



Name: Erika Massa

Student Number: 2011291

Module Title: Data Handling & Business Intelligence LCBB5000

Assignment Title: Assessment 2

Lecturer: Joy Botu

PART A

- **Introduction**

At its core Excel is a spreadsheet program that enables storage, organising and analysing of information. It provides the capacity to do complicated analysis to simple list management and according to Rosenberg (2022), it has possibly become the most essential computer programme in the world's businesses. The crucial role that Excel assumed in the workplace environment, has made it fundamental not only for the company but also for the job market, where having a strong knowledge of Excel can lead to more job opportunities and promotions (Cameron 2021).

The versatility of Excel has made its use pivotal in a wide area of the business e.g. in finance and accounting for analytical and computing features, in marketing to identify the trends, management accounting for forecasting and budgeting, and in human resources managing employee profiles and budgets, as well as arranging working hours (Rosenberg 2022).

Excel can be used for pre-processing the data, the step that prepares the data for data mining and data analysis. To run a proper analysis the data collected needs to be cleaned, which means transforming it into a readable format for the software. The data collected with different formats need to be processed in order to conduct an analysis, eg. data with different currencies need to be converted, or different value descriptors as female and woman need to be corrected with the same value. The step of adjusting and discarding irrelevant data is called Data cleaning and is crucial for pre-processing the data (Luengo et al. 2022).

When the data has been processed, is ready to be analysed. Excel can be used not only to pre-process the data to be used with additional programs and third parts but can itself elaborate analysis that helps businesses with decision-making and problem solving (Luengo et al. 2022). Excel has numerous functions to conduct analysis, and two of the main functions Pivot Table and IF Function, will be presented in the following assessment showing an example of how can be used to analyse the crime rates.

An additional tool provided by Excel to help businesses with decision-making and problem-solving is the ability to create charts to visualise the data. According to Analytik (2020), visualising the data can be extremely valuable for businesses as long spreadsheets with endless numbers can be challenging for the human brain to elaborate the information.

For instance, visualising data helps to determine patterns and elaborate the information delivering information in a way that is universal, faster and more efficient. Excel provides a

wide range of charts (e.g. Histogram, Pie, Bar chart, etc.) that are fast and simple to use and selecting the most appropriate charts is important to obtain the best results (Yi, 2021). The following assessment will provide an example of how incorrect use of the charts can lead to opposite results, providing more confusion than clarity.

- **The use of Pivot Table to determine the crime rate**

The Pivot Table is essentially a dynamic summary report generated from a database. It can help in the transformation of limitless rows and columns of numbers into a comprehensible data presentation and combine items into groups. It is an interactive tool that permits to easily summarize and visualize necessary data. For instance, the term 'Pivot' derives from the facility of rapidly rotating and using the data from different angles and perspectives. Besides, modifying a Pivot Table results rapid and easy and its use is highly flexible, it also has a few limitations, e.g. is not possible to modify the data or to create new rows and columns in the Pivot Table (Walkenbach 2007).

The Crime Rate has been calculated using the Pivot Table function of Excel. Starting with the data that has been provided on the 'Crime_data' sheet of excel (Fig. 1), the data has been selected to press 'Insert – Pivot Table', with the option of creating a new working sheet (Fig. 1).

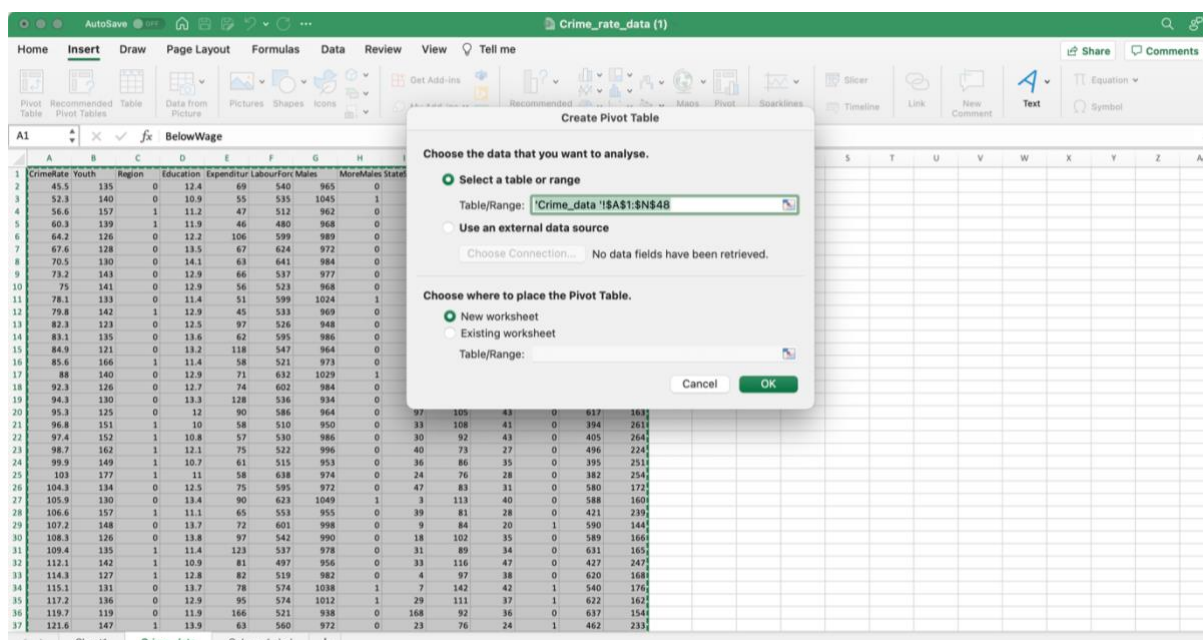


Fig. 1 (Source: Author)

In the new working sheet that has just been created, the PivotTable Fields have been amended, selecting only the data needed to calculate the Crime Rate. Besides selecting the

data label accurately, it is pivotal to place it in the correct fields. As Fig. 2 shows, the PivotTable Fields contain 4 different fields: Filters, Columns, Rows and Values, and using the same data label in different data fields, brings different results. In Fig. 2, the CrimeRate data label has been dragged into the 'Values' field, and the result obtained is shown on the left side of Fig. 2 where the Sum of CrimeRate = 4832. Once the data label is in the field, clicking on the 'i' is possible to access a wide range of options to summarize the data, eg. Sum, Count, Average, Max, etc. According to OAG (2014) the formula to calculate the crime rate is the following: the sum of all crime rates divided by the number of the population and the result is then multiplied by 100.

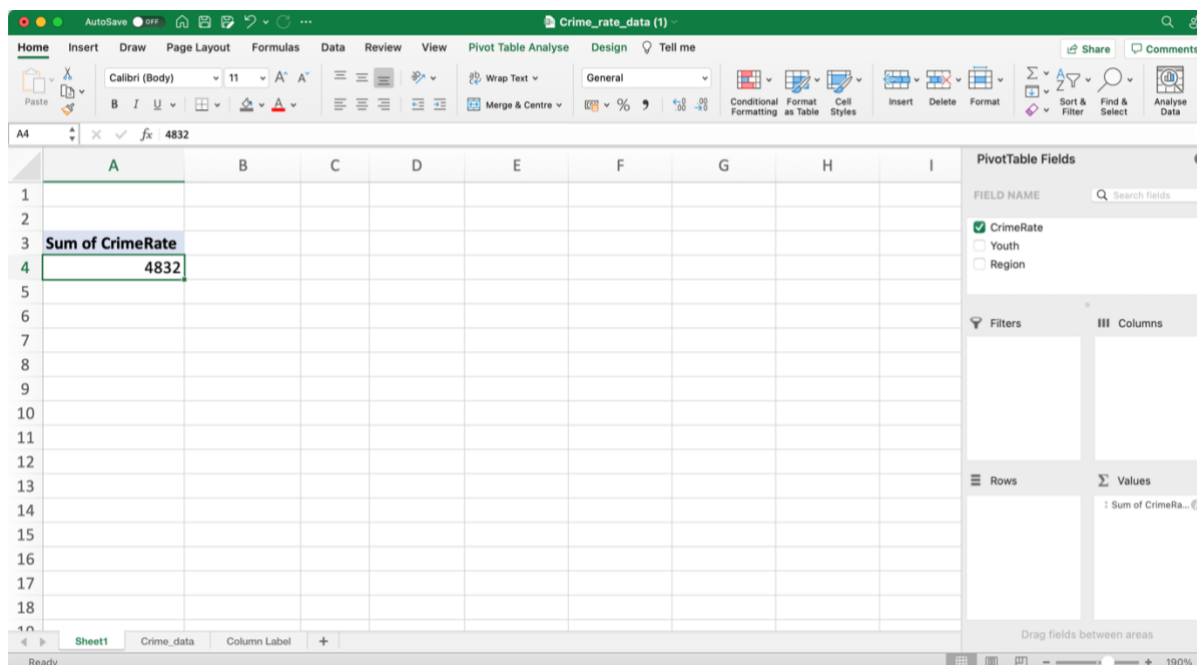


Fig. 2 (Source: Author)

The result 4832 is the sum of all the crimes. In Fig. 3 according to the formula to determine the crime rate, the sum of the entire crimes has been split for the one million of the population. In the cell, B4 has been inserted =GETPIVOTDATA("CrimeRate" and cell A4 has been selected to be entered in the formula. Once the formula was completed has been added /1000000 to determine the CrimeRate of both regions.

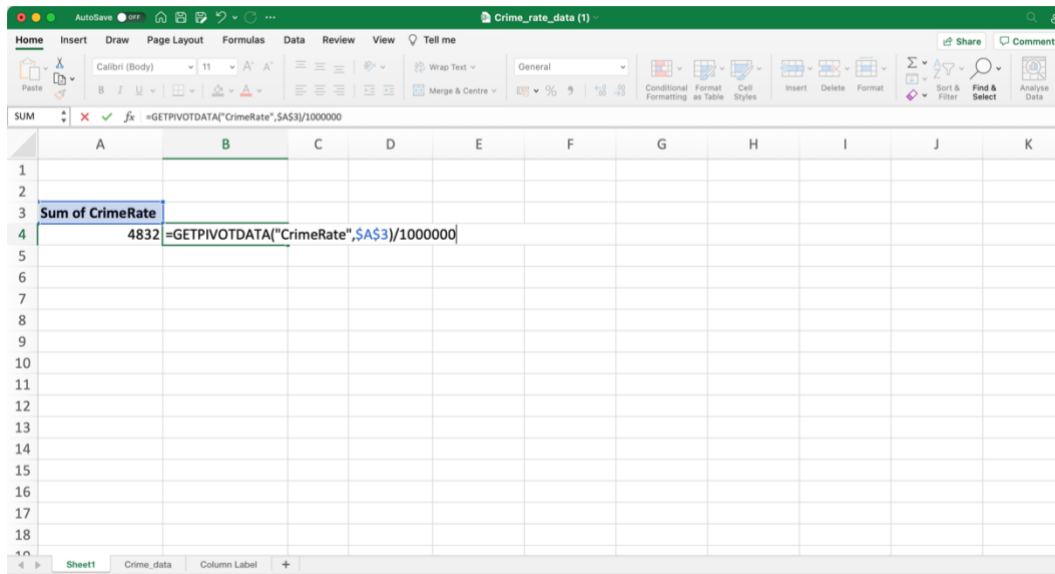


Fig. 3 (Source: Author)

The Crime Rate is 0.004832 (Fig. 4). According to the original formula, this result should have been multiplied by 100. However, as the data already represents the percentage of the crime rate, in this case, is not required to multiply the results by 100.

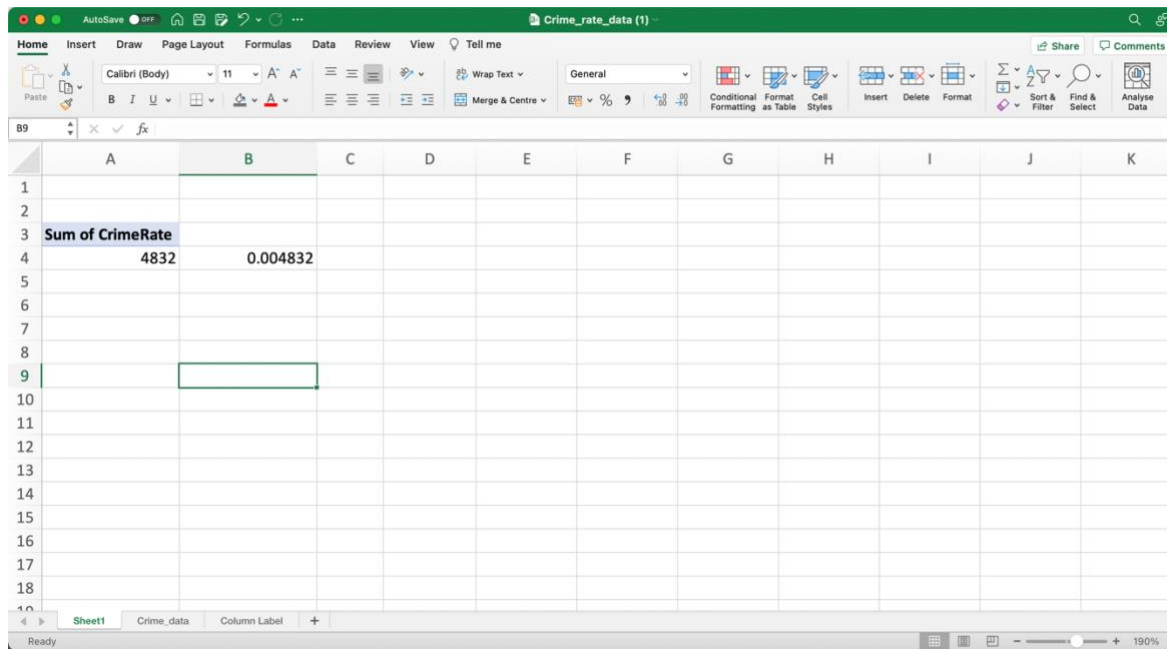
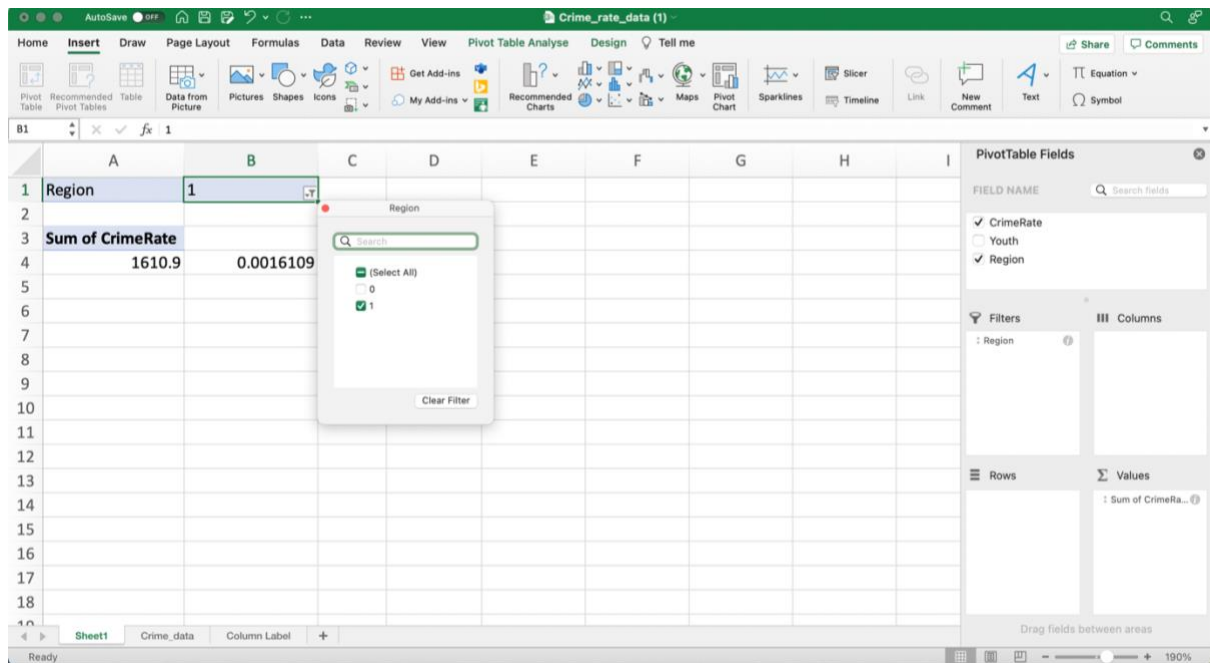


Fig. 4 (Source: Author)

As mentioned before, 4832 is the sum of all the crimes, and 0.004832 is the Crime Rate for all the regions. In Fig. 5 the data label 'Region' has been dragged into the filter to determine the sum and the Crime Rate of each region. The number 0 represents the Northern Region

and the number 1 represents the Southern Region. Filtering the number 1 (Fig. 5), the results show that the sum of Crime for the Southern Region is 1610.9 and the Crime Rate for the 1 million population is 0.0016109. Applying the same procedure but filtering the number 0, the result is that the sum of the Crime Rate for the Northern Region is 3221.1, while the Crime Rate is 0.0032211.



(Fig. 5 Source: Author)

- **The use of the IF function to determine the crime rate**

Excel enables the creation of dynamic models as the *what-if analysis*. These models exploit the formulas inserted to recalculate and alter the value in the cells according to the set of parameters inserted. Therefore, by undertaking a *what-if analysis*, is possible to modify the value in a cell in a systematic method and see the effect on certain formula cells (Walkenbach, 2007). The following is a manual type of *what-if analysis* that has been used to automatically evaluate the level of the crime rate for all regions.

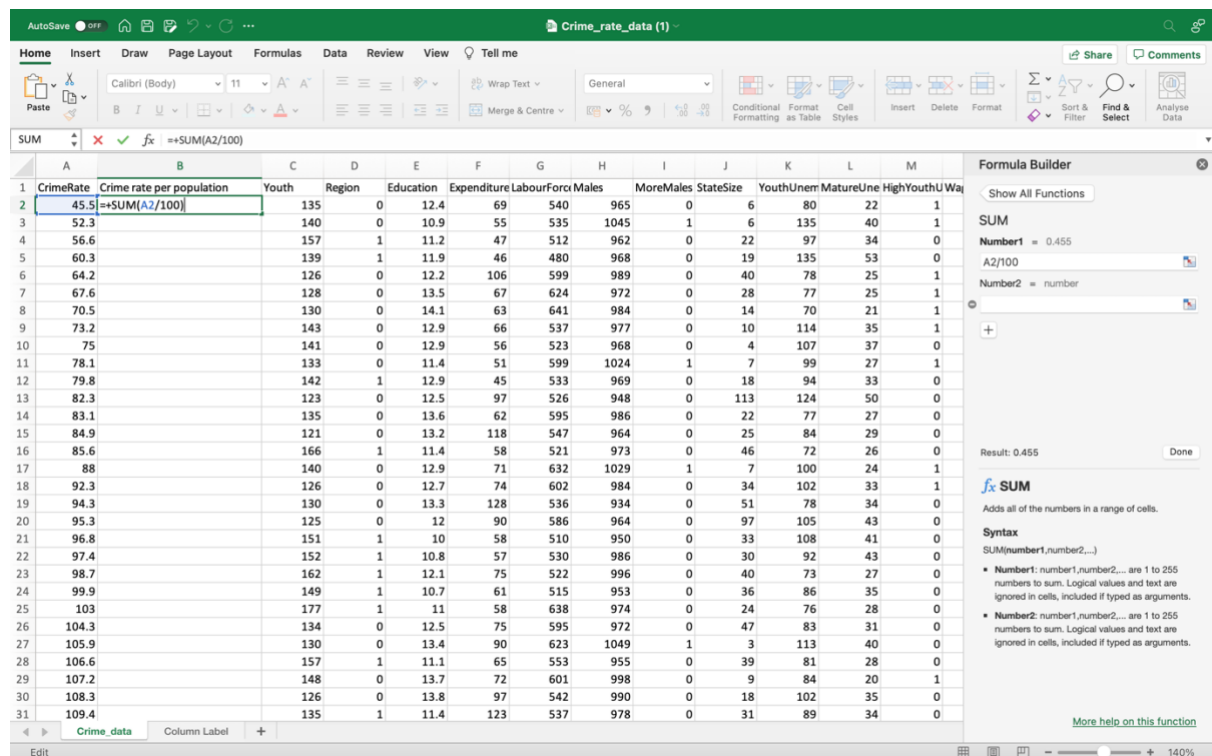


Fig. 6 (Source: Author)

As a first step, has been created another column next to column A. This column B has been labelled 'Crime rate per population' because in each row has been inserted the formula (Fig. 6) to calculate the crime rate for all the areas of the region. In this case, the number of the population used was 100 rather than 1000000 as the areas taken into consideration for the analysis are smaller.

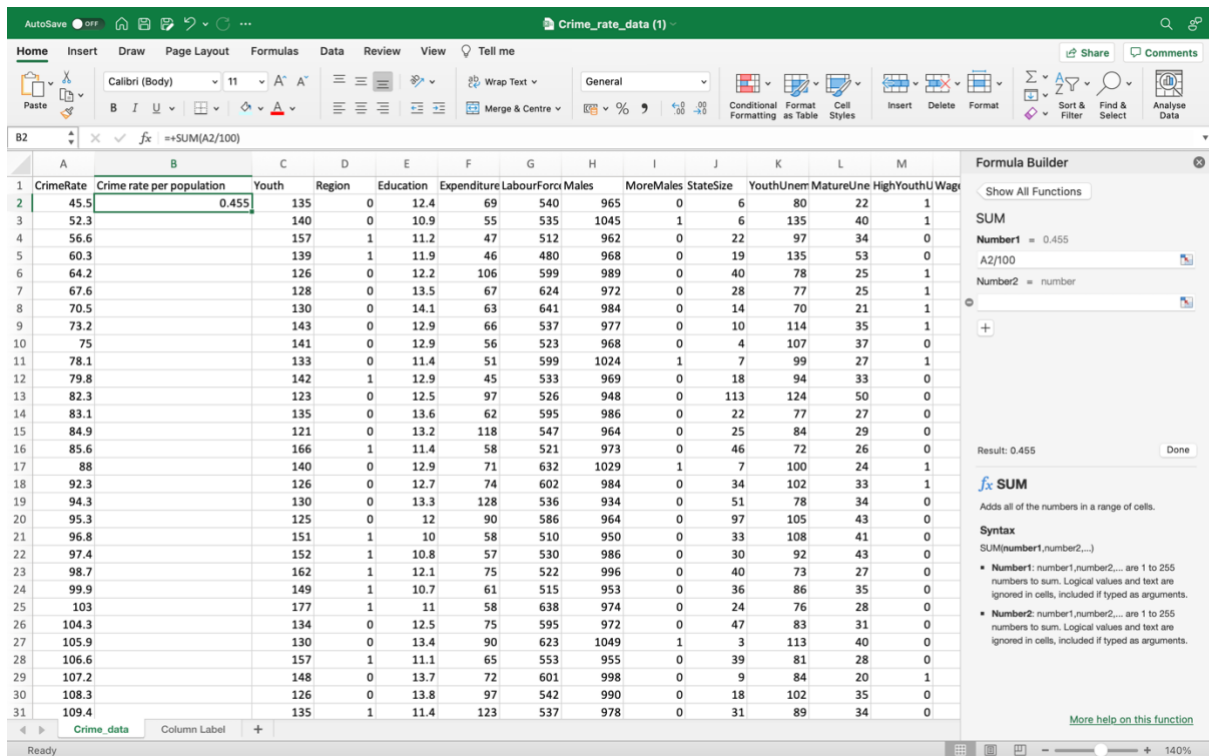


Fig. 7 (Source: Author)

The result is shown in Fig. 7 and is 0.455 for the first row. To know the results for each row the formula has been repeated for all the rows of column B with the following result (Fig. 8):

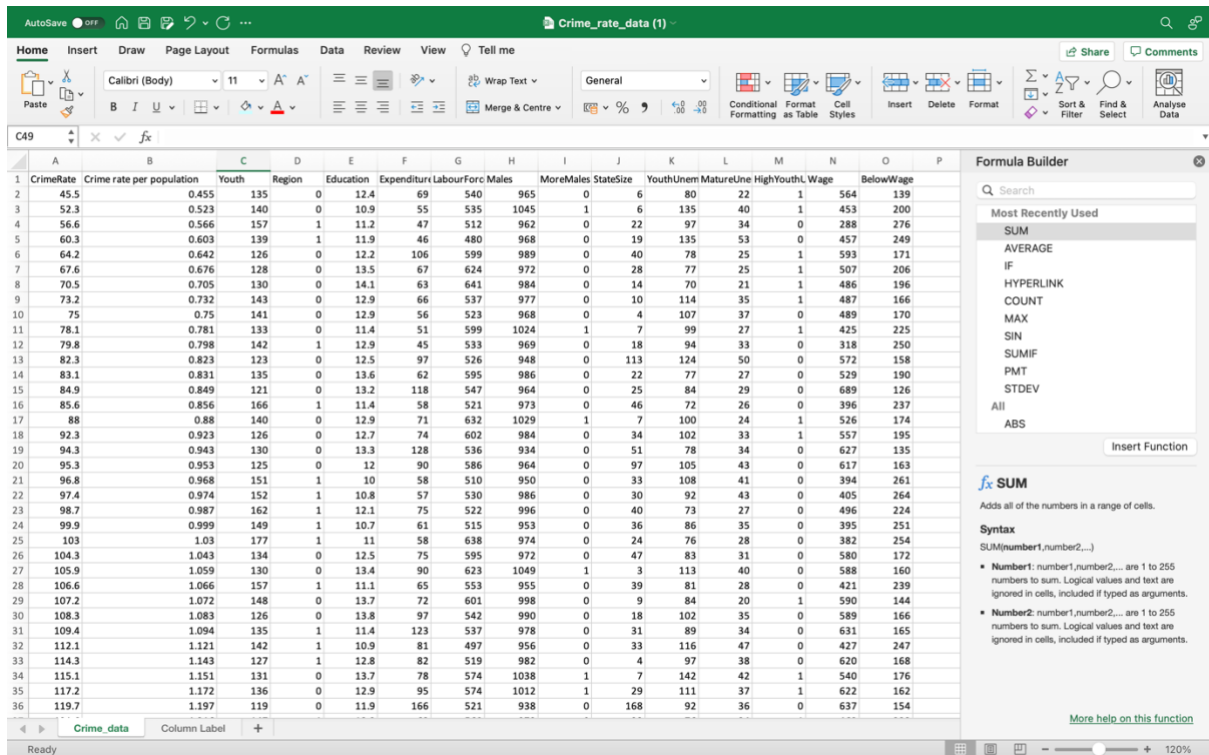


Fig. 8 (Source: Author)

Now that all the data has been calculated is possible to insert the parameters for the *what-if analysis* in the new column 'C'. The formula and the set of parameters that has been used is:
`=IF(B2<0.5,"Very Low",IF(B2<0.8,"Low",IF(B2<1.2,"Medium",IF(B2<1.8,"High","Really High"))))`

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
	CrimeRate	Crime rate per population	Level of Crime Rate	Youth	Region	Education	Expenditur	LabourForc	Males	MoreMale	StateSize	YouthUnen	MatureUn	HighYouth	Wage
2	45.5	0.455	=IF(B2<0.5,"Very Low",IF(B2<0.8,"Low",IF(B2<1.2,"Medium",IF(B2<1.8,"High","Really High"))))												
3	52.3	0.523	Low	140	0	10.9	55	535	1045	1	6	135	40	1	
4	56.6	0.566	Low	157	1	11.2	47	512	962	0	22	97	34	0	
5	60.3	0.603	Low	139	1	11.9	46	480	968	0	19	135	53	0	
6	64.2	0.642	Low	126	0	12.2	106	599	989	0	40	78	25	1	
7	67.6	0.676	Low	128	0	13.5	67	624	972	0	28	77	25	1	
8	70.5	0.705	Low	130	0	14.1	63	641	984	0	14	70	21	1	
9	73.2	0.732	Low	143	0	12.9	66	537	977	0	10	114	35	1	
10	75	0.75	Low	141	0	12.9	56	523	968	0	4	107	37	0	
11	78.1	0.781	Low	133	0	11.4	51	599	1024	1	7	99	27	1	
12	79.8	0.798	Low	142	1	12.9	45	533	969	0	18	94	33	0	
13	82.3	0.823	Medium	123	0	12.5	97	526	948	0	113	124	50	0	
14	83.1	0.831	Medium	135	0	13.6	62	595	986	0	22	77	27	0	
15	84.9	0.849	Medium	121	0	13.2	118	547	964	0	25	84	29	0	
16	85.6	0.856	Medium	166	1	11.4	58	521	973	0	46	72	26	0	
17	88	0.88	Medium	140	0	12.9	71	632	1029	1	7	100	24	1	
18	92.3	0.923	Medium	126	0	12.7	74	602	984	0	34	102	33	1	
19	94.3	0.943	Medium	130	0	13.3	128	536	934	0	51	78	34	0	
20	95.3	0.953	Medium	125	0	12	90	586	964	0	97	105	43	0	
21	96.8	0.968	Medium	151	1	10	58	510	950	0	33	108	41	0	
22	97.4	0.974	Medium	152	1	10.8	57	530	986	0	30	92	43	0	
23	98.7	0.987	Medium	162	1	12.1	75	522	996	0	40	73	27	0	
24	99.9	0.999	Medium	149	1	10.7	61	515	953	0	36	86	35	0	
25	103	1.03	Medium	177	1	11	58	638	974	0	24	76	28	0	
26	104.3	1.043	Medium	134	0	12.5	75	595	972	0	47	83	31	0	
27	105.9	1.059	Medium	130	0	13.4	90	623	1049	1	3	113	40	0	
28	106.6	1.066	Medium	157	1	11.1	65	553	955	0	39	81	28	0	

Fig. 9 (Source: Author)

After the formula has been inserted on cell C2, the function has been repeated for each row of the column. With this function has been possible to automatically calculate if the level of the crime rate in the area is Very Low, Low, Medium, High or Very High.

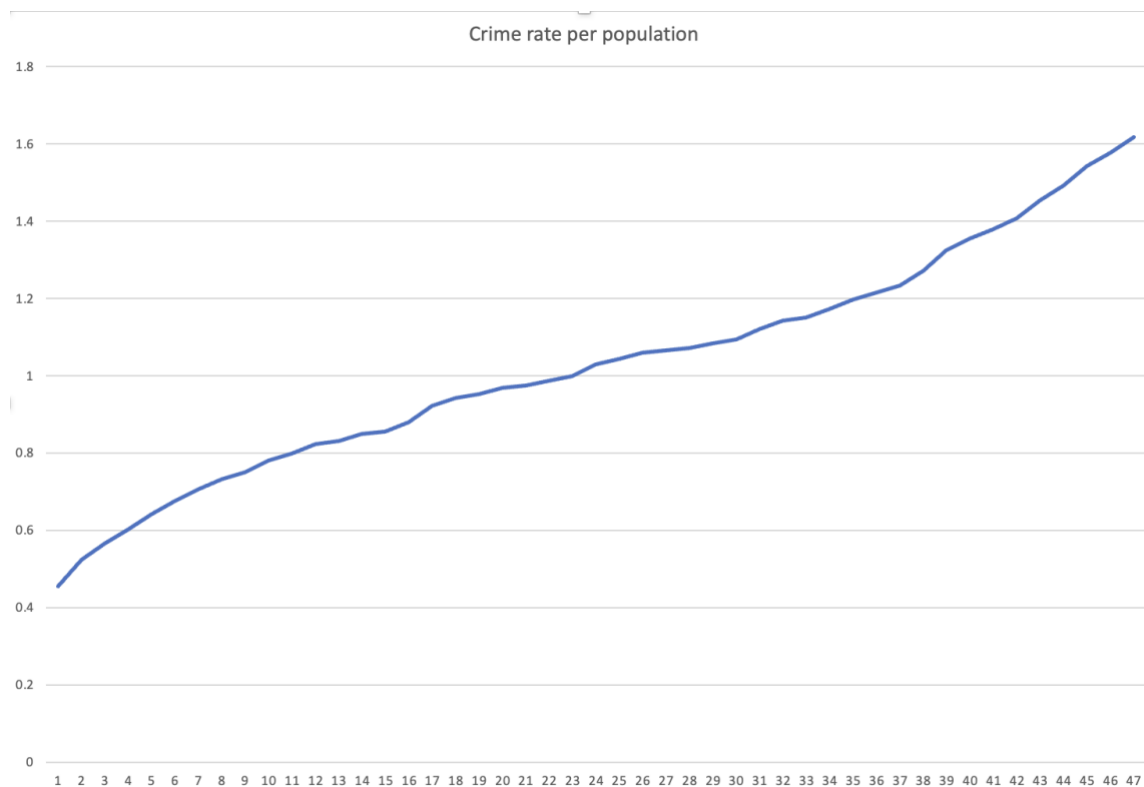


Fig. 10 (Source: Author)

Fig. 10 is the visual representation of the crime rate per population. Graphs can make the data more readable, helping spotting patterns or relationships between multiple data sets (Walkenbach, 2007). However, in Fig.10 has been used only the new data set 'Crime rate per population' as the purpose was to help visualize the different levels of crime rates in the different areas.

There are numerous options of graphs to use in Excel and each of them has different characteristics, advantages, or limitations to serve different goals, for this reason, is recommended to have clear the purpose of the chart before starts creating one (Yi, 2021).

The graph that has been chosen is the Line Graph Chart because as Muth (2018) suggests, it seemed to be the more suitable to represent a high number of values, e.g. employment rate, population and in this case the crime rate. For instance, charts, as Pie charts are not indicated, for representing a large number of different data as the results would be cluttered and unclear (Muth, 2018). To support this statement in Fig. 11 is shown the crime rate per population using Pie graphs charts rather than the Line Graph as in Fig. 10.

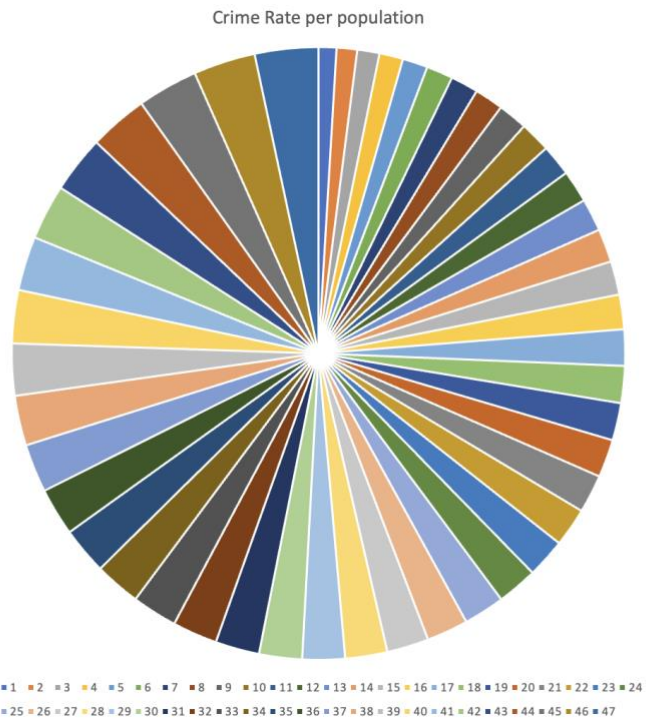


Fig. 11 (Source: Author)

PART B

As already mentioned above, the first step before conducting an analysis is pre-processing the data. The SPSS once opened is composed of two tabs: the Data View and Variable View, where is possible to see additional data. As a first step, the data has been adjusted, and the values for gender and ice cream have been added, otherwise, the analysis would have been incorrect. SPSS, in fact, only recognise numerical data, therefore when grouping variable together is needed to assign a variable number for each category (Hinton 2004). E.g., in this case, have been assigned in the group Gender: 0 for Male and 1 for Female: in the group Ice cream flavour: 1 for Vanilla, 2 for Chocolate and 3 for Strawberry.

- **How many students like vanilla flavour?**

On a total of 200 students, the answer is that 95 students (47.5% of all students) like vanilla flavour. In Fig. 13 in Statistic is possible to see the number of students that have been counted for the frequencies. The N. Valid is 200 and the N. Missing is 0. In the scenario where the N. Missing is different from 0 the analysis is not correct because it means some value is missed.

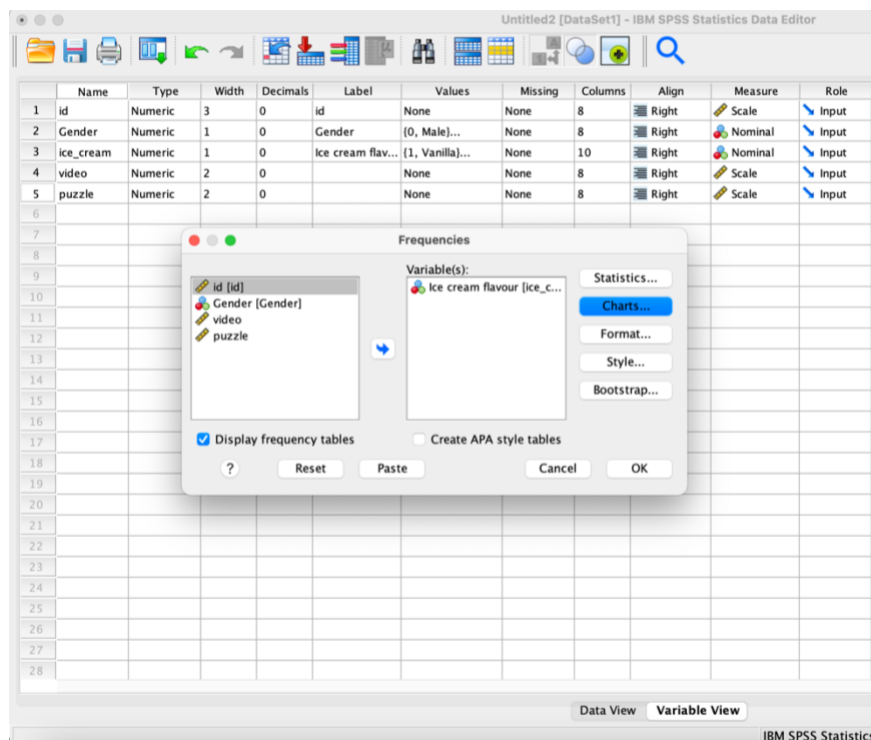


Fig. 12 (Source: Author)

Frequencies

Notes

Output Created		06-MAY-2022 13:52:05
Comments		
Input	Data	/Users/Erika/Desktop/SPSS Ass 2.sav
	Active Dataset	DataSet1
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	200
Missing Value Handling	Definition of Missing	User-defined missing values are treated as missing.
	Cases Used	Statistics are based on all cases with valid data.
Syntax		FREQUENCIES VARIABLES=ice_cream /HISTOGRAM /ORDER=ANALYSIS.
Resources	Processor Time	00:00:00.22
	Elapsed Time	00:00:00.00

Statistics

Ice cream flavour

N	Valid	200
	Missing	0

Ice cream flavour

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Vanilla	95	47.5	47.5	47.5
	Chocolate	47	23.5	23.5	71.0
	Strawberry	58	29.0	29.0	100.0
	Total	200	100.0	100.0	

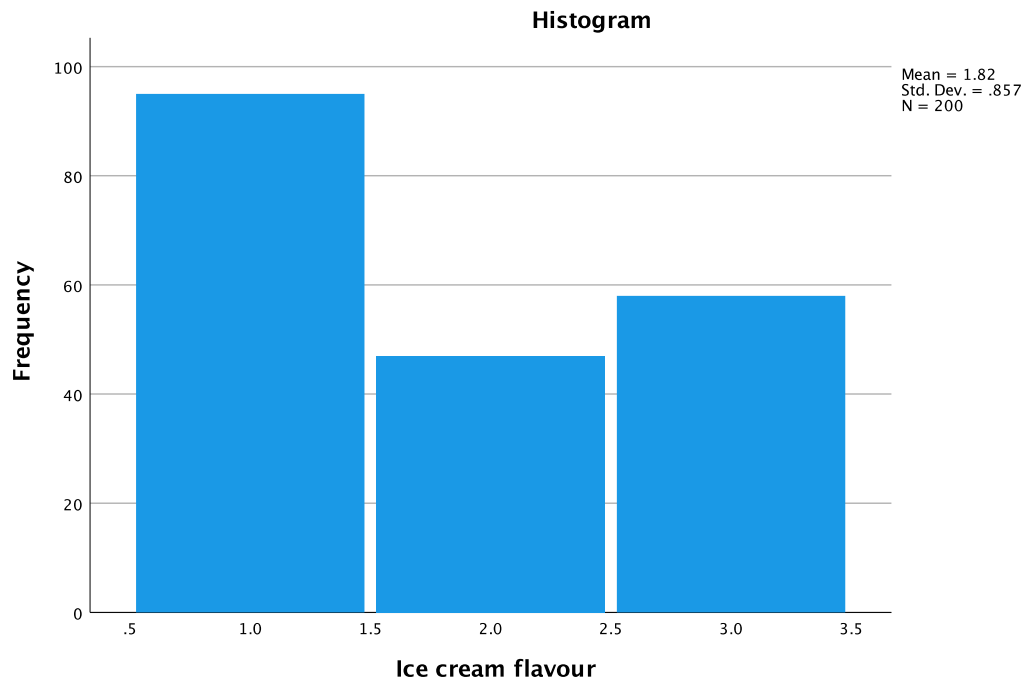


Fig. 13 (Source: Author)

To achieve this result has been selected, from the drop-down menu, 'Analyze' and then 'Descriptive Statistic', and last 'Frequencies'. Descriptive statistics are essentially describing quantitatively the data, therefore are useful to briefly summarize the data. There are three main divisions of Descriptive statistics, one that calculates the central tendency as mean median and average, another one that calculates the variability and the last one is the Frequencies that have been used in Fig. 12, as it counts the amount of times each variable appears (Berkman and Raise, 2016).

To support the statistic, Fig. 13 has also been selected 'Chart' and a Histogram chart has been chosen to show the statistical result. The Histogram chart is useful to view how the variables are distributed.

- **How many students are Male and Female?**

On a total of 200 students, the answer is 91 students (45.5% of all students) are Male, and 109 students (54.5% of all students) are Female.

Frequencies

Notes

Output Created		12-APR-2022 12:38:41
Comments		
Input	Active Dataset	DataSet1
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	200
Missing Value Handling	Definition of Missing	User-defined missing values are treated as missing.
	Cases Used	Statistics are based on all cases with valid data.
Syntax		FREQUENCIES VARIABLES=Gender /STATISTICS=MEAN /BARCHART FREQ /ORDER=ANALYSIS.
Resources	Processor Time	00:00:00.17
	Elapsed Time	00:00:00.00

Statistics

Gender

N	Valid	200
	Missing	0
Mean		.55

Gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	91	45.5	45.5	45.5
	Female	109	54.5	54.5	100.0
	Total	200	100.0	100.0	

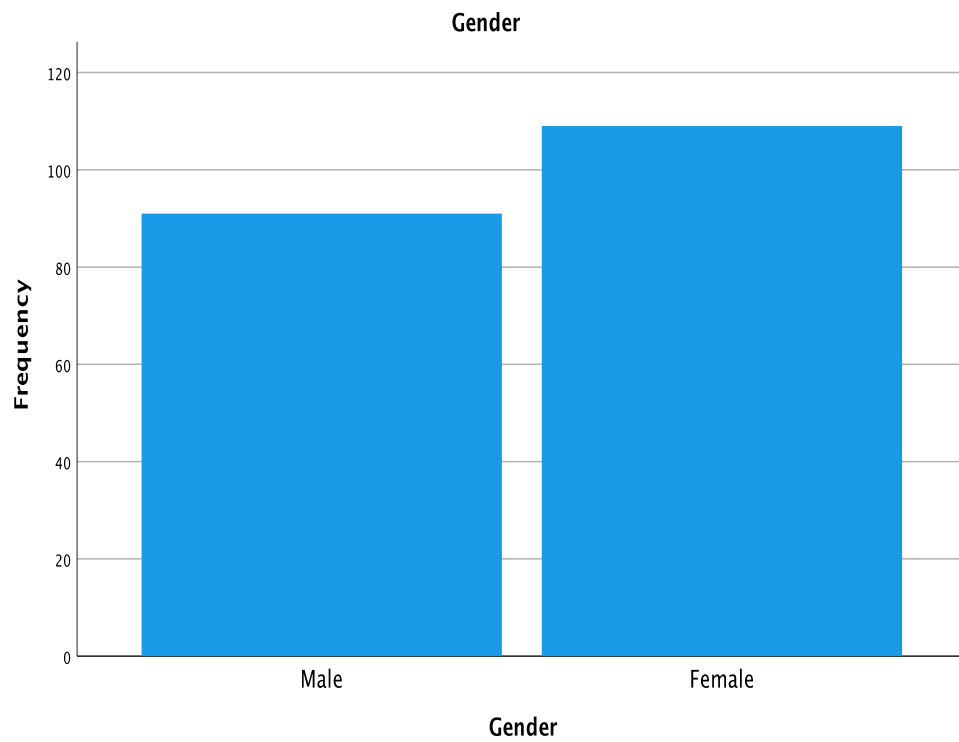


Fig. 14 (Source: Author)

- **What are the Mean and Median of participants who likes chocolate and strawberry flavour?**

The mean is the average of the data set and is measured by summing up all the values and dividing the result by the number of values in the set, e.g. in this case 200 (number of participants). The median corresponds to the value in the middle when all the values are sorted from the lowest to the highest value (Berman and Wang, 20018).

- The Mean of participants who like chocolate flavour is 74.55
- The Median of participants who like chocolate flavour is 63
- The Mean of participants who like strawberry flavour is 109.81
- The Median of participants who like strawberry flavour is 101.50

To draw this conclusion, on the drop-down menu has been selected again 'Analyze', but instead than open 'Descriptive Statistics', this time has been selected 'Compare Means' and 'Means'. In Fig. 15 is possible to see the option that has been dragged in the different tabs to achieve the result in Fig. 16.

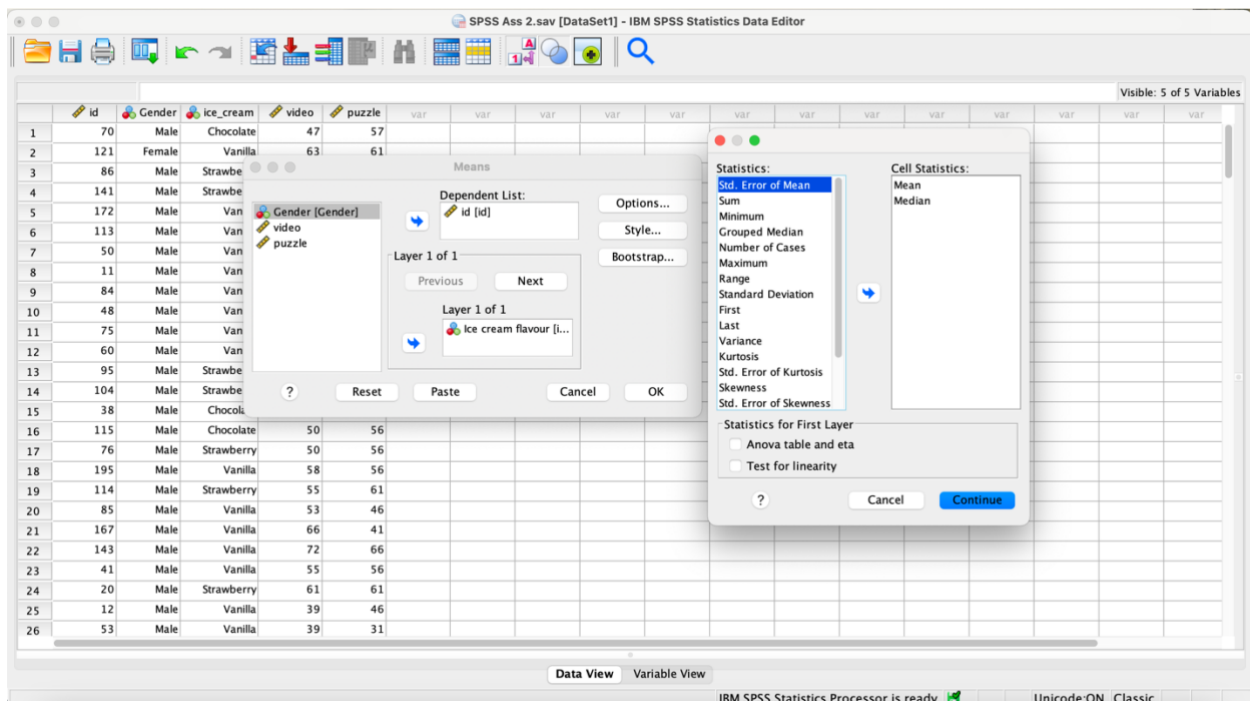


Fig. 15 (Source: Author)

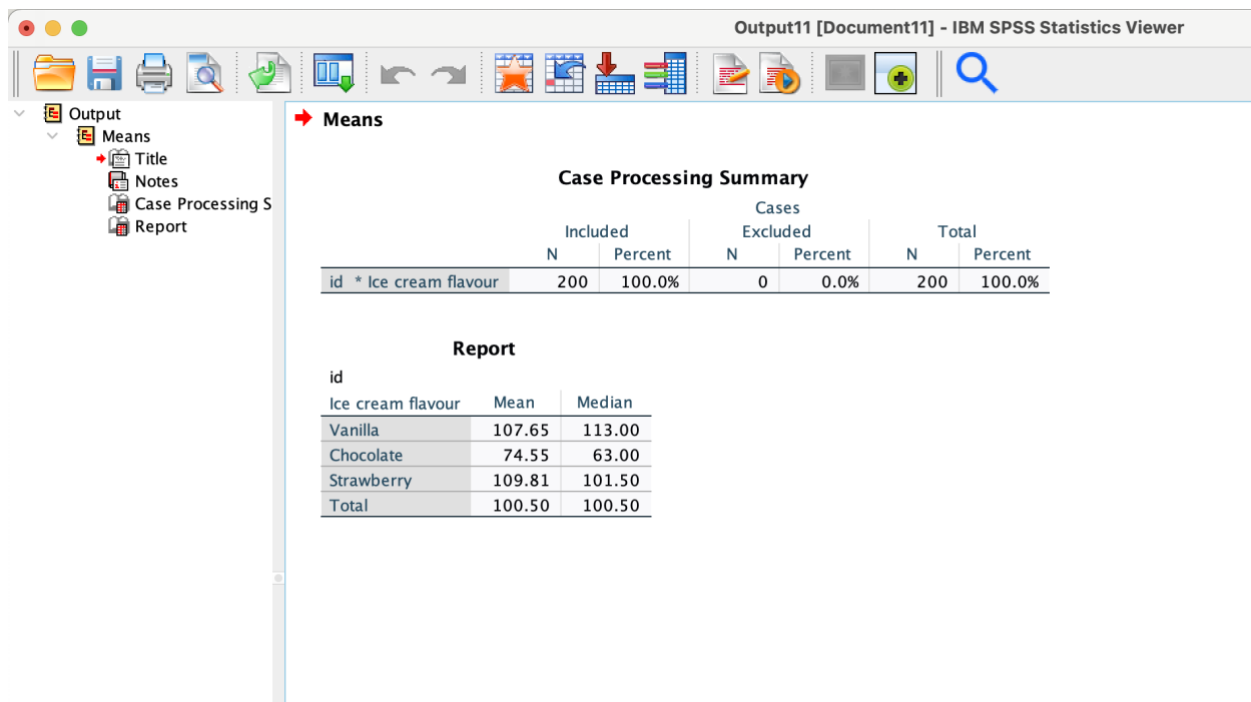


Fig. 16 (Source: Author)

- a) Using the dataset Ice_cream_csv provided in conjunction with SPSS gives a specific example of clustering (ex. K-means). Show your workings with screenshots and clearly explain your results.

Clustery Is a kind of data mining method that groups the data into segments based on their likeness and characteristics. Therefore, the cluster will find the similarity within the selected data sets, e.g. for this case between the video game scores and the puzzle scores. The main purpose of creating a cluster is for businesses to develop efficient marketing strategies as it helps to determine the market segmentation. For instance, businesses should use different marketing strategies for different groups of clustering (Blokdyk, 2021).

The following cluster has been used the data set of video games scores and the puzzle scores, both scale measurements. Nominal data, as gender or ice cream flavour, in fact, cannot be used for cluster analysis (Blokdyk, 2021). During a K-Means cluster analysis, the number of clusters generated can be any, for instance, is recommended to run more analyses to achieve an accurate result (Blokdyk, 2021). In the following assessment, the number of clusters that have been created is two with 10 iterations.

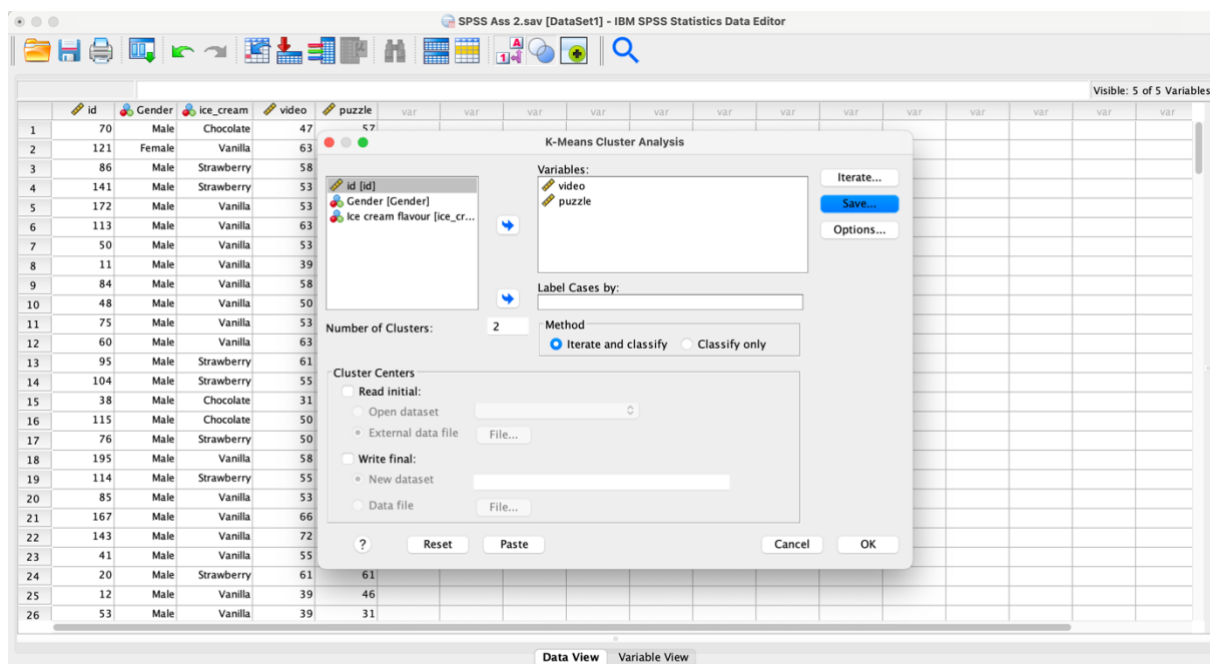


Fig. 17 (Source: Author)

Initial Cluster Centers

	Cluster	
	1	2
video	29	74
puzzle	26	66

Iteration History^a

Iteration	Change in Cluster Centers	
	1	2
1	21.986	18.375
2	.739	.631
3	1.084	.969
4	.215	.219
5	.285	.262
6	.000	.000

Final Cluster Centers

	Cluster	
	1	2
video	45	58
puzzle	44	60

Number of Cases in each Cluster

Cluster	1	95.000
	2	105.000
Valid		200.000
Missing		.000

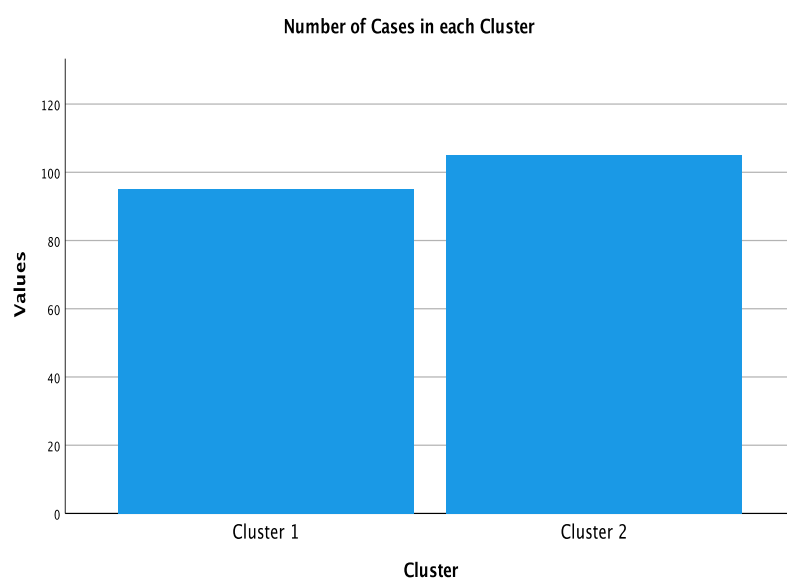


Fig. 18 (Source: Author)

b) Explain the most common data mining and text mining methods that can be used in business with real world examples.

Data mining and text mining help screen the data, narrowing the analysis and research with only pertinent data. Although for businesses, it is pivotal to collect the data, the risk of missing valuable information is high when a large quantity of data is overwhelming.

Data mining is a process of analysing a large amount of data to discover patterns and relationships within the data, turning raw data into useful information for making predictions that help businesses in making decisions and gaining competitive advantage across all industries (Witten and et al. 2016).

Text mining can be considered a sub-category of data mining. While data mining analyses structured data, text mining handles unstructured text data, filtering large amounts of text data to extract meaningful information and transform it into a structured data format to discover knowledge, insights and patterns. Examples of common data sources for text mining are emails, surveys, social media, and library databases.

There are several methods used in data mining, however, the main methods can be considered:

- *Association*: particularly used in marketing as it helps understand consumers' purchasing behaviour. E.g. Tesco has increased its sale by placing the beers next to the diapers after Tesco ClubCards's data showed a correlation purchased with these two items (BBCNews, 1998).
- *Clustering Analysis*: already mentioned above, particularly used in developing marketing strategies, as it helps to segment the consumer.
- *Regression*: is used to identify correlations and relationships between a dependent and an independent variable, in order to predict the output. Regression models, contrary to the Classification models, take into consideration the mean square error, which is not treated as something wrong, but as a deviation from the model. The predicted output values are particularly useful in business to forecast future opportunities and risk, and in forecasting budgets and sales.
- *Classification*: while with the Regression model the output values are continuous, in the Classification model the output values are divided into different categories. Classification models, in fact, divided the data into different groups to predict classes for new data. Classification analyses are particularly useful to develop marketing strategies and in banking to classify the risk of issuing loans.

- *Decision Tree*: can analyse either numerical or categorical data, creating a classification model that, based on the attributes, split the data set into subcategories. As it simplified business decisions by providing the different outcomes for the different decisions, is particularly useful in planning logistics, strategic management and in the law firms.

(Peamkar, 2022a)

b) Discuss the advantages and disadvantages of using SPSS and Excel using relevant practical examples and present your discussions in a table.

- The main difference is that, compared to SPSS, Excel is more limited in the number of statistical tools, in fact, Excel is considered a spreadsheet software, while SPSS is statistical analysis software. Therefore, Excel is more indicated to store the data and reduce the redundancy, while SPSS is more indicated for statistical calculation and manipulation (Hinton, 2004).
- One limitation of Excel is that the data to be analysed, is required in a numeric format, and controversial to SPSS, cannot be linked to the codes used in the data. E.g. when in the paragraph above, has been analysed with SPSS the ice cream data set, the ice cream's flavour and the student's gender were attached to the label 'Vanilla', 'Chocolate', 'Strawberry', 'Female', 'Male'. In Excel, the data would have been inserted as a number only, and a legend should have been attached to describe the meaning of the number.
- SPSS, contrary to Excel, creates and keeps the output in a separate file, which has the advantage of not affecting the data that is stored. Fig. 18 above, is an example of the separate output that has been automatically created in SPSS running an analysis.
- Another limitation of Excel is that for several analytical tools, the data need to be in distinct rows or columns, while in SPSS the cases can be automatically selected. E.g. the mixed data of Females and Males can be automatically processed in SPSS, while Excel requires the data to be sorted by gender.

(Pedamkar, 2022b)

EXCEL	SPSS
Spreadsheet Software	Statistical Analysis Software
Better to store data and reduce redundancy,	Better for statistical calculation and manipulation
The output affect the data stored	The output does not affect the data stored
Only numeric format for analysis	Any format for analysis
Numeric format cannot be linked to the codes	Numeric format can be linked to the codes
Few tools required distincts rows and columns	Can select automatically mixed rows and columns

(Source: Author)

References:

- Analytiks (2020). *Why Data Visualization Is Important*. Analytiks [online] Available at: <https://analytiks.co/importance-of-data-visualization/#:~:text=Data%20visualization%20gives%20us%20a,outliers%20within%20large%20data%20sets>. [Accessed 3rd May 2022]
- BBCNews (1998). *Do diapers drive dads to drink?*. BBC News [online] Available at: <http://news.bbc.co.uk/1/hi/uk/77622.stm> [Accessed 10th May 2022]
- Berkman, E.T. and Reise, S.P. (2016) *A conceptual guide to statistics using SPSS*. London: SAGE.
- Berman, E., Wang, X. (2018) 'Central Tendency', in *Essential Statistics for Public Managers and Policy Analysts*. (4th ed.) Thousand Oaks: CQ Press.
- Blokdyk, G. (2021) *Cluster Analysis: A Complete Guide - 2020 Edition – Practical Tools For Self-Assessment*. (2020 ed.) 5STARCooks.
- Cameron, S. (2021) *The business student's handbook : skills for study and employment*. (7th ed.) Harlow, England: Pearson Education, Limited.
- Hinton, P.R. (2004) *SPSS explained*. London ; Routledge.
- Luengo, J., Jill, D.G., Gallego, S.R., Garcia, S., Herrera, F. (2020). *Big Data Preprocessing: Enabling Smart Data*. (1st ed.) Cham, Switzerland: Springer.
- Muth, L.S. (2018) *What to consider when creating line charts*. Datawrapper news [online] Available at: <https://blog.datawrapper.de/line-charts/> [Accessed on 1st May 2022]
- OAG (2014) *Computational Formulas*. OAG [online] Available at: https://oag.ca.gov/sites/all/files/agweb/pdfs/cjsc/stats/computational_formulas.pdf [Accessed 1st May 2022]
- Pedamkar, P. (2022a) *Data Mining Methods*. EDUCBA. [online] Available at: <https://www.educba.com/data-mining-methods/> [Accessed 10th May 2022]
- Pedamkar, P. (2022b) *SPSS vs Excel*. EDUCBA. [online] Available at: <https://www.educba.com/spss-vs-excel/> [Accessed on 10th May 2022]
- Rosenberg, E. (2022) *The Importance of Excel in Business*. Investopedia [online] Available at: <https://www.investopedia.com/articles/personal-finance/032415/importance-excel-business.asp#:~:text=Microsoft%20Excel%20enables%20users%20to,and%20activities%20of%20their%20workforce>. [Accessed on 1st May 2022]
- Walkenbach, J. (2007) *Excel 2007 bible*. (1st ed.) Indianapolis, IN: Wiley.
- Witten, I.H., Frank, E., Hall, L.A., Pal, C. (2016) *Data mining practical machine learning tools and techniques*. (4th ed.) Amsterdam: Elsevier.
- Yi, M. (2021). *A Complete Guide to Line Charts*. Chartio [online] Available at: <https://chartio.com/learn/charts/line-chart-complete->

[guide/#:~:text=You%20will%20use%20a%20line,\(plotted%20on%20the%20horizontal\).](#)

[Accessed on 1st May 2022]