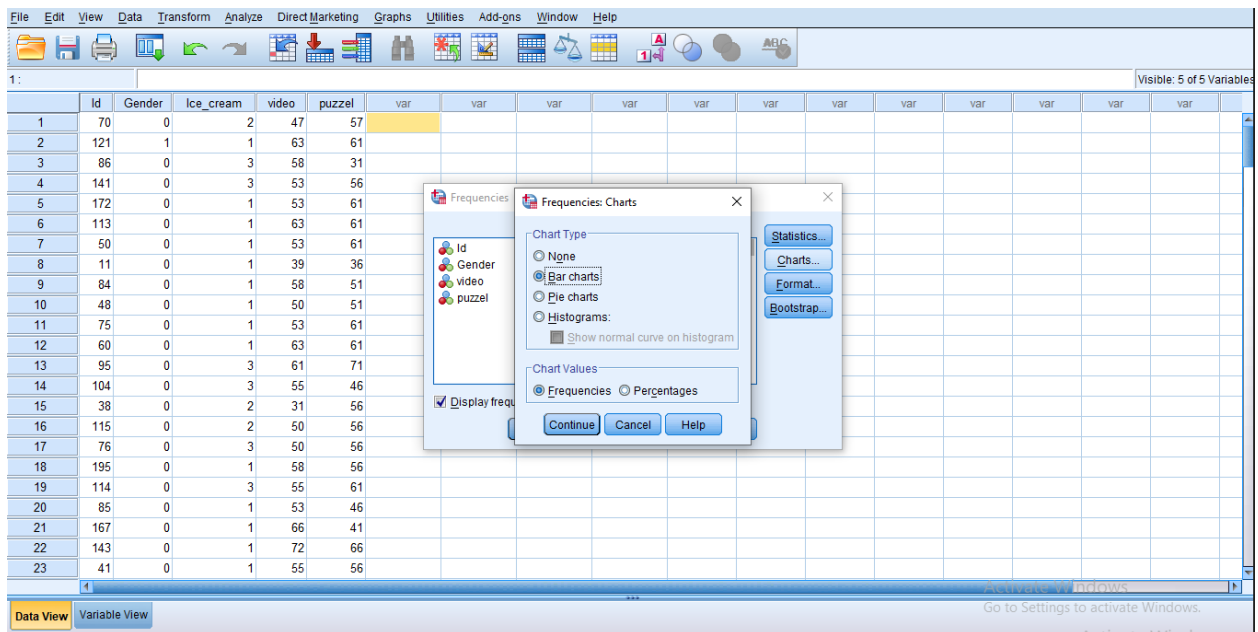
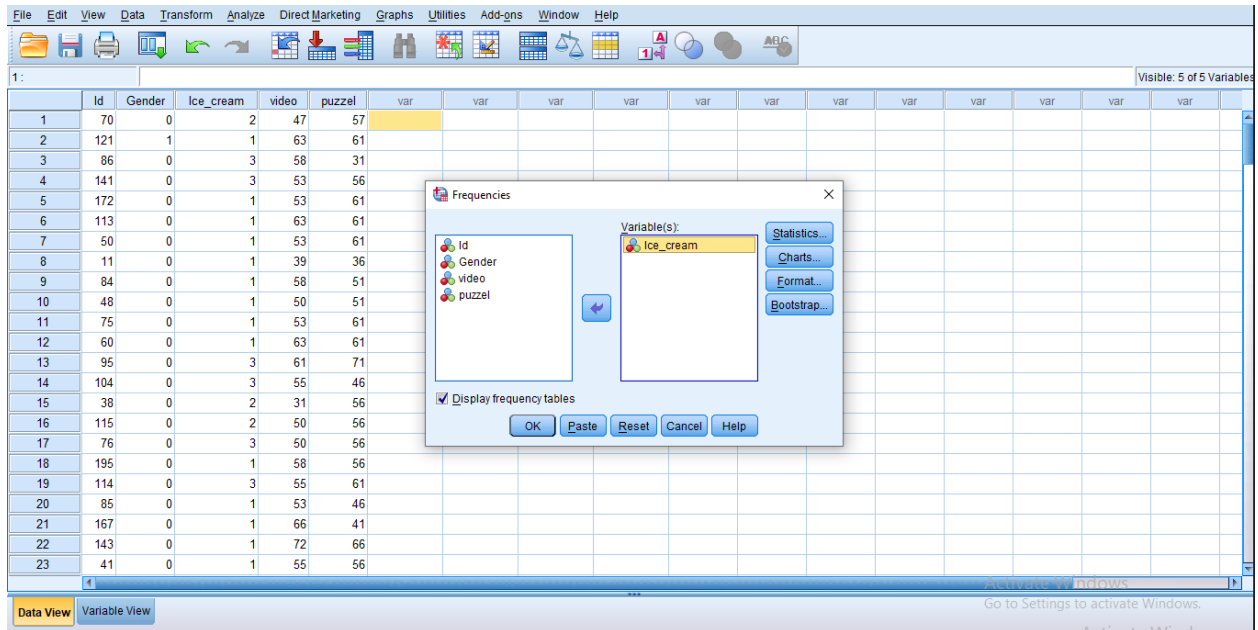
Interpretation: On the basis of the above frequency table, it is identified that the total number of students who like vanilla flavors of ice crème is 95. Out of 200 students, 95 students like vanilla ice crème flavor.

Steps required to follow:

Step 1: Analyse → Descriptive statistics → Frequencies



Step 2: Select variable icecreme → charts → Bar → OK



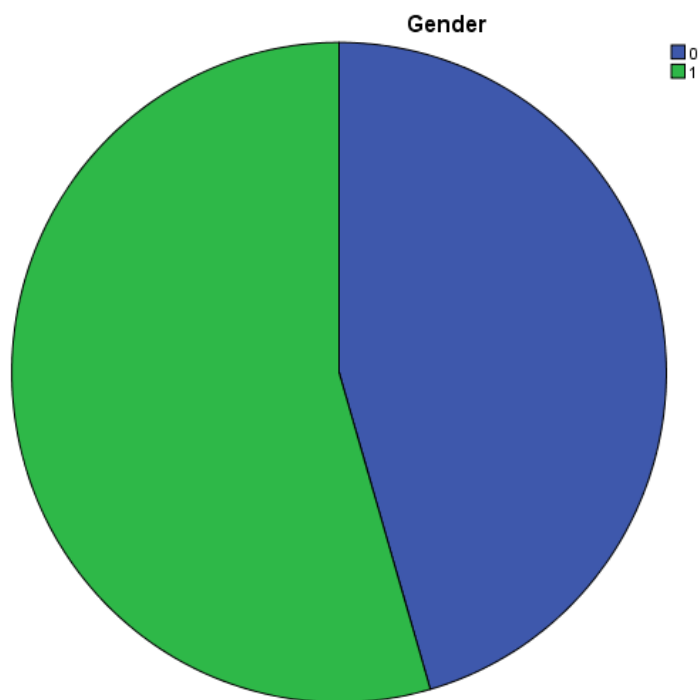
2.2 Presenting the result of number of students who are male and female

Frequencies

Statistics

Gender		
N	Valid	200
	Missing	0

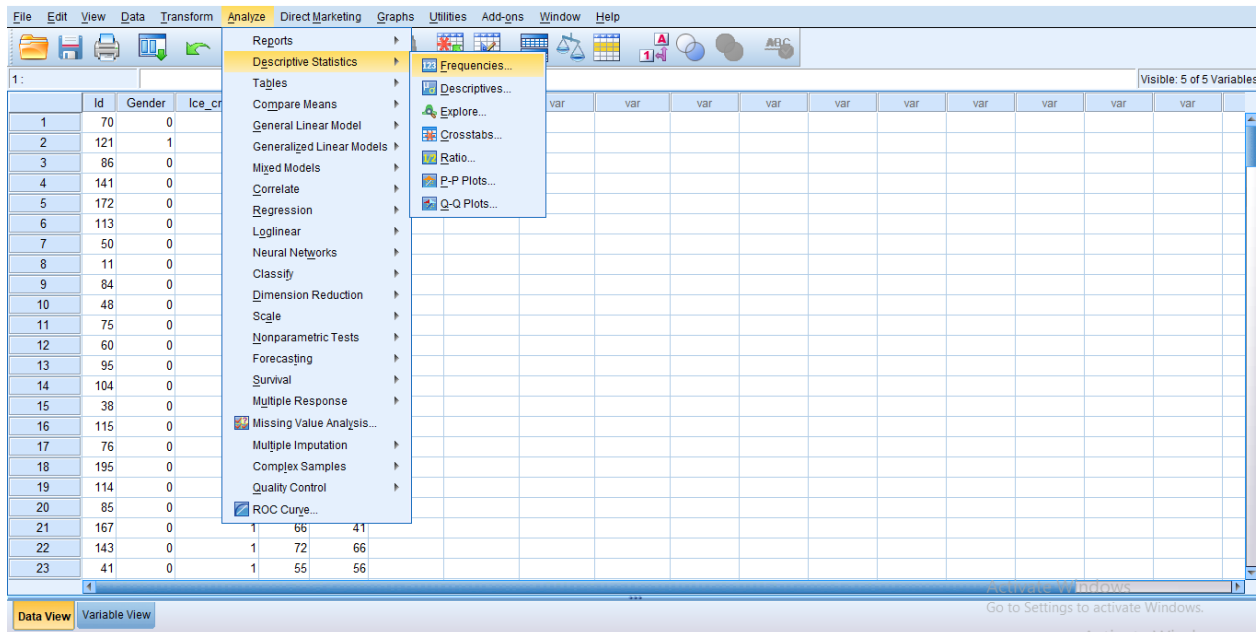
Gender					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	91	45.5	45.5	45.5
	1	109	54.5	54.5	100.0
	Total	200	100.0	100.0	



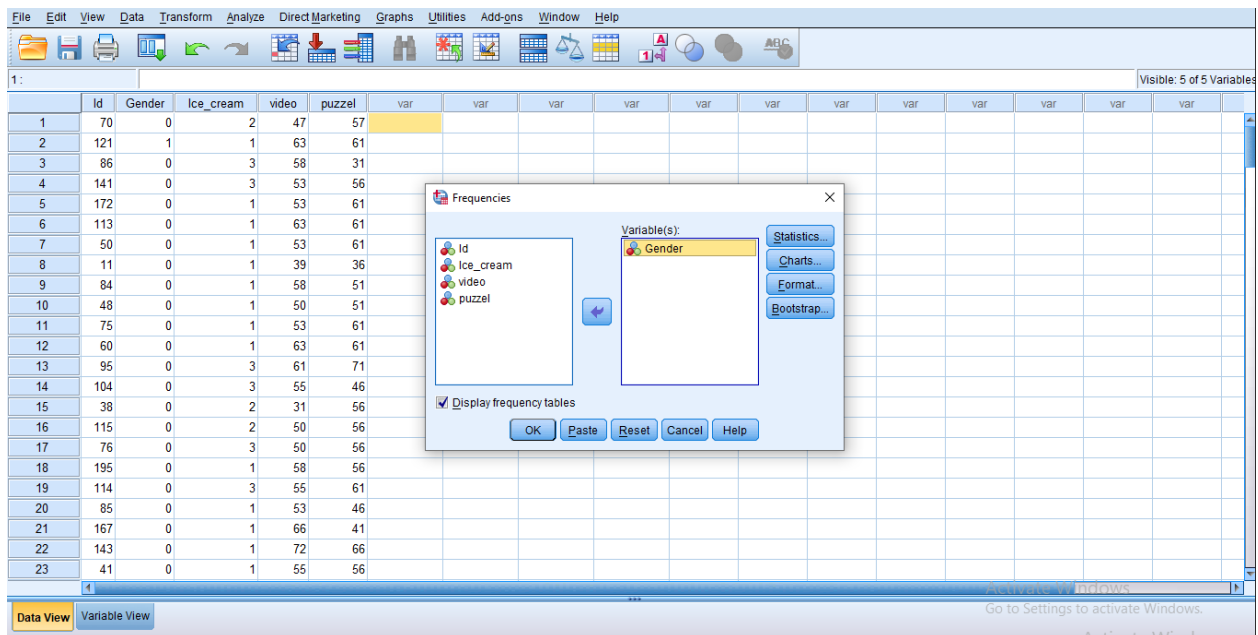
Interpretation: On the basis of the above table and diagram, it is interpreted that out of 200 students, the total number of male is 91 while the total number of female is 109. It means that the total number of female among the 200 high school students is high.

Steps required to follow:

Step 1: Analyse → Descriptive statistics → Frequencies



Step 2: Select variable Gender → charts → Pie → OK

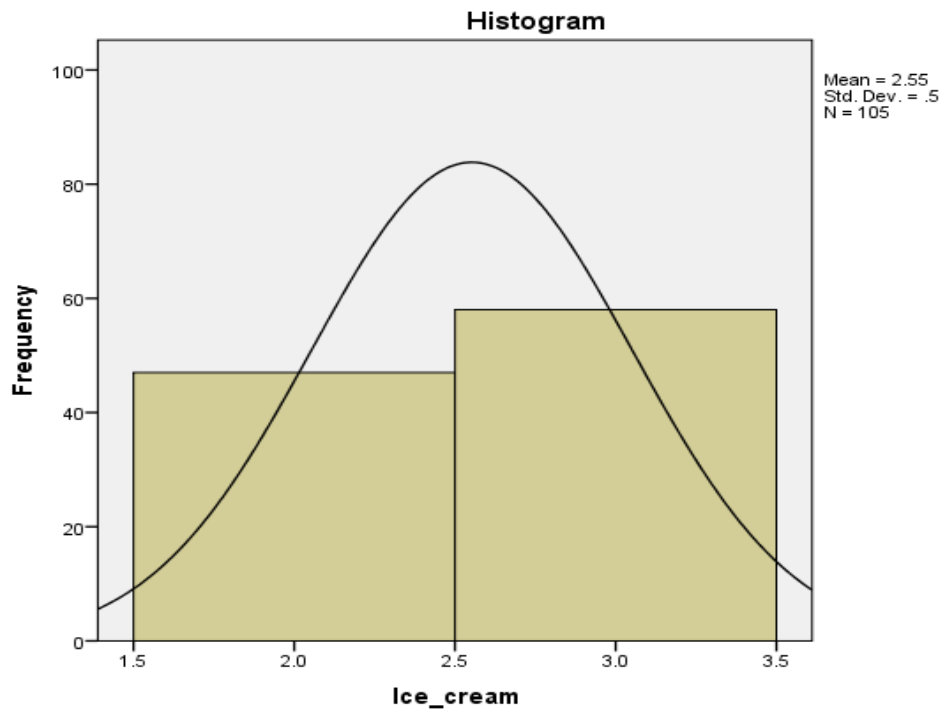


2.3 Presenting the result of mean and median of students like chocolate and strawberry flavor of ice crème

Frequencies

Statistics		
Ice_cream		
N	Valid	105
	Missing	0
Mean		2.55
Median		3.00

Ice_cream					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2	47	44.8	44.8	44.8
	3	58	55.2	55.2	100.0
	Total	105	100.0	100.0	



Interpretation: On the basis of the above diagram and chart, it is identified that the average number of participants or students who like chocolate and strawberry flavor of ice crème is 2.55.

Step required to follow:

File

Edit

View

Data

Transform

Analyze

Direct Marketing

Graphs

Utilities

Add-ons

Window

Help

13:

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

Define Variable Properties...

Set Measurement Level for Unknown...

Copy Data Properties...

New Custom Attribute...

Define Dates...

Define Multiple Response Sets...

Validation

Identify Duplicate Cases...

Identify Unusual Cases...

Sort Cases...

Sort Variables...

Transpose...

Merge Files

Restructure...

Aggregate...

Orthogonal Design

Copy Dataset

Split File...

Select Cases...

Weight Cases...

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

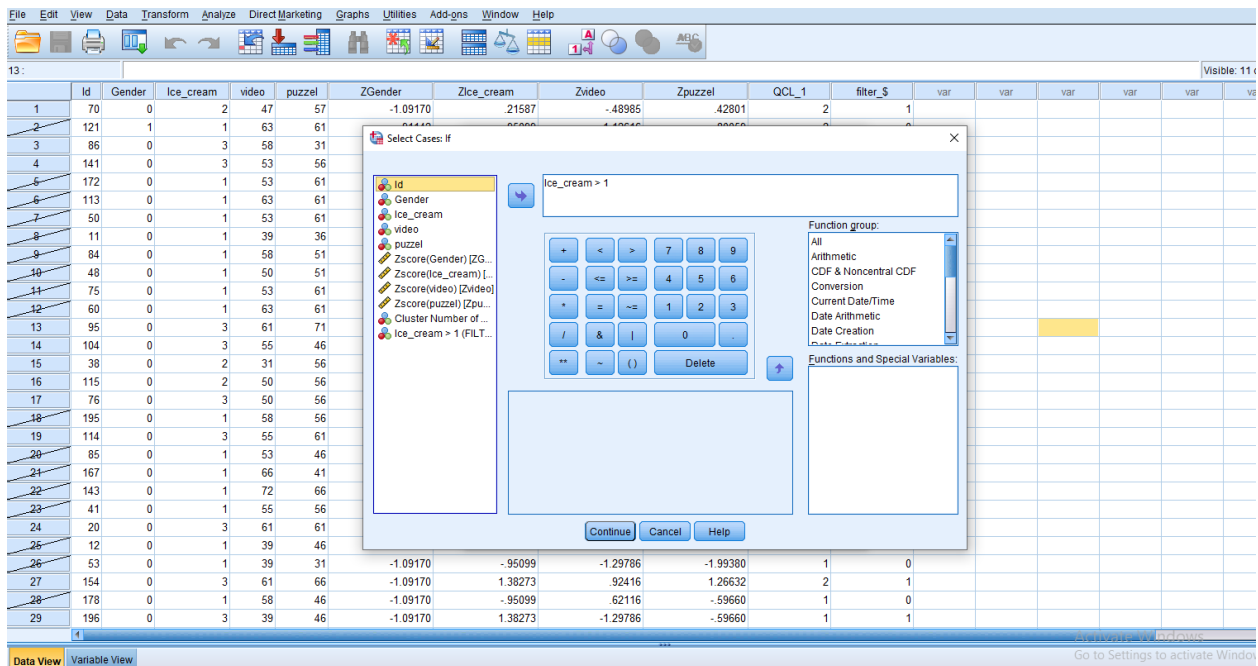
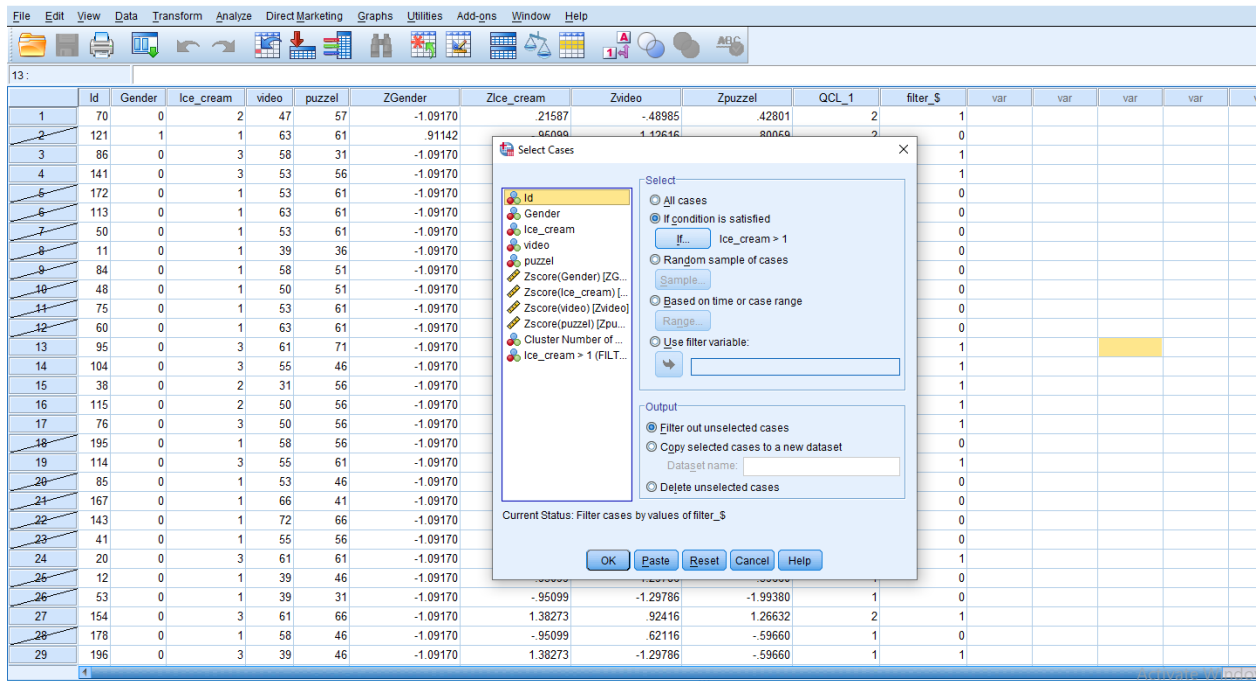
22

23

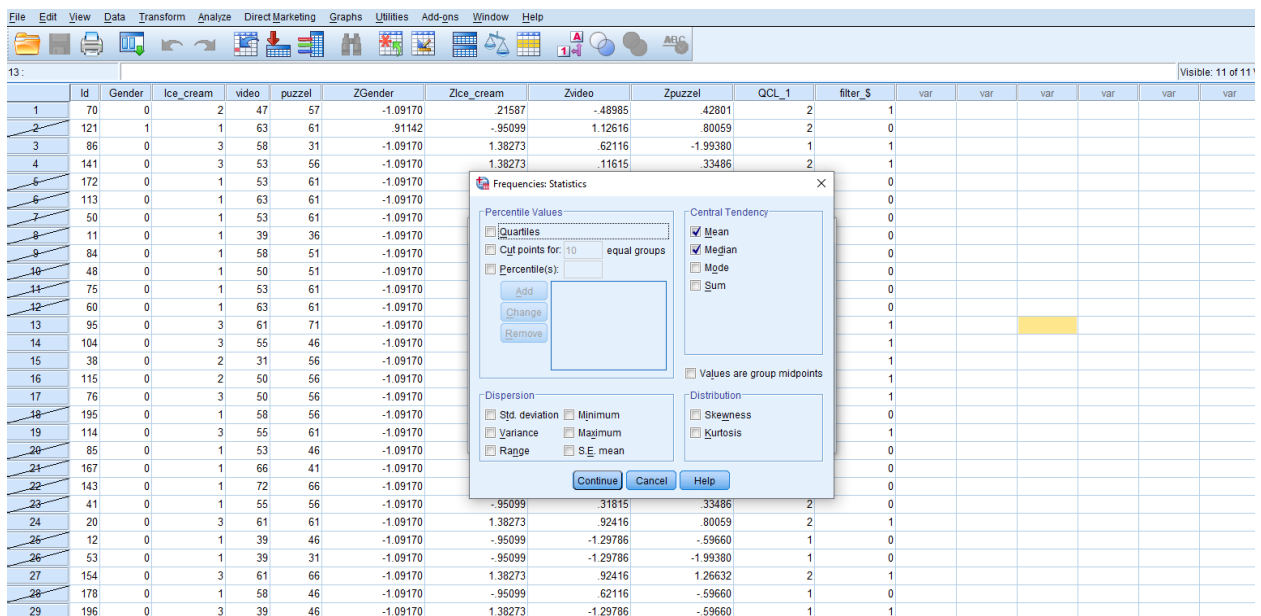
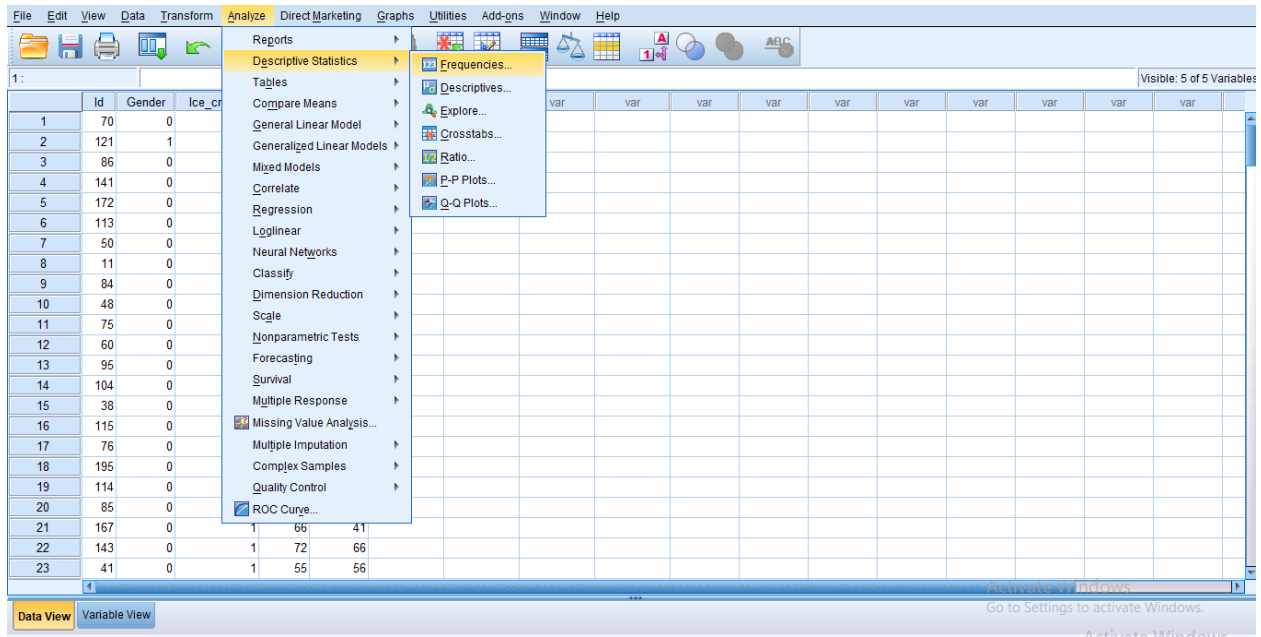
24

25

19



Step 3: Analyze → descriptive statistics → Frequencies → Ice crème → Statistics → Mean & median → OK



2.4 Presenting the result of K-mean clustering along with the step required to follow in the SPSS software

Step 1: Analyze → Descriptive statistics → Descriptive

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
Gender	200	0	1	.54	.499

Ice_cream	200	1	3	1.82	.857
video	200	26	74	51.85	9.901
puzzel	200	26	71	52.40	10.736
Valid N (listwise)	200				

Interpretation: From the above descriptive statistics table, it is analyzed that the majority of students plays video game because the maximum number of students play video game is 74. Further, it is also assessed from the mean value (0.54) is that the average number of students selected for the purpose of data collection and analysis are females (George and Mallery, 2018). Moreover, (1.82) signifies that the average number of students like chocolate flavor of ice crème.

Step 2 Analyze → Classify → K-mean cluster Analysis

The screenshot displays the IBM SPSS Statistics software interface. The 'Analyze' menu is open, and the 'Classify' option is selected, leading to the 'K-Means Cluster...' dialog box. The main data view shows a table with columns: Id, Gender, Ice_cream, Zscore, and several unlabeled columns. The 'K-Means Cluster...' dialog box is visible in the foreground, showing the 'Variables to Cluster' list and the 'Number of Clusters' set to 3. The 'Display' section is also visible, showing options for 'Display of Centroids' and 'Display of Iterations'.

Step 3 Select all Zscore variable → Selected iterate → maximum iteration limit (10) → continue

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

Visible: 9 of 9 Variables

	Id	Gender	Ice_cream	video	puzzel
1	70	0	2	47	57
2	121	1	1	63	61
3	86	0	3	58	31
4	141	0	3	53	56
5	172	0	1	53	61
6	113	0	1	63	61
7	50	0	1	53	61
8	11	0	1	39	36
9	84	0	1	58	51
10	48	0	1	50	51
11	75	0	1	53	61
12	60	0	1	63	61
13	95	0	3	61	71
14	104	0	3	55	46
15	38	0	2	31	56
16	115	0	2	50	56
17	76	0	3	50	56
18	195	0	1	58	56
19	114	0	3	55	61
20	85	0	1	53	46
21	167	0	1	66	41
22	143	0	1	72	66
23	41	0	1	55	56

K-Means Cluster Analysis

Variables:

- Zscore(Gender) [ZGender]
- Zscore(Ice_cream) [ZIce_cream]
- Zscore(video) [Zvideo]
- Zscore(puzzel) [Zpuzzel]

Label Cases by:

Number of Clusters: 2 Method: ☒ Iterate and classify ☐ Classify only

Cluster Centers

☐ Read initial:

- ☐ Open dataset
- ☐ External data file

☐ Write final:

- ☐ New dataset
- ☐ Data file

OK Paste Reset Cancel Help

Data View Variable View

Go to Settings to activate Windows.

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

13:

	Id	Gender	Ice_cream	video	puzzel	ZGender	ZIce_cream	Zvideo	Zpuzzel	QCL_1	filter_\$	var	v
1	70	0	2	47	57	-1.09170	.21587	-.48985	.42801	2	1	1	
2	121	1	1	63	61	.91142	-.95099	1.12616	.80059	2	0	1	
3	86	0	3	58	31	-1.09170						0	
4	141	0	3	53	56	-1.09170						0	
5	172	0	1	53	61	-1.09170						0	
6	113	0	1	63	61	-1.09170						0	
7	50	0	1	53	61	-1.09170						0	
8	11	0	1	39	36	-1.09170						0	
9	84	0	1	58	51	-1.09170						0	
10	48	0	1	50	51	-1.09170						0	
11	75	0	1	53	61	-1.09170						0	
12	60	0	1	63	61	-1.09170						0	
13	95	0	3	61	71	-1.09170						1	
14	104	0	3	55	46	-1.09170						1	
15	38	0	2	31	56	-1.09170						1	
16	115	0	2	50	56	-1.09170						1	
17	76	0	3	50	56	-1.09170						1	
18	195	0	1	58	56	-1.09170						0	
19	114	0	3	55	61	-1.09170						1	
20	85	0	1	53	46	-1.09170						0	
21	167	0	1	66	41	-1.09170						0	
22	143	0	1	72	66	-1.09170						0	
23	41	0	1	55	56	-1.09170						0	
24	20	0	3	61	61	-1.09170						1	
25	12	0	1	39	46	-1.09170	-.95099	-1.29786	-.59660	1	0	0	
26	53	0	1	39	31	-1.09170	-.95099	-1.29786	-1.99380	1	0	0	
27	154	0	3	61	66	-1.09170	1.38273	.92416	1.26632	2	1	1	
28	178	0	1	58	46	-1.09170	-.95099	.62116	-.59660	1	0	0	
29	196	0	3	39	46	-1.09170	1.38273	-1.29786	-.59660	1	1	1	

K-Means Cluster Analysis

Variables:

- Zscore(Gender) [ZGender]
- Zscore(Ice_cream) [ZIce_cream]
- Zscore(video) [Zvideo]
- Zscore(puzzel) [Zpuzzel]

Label Cases by:

Number of Clusters: 2 Method: ☒ Iterate and classify ☐ Classify only

Cluster Centers

☐ Read initial:

- ☐ Open dataset
- ☐ External data file

☐ Write final:

- ☐ New dataset
- ☐ Data file

OK Paste Reset Cancel Help

K-Means Cluster Analysis: It...

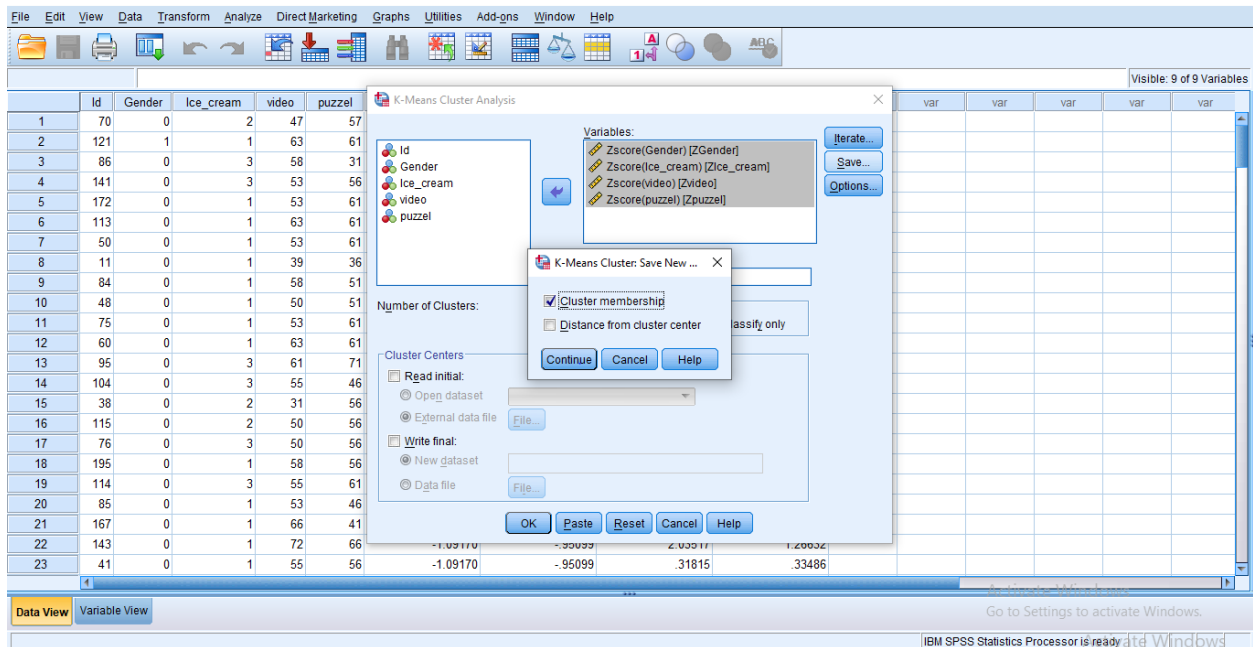
Maximum Iterations: 10

Convergence Criterion: 0

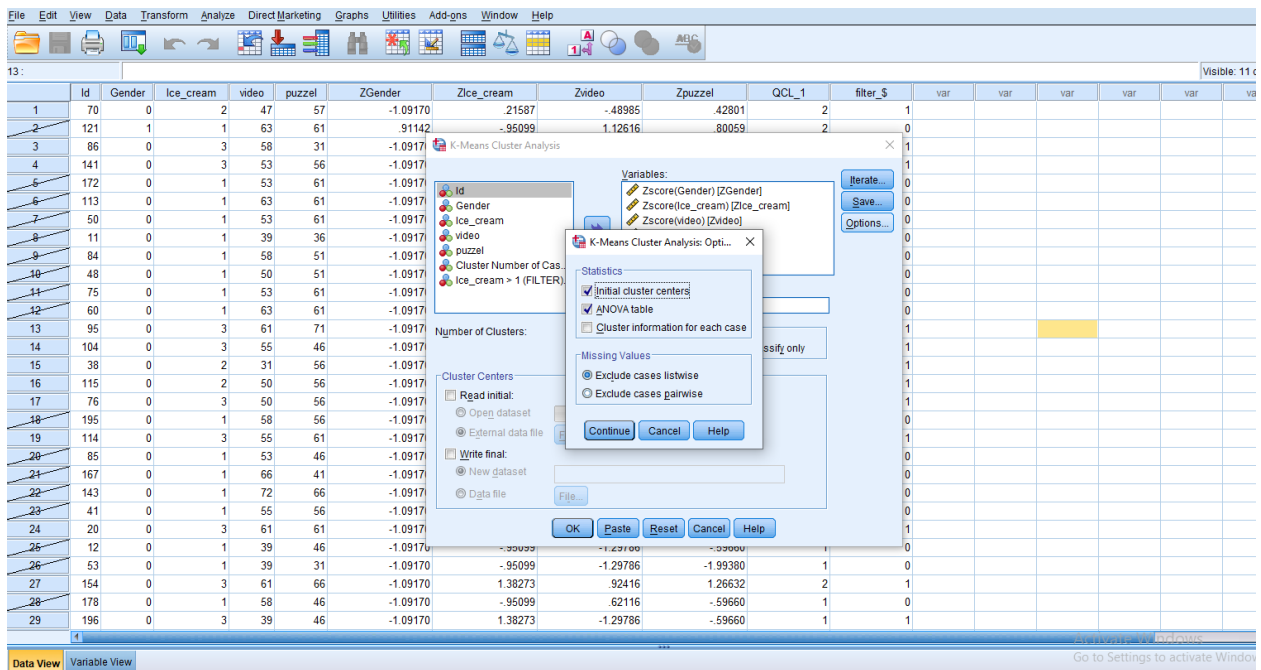
☐ Use running means

Continue Cancel Help

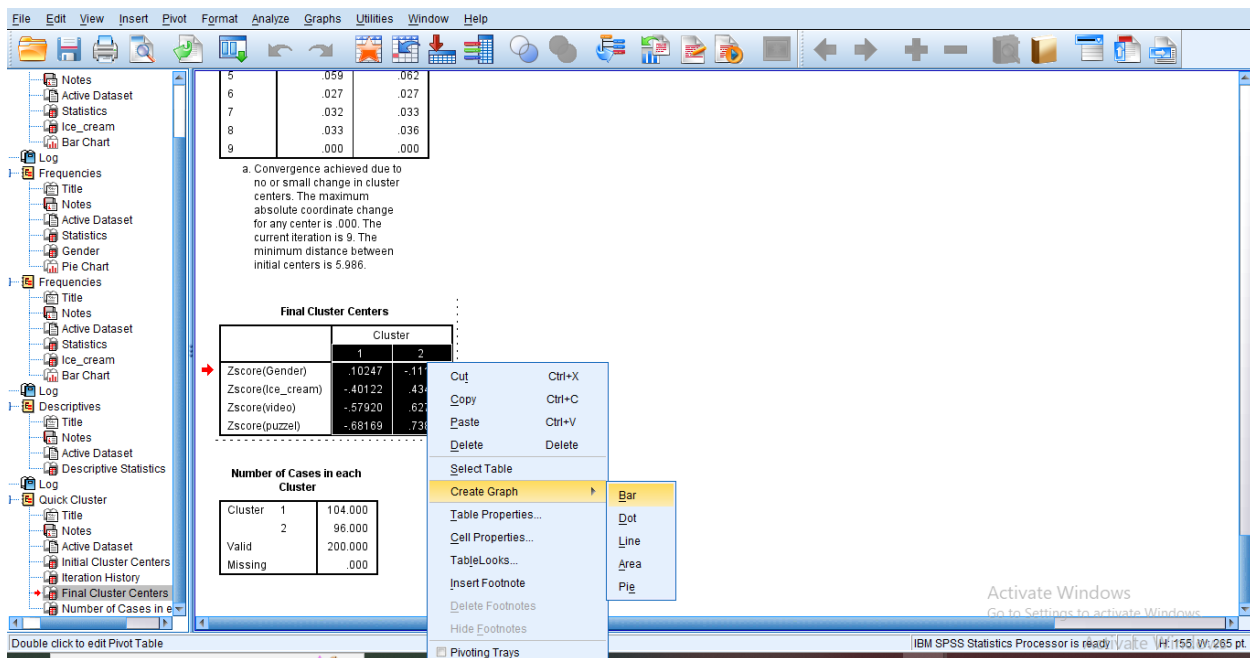
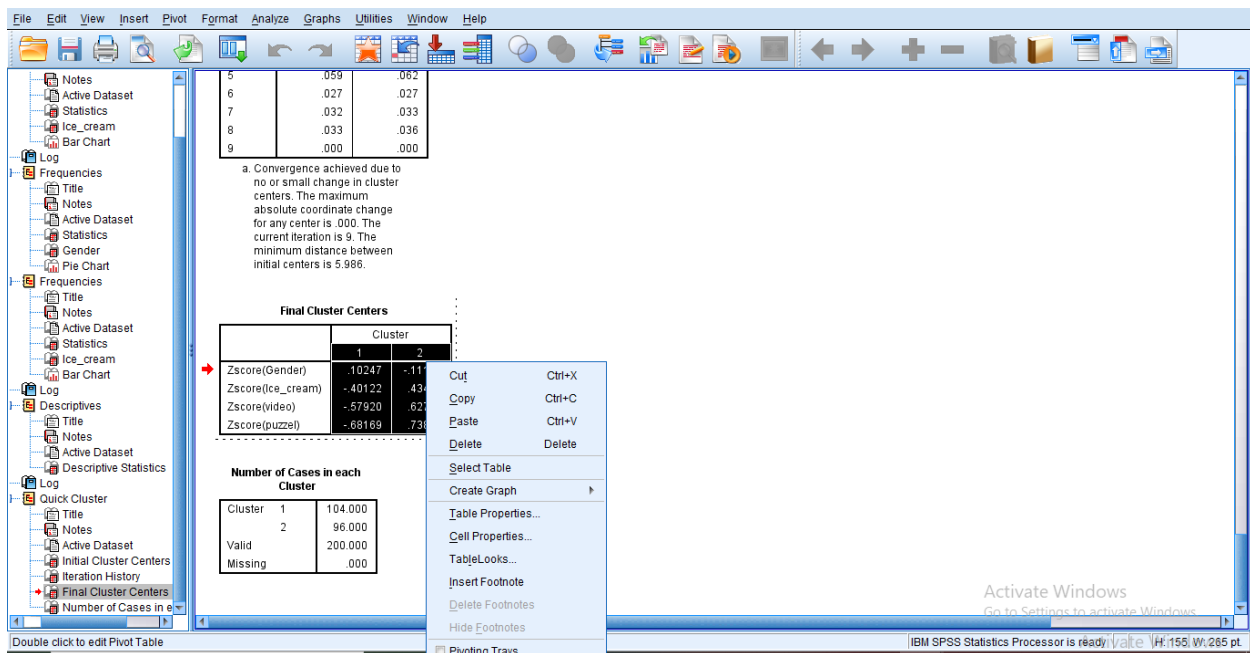
Step 4: Select save → cluster membership → continue



Step 5: Select options → click initial cluster centers and ANNOVA table → continue → OK



Step 6: Go to output file → select final cluster center value → right click → select create graph → Bar



Quick Cluster

Initial Cluster Centers

	Cluster

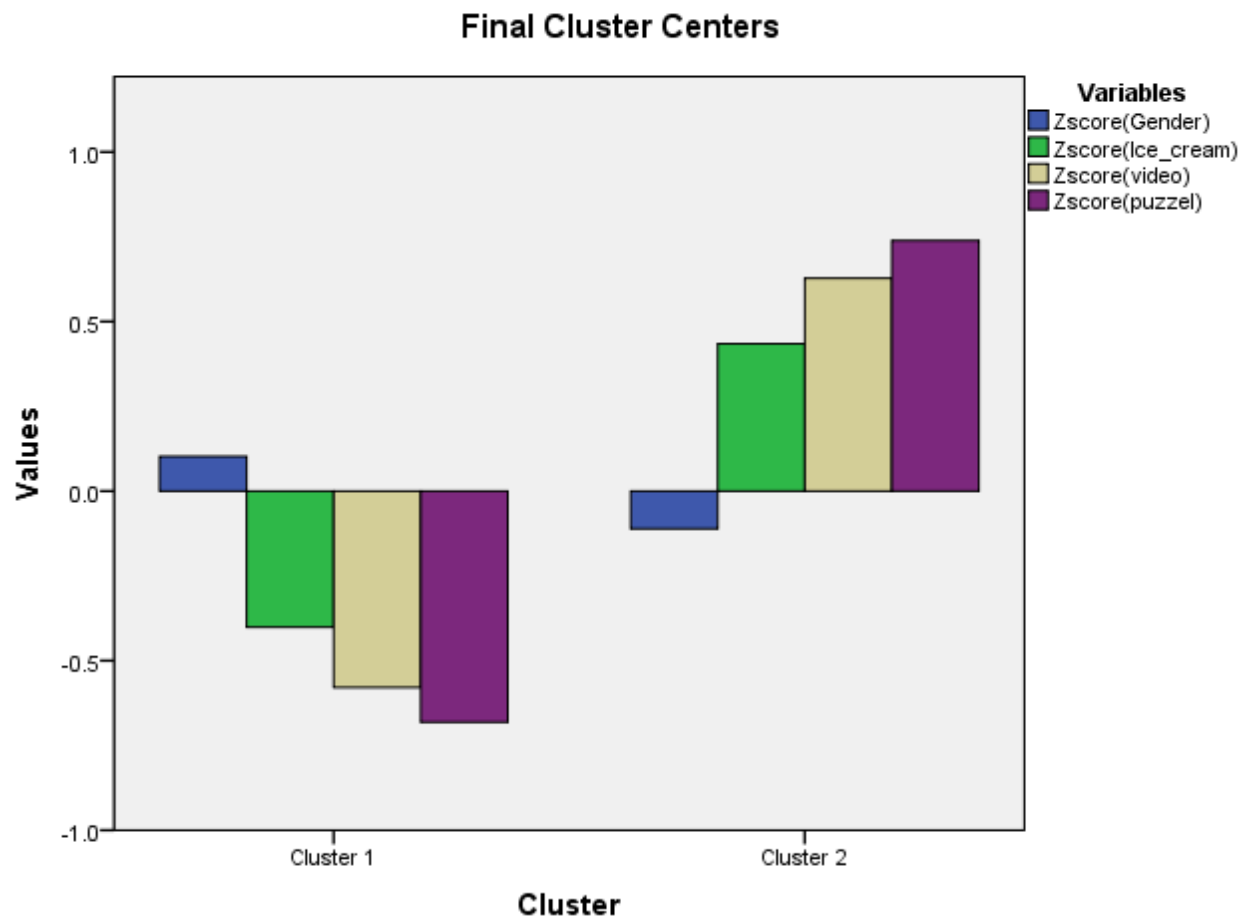
	1	2
Zscore(Gender)	-1.09170	.91142
Zscore(Ice_cream)	-.95099	1.38273
Zscore(video)	-1.80287	1.73217
Zscore(puzzel)	-1.99380	1.73206

Iteration History ^a		
Iteration	Change in Cluster Centers	
	1	2
1	2.087	1.986
2	.125	.132
3	.095	.118
4	.018	.019
5	.059	.062
6	.027	.027
7	.032	.033
8	.033	.036
9	.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 9. The minimum distance between initial centers is 5.986.		

Interpretation: On the basis of the above iteration table, it is analyzed or interpreted that there is clear progress in the iteration process. It is because in the second iteration, a major shift is clearly visible as compared to the first iteration table.

Final Cluster Centers	
	Cluster

	1	2
Zscore(Gender)	.10247	-.11101
Zscore(Ice_cream)	-.40122	.43466
Zscore(video)	-.57920	.62747
Zscore(puzzel)	-.68169	.73850



Interpretation: From the above table and chart, the following points are clearly assessed and interpreted:

- Firstly, it is interpreted that the students in cluster 1 are females. Also, they like more chocolate flavor of ice crème. Not only that, it is also assessed from the cluster 1 that the students of this cluster like playing video games as compared to puzzle game.

- Secondly, it is interpreted that the students in cluster 2 are males. Also, they like chocolate flavor of ice crème. This student prefers playing puzzle game as compared to video games.

Number of Cases in each Cluster		
Cluster	1	104.000
	2	96.000
Valid		200.000
Missing		.000

Interpretation: From the above table, it is identified that the cluster 1 has a value of 104 while on the other hand the cluster 2 has the value of 96. This indicate the significant difference between the cluster 1 and cluster 2.

2.5 Presenting most common data mining and text mining methods used by business with real world examples

In the case of large dataset, the data mining is useful for extracting and discovering patterns and relationships of dataset which further help in solving the business problems through data analysis. Data mining is a key part of data analytics and the most common data mining method is clustering. **Clustering** is the process of grouping a series of different data point on the basis of their characteristics to make informed decision (Ryabko, 2020).

For instance, M&S company uses the clustering method to group their customer based on their age, taste, income etc. This help the company to make decision regarding whether the customer will like their product or not.

The most famous text mining method is **information extraction** which is a process of collecting meaningful information from vast textual data.

For instance, Unilever uses information extraction method to collect the financial performance details of its competitors through their financial statements (Sibuya, 2019).

2.6 Presenting advantage and disadvantage of using SPSS and Excel using practical relevant examples

Particulars	Advantage	Disadvantage
SPSS	• The SPSS software package is time saving and less effort consuming. • With the help of this, the scholar able to work upon large number of data. • The information should be gathered in more conceptual way in the SPSS (Milligan and Cooper, 2019).	• It is quite expensive for the user because of high cost. • It requires providing training to the employees.
EXCEL	• The large number of data has been aligned, arranged in the excel which saves users time. • There are various basic to advance excel formulas and function that help in decision making to users (Norušis, 2018).	• The inserting of manual data in the excel is quite time consuming. • There is a chance of calculation error because of automatic formulas.

REFERENCES

Books and Journals

- Pallant, J., (2020). SPSS survival manual: A step by step guide to data analysis using IBM SPSS. Routledge.
- Betty Jane, J. and Ganesh, E. N., (2020). Big data and internet of things for smart data analytics using machine learning techniques. In International conference on computer networks, big data and IoT (pp. 213-223). Springer, Cham.
- Mihas, P., (2019). Qualitative data analysis. In Oxford research encyclopedia of education.
- Aryadoust, V. and Raquel, M. eds., (2019). Quantitative data analysis for language assessment volume II: Advanced methods. Routledge.
- Fernandes, O., and et.al., (2019), September. Quantitative data analysis: Small unmanned aerial systems at hurricane michael. In 2019 IEEE international symposium on safety, security, and rescue robotics (SSRR) (pp. 116-117). IEEE.
- Chireshe, R., (2020). Quantitative data analysis for information science professionals. In Handbook of research on connecting research methods for information science research (pp. 349-372). IGI Global.
- Winston, W., (2018). Microsoft Excel data analysis and business modeling. Microsoft press.
- Held, B., Moriarty, B. and Richardson, T., (2019). Microsoft Excel Functions and Formulas with Excel 2019/Office 365. Mercury Learning and Information.
- McCullough, B. D. and Wilson, B., (2022). On the accuracy of statistical procedures in Microsoft Excel 2000 and Excel XP. Computational Statistics & Data Analysis. 40(4). pp.713-721.
- George, D. and Mallery, P., (2018). Descriptive statistics. In IBM SPSS Statistics 25 Step by Step (pp. 126-134). Routledge.
- Norušis, M. J., (2018). SPSS 14.0 guide to data analysis. Upper Saddle River, NJ: Prentice Hall.
- Milligan, G. W. and Cooper, M. C., (1987). Methodology review: Clustering methods. Applied psychological measurement. 11(4). pp.329-354.
- Sibuya, M., (2019). A random clustering process. Annals of the Institute of Statistical Mathematics. 45(3). pp.459-465.
- Ryabko, D., (2020). Clustering processes. arXiv preprint arXiv:1004.5194.