



DATA HANDLING REPORT

Module: Data Handling and Business
Intelligence. Module Number: LCBB5000

LECTURER: DR MD NAZMUL HASAN

Student Name:
Student Number:

Contents

Part 1: Superstore Data Analysis	2
1. Introduction	2
2. Superstore Data	3
3. Profitability Analysis	3
4. Analysis Conclusion	7
5. Excel Evaluation and Conclusion	7
Part 2: Smile Clinic Data Mining	9
1. Introduction	9
2. Data Overview and Data Mining	9
3. Clustering and Data Mining Methods	14
4. SPSS Evaluation and Conclusion	15
References	16

Part 1: Superstore Data Analysis

1. Introduction

Business intelligence is the processes, technology and products of manipulating and analysing data in a meaningful way in order to create knowledge and actionable insights on various aspects of the business and it is at the core of decision making in any organisation (Al-Eisawi and Lycett, 2012). Being at the core of most important decisions in a business, tools and technologies have been developed to facilitate the analysis and visualisation in a meaningful way of data. Data gathered through various processes and from multiple areas of business can show how the business is performing and what areas are underperforming. Furthermore, the visualisation of this data gives managers and decision-makers a wider view of the processes which the business is engaged with (Zheng, 2017).

Gathering information from various internal and external area of business can create a base of information, otherwise called data, or database. The data within the database can be further classified by various elements of provenience and chronology to create separate elements of data. These data elements can therefore be combined with other elements which are related in any way to create data sets. The data sets are used by computer programs such as Excel or SPSS to help analyse and extract useful information (Alamaki, Rantala and Valkokari, 2018).

While there are many other tools to manipulate data, Excel from Microsoft is one of the most popular due to accessibility and ease of use (Winston, 2016). Berk and Carey (2007) stated that Excel is a tool for learning basic statistical analysis and beyond, with applications far more reaching than education. Most small businesses resort to Excel to anything from managing departmental data, to payroll or even accounting.

This report will look at how Excel can be used to determine the cause and degree of decline in profits of Superstore by looking at Sales data set and manipulating the data to create useful information.

2. Superstore Data

The Superstore Sales Data set provided for this assignment contains 8399 rows of information. Each row contains a row ID, order ID, order date, order priority, order quantity, a sales figure, a flat discount, shipment method, profit, unit price, shipping cost, customer name, province, region, customer segment, Product category, product sub-category, product name and a shipment date.

At the first glance this data set contains elements appropriate to sales department and can be seen that the data has been recorded from 1st January 2009 to 30th December 2012.

Other key figures at first glance are:

- ◆ Maximum order size is of 50 items, with an average at 25.07
- ◆ The maximum sale was for 13 items valued at 89061.05 with a profit of 27220.69
- ◆ Data is contained in 158,441 cells.
- ◆ Data seems to be of sales department origin, containing only data relevant to invoices and order tracking, rather than organisational performance.

3. Profitability Analysis

There are several ways to determine the profitability in the sales data. While writing new tables using function commands and formulas is most common when dealing with a relatively small data set, manual handling of large data sets is counterproductive and tedious. Excel has integrated automated data analysis tools which can substitute the classical function processing. Bissett (2020) recognises that in order to understand these automated analysis tools, a basic understanding of various data processing functions must be maintained.

To establish the behaviour of profit data, a correlation function can be used. First, the profit data can be looked if it has any correlation with other elements. This is done by entering the correlation function as =CORREL to determine if, and how different data sets affect each other. Calculations has shown a $r=0.58$ correlation between sales

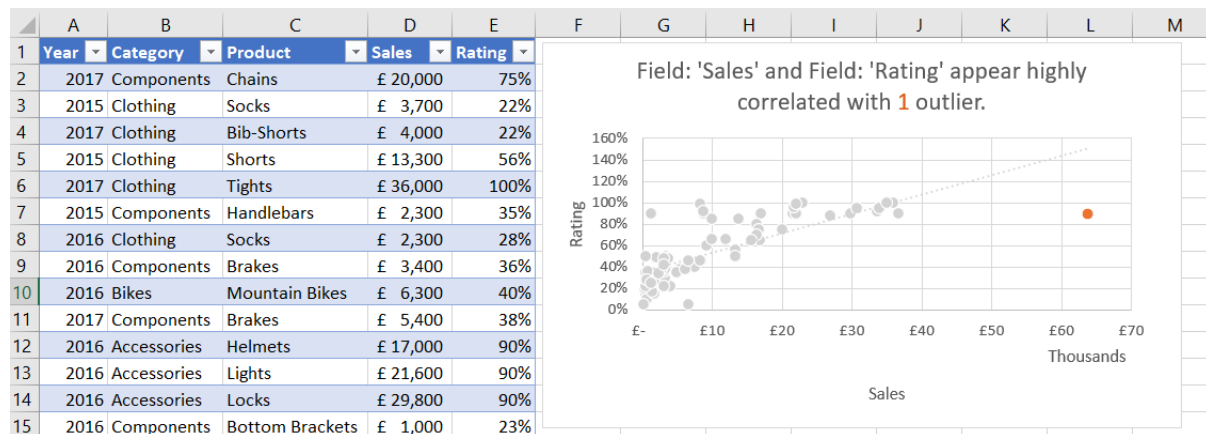
and profits, which is statistically significant. This show that the increase in sales directly affects positively profit amounts.

Using =SUM function we can extract the total values for profits and sales and determine that the sales stand at 14,915,600.82, yet Profits stand at 1,521,767.98. This means that the profit margin is 10.20% for the period.

A correlation calculation between profits and order date shows that there is no significant correlation at $r = -0.007$. The same correlation result is obtained for shipment dates too, showing that date of order and shipment do not correlate with profits. Therefore, the strongest factor when analysing profitability must be Sales data set.

A correlation between Sales data set and the Rating of sale for various items is found with only 1 outlier which can be regarded as an exception. This means that Sales are directly correlated with the items sold as seen in Figure 1.

Figure 1, Sales and Products correlation. Source: Produced by author.

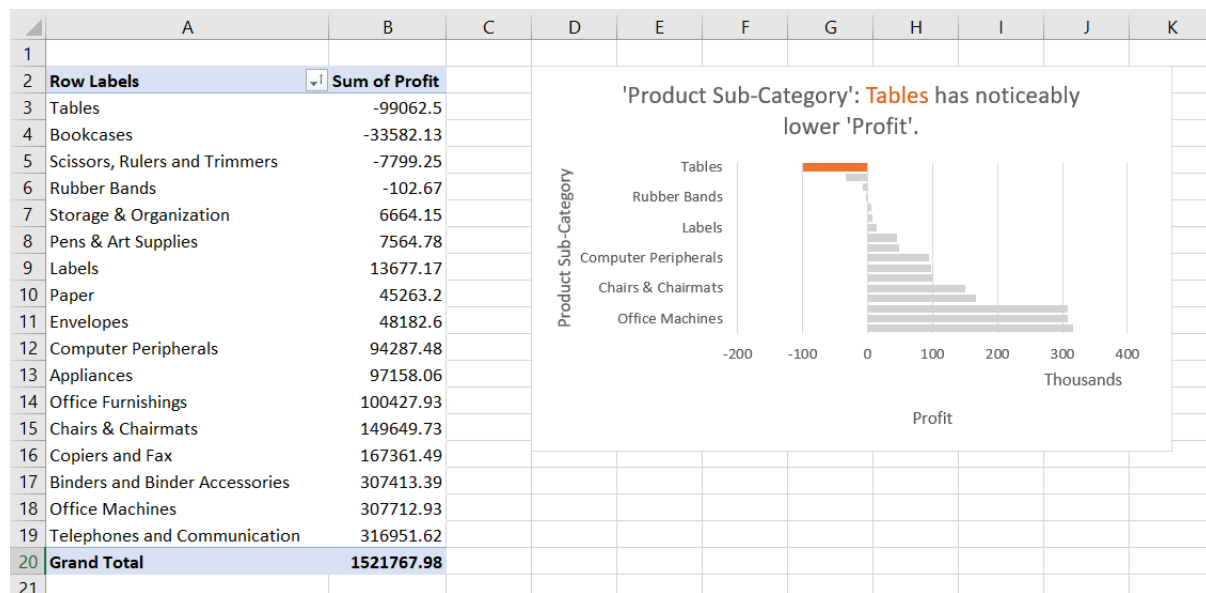


A further analysis of the data using the sum (=sum) of sales over the categories sold show that Accessories and clothing categories of product sells best. Moreover, products such as tables, bookcases, scissors, rulers, trimmers and rubber bands are unprofitable, and even causes a loss for the business. These are shown in Figure 2 and Figure 3. These are evidence of the most performant products, and worst performing sales. It must be considered that data such as breakeven point is not included to make a more detailed analysis of product market performance.

Figure 2, Most profitable category. Source: Produced by author.



Figure 3, Most unprofitable products, Source: Produced by author.



A further step to identify profitability issues is to look at the customer segment. Identifying areas of the market that are most profitable allows the business to focus on the segment in which it performs the best (Porter and Advantage, 1985). The data allows for a segment analysis by checking profitability of each market segment. The data shows that corporate customers are responsible for the most amount of profit as seen in Figure 4. Furthermore, the number of ordered items can be analysed in the same way and data shows that corporate customers order twice as much as the next in ranking (Figure 5).

Figure 4, Profit by Customer Segment, Source: produced by the author

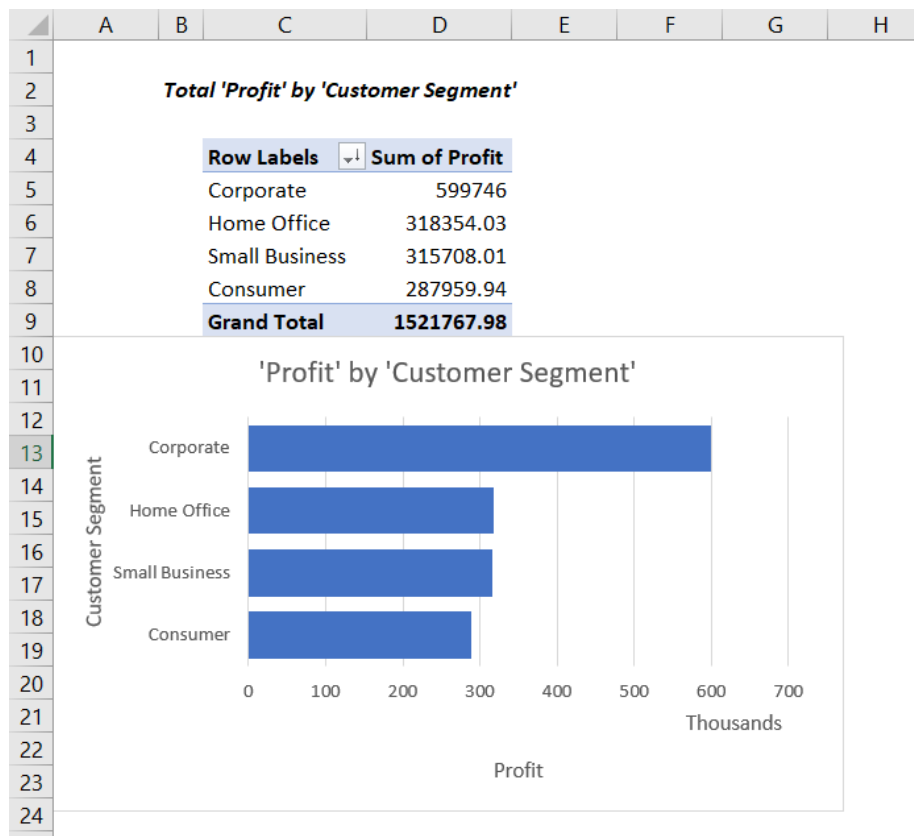
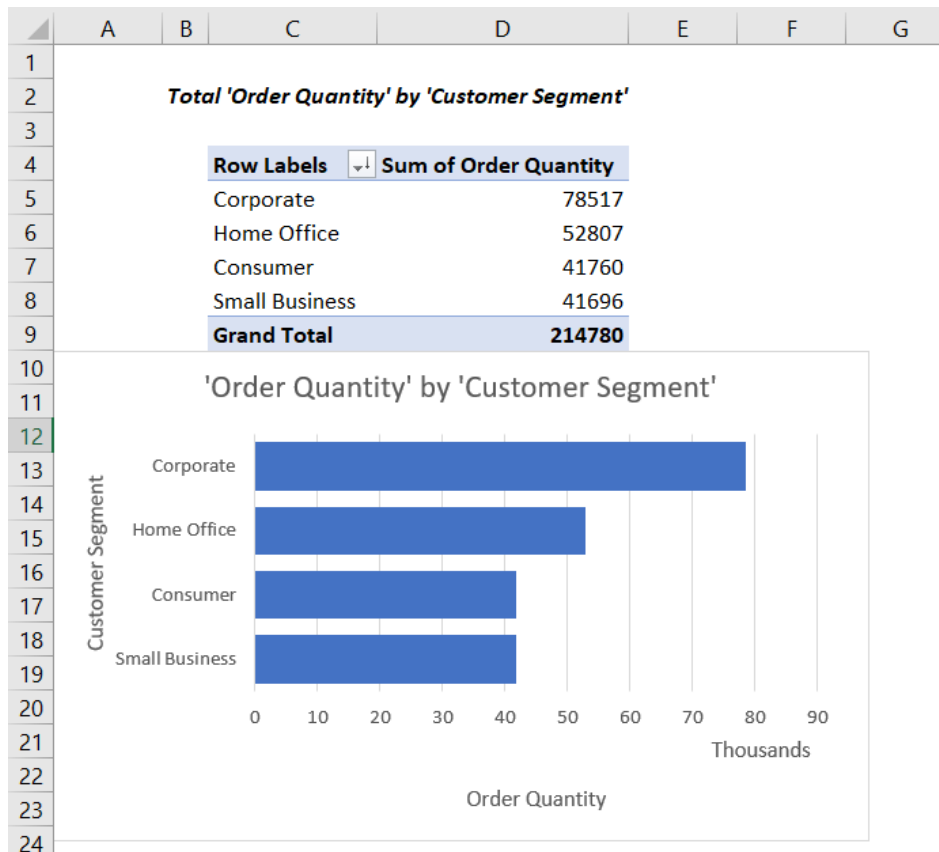


Figure 5, Order Quantity by Customer Segment. Source: produced by author (SUMIF)



So far, the data has revealed that profitability of Superstore can be traced to products sold and customer segments. Seasonal or chronological variations can be excluded due to a lack of correlation between profits and order timings. This means that a diversification strategy is in place for the business, yet differentiation and market segment targeting is lacking. Furthermore, the business offers products that are incurring significant losses, which are not counted towards a breakeven point or justified in other meaningful way within the given data.

4. Analysis Conclusion

The analysis has shown that the business is failing to manage unprofitable products, and does not use a focused product line-up. A review of sold products and a closer management of unprofitable sections is advised. Moreover, differentiation and segmentation strategy must be enforced to increase the number of high-volume, paying corporate clients. If implemented, these will increase the profitability of the business.

The business needs to adopt a development plan for a product-focus strategic approach which is compatible in implemented in conjunction with a marketing campaign which aims to expand the most profitable market segment. By adopting specialisation, differentiation and focused strategies the business can increase the profitability at an acceptable rate.

Considering the given information and the current trend in the profitability of the business, a streamlined implementation of innovative and new management practices to improve operational management efficiency is needed. Due the lack of further data addressing sub-structure of profit data set such as overheads expenditure, actual cost of sales and other costs, the recommendations have been limited to sales and marketing.

5. Excel Evaluation and Conclusion

Using Excel as a data manipulation and analysis tool is common. This is mainly due to simplicity and ease of use of the systems, which automates tedious calculations and data manipulation processes to arrive at useful results fast and efficiently.

Dealing with small amounts of data, Excel can be the perfect tool for small businesses and sole traders. This is because it offers a clean, easy to use interface, and the resources available for the support and training are at easy to access. Furthermore, databases can be stored in various formats and archived for record keeping and preservation in a very easy manner.

Excel can run into some problems when processing large data sets or big data. Most problems are around performance and stability of the program. While small problems can be fixed without requiring much technical expertise using built-in troubleshooting, bigger data sets can create unsolvable errors. Excel is limited to a worksheet of 1,048,576 rows by 16,384 columns, offering over 17 billion cells to work with. While an impressive number, large corporations might encounter data which would not fit in a worksheet. As seen in previous paragraph, Excel can handle practical tasks of data analysis, either by using automated processes, or implementing function commands within a worksheet to manipulate data sets. To learn syntaxes takes logical and technical skill (Baier and Neuwirth, 2007).

A serious disadvantage is the lack of security for a spreadsheet. With global increased awareness in cybersecurity, business intelligence security, corporate data security is of vital importance for any organisation which rely on data to operate. Features such as multiple user access and easy of access makes storing data on spreadsheets quite vulnerable to unauthorised access. This is a very big problem for big corporation which resort to other data manipulating tools to encrypt the data at the same time as it records and analyses it.

In conclusion, Excel can be a versatile and easy to use tool for small and micro businesses, which benefit from the flexibility and versatility of the program, and availability of resources for manipulating the data. Selwyn (2020) argues that due to the easy of use and access by educational institutions, excel has been used to manipulate fictive “bad” data to create meaningful information, thus sometimes provoking counterarguments. Nevertheless, usage of small data, even if “bad” to simulate manipulation of big data offers students with a chance to practice concepts and techniques to work with information, business intelligence and data.

Part 2: Smile Clinic Data Mining

1. Introduction

This section of the report will look at how SPSS from IBM can handle data to produce reports. SPSS stands for Statistical Package for Social Sciences, while this name has become redundant due to the expansion of uses for the software package well beyond its intended purposes at launch. Now SPSS can manage a very diversified list of data mining, analysis, manipulation and visualisation.

For the purpose of this report, the author will use SPSS to show examples of data analysis and mining, clustering and frequency analysis. Data used for this section is provided.

2. Data Overview and Data Mining

The data as seen in Excel is composed of 100 rows by 5 columns. These contain interview number, name of respondent, gender, age and if or not they have purchased rice. This data set is constant with a basic survey data set. The number of respondents is 100 and number of questions, or variables are 4. The 5th variable is the identifier of interview line, numbered 1 to 100, this is a categorical variable in SPSS. The first variable is the name of the respondent. This has no weight statistically speaking yet serves as a second identifier for responses. The second variable and first of significance is the gender of the respondent. Since SPSS needs data encoded into a binary, nominal, ordinal or interval data type, the variable of gender has been encoded to represent male as 1 and female as 2. The third variable is the age of the respondent, shown as a whole number. The last variable is the rice consumption. This is encoded as such: 1 is for respondents which have eaten rice, and 0 is for those who did not.

When working with SPSS the simplest place to start is running a two-step cluster analysis. A cluster analysis in SPSS runs several tools to examine the data and reveal simple relations and functions. Using the provided data sample, we can open

the Two Step Cluster Analysis tool (Figure 6) and set our variables categories. As categorical variable we have InterviewID and Name, the rest are Continuous Variables. This can be changed as needed for analysis.

Figure 6. Two Step Cluster Analysis Setup

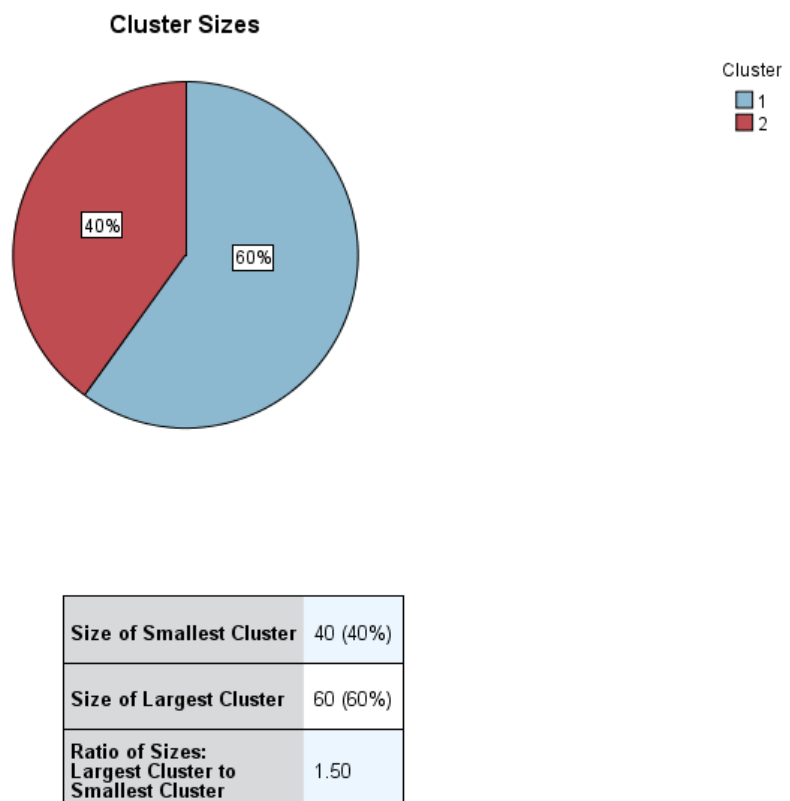
The screenshot shows the 'Two Step Cluster Analysis Setup' dialog box. It features a large empty box on the left for variable selection. To the right, there are two lists: 'Categorical Variables' containing 'InterviewID' and 'Name', and 'Continuous Variables' containing 'Rice', 'Age', and 'Gender'. Below these lists are four configuration sections: 'Distance Measure' with radio buttons for 'Log-likelihood' (selected) and 'Euclidean'; 'Count of Continuous Variables' with 'To be Standardized' set to 3 and 'Assumed Standardized' set to 0; 'Number of Clusters' with radio buttons for 'Determine automatically' (selected) and 'Specify fixed', with a 'Maximum' value of 15 and a 'Number' of 5; and 'Clustering Criterion' with radio buttons for 'Schwarz's Bayesian Criterion (BIC)' (selected) and 'Akaike's Information Criterion (AIC)'. At the bottom are buttons for 'OK', 'Paste', 'Reset', 'Cancel', and 'Help'. On the top right are 'Options...' and 'Output...' buttons.

To run a cohesive cluster analysis, specific clustering must be chosen, then, analysed to see if there are any significant important data.

To accomplish the aim of this report, we need to cluster data to give meaningful information which can respond to questions asked. For example, the first question is how many customers of Smile Clinic eat rice?

To find out, a two-step clustering can be used to produce reports on the data sets. We use InterviewID as a continuous variable and rice as a categorical variable to obtain the following cluster report shown in Figure 7:

Figure 7. Cluster analysis of how many people eat rice



From the cluster we can understand that 60% of the respondents have eaten rice. This is a simple example of clustering function within SPSS. It produces significant data based on the information contained within the data sets.

In Figure 8, we can observe that SPSS created a short cluster analysis report, based on the size of the clusters. The Two Step Cluster shows a table of clustering containing fields for label, and description, which can be used to format the data and 'inject' more information in the database. The Size and Inputs fields are related directly to the percentages of cluster responses frequency over the number of cluster elements.

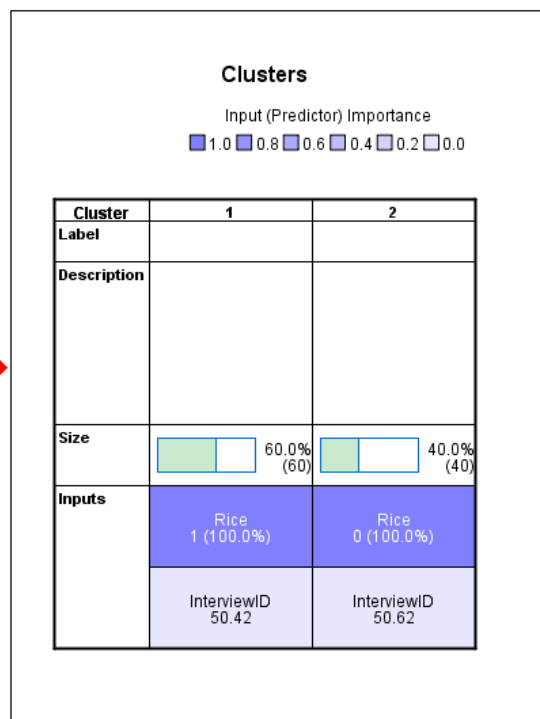
The second information of interest would be how many of the responders are male and how many females.

For this we can use the same method yet change the categorical variable from rice to gender.

Figure 8. Cluster report

```
/DISTANCE LIKELIHOOD
/NUMCLUSTERS AUTO 15 BIC
/HANDLENOISE 0
/MEMALLOCATE 64
/CRITERIA INITHRESHOLD(0) MXBRANCH(8) MXLEVEL(3)
/VIEWMODEL DISPLAY=YES.
```

TwoStep Cluster



Therefore, when we run the analysis, we will get a model summary showing number of clusters, Inputs and the quality of cluster displayed in a silhouette measure of cohesion and separation (Figure 9). This is important for determining the quality of cluster analysis and the cohesion of the data we used for this. By opening the model summary, we can see the model viewer (Figure 10), which shows different visualisation models for the clusters.

Figure 9, Model Summary

→ TwoStep Cluster

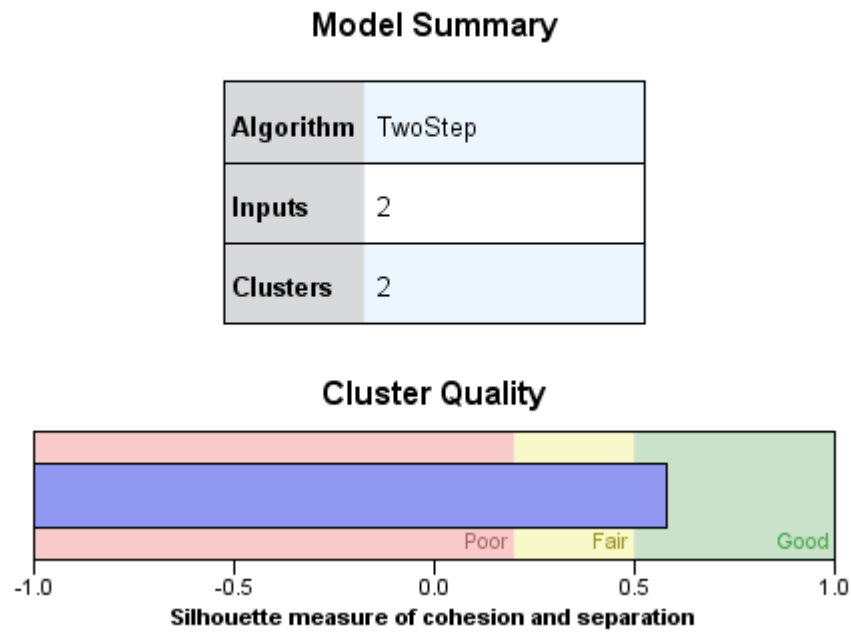
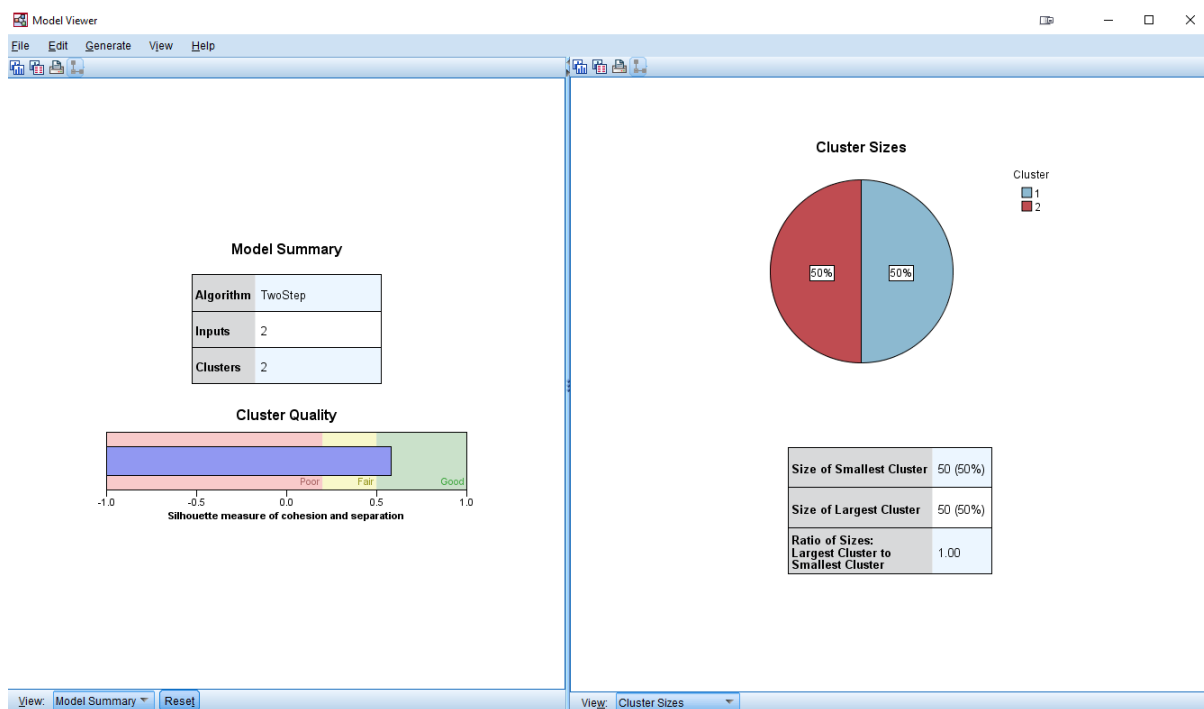


Figure 10, Model Viewer



On the right side we can observe that the clusters are distributed in equal measure, telling us that the data contains an equal number of male and female respondents.

The next data we need to extract is the mean and median of age of respondents (customers). To do this, we will use descriptive statistics (Figure 11). This generates a mean and variable table. We can do this for both of data sets for age and gender at once. The result shows the mean of the data in Figure 12.

Figure 11. Descriptive Tool

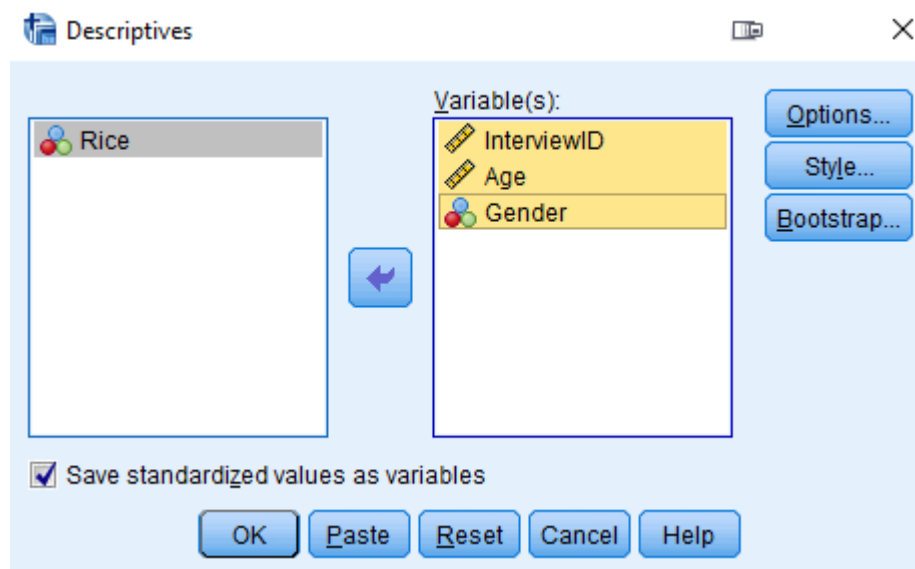


Figure 12. Mean and Variable for age and gender

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
InterviewID	100	1	100	50.50	29.011
Age	100	13	26	20.35	3.560
Gender	100	1	2	1.50	.503
Valid N (listwise)	100				

3. Clustering and Data Mining Methods

Data clusters are sets of data that can relate to each other in various ways to create meaningful information which can be acted upon. Clustering in SPSS can be done in three main ways, TwoWay Clustering, K-Mean Clustering and Hierarchical Clustering. Each of these methods employ different frameworks in order to analyse

the data. Beside the clustering method, analysts can use a large number of tools. These tools include data mining methods and techniques. Some of them are pattern tracking, classification, association, regression, prediction, correlation, and more. Each of these methods are based on approved and proven mathematical models for statistical and data analysis developed over the years (Wendler and Grottrup, 2016). Dixon (1951) stated that some of these data mining methods are fundamental to the process which transform statistical information into insightful information. With time, models and methods have been developed to analyse any kind of data, some of which use mathematical models which are interchangeable to cover even a larger sector of data analysis. It is common that many researchers would use some form of data analysis method during their research (Ott and Longenecker, 2015).

4. SPSS Evaluation and Conclusion

As seen in the previous paragraph, SPSS can handle data in a more powerful and complex way than Excel. SPSS can transform and manipulate data in clusters and through usage of automated functions of the software. In contrast, Excel makes it easy for data recording, quick visualisation and has some statistical analysis potential. It is recognized that SPSS is far superior to Excel when processing data, especially big data banks and sources. To do the same, in Excel, analysts must work harder. This is because analysts must implement function, analytical frameworks and formulas by writing function manually and then manually selecting, sorting and preparing the data.

Therefore, in conclusion, Excel is a simple tool for data management, whereas SPSS is a powerful statistical analytical tool. Both of them have uses. A business might record data with Excel due to accessibility and later analyse the data using SPSS. If the data is correctly formatted, the switch between the two software can be easy and interchangeable.

References

Part 1:

- Al-Eisawi, D. and Lycett, M., 2012. Business Intelligence. In *Definitions, managerial effects and aspects: A systematic literature review,[w:] Proceedings of the 14th International Conferences on Enterprise Information Systems* (Vol. 28, pp. 209-214).
- Alamäki, A., Rantala, T., Valkokari, K. and Palomäki, K., 2018, September. Business roles in creating value from data in collaborative networks. In *Working Conference on Virtual Enterprises* (pp. 612-622). Springer, Cham.
- Baier, T. and Neuwirth, E., 2007. Excel::COM:: R . *Computational Statistics*, 22(1), pp.91-108.
- Berk, K.N. and Carey, P., 2007. *Data Analysis with Microsoft Excel*. Brooks/Cole, Cengage Learning.
- Bissett, B.D., 2020. *Automated data analysis using excel*. Chapman and Hall/CRC.
- Porter, M.E. and Advantage, C., 1985. Creating and sustaining superior performance. *Competitive advantage*, 167, pp.167-206.
- Selwyn, N., 2020. 'Just playing around with Excel and pivot tables'-the realities of data-driven schooling. *Research Papers in Education*, pp.1-20.
- Winston, W., 2016. *Microsoft Excel data analysis and business modeling*. Microsoft press.
- Zheng, J.G., 2017. Data visualization in business intelligence. In *Global Business Intelligence* (pp. 67-81). Routledge.

Part 2:

Dixon, Wilfrid J., and Frank J. Massey Jr. "Introduction to statistical analysis." (1951).

Ott, R.L. and Longnecker, M.T., 2015. *An introduction to statistical methods and data analysis*. Cengage Learning.

Wendler, T. and Gröttrup, S., 2016. *Data mining with SPSS modeler: theory, exercises and solutions*. Springer.