

Student Name: Hena Rodriques

Student ID: 1921350

Module Code: LCBB5000

Programme: BA Business Management

Module Title: Data Handling & Business Intelligence

Assignment Title: Assignment-2

Module Lecturer(s) Name: Amitave Banik-GP2

Table of Contents

Index of Figures	3
Part 1: Data Analysis in Excel	4
Use of PIVOT Table.....	4
Use of	4
Dataset Description.....	4
Dataset Pre-Processing in Excel	4
Data Visualization	4
1. Customer Segmentation	4
2. Product Category	6
3. Sales and Profit Vs Year.....	7
Scatter Plot Analysis.....	8
Descriptive Analysis of Numerical Variables.....	9
Part 2: Ice-Cream Data Analysis	11
Data Description and Visualization	12
2.1 Descriptive Summary of Ice_Cream.....	12
2.2 Gender Analysis	14
2.3 Mean and Median.....	14
2.4 K-Means Clustering	14
2.5 Data Mining and Text Mining Methods	15
2.6 Advantages and Disadvantages of Using SPSS and Excel.....	15
SPSS as an Analytical Tool	15
Excel as an Analytical Tool	16

Index of Figures

Figure 1: Profit from Different Customer Segment	6
Figure 2: Pie Chart for the Average of Profit.....	7
Figure 3: Total Sales Vs Year	8
Figure 4: Line Chart for Average Profit Value	9
Figure 5: Scatter Plot between Profit and Sales	9
Figure 6: Ice_Cream Types.....	13
Figure 7: Count of Gender in the Given Dataset.....	14
Figure 8: Descriptive Table for Attributes.....	15
Figure 9: Initial Cluster Center and Iteration History.....	15
Figure 10: Final Cluster Center, ANOVA and Cases in Each Cluster	16

Part 1: Data Analysis in Excel

The data analysis toolpak is used for performing quantitative analysis in Excel. Excel is also used for creating different types of visualization. The visualization in Excel is done by selecting the appropriate charts from the list of the plots available in the Excel. For smaller dataset, there is no need of using the advanced excel function such as PIVOT, IF and other functions such as LOOKUP. However, for the larger dataset, it is required to use either PIVOT or existing sorting function in excel to summarize the big dataset. There are more than 8000 observations are present in the provided dataset for the superstore (Khadka, 2019). The total numbers of attributes in the given dataset are 21. The dataset contains all sorts of variable such as categorical, continuous and textual. The categorical variables in the given dataset are “Product Category”, “Customer Type” and “Sub-Category of Product”. The continuous variable in the given dataset are “Discount”, “Profit”, “Sales” and “Order Quantity”.

In this work, the major objective is to assess the decline in sales or profit of the superstore. The analysis is performed by using the PIVOT and the Visualization method.

Use of PIVOT Table, and IF in Excel

PIVOT in excel is used for a excel dataset to make the filtering operation simpler. The PIVOT TABLE is an interactive and simple method for summarizing the large tabular dataset. For using the PIVOT TABLE, the reference dataset is selected and a new excel sheet is generated for same. Advantages of using the PivotTable are as follows

1. User interactive charts and table can be generated by using the PivotTable based on appropriate selection of attributes.
2. Different aggregating methods can be employed for a quantitative attribute in the dataset.

For establishing the logical comparison between the expected and the actual value in a data column, the IF function of excel is used. There are also numerous sorts of the IF function in excel that can be used based on the application type.

Dataset Description

The dataset used for the analysis work is the superstore dataset. The numbers of attributes in the dataset are 21 and total counts of observations in the dataset are 8399. The dataset includes a variety of variables, including continuous, textual, and categorical. "Product Category,"

"Customer Type," and "Sub-Category of Product" are the categorical variables in the dataset provided. Discount, Profit, Sales, and Order Quantity are the continuous variables in the dataset.

Dataset Pre-Processing in Excel

The Excel can be used for the data pre-processing with use of some excel function. Excel can be used for pre-processing the dataset and is frequently used by the individual who are not from the programming background (Mark, 2018). The functions in excel assists in transforming a raw dataset into structured data. Some Common functions that are used for the data pre-processing in excel are listed below

1. IF
2. COUNTIF
3. COUNTA
4. IFERROR
5. FIND
6. ISNUMBER

Data Visualization

Data visualization is used for summarizing the large tabular dataset in interactive plots. The interpretation of insights is simpler with use of the visualization. Different type of visualization is created in this work for the superstore dataset and for creating the visualization, first the PIVOT table has been created and the PIVOT charts have been generated.

The visualization and their interpretation for the super-store dataset are as follows.

1. Customer Segmentation

In the very first step, the analysis is done on what type of customers are there in the supermarket and about what profit the owner gets from every type of customer, and also the amount of discount to be given to the corresponding customer type from the analysis that was on the given dataset the results are:

Row Labels	Average of Profit	Sum of Discount
Consumer	174.6270103	82.29
Corporate	194.9759428	153.31

Home Office	156.6702904	100.47
Small Business	192.270408	81.12
Grand Total	181.1844243	417.19

From the above table, there are many insights are found. The highest amount of the average profit that comes from the corporate customer is about 194.975 and also the highest discount is also given to the corporate customer which is about 153.31. And there is not so much difference between the profit of the corporate and small businesses it's about 192.27 but there is seen a huge difference in the sum of discounts for small businesses it's about 81.12. There is seen as a difference between corporate, small business, and consumer, home offices whereas the average profit in consumers is close to 174.62 and the lowest profit came from home office its about 156.670 while the sum of discount for a home office is second highest which is about 100.47. The sum of the discount for consumers is about 82.29 while the lowest sum of the discount is for small businesses. From the graph the visualization is clearer:

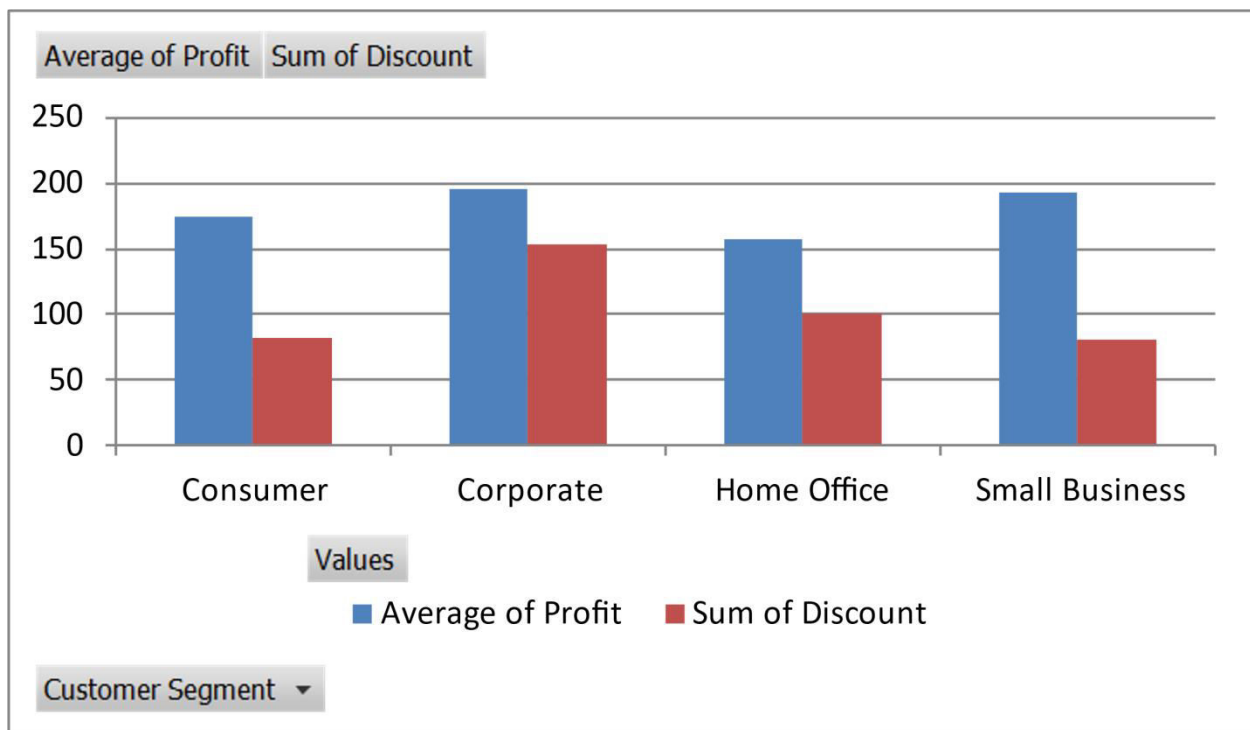


Figure 1: Profit from Different Customer Segment

The comparative bar plot shown above illustrates the average profit made by the superstore and the discount provided to each customer segment. The super store generates highest revenue from

the “Corporate” customer segment. The lowest profit is generated by the customer segment labeled as “Home Office”.

2. Product Category

There are three product categories in the given dataset. The three categories of the product are “Furniture”, “Office Supplies”, and “Technology”. The average sales and the average profit value generate by the superstore for each type of product category is shown in the below table (Sall, and Lehman, Berk & Carey, 2019). The table is created by using the PIVOT option in Excel. The pie distribution plot is also shown.

Row Labels	Average of Profit	Average of Sales
Furniture	68.11660673	3003.82282
Office Supplies	112.3690738	814.0481779
Technology	429.2075157	2897.941008
(blank)		
Grand Total	181.1844243	1775.878179

The above table demonstrates that the average highest profit is generated by selling the “Technology Products”. However, the average sales of technological products are not highest. The average highest sales are found for the product category “Furniture”. The superstore is selling the highest number of furniture products. However, average profit value generated by the firm is lowest for the “Furniture” products.

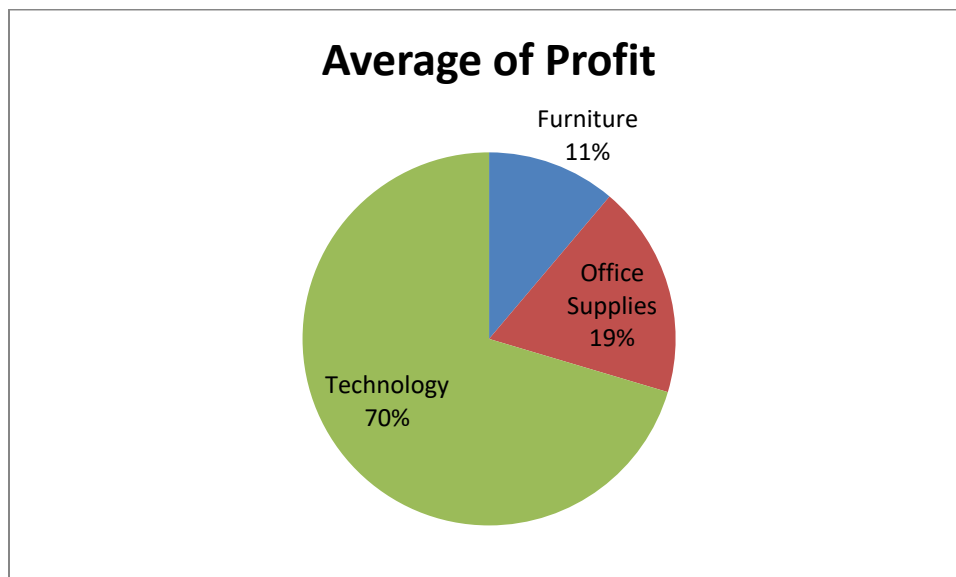


Figure 2: Pie Chart for the Average of Profit

Approximately, 70% profit generated by the superstore is from the selling of the “Technology” products. The percentage of profit share for the “Furniture” and “Office Supplies” are 11% and 19% respectively.

3. Sales and Profit Vs Year

Using the PIVOT table option in excel, the sales and the profit versus year has been determined and shown in the picture below.

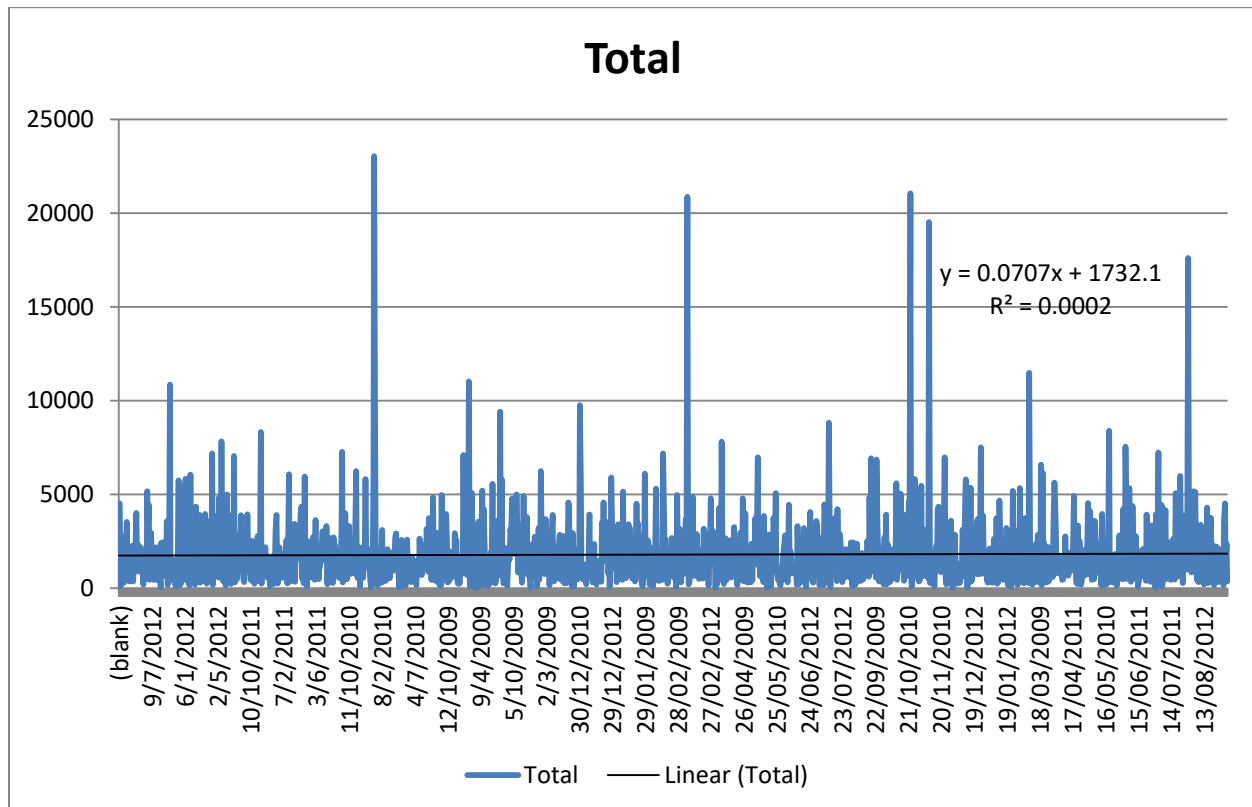


Figure 3: Total Sales Vs Year

There is large variation in the total sales value of the column. The trend line shown in the above picture demonstrates a positive trend in total sales of the superstore.

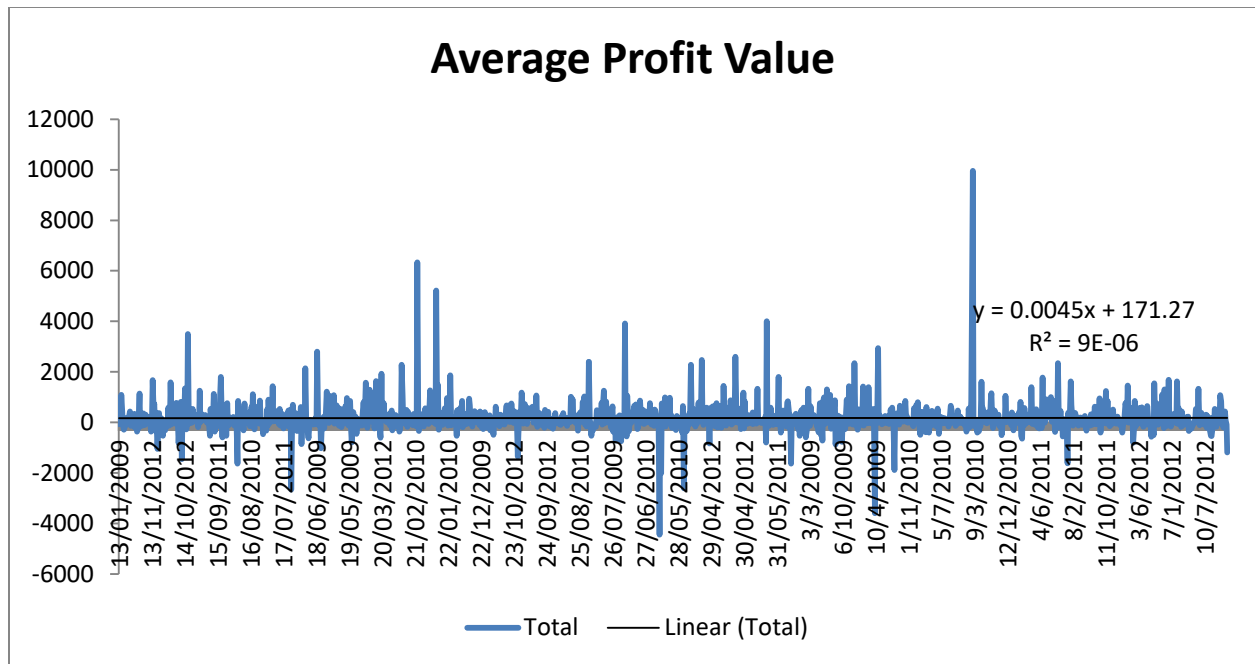


Figure 4: Line Chart for Average Profit Value

The column's total profit value varies significantly. The trend line seen in the image above shows a rise in the superstore's overall profit.

Scatter Plot Analysis

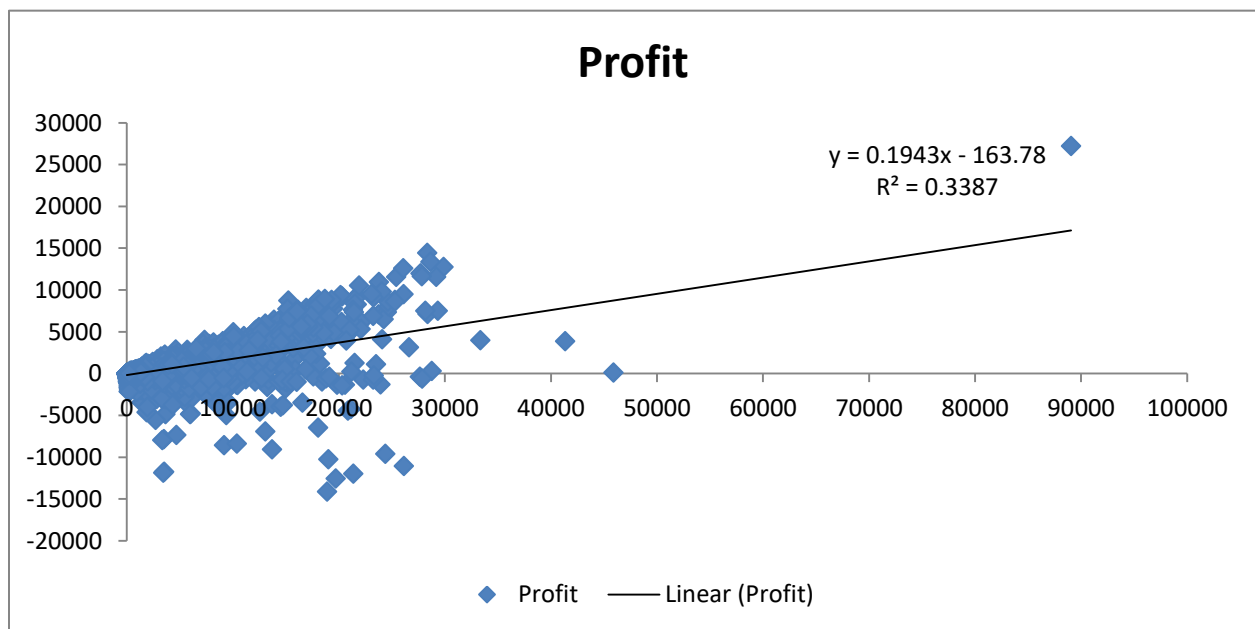


Figure 5: Scatter Plot between Profit and Sales

Scatter plot visualization are used for assessing the association between two or more variables and also the type of association (Positive or Negative) is examined by the trend line in the scatter plot. The above shown scatter plot is captured between the two attributes “Profit” and “Sales”. A positive trend line is obtained between the profit and the sales attributes. The trend line equation and the R-Squared value for the above scatter plot are shown in the above plot. The trend line equation is expressed as

$$y = 0.1943 * X - 163.78$$

The coefficient of the trend line equation is positive. The R-square value shown in the above scatter plot is 0.3387. This is a poor value of the R-Square. The value demonstrates the how much the variation of the response variable is explained by the independent variable.

Descriptive Analysis of Numerical Variables

The numeric variables in the given super store dataset are “Discount”, “Profit”, “Sales” and “Order Quantity”. The statistical description of the numeric variable in the proposed super store dataset are shown in the table below

Statistical Parameters	Qrder Quantity	Sales	Discount	Profit
Mean	25.57209192	1775.878179	0.049671389	181.1844243
Standard Error	0.15800598	39.11843664	0.000347238	13.05733595
Median	26	449.42	0.05	−1.5
Mode	31	74.02	0.01	−969.05
Standard Deviation	14.4806253	3585.050525	0.03182302	1196.653371
Sample Variance	209.688509	12852587.27	0.001012705	1431979.291
Kurtosis	−1.207991009	60.92837614	−0.959411063	67.34970524
Skewness	−0.01727784	5.386982374	0.073916963	3.647238938
Range	49	89058.81	0.25	41361.39
Minimum	1	2.24	0	−14140.7
Maximum	50	89061.05	0.25	27220.69
Sum	214780	14915600.82	417.19	1521767.98
Count	8399	8399	8399	8399

The above table includes the statistical parameter values such as “Mean”, “Median”, “Skewness”, “Kurtosis”, “Range”, “Minimum” and “Maximum”. The standard deviation of each quantitative variable is also shown in the above table. The average value of the “Ordered Quantity” is 25.57 and the median value is 26. The average value of the sales is 1775.878 and the

value of the average profit is 181.184. There is large difference exists between the minimum and maximum values of the sales data column. The minimum value of sales column is 2.24 and the maximum value of sales is 89061.05 (Stančin, and Jović, 2019). Hence, the range of the “Sales” is 89058.81. In the “Profit” column there are negative values present for some months. The minimum value of the profit is negative and equals to -14140.7 and the maximum value of the profit is 27220.69. The range of the profit column is “41361.39”. The total count of valid observations in the dataset is 8399.

Part 2: Ice-Cream Data Analysis

The survey dataset that includes the observations of 200 students of high school have been imported to the SPSS analytical tool. The total numbers of variables in the dataset are 5 including the id. However, only 4 attributes are major data columns that can be used in this work for analysis. The attributes in the dataset are “Female”, “Ice-Cream”, “puzzle” and “video”. The complete analysis of the dataset has been performed by using the SPSS analytical tool. The description of the attributes in the dataset is demonstrated in the table below

Attribute Name	Variable Type	Description
Female	Categorical	This variable has two sorts of observations 0 and 1. Where “1” represents a female participant.
Ice_Cream	Categorical	There are three types of ice creams in the dataset. The types of the ice cream (Strawberry, Chocolate, and Vanilla) are demonstrated by this variable.
Video	Continuous	The score of the video game is represented by this attribute.
Puzzle	Continuous	The score of the puzzle game is represented by this attribute.

Data Description and Visualization

2.1 Descriptive Summary of Ice_Cream

The descriptive summary of the all the attributes in the given dataset is shown in the above table. The attributes “Female” and “Ice_Cream” are categorical attributes (Vivaudou, 2019). However, the attributes “video” and “puzzle”.

The pie and bar plot shown below demonstrates the count of each type of ice cream in the given dataset.

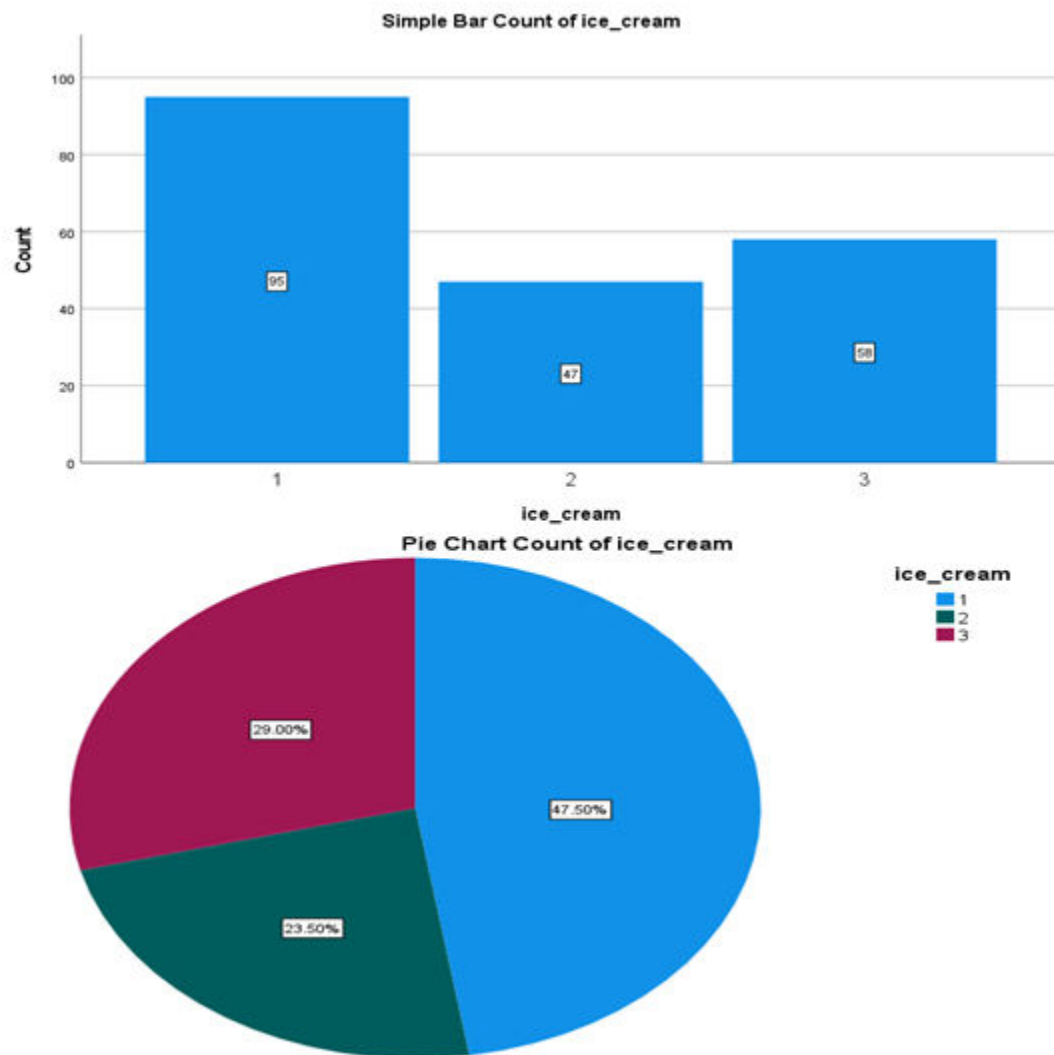


Figure 6: Ice_Cream Types

There are three types of ice creams are given in the dataset. Total number of ice creams in the dataset is 200. The ice cream types are labeled with the numeric numbers as “1”, “2” and “3”. Where,

- 1: Vanilla Flavor
- 2: Chocolate Flavor
- 3: Strawberry Flavor

Among 200 students, 95 students like the Vanilla flavor ice cream. A total of 47 and 50 students like the chocolate and strawberry flavored ice-cream.

2.2 Gender Analysis

The bar plot shown below demonstrates the count of male and female students in the dataset. The variable “Female” in the dataset represents where the student is male or female.

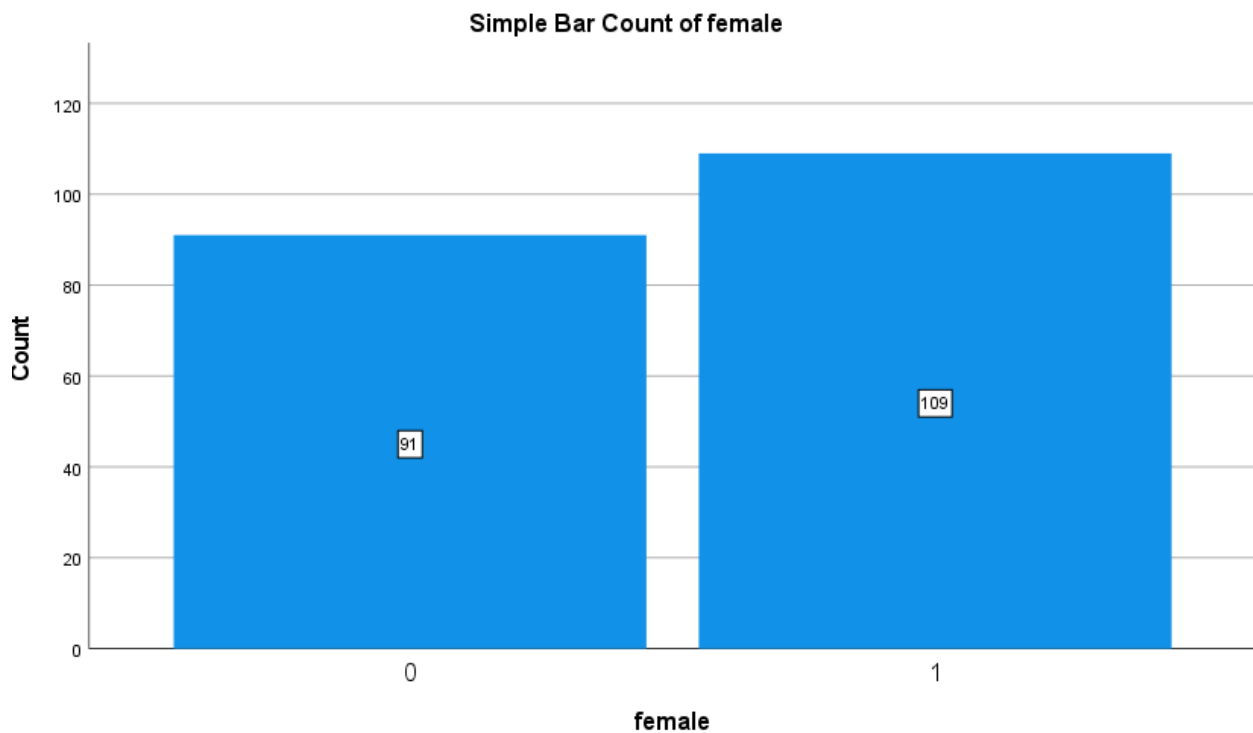


Figure 7: Count of Gender in the Given Dataset

The class label “0” represents a male student and the class label “1” depicts a female student in the given dataset. The count of the female student is 109 as shown in the above bar plot (López-Zambrano, et al 2018). The number of male students in the above dataset is lower as compared to the female students.

2.3 Mean and Median

Descriptive Statistics											
	N Statistic	Range Statistic	Minimum Statistic	Maximum Statistic	Mean Statistic	Std. Deviation Statistic	Variance Statistic	Skewness		Kurtosis	
female	200	1	0	1	.54	.499	.249	-.182	.172	-1.987	.342
ice_cream	200	2	1	3	1.82	.857	.734	.366	.172	-1.545	.342
video	200	48	26	74	51.85	9.901	98.028	-.189	.172	-.556	.342
puzzle	200	45	26	71	52.40	10.736	115.257	-.382	.172	-.525	.342
Valid N (listwise)	200										

Figure 8: Descriptive Table for Attributes

The descriptive statistics table shown in the above table illustrates the mean and median values.

2.4 K-Means Clustering

Clustering techniques are used for completing the classification tasks. The K-means clustering technique is an unsupervised classification method and used for creating different groups of unlabeled data instances based on associated feature variables (Kadra, et al 2021). A parameter called “K” is defined for creating different clusters and the value of the “K” represents the expected number of clusters.

The clustering analysis is performed using the K-Means approach in the SPSS. The obtained clustering result is as follows

Initial Cluster Centers			
	Cluster		
	1	2	3
video	29	55	72
puzzle	26	71	31
female	1	1	0

Iteration History ^a			
	Change in Cluster Centers		
Iteration	1	2	3
1	16.285	12.782	17.573
2	3.070	1.812	1.038
3	1.793	.967	.328
4	1.400	.798	.374
5	1.097	.683	.472
6	.406	.114	.616
7	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 7. The minimum distance between initial centers is 43.301.

Figure 9: Initial Cluster Center and Iteration History

The initial cluster center generated as a result and the iteration history are shown in the above picture. There are three types of Ice cream in the given dataset. The initial cluster centers for the

ice cream labels 1, 2, and 3 for video games are 29, 55, and 72 respectively. Similarly, the cluster center for puzzle score are 26, 71, and 31 respectively for 1, 2, and 3. The total numbers of iterations are 7.

Final Cluster Centers			
	Cluster		
	1	2	3
video	41	58	56
puzzle	46	61	39
female	1	1	0

ANOVA						
	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
video	5889.952	2	39.226	197	150.153	<.001
puzzle	7430.226	2	40.994	197	181.253	<.001
female	.403	2	.248	197	1.628	.199

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.

Number of Cases in each Cluster		
Cluster	1	67.000
	2	101.000
	3	32.000
Valid		200.000
Missing		.000

Figure 10: Final Cluster Center, ANOVA and Cases in Each Cluster

The final cluster center for video and puzzle game scores are as shown in the above picture. The number of class in each cluster generated by K-Means approach for ice cream types 1, 2, and 3 are 67, 101, and 32 respectively (Yuan, and Yang, 2019).

2.5 Data Mining and Text Mining Methods

The process of data mining is used to find the patterns in the data set and the building predictive model using analytical tools. The major methods of the data mining are building the predictive models and visualization. Visualization is known for plotting data using graphs and charts. The modeling is further classified into two types' classification and prediction. Classification is a concept for learning to make a supervised machine so that it can easily classify the dataset into various classes according to the type. On the other hand, regression is used for finding the correlation or it can be said to find the relation between independent and dependent variables. Data mining is used in the stock market suppose the officials want to find the price of gold for the next day so for this first is to find all the stocks that are directly dependent on the price of gold so that on the basis of the past data the prediction can be done.

The process of text mining is used to convert unstructured text into a structured format to find the patterns that give meaning to the text which helps to find the important insights. The textual analysis is performed using the standard analytical tools such as “Python”, and “R”.

Real Life Examples

1. **Predicting the Price of the Stock:** The prediction of the stock price is done by using the time series forecasting. The data mining methods are used for predicting the stock price. The process of data collection and preparation is also required for the stock dataset.
2. **Sentiment Analysis of Tweet Dataset:** Sentiments are the opinions that are classified based on the text dataset. There are three sorts of sentiments positive, neutral and negative. The sentiment analysis is performed by using standard analytical tools with textual dataset. The sentiment analysis of the tweet dataset can be performed for analyzing the sentiment of the tweets posted by individuals.

2.6 Advantages and Disadvantages of Using SPSS and Excel

SPSS as an Analytical Tool

SPSS and Excel, both are used for the analysis of the dataset. There are also numerous tools are available for the analysis of the dataset. The selection of the analytical tool depends on the computation speed and the ease of implementing the statistical methods on the dataset. SPSS is used for analyzing the dataset having large number of instances. Visualization as well as the

analysis of the data in SPSS is simpler. The major application of the SPSS is found for the data related to the social studies, survey and research works. The SPSS software is capable for performing statistical analysis, predictive modeling, visualization and textual analysis. Advantages and the disadvantages of the SPSS analytical tool used for the analysis is as follows

Advantages

The SPSS tool assists in importing the dataset from other analytical tool as well as supports different sorts of data file type such as xlsx, csv, text and sas. The data handling capability of the SPSS software is excellent. There is no requirement of syntax for assessing a dataset (Jahwar, and Abdulazeez, 2020). However, for every analysis performed a syntax file is created that can be changed for getting the expected results. Some major advantages of using the SPSS software is shown in below points

- The data handling capability of the SPSS software is excellent.
- Excellent charts and summarized table can be easily created without use of any syntax code.
- The SPSS software is available in multiple languages and hence, a lot of individual can use the tool for performing the analysis.
- Anyone with some sort knowledge about the statistical analysis can easily use to the software for performing analysis and creating model.
- The likelihood of generating erroneous results is very little.
- An individual can select the type of chart that is appropriately fits for the data attribute.

Disadvantages

There are numerous advantages of using the SPSS software. However, there are some limitations of using the SPSS software. These limitations are as follows

- Analysis of the large datasets (Big Data) is not possible by using the SPSS tool as the computation is very high.
- The analysis performed by using the dataset that is collected using the biased method will generate poor results.
- Only simple summary of the models are obtained with the SPSS tool.

Excel as an Analytical Tool

For performing fundamental analysis creating basic models, the data analysis toolpak option of Excel can be used. Also, Excel support different types of visualization that can be used for capturing different sorts of charts. Some major advantages and disadvantages of using EXCEL data analysis toolpak are as follows

Advantages:

1. The analysis of data is done by using the natural queries. This will assists in understanding the models and results easily.
2. There is no need of any code script to perform the analysis.
3. Different sorts of visualization can be captured by using the Excel tool.

Disadvantages:

1. Performing analysis on unstructured data is a complex task in excel.
2. Analysis become complex with large dataset files.
3. Creating the interactive charts is a challenging task with use of Excel.

References

- Jahwar, A.F. and Abdulazeez, A.M., 2020. Meta-heuristic algorithms for K-means clustering: A review. *PalArch's Journal of Archaeology of Egypt/Egyptology*, 17(7), pp.12002-12020.
- Kadra, A., Lindauer, M., Hutter, F. and Grabocka, J., 2021. Regularization is all you need: Simple neural nets can excel on tabular data. *arXiv preprint arXiv:2106.11189*.
- Khadka, B., 2019. *Data analysis theory and practice: Case: Python and Excel Tools*.
- López-Zambrano, J., Martinez, J.A., Rojas, J. and Romero, C., 2018, July. A tool for preprocessing moodle data sets. In *Proceedings of the 11th International Conference on Educational Data Mining*, Buffalo, NY, USA (pp. 15-18).
- Mark, K., 2018. *Introduction to Data Analysis Using Excel and Lab Report Writing Using LaGCC Institution Data*.
- Sall, C. and Lehman, J.M.P., Berk & Carey, 2019. *Data Analysis with Microsoft Excel, Updated for Office XP Carver, Doing Data Analysis with Minitab 14 Carver & Nash, Doing Data Analysis with SPSS Version 12 Casella & Berger, Statistical Inference Derr, Statistical Consulting: A Guide to Effective Communication*.
- Stančin, I. and Jović, A., 2019, May. An overview and comparison of free Python libraries for data mining and big data analysis. In *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (pp. 977-982). IEEE.
- Vivaudou, M., 2019. eeFit: a Microsoft Excel-embedded program for interactive analysis and fitting of experimental dose–response data. *Biotechniques*, 66(4), pp.186-193.
- Yuan, C. and Yang, H., 2019. Research on K-value selection method of K-means clustering algorithm. *J*, 2(2), pp.226-235.