



DATA ANALYSIS & VISUALISATION

ASSESSMENT 2

Module: Data Handling & Business Intelligence (LCBB5000)

Prepared by:

Mostapha Boulkheir (2114677)

Contents

Part 1.....	2
Introduction	2
What is Excel	2
Analyze the data	3
Process the data.....	3
Data visualization	3
Superstore dataset analysis	4
Conclusion.....	8
Part 2.....	8
SPSS.....	8
Question 1 : How many students like vanilla flavor of ice-cream?.....	9
Question 2: How many students are Male and Female?.....	10
Question 3: What are the Mean and Median of participants who likes chocolate and strawberry flavor ice-cream?	10
Question 5: Explain the most common data mining and text mining methods that can be used in business with real world examples.....	12
Question 6:Discuss the advantages and disadvantages of using SPSS and Excel using relevant practical examples and present your discussions in a table.....	13
References	14

Part 1

Introduction

The availability of a variety of information is now part of the daily life of all organizations. This information does not come in a unique form and can be used immediately. There are various techniques and methodologies which, depending on the business area and needs, can transform these into a form of data that can be used by organizations. These transformations have the task of converting data from a raw form to a structured form, which can then be used for decision-making or control processes.

There are many tools for these transformations. In this paper he will focus on the Excel tool, which has unique properties that allow you to go and view anomalies and problems.

The Excel file under analysis, presents the information of a Superstore in the United States that has a dataset that contains different information presented in four years, between 2009 and 2012. As an example of data, we can have what are the types of products sold, the profit and the geographical areas where they are mostly sold, the type of clientele and the quantities in correlation to the profit.

With Excel we will highlight where the decline in profits occurs, also highlighting the period where the declines occur.

What is Excel

The Microsoft Office package has developed specific applications for each type of computer functionality: e-mail, text editing, video creation and data analysis and calculation. Particularly, such as Excel, a spreadsheet that allows you to efficiently process large amounts of data and offers powerful tools with which to communicate the results (Walkenbach, 2007).

Excel is a powerful and flexible tool that can be used for example to analyse data and visualize the data to get results in a professional way. Its main function can be:

- record data,
- check data,
- process data,
- perform database analyses,
- create summary tables and graphs,
- share work and results,
- create reports,
- create control tools.

The main feature of Excel is that it easily adapts to any type of analysis that you want to do, you can use many functions that it contains within it, giving you to quickly create a graph, tables (Walkenbach, 2007).

This means that the ability to organize, calculate and evaluate important data becomes very fast, thus giving companies some skills that can be used in business management, decision, and monitoring processes.

Analyze the data

A definition of data analysis can be as follows: "the collection and processing of large amounts of data to extract hidden information" (Azevedo & Santos, 2015).

One of the purposes of this analysis is to highlight all the relevant information for the decisions that need to be made at the level of strategies. So, data analysis is a complex tool with which decisions relevant to professional or business activities are supported.

To carry out the simplest analyses, it is sufficient to use a spreadsheet such as Excel (Walkenbach, 2007). Excel's Data Analysis Tools are a set of commands that allow you to simplify and reduce the steps necessary to carry out analyses, which can also involve a very large number of pieces of information.

Process the data

Freshly collected data, i.e., unprocessed, is called raw. But as soon as they are sorted and rearranged, they are called "processed data" (Witten, 2011).

Data processing helps companies in the ability to access the huge amount of data they hold. This is to help them acquire advantageous competitive strategies and have in-depth and fundamentally important information on market trends.

For data processing, one of the tools used is Excel, which performs operations and calculations on values and using formulas and Pivot functions. Both of these are useful to always have in mind in order to manipulate data. The pivot function, an advanced and apparently complex tool, allows you to process large amounts of data to perform relatively simple calculations. As for the formulas, Excel includes a vast number, these based on the type of processing and research you have to select the appropriate ones (Walkenbach, 2007).

Data visualization

Second editors & Engbretsen (2020), Data visualization is defined as the visual presentation of data or information. The objective of this is to communicate in a clear and effective way, to those who observe them, the data and information in their possession. Particularly, this happens presented the data in the form of graphs, charts, or maps.

With the use of Excel, we can think that after processing and analysing the data, results are obtained that now need to be communicated in some way. This is where the many fundamental tools of Excel come into play, which have the purpose of effective communication. The fundamental tools are graphs, which allow you to visually communicate your results. This allows you to create professional

graphics in just a few minutes, thanks to the predefined styles, the set of configuration tools and a series of automations that make creating and managing graphics quick and easy (Walkenbach, 2007).

Thanks to these, for example, you can go to highlight trends and outliers within a dataset or highlight interested parts of the data that could be used for greater business efficiency.

Superstore dataset analysis

Super Store is a retail business where furniture, office supplies and technology products are sold, and their customers are mass consumers, businesses and home offices.

The dataset contains 8400 rows and 21 columns, where inside we find different types of information, some are listed here:

- Sales prices and profit for each individual product.
- Geographic information (city, region and state);
- Sale date and delivery date.
- Categories and sub-categories of the various products.
- Shipping type and shipping price.
- Quantity sold and the unit price.

The aim is to analyse and identify all areas of weakness and where there are no profit margins of this Superstore during the four years between 2009 and 2019. This analysis will be carried out on the following points:

- Profit trend between 2009 and 2012, analysing both years and months with low profits.
- Analyse in detail the regions and then the provinces with low profits over the years.
- Non-profitable customer segment.
- Analyse which products or by-products lead to profit losses.

For the analysis of this report, we will use Excel, in particular we will use the Pivot and VLOOKUP.

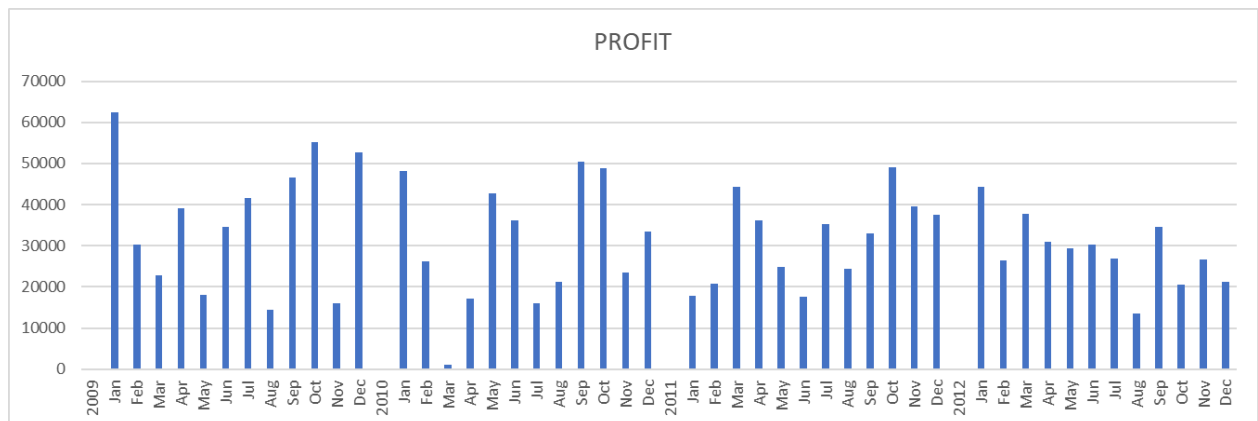
Decline in profits from 2009 to 2012

The Superstar presents a series of information on its profiles from 2009 to 2012. The first approach is that, using the Excel Pivot function, go and see the profit trend over the years.



This immediately showed a trend of decline. If it is noted that the year 2012 is the one that presents a lower profit than all the other 3 years. Year 2010 also presented a decline, but soon after year there was an immediate recovery. From an initial analysis referring to the years, it was immediately noted that the year with the most loss of profit is 2012.

Always using the Pivot function, we extrapolated the profit trend in the various months of the real years. As a first visible result is that in the last months of 2012, especially the month of September and November of 2012 have a lower profit than the same months in previous years.

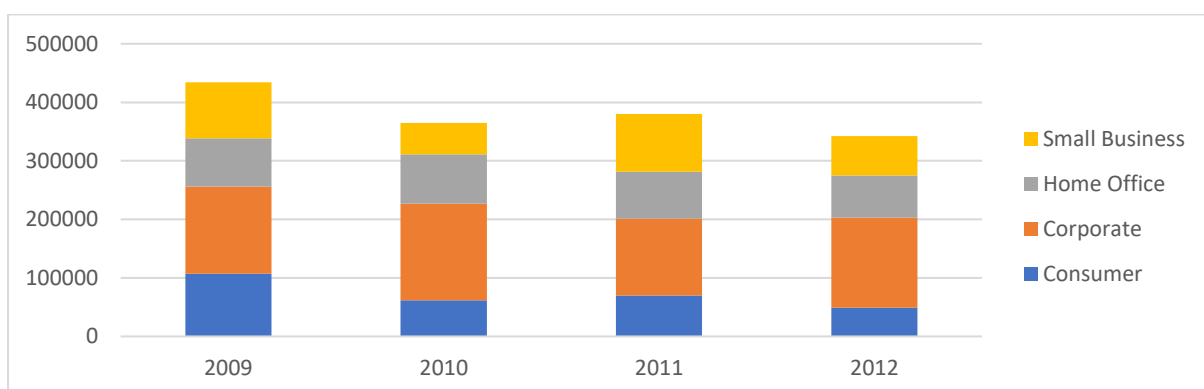
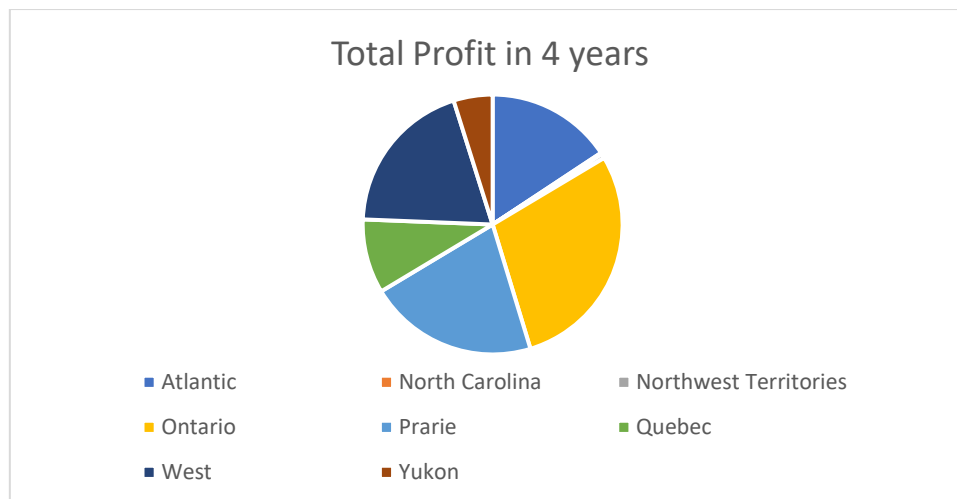


Another very interesting data, the fact of the month of March 2010 which presents a low profit figure, ie the lowest profit ever recorded in these 4 years.

Here we have used the bar graphs, extrapolated from the pivot function, which allow one or more data series to be represented by displaying them as a series of bars which in this case indicate the various levels of profits with the specific year or month. This allows you to view progress and immediately highlights where there are anomalies (Walkenbach, 2007).

Decline in profiles by region (year, month)

Analysis by region is based on total profit in the years 2009-2012, quickly showing that North Carolina has the lowest value, followed by the Northwest Territory, compared to all other regions.



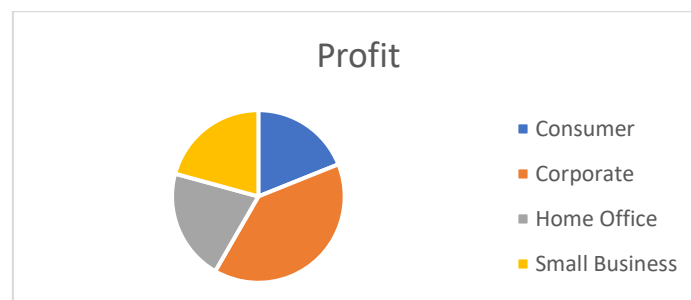
Now going to analyze the trend of the profiles in the regions by year, it is noted that always for all four years we have that the North Carolina and Northwest region always have low profits. Analyzing the trend over the years, it can be seen that 2012 has even negative values for 2012.

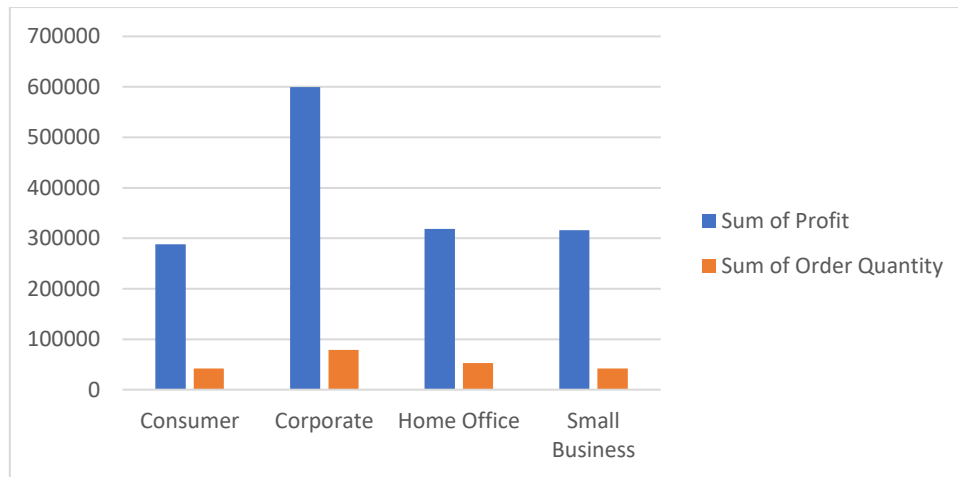
Using the Vlookup function, we went to extract the negative profit values. Highlighting that the Northwest Territories 2019 profit figure has a negative profit on record in all 4 years.

region	Years	PROFIT
North Carolina	2009	-1282.19
Northwest Territories	2012	-1444.58

Decline in profiles by category (year, month)

The categories sold by the super store are furniture, office supplies. It is immediately evident that the category with the most profit is that of technologies, while furniture is in last place.





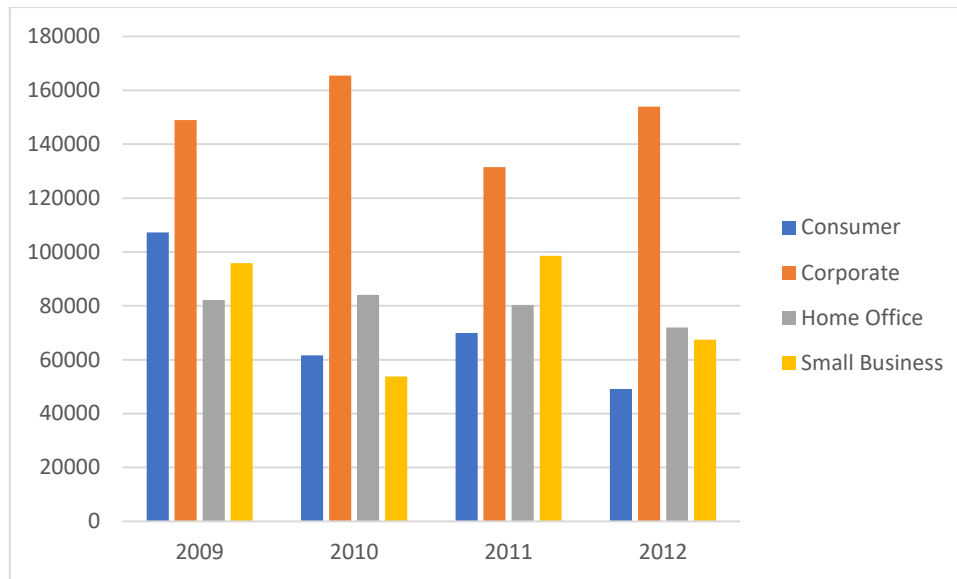
Analyzing the profit with the quantities sold, it is confirmed that the quantity of furniture is the lowest of the three over the years. Highlighting a very significant drop in profits in 2012.

Using the Vlookup function, we went to highlight that there is a negative profit in the year 2012 for furniture. This affected the total profit for the year 2012 creating a very significant loss.

Yers	Total Profit Consumer
2009	61804.53
2010	9397.4
2011	50422.45
2012	-4191.35

Decline in profiles by customer segment (year, month)

Here it is highlighted that the Consumer segment is the one with the lowest profit in the total profits of the 4 years. The quantities sold are also smaller than in all other categories.



Analyzing the profit trend over the years, it has been shown that the profit trend has also decreased in other customers.

Conclusion

Importance of data processing, analysis and visualization processes has a fundamental importance in organizations. This causes the data to pass from a raw form to a form that can be analyzed and visualized to implement decision processes and study strategies to be competitive. Excel was used for visualization.

Here we have seen the importance of Excel, it means that with a few commands it is possible to obtain results in a short time, such as for example the extraction of key information with respect to a mountain of data.

This is thanks to his vspeed of analysis, which leads to the visualization of the data in a clear and simple way, immediately highlighting any anomalies that come immediately to the eye. In this quick analysis, we performed a standard and simple data manipulation to understand and analyze a dataset.

In this case we went to analyze and visualize the dataset of the superstore. We got some in-depth results on profits, by period, geographic location, customer segmentation and various product categories. Highlighting the problems of the decline in profiles and trying to highlight the causes and factors that led to it.

Part 2

SPSS

SPSS is software that is used for statistical data analysis. It is a very complete program, relatively simple to use and very versatile and for this reason it is widely used in the academic field and in many business sectors.

SPSS stands for Statistical Package for Social Science. In fact, this software was initially developed for the social sciences. Over time its potential then extended and today it is also widely used in other sectors. In fact, SPSS is used in medicine, psychology, economics, marketing and other fields both in companies and universities (Pallant, 2010).

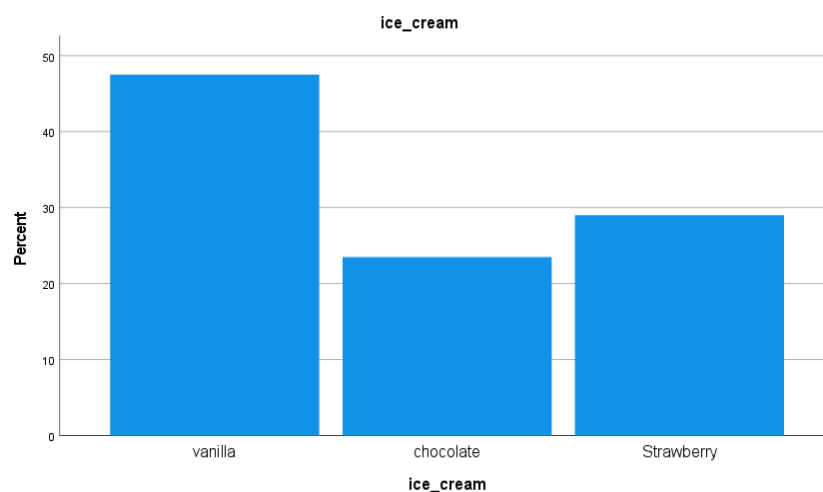
This software allows you to perform comprehensive statistical analysis to summarize, describe and understand the information contained in the data. These analyses, even multivariate, can be performed both through a graphical interface and through code for more in-depth analyses.

In this report, a collection of student data will be analyzed using SPSS and the following questions will be answered:

Question 1 : How many students like vanilla flavor of ice-cream?

The results showed that out of a group of 200 people, 95 like vanilla ice cream, with a percentage of 47.5%. It can also be inferred that it is the most popular flavor among this group of people.

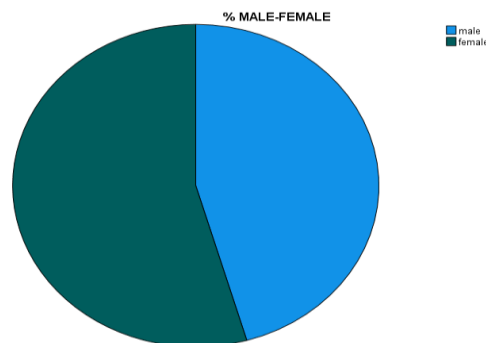
		ice_cream			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	vanilla	95	47.5	47.5	47.5
	chocolate	47	23.5	23.5	71.0
	strawberry	58	29.0	29.0	100.0
	Total	200	100.0	100.0	



Question 2: How many students are Male and Female?

Out of the total of 200 students, we obtained that 91 students are male and 109 female. In addition to highlighting that we have a percentage of 45.5% males and 54.5% females, noting that the female gender percentage is greater than the male one.

		STUDENTS			Cumulative
		Frequency	Percent	Valid Percent	Percent
Valid	bad	91	45.5	45.5	45.5
	female	109	54.5	54.5	100.0
	Total	200	100.0	100.0	



Question 3: What are the Mean and Median of participants who likes chocolate and strawberry flavor ice-cream?

Mean - Median

ice_cream	No	mean	Median
vanilla	95	107.65	113.00
chocolate	47	74.55	63.00
strawberry	58	109.81	101.50
Total	200	100.50	100.50

The mean and the median are important indicators of descriptive statistics, which are used to describe the phenomenon of a population or a sample (Pallant, 2010).

From the obtained statistics we see that:

- Participants who like chocolate ice cream have a mean of 74.55 and a median of 63.

- Participants who like strawberry ice cream have a mean of 103.81 and a median of 101.5.

Issue 4: Using the dataset Ice_cream_csv provided in conjunction with SPSS, give a specific example of clustering (ex. K-Means). Show your workings with screenshots and clearly explain your results.

Cluster analysis consists of a set of statistical techniques aimed at identifying groups of similar units with respect to a set of variables taken into consideration and according to a specific criterion. The objective that is set is substantially that of bringing together heterogeneous units into several basically homogeneous and mutually exhaustive subsets (Walkenbach, 2007).

Here the K-means cluster method was chosen, with the aim of grouping data points similar, discovering hidden patterns, dividing the dataset into a number of clusters.

Analysis with SPSS K-Means

Number of Cases in each Cluster		
clusters	1	76,000
	2	59,000
	3	62,000
Valid		200,000
missing		.000

From the first table obtained there are 3 clusters, to which respectively 76, 59 and 62 objects belong.

Iteration History			
Change in Cluster Centers			
iteration	1	2	3
1	1,634	1976	2,320
2	.233	.167	.472
3	.098	.215	.280

Now knowing that the "final cluster centers" of a group are made up of the means of each variable within the group and help us understand the characteristics of the objects belonging to each group. These reflect the characteristics for each case and for each cluster.

The three groups seem distant from each other, the greatest distance is observed between the first and the third, while the first and second seem closer.

ANOVA						
	clusters		Error			
	mean square	df	mean square	df	f	Mr
Zscore(ice_cream)	75,184	2	.247	197	304,553	<.001
Zscore(ice_cream)	75,184	2	.247	197	304,553	<.001
Zscore(id)	36,971	2	.635	197	58,238	<.001
Zscore(female)	3.101	2	.979	197	3,169	.044
Zscore(video)	42,210	2	.582	197	72,574	<.001
Zscore(puzzle)	28,951	2	.716	197	40,422	<.001

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Question 5: Explain the most common data mining and text mining methods that can be used in business with real world examples.

Data mining is the analytical process of analysing large amounts of data in order to extract strategic information that is accurate and useful for decision making.

Data mining activities are divided into two types: predictive methods that use some variables to predict the future value of other variables; descriptive methods that consist in identifying understandable patterns that describe the data (Agrawal, 2022).

Some examples can be:

- Scoring system: This technique involves assigning scores based on the probability that a certain event occurs. One of the most "famous" examples of this application is called credit scoring. The score assigned in this case represents the applicant's ability to repay a debt.
- Customer segmentation: starting from the behavioural and socio-demographic information of the users, a clustering algorithm will be able to divide the subjects into homogeneous groups. At this point the marketing strategies can be defined according to the characteristics of the different groups to maximize their effectiveness such as placing the products on the shelves.

Text mining, also called text data mining, is the process of extracting insights or meaningful information from unstructured text data (Agrawal, 2022). It is a subtype of data mining involving text, one of the most common data types within databases. Like data mining, it seeks to extract useful insights from data sources by identifying and exploring patterns in the data. In text mining, however, data sources are limited to text. Filters large amounts of text data and extracts what you need.

An example of text mining is the natural language processing. Natural language processing applications are used to derive insights from unstructured text data and allow access to the extracted information to generate new understanding of that data.

Question 6: Discuss the advantages and disadvantages of using SPSS and Excel using relevant practical examples and present your discussions in a table.

SPSS is provided for formulating statistical processes and data manipulation techniques, with a few exceptions used for professional manipulation of large blocks of data. It has the nice feature of being open source and never expires, so you can keep using it for as long as you like (Pallant, 2010).

SPSS offers charting with greater analytical capabilities. After creating a standard chart in SPSS, you can highlight dates, ranges and more. SPSS can also output graphics as files.

Excel allows users to store information in tabular form and manipulate the data in endless ways. The most common is sorting and filtering data, using formulas and pivot tables to manipulate data and create new information (Walkenbach, 2007).

Excel also offers several ways to import and export data so you can integrate it into your workflow. It also has programming capabilities to automate steps or create your own custom functions (Admin, 2020).

Excel is a data entry and dataset creation tool that can be used and manipulated for more than required. And whether it's data manipulation or data management, Excel is a great tool. It also gives you the flexibility to run analysis and generate reports using an external database, saving you a lot of time. Modern versions of Excel have great graphics tools and visualization techniques.

References

- Admin, W. (2020, 03 30). *WHICH SOFTWARE IS BETTER FOR RESEARCH? SPSS or MS-EXCEL*. Retrieved 01 12, 2023, from writeneed: <https://www.writeneed.com/post/which-software-is-better-for-reasearch-spss-or-ms-excel>
- Agrawal, V. (2022, 05 20). *Data Mining vs Text Mining vs Web Mining: 3 Critical Differences*. Retrieved 01 12, 2023, from hevodata: <https://hevodata.com/learn/data-mining-vs-text-mining-vs-web-mining/>
- Azevedo, A., & Santos, M. F. (2015). *Integration of data mining in business intelligence systems*. Hershey: Business Science Reference.
- editor, H. K., & Engbretsen, M. (2020). *Data Visualization in Society*. Baltimore: Project Muse.
- Pallant, J. (2010). *SPSS survival manual : a step by step guide to data analysis using SPSS* (4th ed.). Maidenhead: McGraw-Hill.
- Walkenbach, J. (2007). *Excel 2003 bible*. Indianapolis : Wiley-Blackwell.
- Witten, I. H. (2011). *Data mining [electronic resource] : practical machine learning tools and techniques* (3rd ed. ed.). Amsterdam: Elsevier/Morgan Kaufmann.