Alina-Valeria Buzuloiu

2029963

Assessment 2

DATA ANALYSIS & VISUALISATION

LCBB5000

DATA HANDLING & BUSINESS INTELLIGENCE

Lecturer: Amitave Banik

October 2022

CONTENTS

Part 1. Superstore Data Analysis

Introduction ……...……. 3

Data analytics using Excel ……..……...…… 4

Pre-processing of the Superstore Data ….……...……. 6

Data Clean Up …….…….. 8

Data Analysis ……..….. 11

Data Visualisation ……..….. 12

PART 2: Data Analysis Using SPSS ……. 18

Introduction …….. 18

Descriptive Statistics ……. 19

Data mining and Text Mining Techniques …… 24

The Advantages and Disadvantages of Using SPSS and Excel ……. 26

References …….… 28

PART 1: Superstore Data Analysis

Introduction

Data forms one of the most valuable resources that any organization can have at their disposal. The value that the data contains depends on the informational requirement of the organization with the data. Insights about the raw dataset can be obtained using different data analytics and visualization tools that are available in the market. According to Nigam (2020) among the most common types of data visualization tools that are widely used by most organizations include those provided by Microsoft like Excel and Power BI, Google Sheets provided by Google, among others. The versatility that each of these tools have towards data presentation and analysis also depends on the features that are installed by the developers of the tools.

The focus of this paper goes to the use of Microsoft Excel in data analysis and visualization. The Microsoft Excel comes as one of the packages contained in the Microsoft Office tool pack. Basically, Excel comes as a free to use data analysis and visualization tools once the package has been installed on a computer that runs on a Windows operating system. While excel was squarely built as a visualization tool, it contains some of the most powerful and versatile data analysis tools that are used by professionals at different levels to analyse and at the same time illustrate datasets (Linoff, 2015). The analysis that the paper covers contain several analysed tables

using the powerful pivot charts in Excel and a number of visual representations of the tables created in the analysis of the sales performance and decline over the years within the Superstore dataset.

Data Analytics Using Excel

Data analysis involves the process of data leaning, transformations of the data, and analysing the dataset with the main objective of attaining usable information that can be relied upon for decision making within an organization. Applying the use of data analysis for decision making and utilizing the relevant visualization tools give insights to areas that might have been difficult to see without the tools. The visualizations and analytics therefore assist in reducing the risks of making wrong or biased decisions concerning a business or an organization. Data analytics involves not only the analysis of data but also an array of several tasks that culminate in the final processed data (Linoff, 2015). The tasks involved in data analytics include: data collection, organization of the data, its storage, and the software, tools and techniques that will be used to bring out the final output of the analysed data. The data analysis process that this paper utilizes involves the transformation of the raw Superstore dataset into statistical information with relevant explanations on the findings.

Data visualization on the other hand is mainly concerned with the depiction of data through graphical representations like charts, graphs, and tables. For instance, for a very large dataset, the best way to convey a trend within the dataset is through the use of time series line graph or bar charts (Agarwal, 2021). The best way to represent the data through a visual representation id through the use of data mapping. Mapping of the cleaned-up data informs how the various characteristics of the components within the data can change in regard to the overall data. Mapping tools like bar chart can be looked at through the magnitude of the variables that it represents. Information therefore is obtained through the mapped data visualization tool and therefore has to be presented in the correct manner.

Basically, the iterative data analysis process contains several phases that have been also heighted before. The process includes the following phases: the lists and specification of the requirements for the dataset, the process of data collection, Processing of the data collected, Cleaning of the data, The actual analysis of the data, and finally the communication of the data through the various visualization tools and reports. The figure below offers a graphical representation of the data analysis process as discussed above:

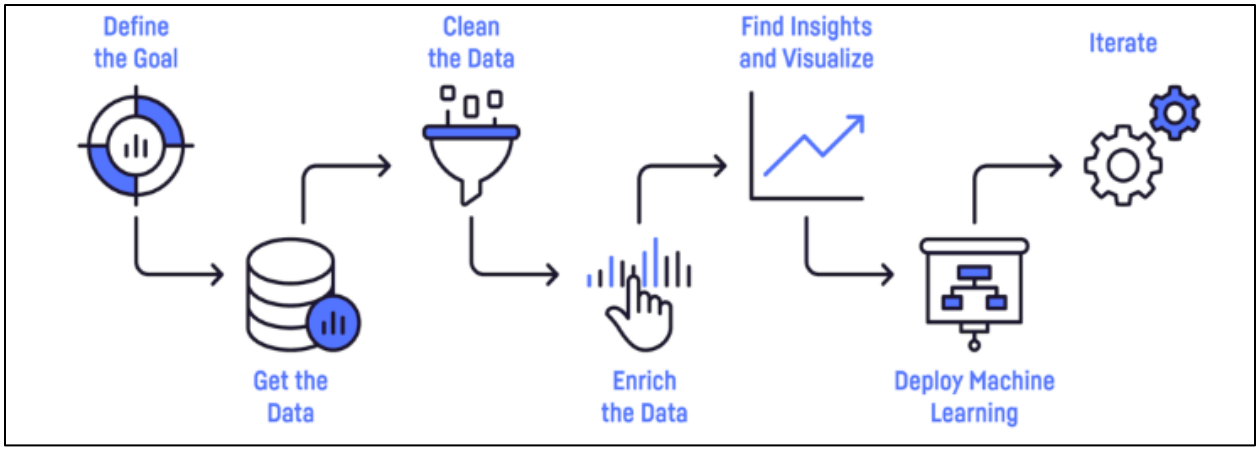


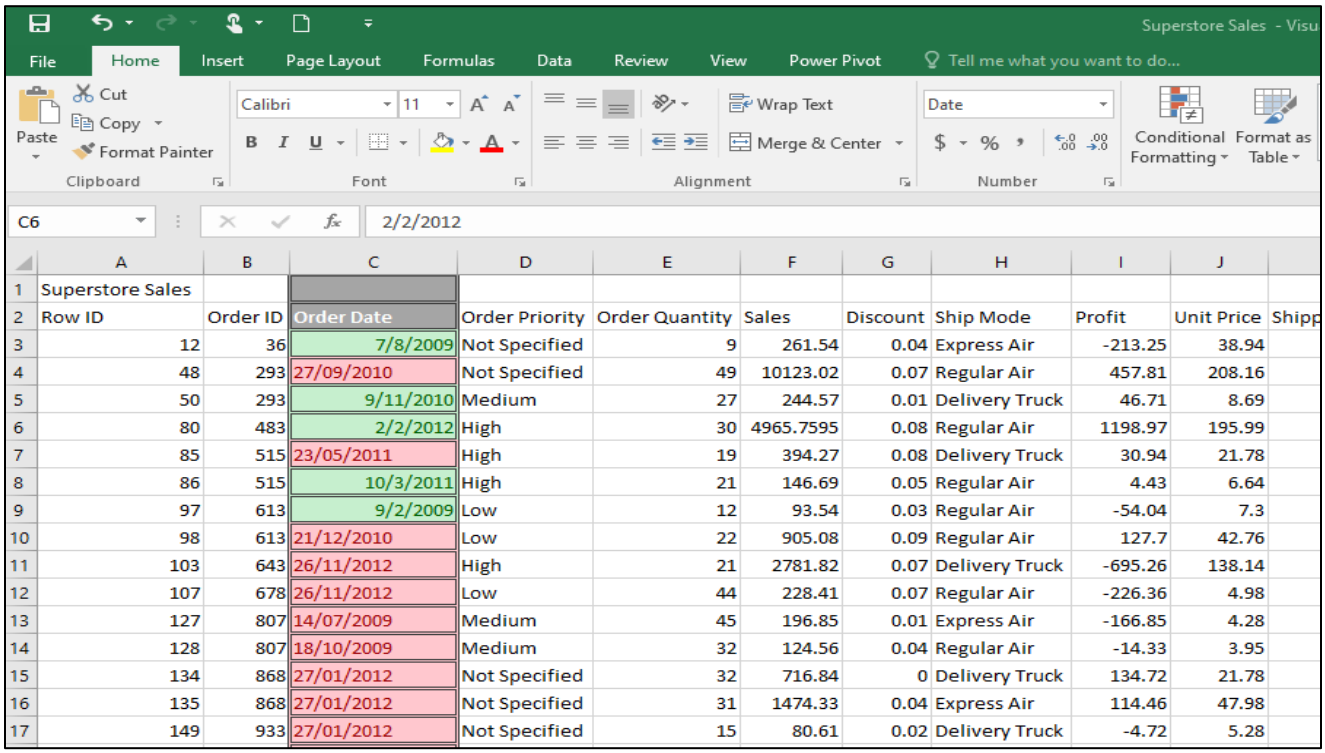
Figure 1: The data analysis process (Agarwal, 2021)

Data analysis using Microsoft Excel offers the ability to explore data in a number of ways as well as their interpretation. The information used for this purpose might come from different variables and fields within the data. Excel allows for the use of several different formats and data conversions. Tools and actions such as power pivots, Data models, Conditional formatting, Text functions, Date functions, Formula, Solver, Data Analysis toolkit, among others allow for the cleaning and formatting of the raw data obtained, its subsequent preparation and analysis (Goldwater, 2007).

Pre-processing of the Superstore Data

The first process that was involved in the processing of the data was to obtain it from the file provided by the university. The data was then searched for any missing values or fields that had mismatching values. Missing values within a

dataset has the potential to give out results output that are biased or contain the wrong information and therefore have to be identified beforehand and either changed or removed completely from the data. For the Superstore dataset, there were no missing or mismatching values that were identifiable. The dataset looked all cleaned up although the date values contained a mixture of Number Functions and the date Functions as shown in Figure 2below. The green highlighted date values were in the correct date format while the red highlighted values were in the general format and needed to be changed to the correct format.



	A	B	C	D	E	F	G	H	I	J
1	Superstore Sales									
2	Row ID	Order ID	Order Date	Order Priority	Order Quantity	Sales	Discount	Ship Mode	Profit	Unit Price
3	12	36	7/8/2009	Not Specified	9	261.54	0.04	Express Air	-213.25	38.94
4	48	293	27/09/2010	Not Specified	49	10123.02	0.07	Regular Air	457.81	208.16
5	50	293	9/11/2010	Medium	27	244.57	0.01	Delivery Truck	46.71	8.69
6	80	483	2/2/2012	High	30	4965.7595	0.08	Regular Air	1198.97	195.99
7	85	515	23/05/2011	High	19	394.27	0.08	Delivery Truck	30.94	21.78
8	86	515	10/3/2011	High	21	146.69	0.05	Regular Air	4.43	6.64
9	97	613	9/2/2009	Low	12	93.54	0.03	Regular Air	-54.04	7.3
10	98	613	21/12/2010	Low	22	905.08	0.09	Regular Air	127.7	42.76
11	103	643	26/11/2012	High	21	2781.82	0.07	Delivery Truck	-695.26	138.14
12	107	678	26/11/2012	Low	44	228.41	0.07	Regular Air	-226.36	4.98
13	127	807	14/07/2009	Medium	45	196.85	0.01	Express Air	-166.85	4.28
14	128	807	18/10/2009	Medium	32	124.56	0.04	Regular Air	-14.33	3.95
15	134	868	27/01/2012	Not Specified	32	716.84	0	Delivery Truck	134.72	21.78
16	135	868	27/01/2012	Not Specified	31	1474.33	0.04	Express Air	114.46	47.98
17	149	933	27/01/2012	Not Specified	15	80.61	0.02	Delivery Truck	-4.72	5.28

Figure 2: Screen shot of the Date Field with mixed functions

After identifying the issues within the dataset provided, the next course of action was to copy the original data file and save it under a new name for data clean-up and analysis. The original file was in a CSV text format and therefore the data used for the analysis had to take the format of an Excel Worksheet.

Data Clean Up

The next process before the analysis of the data after was to initiate the data clean-up process. The new saved dataset was utilized for this purpose, starting with removing the header in row A1 and keeping the column titles in row 2 of the data.

-

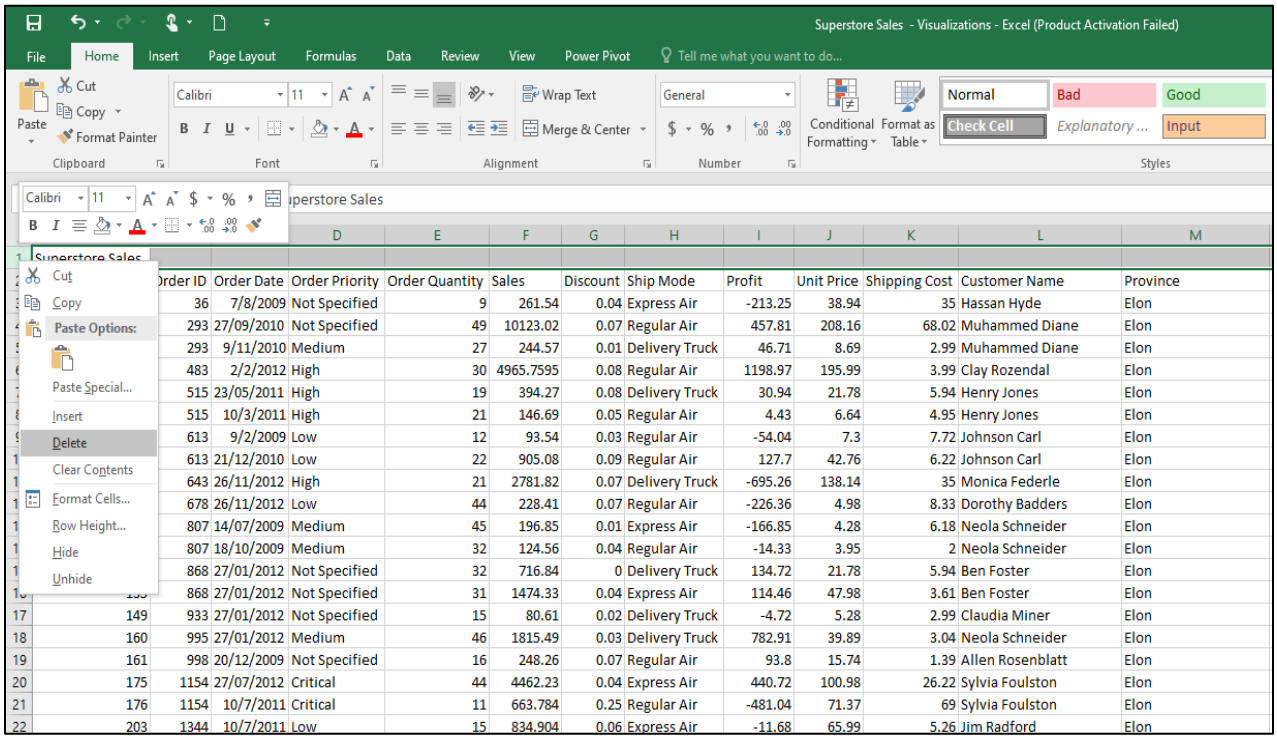


Figure 3: Data clean up – Removal of the upper header

The previously identified issues in the original dataset was the issue of dates having different formats that could not allow the data to be analysed. To correct the anomaly, a new column was inserted right after the date in column D and then the function Text-to-column from the data tab was used to reformat the date values correctly as shown in the figures below.

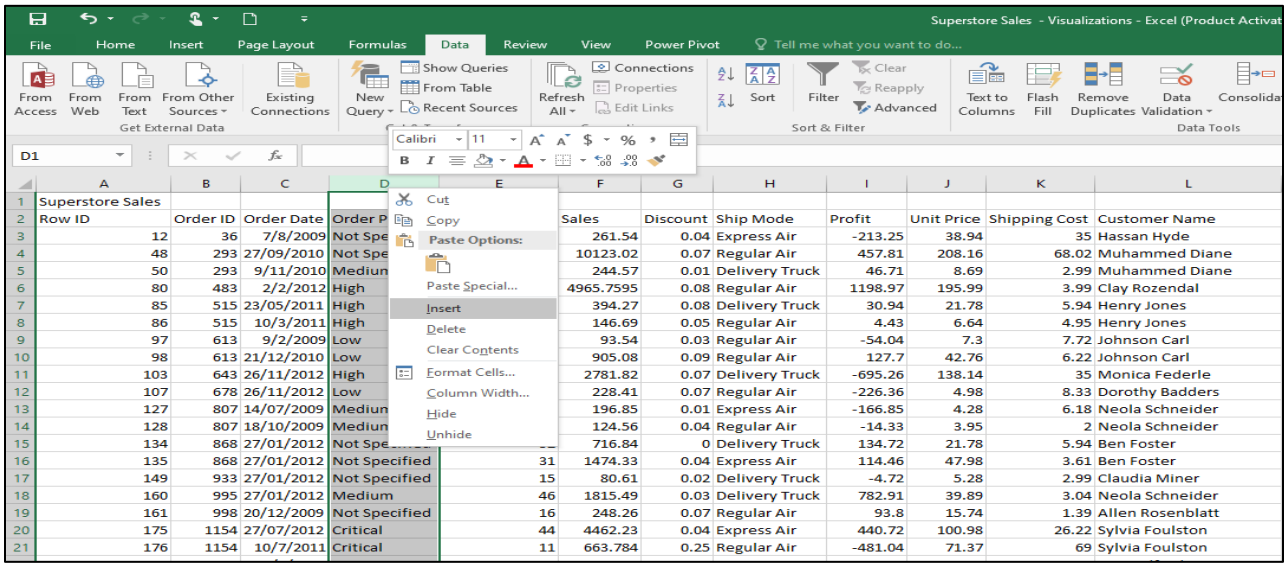


Figure 4: Data clean up – Formatting of the Date values

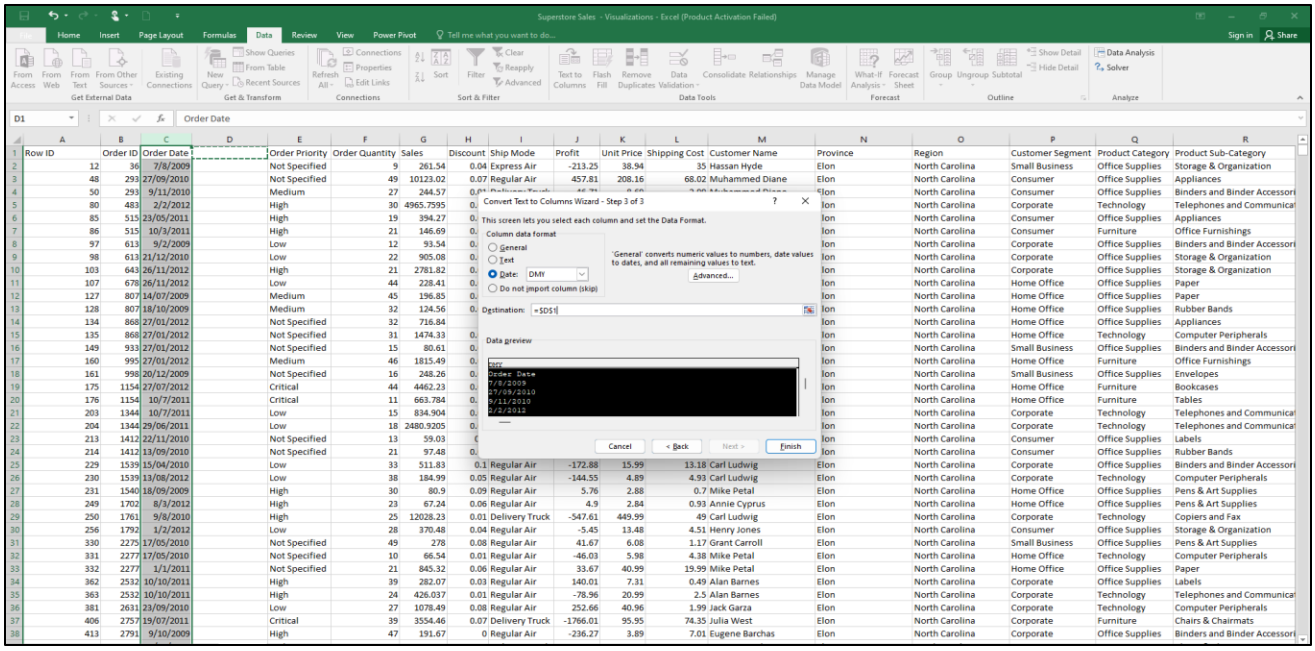
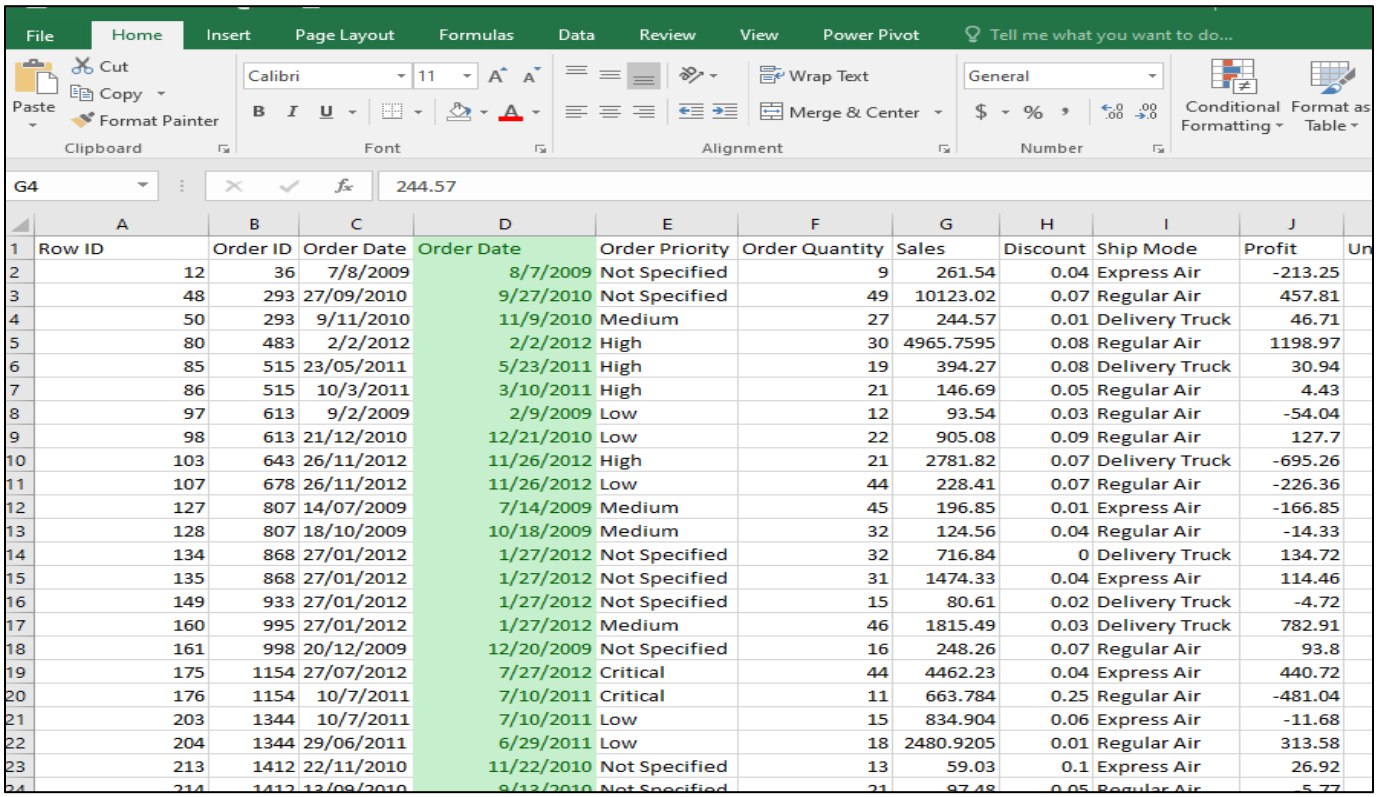


Figure 5: Data clean up – Date formatting using text to column query wizard

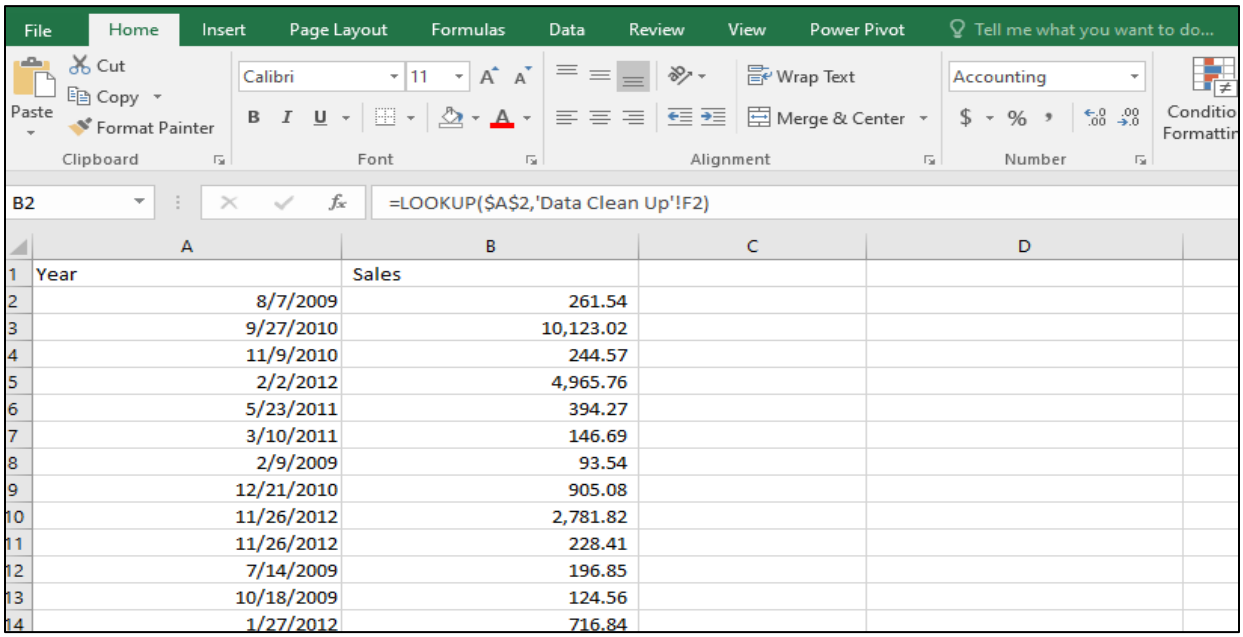


Row ID	Order ID	Order Date	Order Date	Order Priority	Order Quantity	Sales	Discount	Ship Mode	Profit
12	36	7/8/2009	8/7/2009	Not Specified	9	261.54	0.04	Express Air	-213.25
48	293	27/09/2010	9/27/2010	Not Specified	49	10123.02	0.07	Regular Air	457.81
50	293	9/11/2010	11/9/2010	Medium	27	244.57	0.01	Delivery Truck	46.71
80	483	2/2/2012	2/2/2012	High	30	4965.7595	0.08	Regular Air	1198.97
85	515	23/05/2011	5/23/2011	High	19	394.27	0.08	Delivery Truck	30.94
86	515	10/3/2011	3/10/2011	High	21	146.69	0.05	Regular Air	4.43
97	613	9/2/2009	2/9/2009	Low	12	93.54	0.03	Regular Air	-54.04
98	613	21/12/2010	12/21/2010	Low	22	905.08	0.09	Regular Air	127.7
103	643	26/11/2012	11/26/2012	High	21	2781.82	0.07	Delivery Truck	-695.26
107	678	26/11/2012	11/26/2012	Low	44	228.41	0.07	Regular Air	-226.36
127	807	14/07/2009	7/14/2009	Medium	45	196.85	0.01	Express Air	-166.85
128	807	18/10/2009	10/18/2009	Medium	32	124.56	0.04	Regular Air	-14.33
134	868	27/01/2012	1/27/2012	Not Specified	32	716.84	0	Delivery Truck	134.72
135	868	27/01/2012	1/27/2012	Not Specified	31	1474.33	0.04	Express Air	114.46
149	933	27/01/2012	1/27/2012	Not Specified	15	80.61	0.02	Delivery Truck	-4.72
160	995	27/01/2012	1/27/2012	Medium	46	1815.49	0.03	Delivery Truck	782.91
161	998	20/12/2009	12/20/2009	Not Specified	16	248.26	0.07	Regular Air	93.8
175	1154	27/07/2012	7/27/2012	Critical	44	4462.23	0.04	Express Air	440.72
176	1154	10/7/2011	7/10/2011	Critical	11	663.784	0.25	Regular Air	-481.04
203	1344	10/7/2011	7/10/2011	Low	15	834.904	0.06	Express Air	-11.68
204	1344	29/06/2011	6/29/2011	Low	18	2480.9205	0.01	Regular Air	313.58
213	1412	22/11/2010	11/22/2010	Not Specified	13	59.03	0.1	Express Air	26.92
214	1412	13/09/2010	9/13/2010	Not Specified	21	97.48	0.05	Regular Air	5.77

Figure 6: Formatted and cleaned up data (values shown in Green highlighted column)

Data Analysis

After the data had been clean up and propped up for use, the next step of the process involved data analysis. At this level, several tools that are provided within the Excel were utilized to analyse the raw data and process the information that would be used for decision making. The appropriate tools for in this case were the Power Pivot, and the formula functions. The data used were obtained from the cleaned data and saved in a new worksheet. The formula applied to retrieve the important data values for the analysis of the sales revenue decline over the years was the LOOKUP function as shown in figure 7 below. Once these values were obtained, the Power Pivot table was deployed to make appropriate tables and charts for visualization and interpretation of the data.



	A	B	C	D
1	Year	Sales		
2	8/7/2009	261.54		
3	9/27/2010	10,123.02		
4	11/9/2010	244.57		
5	2/2/2012	4,965.76		
6	5/23/2011	394.27		
7	3/10/2011	146.69		
8	2/9/2009	93.54		
9	12/21/2010	905.08		
10	11/26/2012	2,781.82		
11	11/26/2012	228.41		
12	7/14/2009	196.85		
13	10/18/2009	124.56		
14	1/27/2012	716.84		

Figure 7: Data analysis using the LOOKUP Function

Data Visualisation

The data visualisation for the Superstore dataset was displayed through the use of visual charts graphs and tables. The power pivot was a powerful tool that was used to establish the sales trends that happened over the period between 2009 and 2012. To further illustrate the information contained within the dataset, charts were also added to the pivot tables obtained.

The following visualizations regarding the sales information were derived from the dataset and presented as follows:

a) Sales Revenue by Product Category per Year

The values of sales by product category were retrieved using the pivot tables and charts and obtained as shown in the table below.

Table 1: Sales Revenue by Product Category

Year	Furniture	Office Supplies	Technology	Grand Total
2009	**1,472,671.724**	**1,035,399.64**	**1,701,825.48**	**4,209,896.846**
2010	**1,252,518.416**	**910,359.95**	**1,397,208.67**	**3,560,087.045**
2011	**1,268,656.078**	**796,383.79**	**1,364,905.11**	**3,429,944.981**
2012	**1,184,744.324**	**1,010,618.72**	**1,520,308.90**	**3,715,671.953**
Total	**5,178,590.542**	**3,752,762.1**	**5,984,248.182**	**14,915,600.82**

The above table shows the total sales revenues that were earned by Superstore over the period of four years, between 2009 and 2012. The trend shows that the sales value in the first year, 2009, were the highest in all the three categories which was at $ 4,209,896.846. The year 2010 and 2011 had the least sales recorded, at $3,560,087.045 and $ 3,429,944.981 compared to the first year and the last year, 2012 which recorded a total of $ 3,715,671.953 in sales. Of the three product categories, technology had the highest sales revenues returns followed by Furniture at $ 5,984,248.182 and $5,178,590.542 respectively. The overall sales recorded was $14,915,600.82 for the whole period.

The data was represented in a bar graph as shown below:

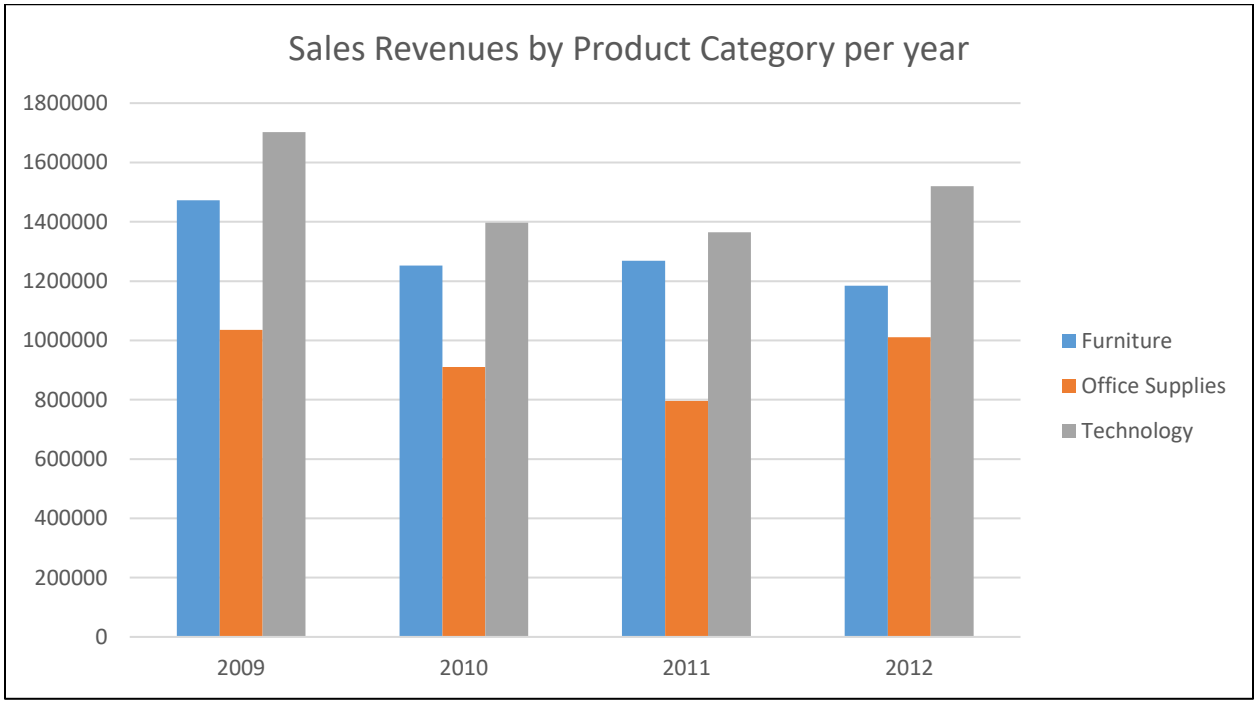


Figure 8: Sales Revenue by Product Category

This information was further broken down and analysed by the product subcategories that were under the three different categories. The sales revenues were presented per year of performance throughout the period. The results are shown in the table 2, surface Chart and bar chart below.

Sum of Sales	Column Labels				
	2009	2010	2011	2012	Grand Total
Product Category					
Furniture	**1,472,671.72**	**1,252,518.42**	**1,268,656.08**	**1,184,744.32**	**5,178,590.54**
Bookcases	253,942.12	263,974.29	140,925.43	163,810.20	822,652.04
Chairs & Chairmats	518,943.44	391,712.18	457,000.35	394,180.58	1,761,836.55
Office Furnishings	220,564.95	165,111.94	163,254.28	149,162.64	698,093.81
Tables	479,221.21	431,720.01	507,476.02	477,590.90	1,896,008.14
Office Supplies	**1,035,399.64**	**910,359.95**	**796,383.79**	**1,010,618.72**	**3,752,762.10**
Appliances	170,657.15	216,047.30	169,051.10	181,235.99	736,991.54
Binders and Binder Accessories	360,193.28	178,343.30	184,452.78	299,968.23	1,022,957.59
Envelopes	31,376.02	55,584.16	32,754.12	54,371.50	174,085.80
Labels	10,363.59	7,392.61	12,919.18	8,306.17	38,981.55
Paper	118,290.23	116,873.31	99,929.73	111,359.59	446,452.86
Pens & Art Supplies	42,421.39	46,866.65	42,603.72	35,215.46	167,107.22
Rubber Bands	3,188.07	3,629.32	4,219.94	3,969.30	15,006.63
Scissors, Rulers and Trimmers	16,653.41	6,877.25	10,644.95	46,820.70	80,996.31
Storage & Organization	282,256.50	278,746.05	239,808.27	269,371.78	1,070,182.60
Technology	**1,701,825.48**	**1,397,208.68**	**1,364,905.11**	**1,520,308.91**	**5,984,248.18**
Computer Peripherals	215,565.07	175,393.28	190,112.38	214,805.21	795,875.94
Copiers and Fax	273,286.86	339,712.51	280,821.12	236,540.81	1,130,361.30
Office Machines	732,505.45	446,780.14	426,103.41	563,308.14	2,168,697.14
Telephones and Communication	480,468.10	435,322.75	467,868.20	505,654.75	1,889,313.80
Grand Total	**4,209,896.85**	**3,560,087.04**	**3,429,944.98**	**3,715,671.95**	**14,915,600.82**

Table 2: Sales revenue by Subcategory

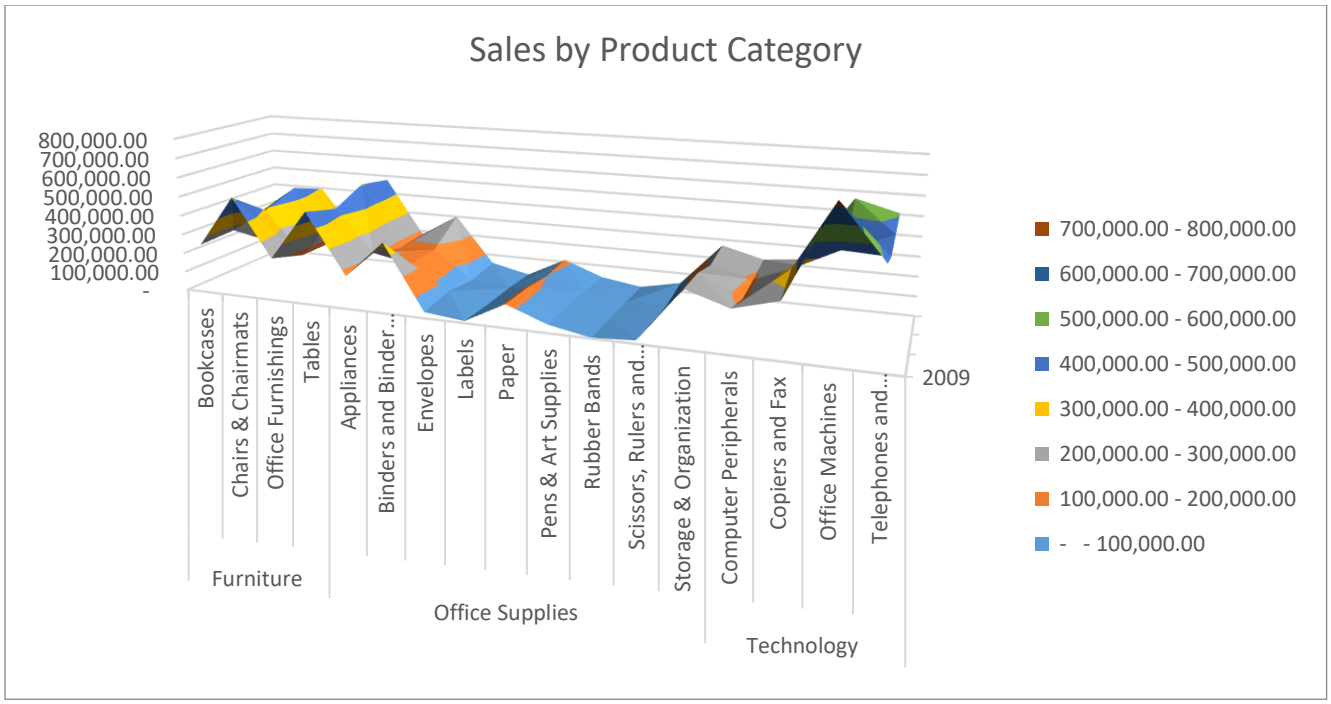


Figure 9: Surface chart for sales by product subcategory

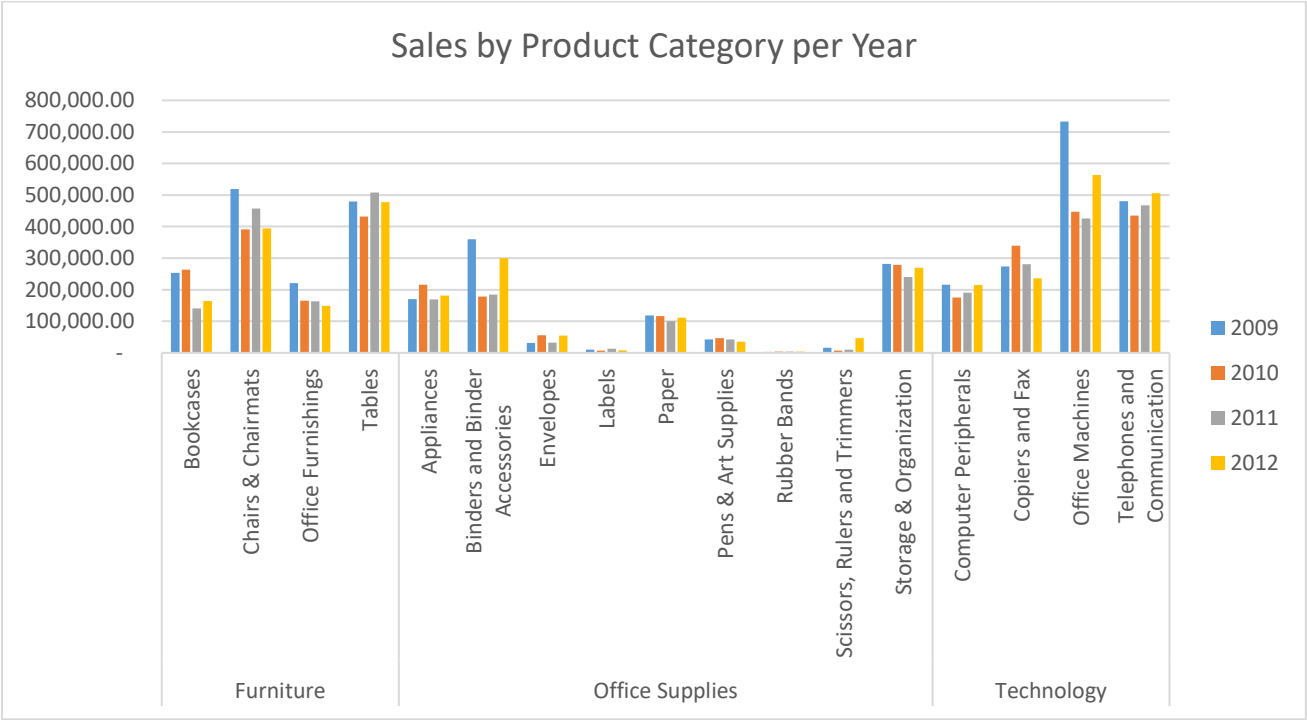


Figure 10: Bar chart for Sales by Product subcategory per year.

The information reveals that the best performing subcategory was the office machines that had the highest sales records in 2009 at a value of $732,505.32, while the subcategory with the least performance record was Rubber bands that had the lowest sales record of $3,188.07 in the year 2009. In the overall performance, 2009 saw the best sales recorded for all the subcategories, followed by the year 2012. The worst year 2011 as it was also depicted for the overall performance of the product categories.

b) Sales revenue by Customer segment

Employing the use of Pivot tables, the values of the sales revenues by the customer segment was retrieved and the table below was obtained.

Sum of Sales	Column Labels				
	2009	2010	2011	2012	Grand Total
Customer Segment					
Consumer	962821.7565	697535.7225	715860.2585	687393.3445	3063611.082
Corporate	1518609.628	1242809.044	1243534.101	1493952.106	5498904.878
Home Office	980717.1355	1011872.409	786500.3265	785674.004	3564763.875
Small Business	747748.326	607869.8695	684050.295	748652.499	2788320.99
Grand Total	**4209896.846**	**3560087.045**	**3429944.981**	**3715671.953**	**14915600.82**

Table 3: Sales by the Customers Segment.

The above information was presented using a bar chart graph as shown in the following chart;

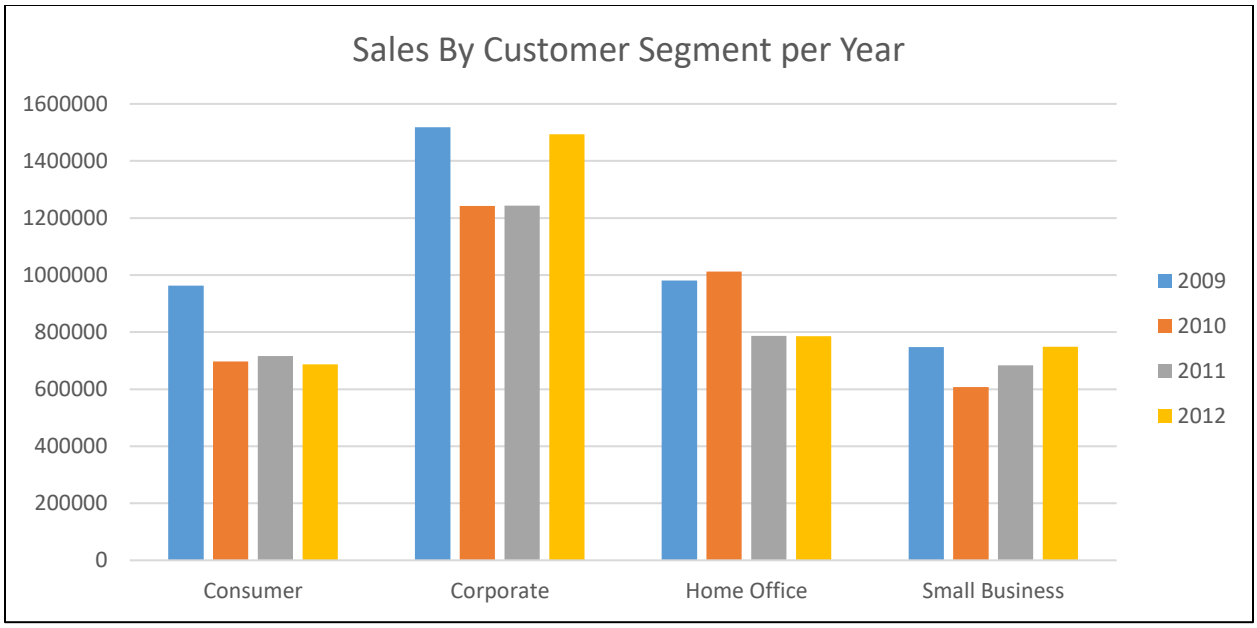


Figure 11: Bar chart for sales by customer segment per year

The information contained in the two visualizations reveal that the corporate customer segment had the highest sales values over the period, with 2009 and 2012 recording the highest sales revenues. The other customer segments had almost equal distributions over the periods. The consumer segment recorded a total of $3,063,611.08, the corporate segment had the highest sales revenues earned at $5,498,904.88, while the home office segment had a total of $3,564,763.86. Finally, the small business segment recorded total sales revenues of $2,788,320.99, which was the lowest amount of sales revenues among all the customer segments.

PART 2: Data Analysis Using SPSS

Introduction

The following part analyses the Ice cream data provided for the assignment and utilizes several descriptive statistical analysis tools that are found with SPSS. The data analysis process follows the general rules of acquiring the data, performing data clean-up, analysing the data using the data analysis tool provided, and then summarizing the output in visualizations and report format. The data sets provided were in excel format and CSV text file, both of which can be read by the IBM SPSS statistical analysis tool. The data analysis therefore was performed in two parts, the first part being the descriptive statistical analysis that sought to find the mean, median and proportions of the values within the given variables, and the second part being the inferential data analysis that sought to find the clustering within the dataset provided in CSV file format. The section further analyses data mining and text mining methods and their applications in the real world. A discussion on the differences between the IBM SPSS statistical tool and Microsoft Excel follows with the conclusion about the two data analysis tools.

Descriptive Statistics

a) Ice Cream Flavour choice

The first descriptive statistics for the analysis was to find the number of students who liked vanilla ice creams. From the below output, there were 95 students who preferred to have the vanilla ice cream flavour, that also represented 47.5 percent of the total students observed.

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	

Table 4: SPSS output on the Ice cream descriptive statistics

The data was presented on a graph as shown below:

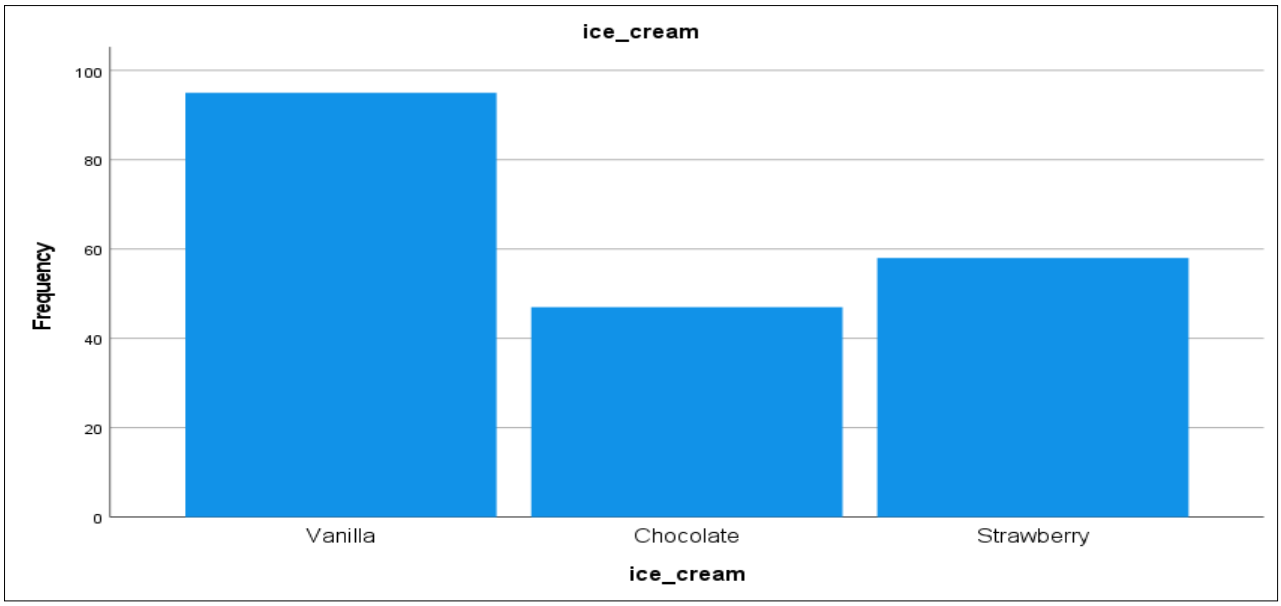


Figure 12: Bar graph showing the frequency of the Ice cream flavour preferences

b) Gender

The second descriptive statistics was on the gender, the number of male and female students who took part in the Ice cream survey. The following data was retrieved from the SPSS output.

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	

Table 5: SPSS output of the Gender Frequency Table

The number of Males in the Ice cream survey was 91 which represented 45.5 percent of the total respondents. The number of female respondents were 109, representing 54.5 percent of the total respondents in the survey. The insights that can be derived from the above data indicate that Ice cream is mostly preferred by female students compared to their male counterparts. The total number of respondents were 200.

The above data was also presented in a bar chart and the result was as shown below:

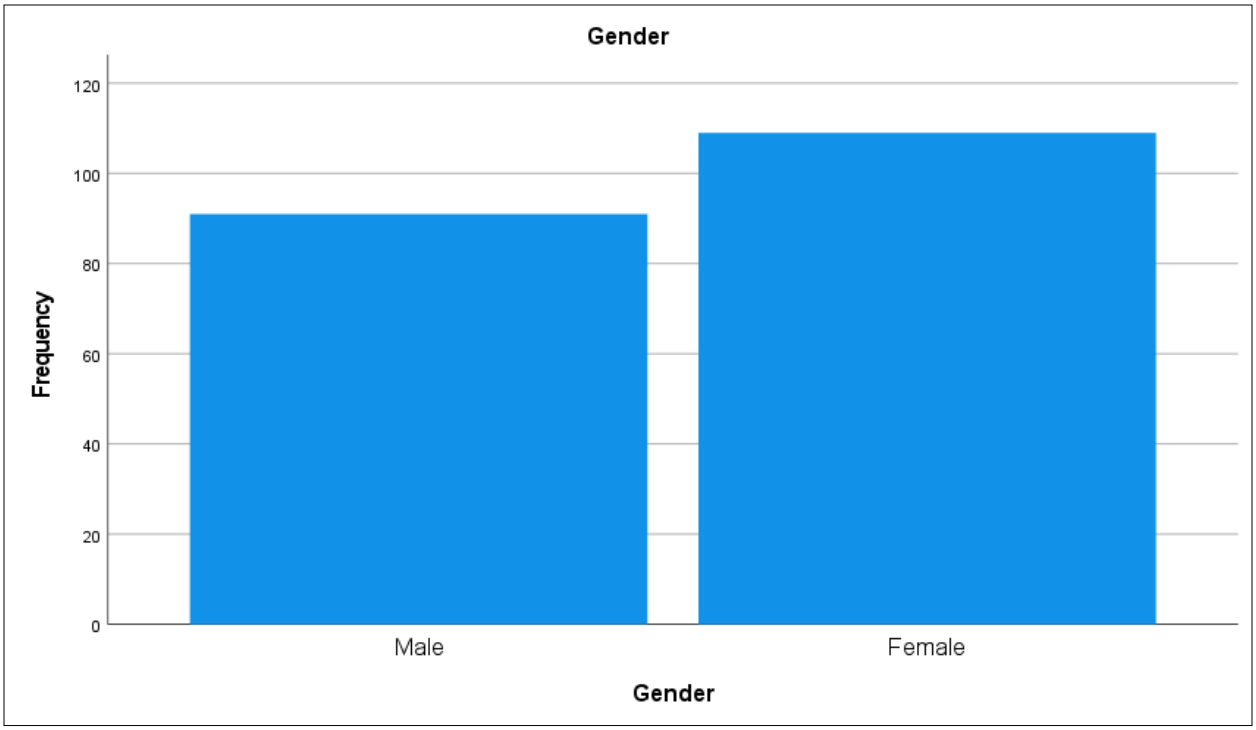


Figure 13: frequency bar graph of the gender participation in the survey

c) Mean and median of the chocolate and Strawberry flavour's preferences

The mean and median of the students who preferred chocolate and strawberry flavours was as follows:

Statistics		
ice-cream		
N	Valid	200
	Missing	0
Mean		1.82
Median		2.00

The total means of the ice cream variable was 1.82 and the median recorded was 2. There was no other information regarding the data variable was availed since the data was nominal in nature.

d) Clustering

To perform the clustering, a model that took the two common variables was establish and the Nearest Neighbour clustering was performed. The results of the output were presented and shown in Figure 14 below. The model took the two and data variables and divided them into two, with a training set and hold out set. The training set had 132 values and the hold out set had the remaining 68 values. The model has 1 variable selected with the predictor targets of 3, relating to the ice cream flavour: Vanilla, Chocolate and Strawberry. There was a closer link between the training model and the Holdout model as indicated on the hart. The spaces between the predictor models were closer on both the gender (denoted by the female variable set).

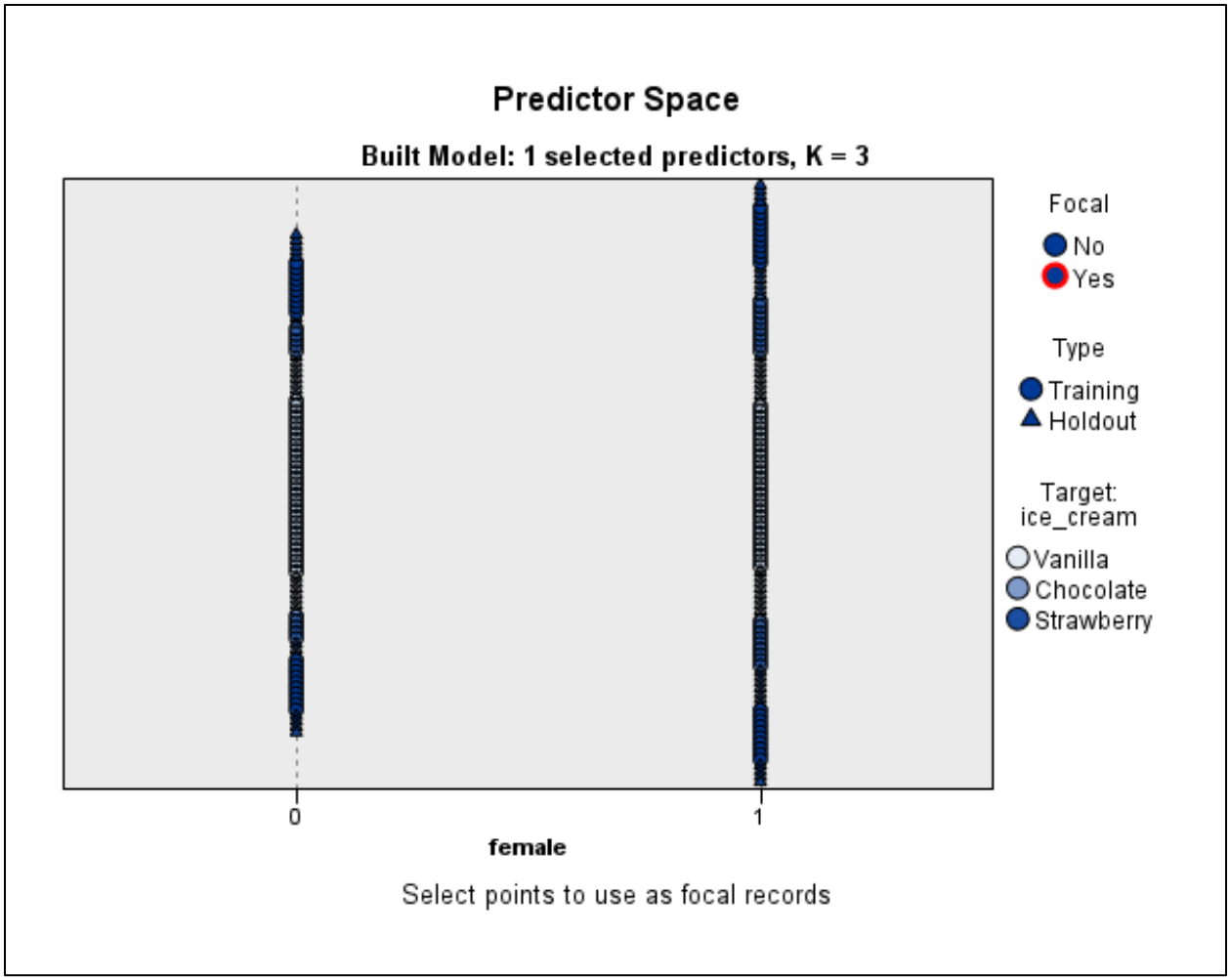


Figure 14: Nearest Neighbour Analysis

The other clustering technique that was performed was the K – mean cluster analysis. The distances between the two variables were 1.544, with the ANOVA output indicating mean square values of 120.193 for Ice cream and 0.286 for the gender variables. The Mean Error for the two sets were 0.131 and 0.249 for Ice-cream and Gender (female) variables respectively.

Distances between Final Cluster Centers

Cluster	1	2
1		1.554
2	1.554	

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
ice_cream	120.193	1	.131	198	916.660	<.001
female	.286	1	.249	198	1.147	.285

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.

Figure 15: ANOVA output for the K-mean cluster

Data mining and Text Mining Techniques

The most commonly used data mining and text mining methods across the major industries and professions across the world include: clustering, Prediction, classification and Association methods. These techniques each have their own objectives and are used to solve different business problems, applying different data mining tools (Tyagi, 2021). Below is a brief discussion on each of the highlighted data mining methods and their application to solving the business objectives/ problems in the real world.

a) Prediction: Prediction refers to the method of forecasting certain outcomes within a dataset provided, with a number of pre-set conditions expected on the outcome. The datamining method is mainly used in the prediction of prices and employed

to calculate future prices of commodities based on the past price performance and prevailing circumstances at the time. The tools that are used to perform this type of data mining technique is the oracle data mining software.

b) Clustering: this involves the identification of certain variables within a dataset and then grouping variables with similar characteristics into the same clusters. The method is mainly used in the classification of dataset variables and retrieval of data while assigning them into the clusters that they fit into to ease the process of data analysis.

c) Association: involves the establishment of the relationship that can be observed between two or more sets of items in a dataset. The aim of this method of data mining technique is to identify certain patterns hidden within a dataset containing large volume of data for a business. The most probable data analytics tools to use in association is the use of R programming tool.

d) Classification: the method involves identifying, selecting and segmenting data into several classes that fit their characteristics with the main aim of retrieving the data and then assigning them into their classes. The classification of the data into groups as such helps in the easy manipulation of data and the subsequent calculations for optimal results.

The Advantages and Disadvantages of Using SPSS and Excel

The similarities that exist between the two data analysis tools, SPSS and Excel, are numerous. However, the two tools differ greatly when it comes to the in-depth statistical analysis of data. The following are the advantages and disadvantages of using each of the data analysis tools.

Advantages

SPSS	Microsoft Excel
1. Data management: SPSS offers a number of tools for the control of data. It also offers much quicker and accurate analysis of data that is fed through it	Excel also offers an easier data management on its platform and anyone can use it without the need for deeper training. It is easier even for learners to manage through excel.
2. Ease of data manipulation: Performing tasks like creating chart and graphs is much easier in SPSS.	Excel offers a great platform for data visualization and creating charts and graphs for such is also easier and much quicker
3. Statistical functions in SPSS are much easier to perform and are	Excel offers data analysis tool that contain the statistical functions, together with an array of formulas that

offered in a wide variety compared to the other data analysis tools	can be applied to find statistical functions

Disadvantages

In SPSS, to manipulate the data one has to have an extensive knowledge of statistics and also invest time in learning the tool	Excel too have some functions that one has to invest time and vast knowledge to operate, especially the statistical functions built within it
SPSS does not come cheaply and requires some financial resources to have the software installed.	Excel comes as a package of the Microsoft Office and that too requires some financial resources to have in it in place
	Excel doesn't have all the necessary statistical packages/ tools required to run functions for deep data analysis functions.

References

Agarwal, G. 2021. *A comprehensive guide to Microsoft Excel for data analysis*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/11/a-comprehensive-guide-on-microsoft-excel-for-data-analysis/, accessed on 02/01/2023

Goldwater, E. 2007. *Using Excel for data analysis*. OIT web hosting. https://people.umass.edu/evagold/excel.html, accessed on 04/01/2023

Linoff, G. S. 2015. *Data analysis using SQL and Excel*. John Wiley & Sons.

Nigam, M. 2020. *Advanced analytics with Excel 2019: Perform data analysis using Excel's most popular features*. BPB Publications.

Tyagi, N. 2021. *7 text mining techniques*. Analytics Steps - A leading source of Technical & Financial content. https://www.analyticssteps.com/blogs/top-7-text-mining-techniques, accessed on 07/01/2023