UNDIRECTED KNOWLEDGE DISCOVERY PROJECT

Sukhjeet Kaur

STUDENT ID: 45665761 Masters of applied statistics

Contents

Executive Summary

Currently, Coles has a declining market share compared with new competitors such as ALDI which is gaining momentum in the competitive market, because of Australians' insatiable taste for lower prices. Moreover, its parent company Westfarmers is deciding to spin it off as a separate company as per recent news. Considering all these aspects this report is designed to assist in marketing of certain products and identify target markets.

Two approaches are applied to large set of Coles customers and transactional data.

- The application of k-means clustering algorithm allows to differentiate the customers by assigning them a particular cluster or segment based on their income or value of purchases and other demographic variables.
- Market basket analysis by association rule mining, using apriori algorithm, allows to discover pattern of purchases by assessing through various metrics. The application of this technique yields the following recommendations:
    i. Fish and cleaning products are often purchased together, therefore could be placed separately at a distance within stores, to encourage customers to traverse the store more in order to increase the sales of products.
    ii. Other items which are very often purchased together are Nappies and baby food. These items can be collocated to yield increase in the sales as well as customer satisfaction.

Introduction

The amount of simulated Coles data is large enough to look for their customers and products' sales. In order to maintain the competitiveness businesses should utilise this data to drive decision making on marketing, planning and customer service.

The exploration of this transactional data containing customers' demographics uses unsupervised learning techniques to aim the following questions:

- Whether Coles have distinct group of customers, depending on their spending habits and demographic traits?
- Does the transactional data help to develop customers' purchasing patterns, if so, then how the information obtained by analysing patterns can be used to improve sales, customers' satisfaction and other business aims?

Description of the data set

Original dataset

The Coles dataset used for analysis in this project involved 58,100 transactions with 53 attributes. These attributes could be classified into 3 categories namely: transactional variables (e.g. pmethod, value - which records information about the transaction), demographic variables (sex, homeown, number of children etc. – containing the information about the customer making particular transaction) and product variables (eg. milk, bread, fish, juice etc – containing information about which items were purchased in the transaction) which are 2, 6 and 44 in number respectively.

Table 1: Dataset attribute description

Variable	Description	Type	Value
Receipt ID	Unique identification for each transaction	Nominal	Sequentially increasing unique ID
value	Amount spent in the transaction	continuous	Positive real numbers
Pmethod	Method of payment	Nominal	1= cash,2= credit card, 3= eftpos, 4=other
Sex	Customer's sex	Nominal(binary)	1= Male, 2= Female
Home own	Whether customer owns a home	Nominal	1= yes, 2= no, 3= unknown
Income	Annual income of the customer	Continuous	Positive real number
Age	Age in years	Continuous	Positive integer
Postcode	Postcode of customer (where the customer lives)	Nominal	Four numeric characters
nchildren	Number of children	Continuous	Positive integer

The remaining 44 product variables record the presence of a particular product in a transaction, which are encoded by 0 or 1 indicating the absence or presence of a product in a transaction respectively.

1) Data pre-processing

Missing and non-conformant data

Each attribute was assessed to check whether there is any missing data, outliers (- values beyond constraints) or some kind of erroneous data in order to carry out the fruitful analysis. (refer figure1 and 2 of appendix1 to look for missing values)

ReceiptID: The data has 9 duplicate ReceiptIDs but their transactions were completely different so it is considered that there are no duplicate transactions, hence not removed.

Value: In this attribute, no missing values were found but three highly unlikely values (extreme outliers i.e. 1967.7,1243.0,802.1) were detected which were plausible so were not removed from the data. (fig4 of Appendix1 shows outliers)

Pmethod: On investigation, it was discovered that there were 7358 values (12.66%) coded as 4 or other number which were recorded as missing and imputed with the mean of that column by taking only 1 significant digit.

Sex: There were no missing values in this variable, and the values were consistent with the meta data.

Postcode: This variable had 9887 missing values and some of the values coded with more than 4 digits, so this variable was not considered for further analysis.

Homeown: There were 1549values (2.66%) that were coded as 3 for unknown and other in the dataset which were treated as missing values, and were imputed with the mean of the column by taking value of one significant digit.

Income: There was 1 missing value in this variable, and was imputed with the mean of incomes. In terms of outliers, there was one extreme value of income which was recorded as $650235.00, although it is plausible income but this value was not considered for cluster analysis. (fig5 of Appendix1 with outliers)

Age: The age variable has 1missing value which was imputed by mean of the age variable and was rounded off to its nearest integer. The minimum and maximum ages were 10 years and 95 years respectively, these ages are also entirely plausible so no adjustment to these values were made.

nchildren: In this attribute,2 missing values were recorded which were replaced with mean by taking only 1 significant digit. The maximum number of children were considered to be 10 remaining values were also imputed with mean of up to one significant digit.

In terms of product variables, fruit had 10 values coded as 11,3,5,6,7," o" and 2 missing values which were replaced by 1 except 0 for "o" by assuming that 0 was mistakenly coded as "o" and others as numbers greater than 1 and mean respectively.

Fruitjuice variable had 10 values coded as 2 which were imputed with mean by taking up to 1 significant digit. cereal had 9 missing values replaced by mean, cannedveg, Pizzabase, milk, confectionery had 1 missing value each and no imputation was done as these were from same transaction.

The distribution of the continuous variables are shown in figure3 of Appendix1

Who are the customers?

Some observations of the customer's demographics and transactional values were noted to describe what majority of the customers are: (see figure A)

- The customers were 40.19% males and 59.80% females.
- The average value of the transaction recorded on all the transactions was $63.58
- Majority of the Coles customers own a home (i.e.74.90%)
- The average income of the customers was $70169
- On average the age of customers was 38years
- The popular choice of payment method among customers was credit card (i.e.55.21%)
- Majority of the customers are having at least one child.

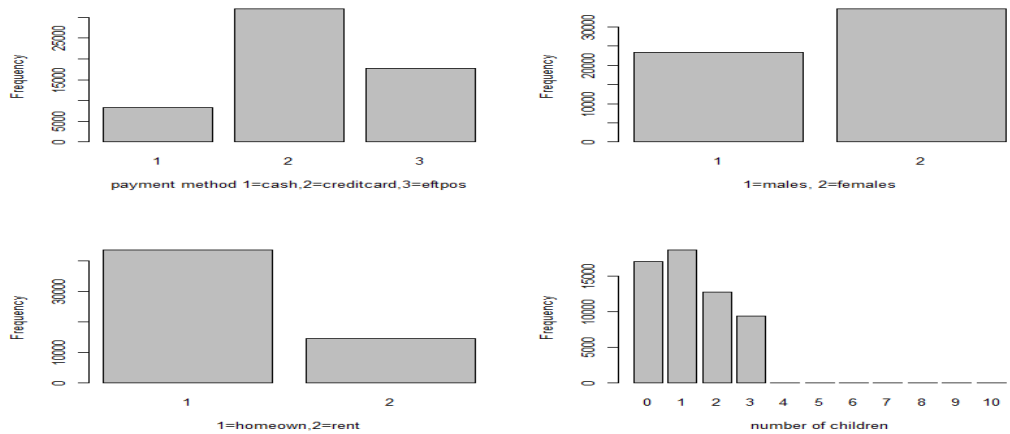

Figure A: Proportion of customers' demographic attributes.

What are the customers' purchasing?

The simulated dataset contains purchase information of 44 common food items and household goods. The most frequently purchased items were bread (purchased in 48058 transactions, or 82.7%), followed by milk (purchased in 47239 transactions, or 81.3% and cereal which was found in 44278 baskets (76%). In contrast, least sold items are energy drinks found in just 1112 transactions (2%), frozen fish (found in 1755 baskets or 3%), tea towels found in 3% baskets as well. The frequency plot of items purchased is given in Appendix2 Figure6

Methodology

I. Clustering of customers

To find distinct group of Coles customers, k-means clustering algorithm was applied to customers' transactional variables such as age, income, value of purchases, number of children. k-means partitioning technique is most popular method as it is highly efficient and suitable for large datasets and it is only used for continuous variables. For this reason, kmeans algorithm is chosen. This method divides dataset into clusters based on Euclidean distance over the sample space. Alternative methods such as hierarchical clustering was not chosen as it is not as efficient as k means for large datasets.

II. Market basket Analysis

Association between frequently purchased items were mined using apriori algorithm, implemented by the R package arules. This algorithm provides easily interpretable association rules and values to assess those rules. Also, this is computationally efficient algorithm and very concise to read effectively.

The threshold for rule support was set low, at 0.1 as dataset is very large, because most interesting associations would occur with low frequency. Support of 0.1 means, itemsets occur 5810 times in the dataset. Rule confidence was set to 0.8 to make sure that there is strong association between rule antecedents and their consequents. The minimum and maximum length of itemsets is set to 2 & 5 respectively, to ensure less rules are formed for better analysis. Rules with lift greater than 1 were considered a standard cut off with lift greater than 1 indicating that the rule occurrence is better than chance,

The overall aim is to predict the rules with high support, high confidence and high lift. The key thing is to ensure consideration of actionable and valuable or viable rules. Data was available in indicator format, and so new file was created containing only the 44 product variables and was loaded into R as item matrix to get the result.

Results

Cluster Analysis

Four distinct customer groups were formed through cluster analysis (see figure7 of Appendix2) There are clear differences between spending habits among these groups with transaction values ranging from $73.59 to $102.14

	Group.1	Value	income	age	nchildren
1	1	79.94001	134035.99	40.30674	1.386515
2	2	102.14736	30285.35	49.69060	1.078632
3	3	78.49783	78696.14	40.64480	1.238246
4	4	73.59546	66775.91	38.97376	1.277311

Table B: Identified groups based on cluster analysis

Young customers with moderate income spending on average $73.59 and have children. Old adults with low income spending on average high amount of $102.14 and have children. Adults with very high income make purchases of $80 on average and have kids. Another group of Adults with income higher than median make purchase of $78 have kids.

Association Rules

In total,6550 rules were generated. Out of those 3 rules which are actionable and viable are explained below: (Tables1,2,3 given in Appendix2 contain the top 5 rules sorted for support, confidence and lift)

Rule1: (Highest confidence) Tomato sauce, bread, vegetables, olive. Oil => banana. This rule had a confidence of 99.5% which means that 95.5% of the times customer who buys fruit, frozen meal, tomato sauce, vegetables tend to buy banana also. In fact, all the top rules contain bananas as a consequent. This combination of items occurs in 5885 transactions which is only 10% of total but it is quite significant. Moreover, lift is also high for this rule which indicates the occurrence of this combination is not by chance. Given that this combination includes range of grocery items, rather than placing these items next to each other, one possibility is to sell these as a deal bundle.

(Note: The rule with highest confidence was not explained here as it contains fruit as antecedent for bananas and was deemed obvious)

Rule2: (Highest lift) Fish, vegetables, bananas =>household cleaners. This is another combination that occurs in 10% of the baskets, however has high confidence (91.8) and highest lift of 2.4. This association rule is therefore quite strong as around 92% of the customers purchase household cleaners when they purchase fish, vegetables, bananas. Hence, it suggests that house hold cleaners are very likely bought with general grocery items. Therefore, discounts for cleaning products can yield high sales and customer satisfaction.

Rule3 (Highest support) milk=>bread. With 0.827 confidence, 82% of the customers purchase bread when the buy milk. The support for this combination is very high 0.672 indicating that these items occurred in 67% of all transactions-a huge amount with such a large dataset. The lift of .99 also indicates that this rule is slightly better than chance. These items could be collocated to increase sales.

Conclusion

The aims of the project were met through the data mining techniques: association rules and kmeans clustering algorithm. Four distinct group of customers were identified, while useful association rules were mined that assist the following suggestions. After

implementation of these recommendations Coles can be a competitive supermarket chain in future years.

Recommendations:

- Bananas were most dominant item as a consequent of combinations with high support, confidence, lift. Thus, this item should be sold by providing some deal or discounts.
- Cleaning products are often purchased with other grocery items; therefore, these products should be sold at a discount or in deal bundle.

Limitations and future research

- The quantity of products within each transaction is unknown. This data could put great insights into which items are purchased in bulk at one point and which are bought individually over time. This could then be fed to marketing campaigns e.g. targeting items purchased in bulk.
- Future research should also discover specific purchasing patterns among the different groups found in cluster analysis.
- Dataset contained insufficient amount of postcode data. Future data should be provided with customers from vast range of areas.
- Online shopping patterns should be investigated separately to determine if shopping behaviour differ among different shopping methods.

APPENDICES

Appendix1

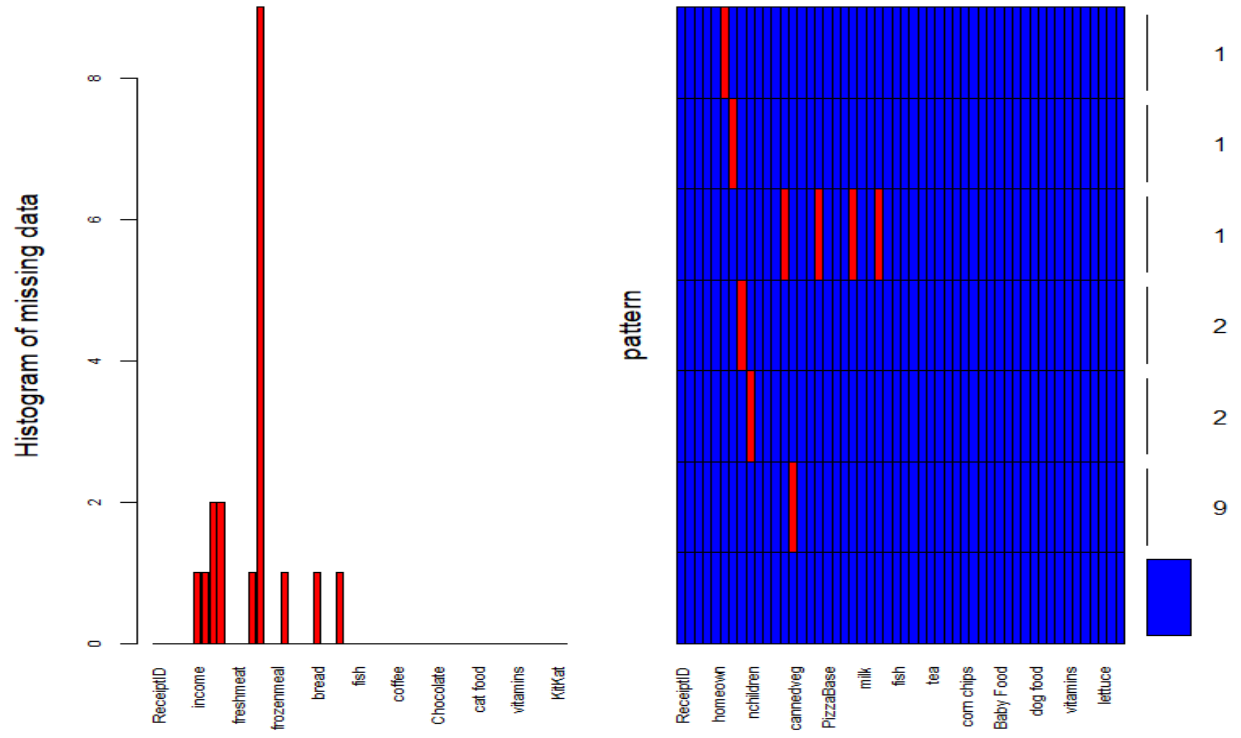

Figure1: Missing values pattern in dataset excluding postcode.

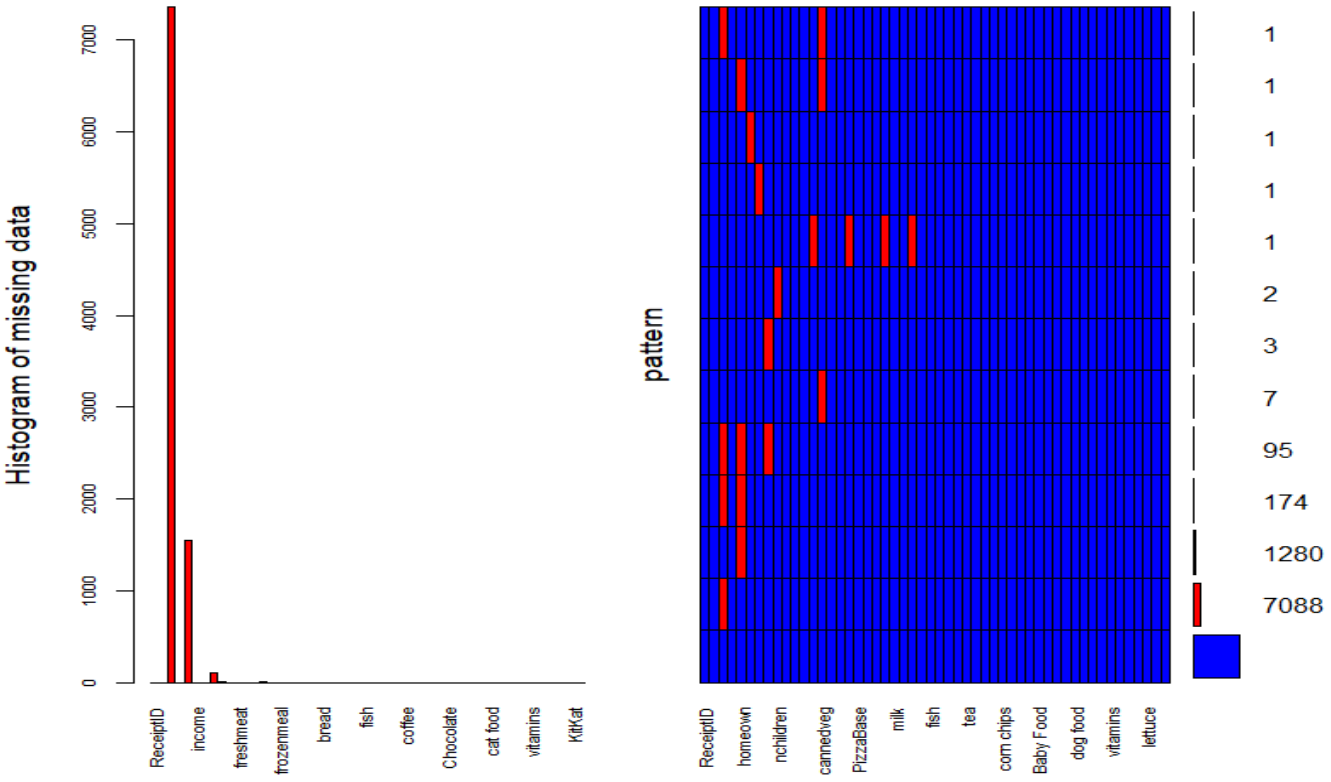

Figure2: Missing values pattern before imputing missing values.

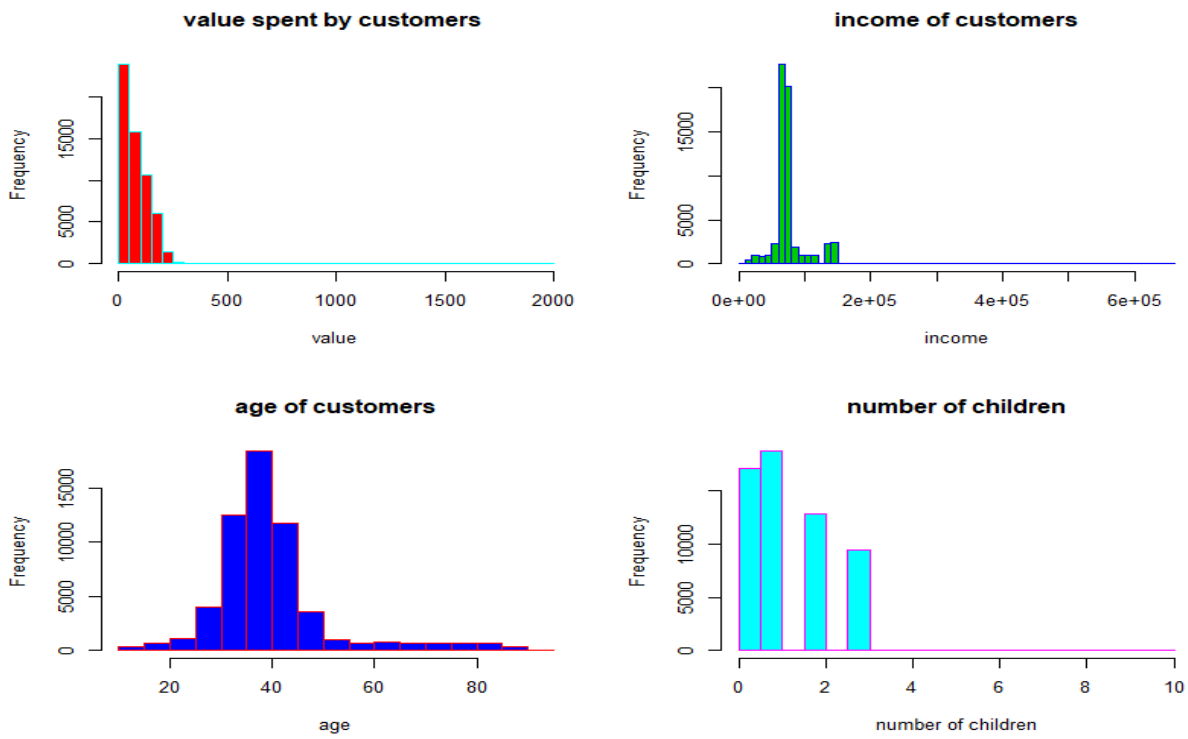

Figure3: Histograms of unadjusted continuous variables.

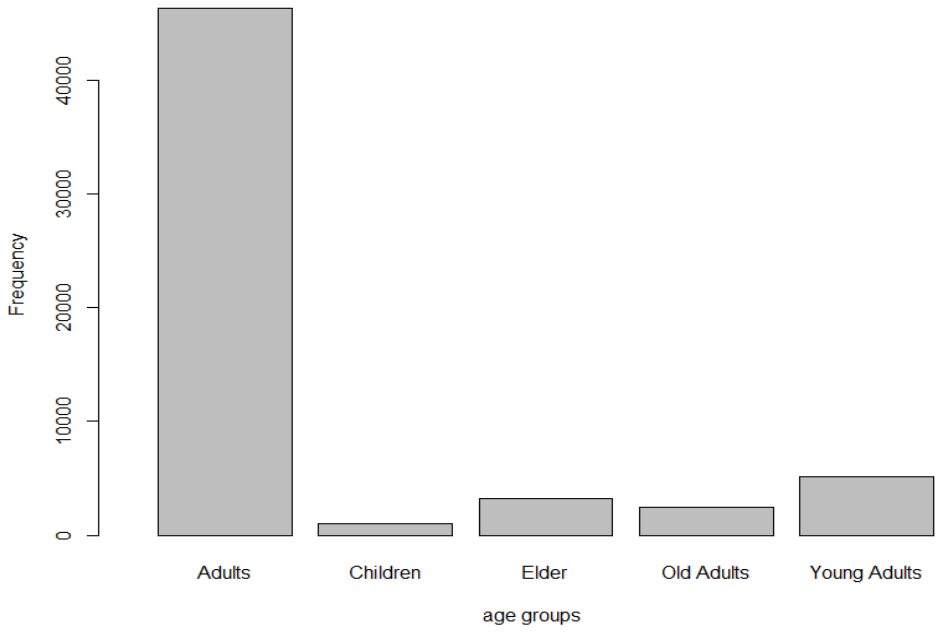

Figure a: Frequency of ages based on groups

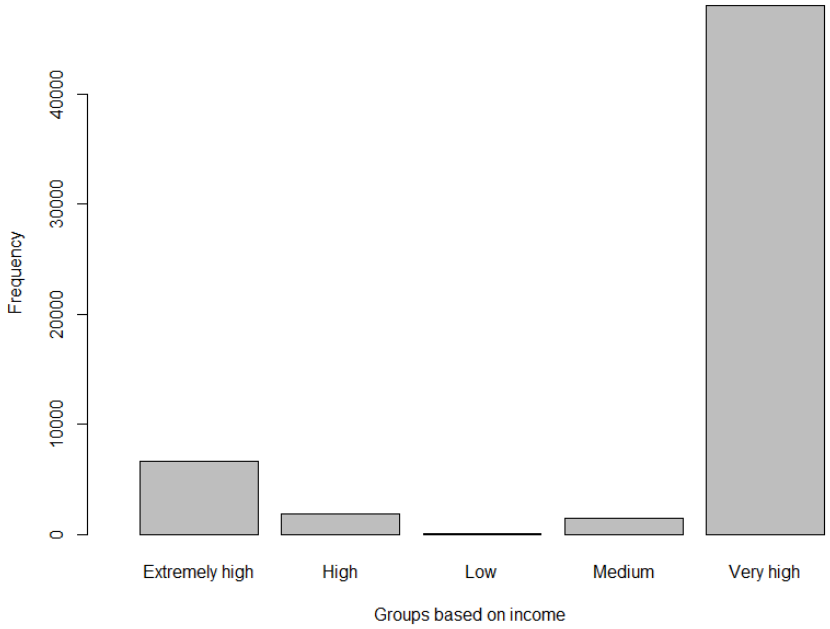

Figure b: Frequency of customers based on group of incomes.

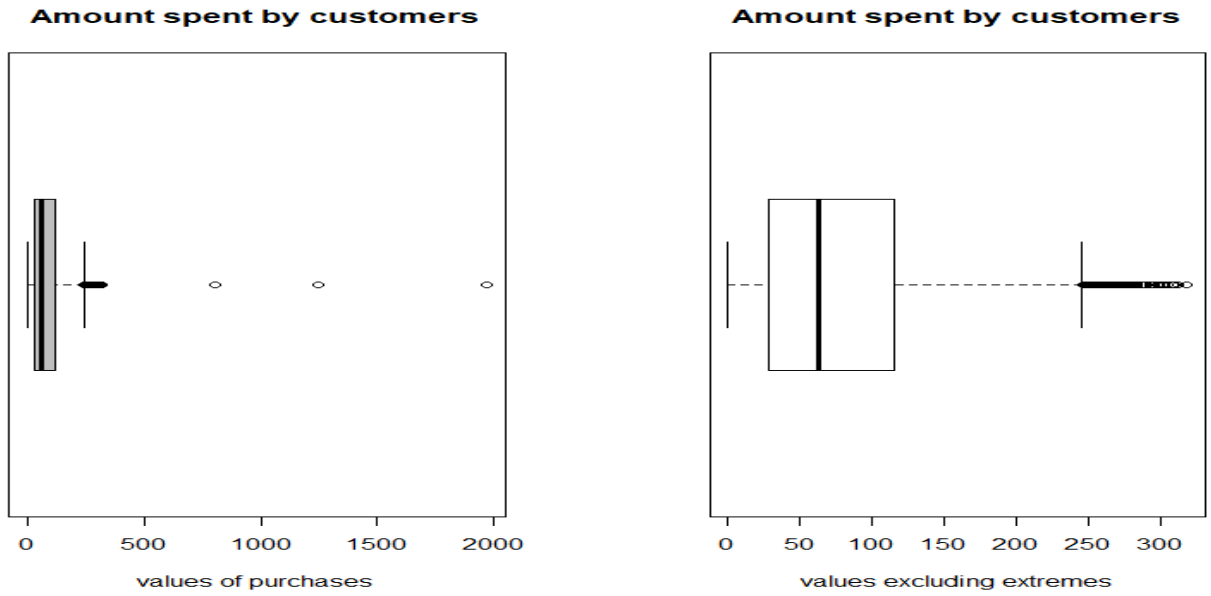

Figure4: Boxplot including Extreme values of purchases and without extremes.

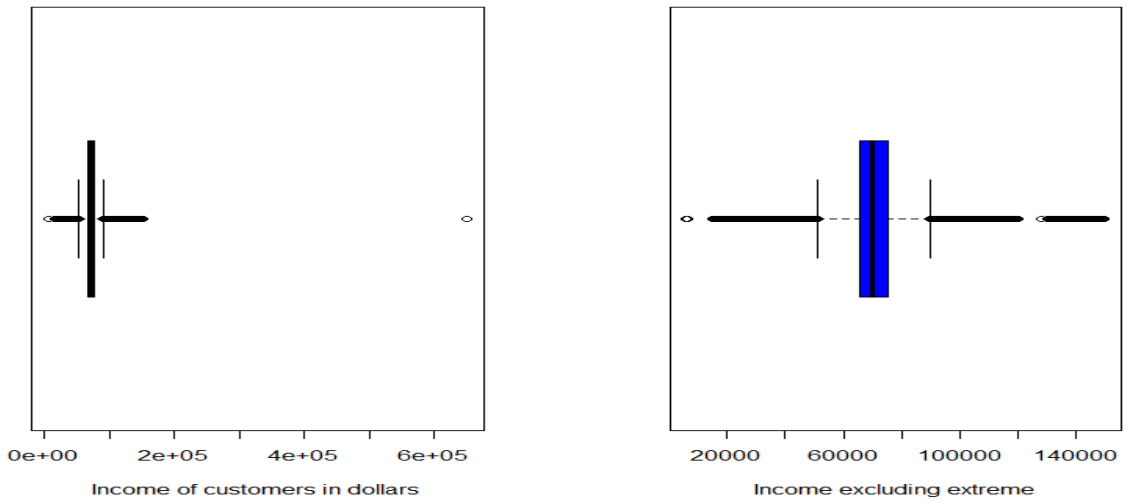

Figure5: Boxplot of income before and after excluding highest income.

Appendix2

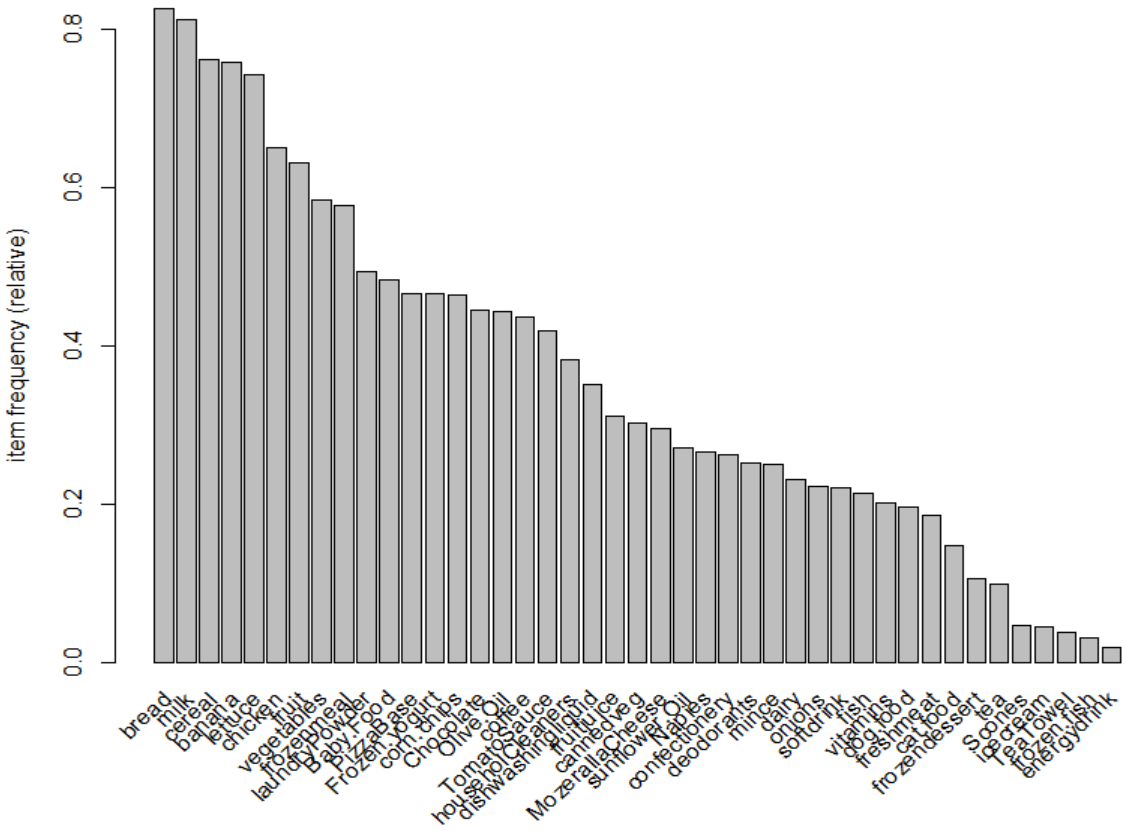

Figure6: Relative frequency plot of items

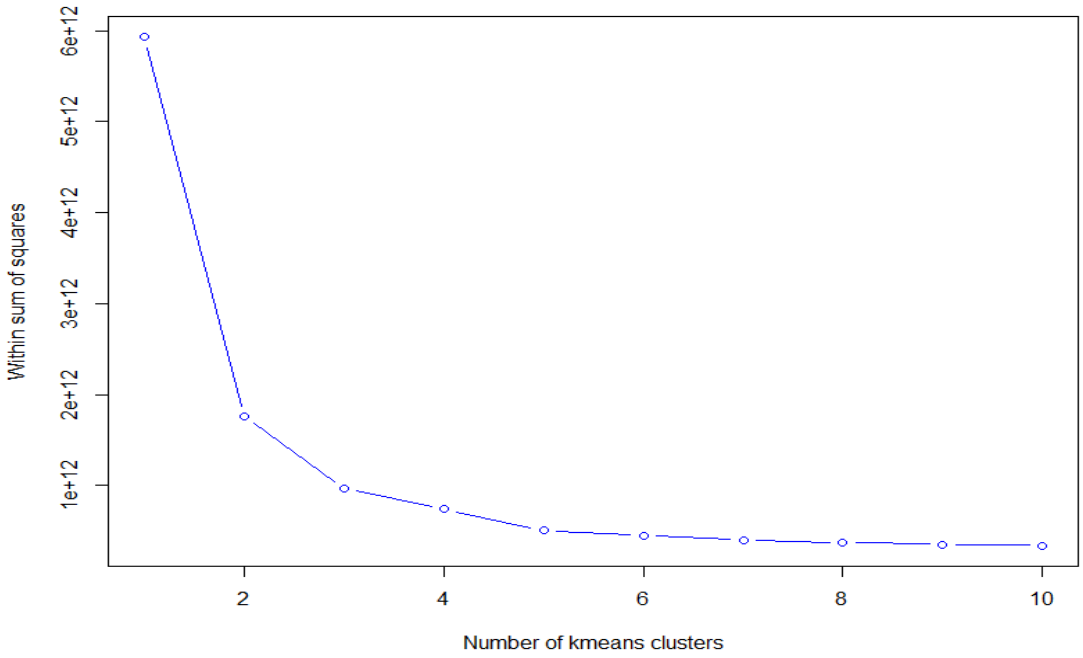

Figure7: Number of clusters formed

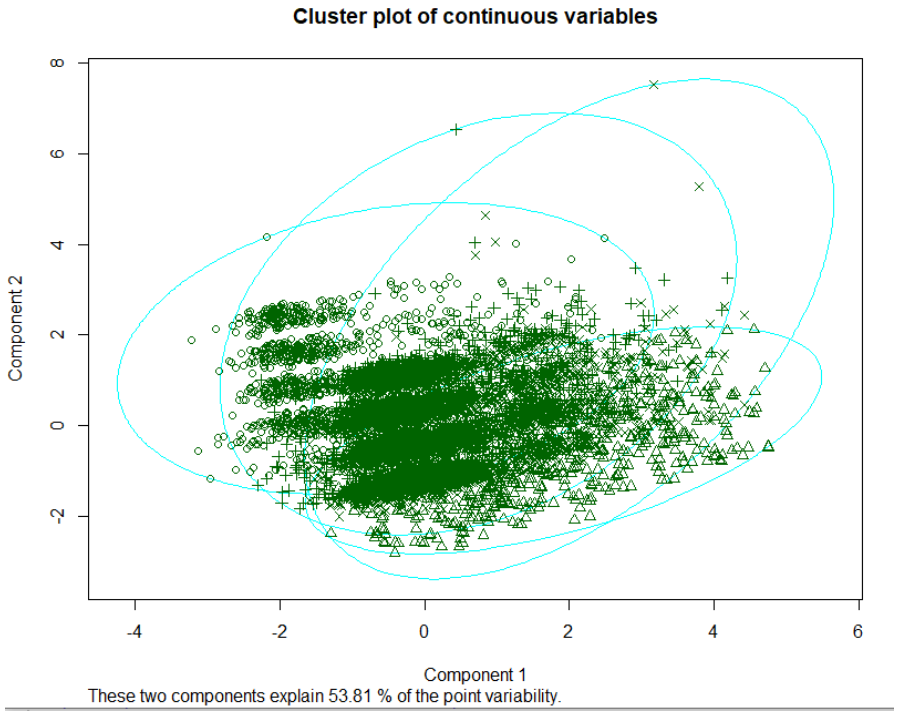

Figure8: Graphical representation of clusters formed.

Table1: Top 5 rules with highest confidence

	lhs		rhs	support	confidence	lift	count
[1]	{fruit,frozenmeal,TomatoSauce,vegetables}	=>	{banana}	0.1045285	0.9955738	1.311071	6073
[2]	{TomatoSauce,bread,vegetables,Olive.Oil}	=>	{banana}	0.1012926	0.9950964	1.310443	5885
[3]	{fruit,TomatoSauce,vegetables,coffee}	=>	{banana}	0.1019811	0.9947952	1.310046	5925
[4]	{TomatoSauce,bread,vegetables,Chocolate}	=>	{banana}	0.1035646	0.9940525	1.309068	6017
[5]	{TomatoSauce,vegetables,Olive.Oil}	=>	{banana}	0.1163531	0.9938253	1.308769	6760

Table2: Top 5 rules with highest support

	lhs		rhs	support	confidence	lift	count
[1]	{milk}	=>	{bread}	0.6724556	0.8270497	0.9998493	39069
[2]	{bread}	=>	{milk}	0.6724556	0.8129552	0.9998493	39069
[3]	{cereal}	=>	{bread}	0.6370161	0.8358553	1.0104947	37010
[4]	{banana}	=>	{bread}	0.6360006	0.8375493	1.0125427	36951
[5]	{cereal}	=>	{milk}	0.6207164	0.8144677	1.0017096	36063

Table3: Top 5 rules with highest lift

	lhs		rhs	support	confidence	lift	count
[1]	{fish,vegetables,banana}	=>	{householCleaners}	0.1013443	0.9182782	2.400929	5888
[2]	{fish,vegetables}	=>	{householCleaners}	0.1078332	0.8524969	2.228937	6265
[3]	{cereal,fish,banana}	=>	{householCleaners}	0.1017573	0.8379872	2.191000	5912
[4]	{bread,fish,banana}	=>	{householCleaners}	0.1085389	0.8328051	2.177451	6306
[5]	{milk,fish,banana}	=>	{householCleaners}	0.1049244	0.8183649	2.139696	6096

References

Berry, M. and Linoff, G. (2004). Data mining techniques. 2nd ed. Indianapolis: Wiley, Chapter 9-11.

Hastie, T. (2007). The elements of statistical learning. 3rd ed. New York: Springer, Chapter 14.

Han, J. and Kamber, M. (2006). Data mining. 2nd ed. Amsterdam: Elsevier.

https://www.news.com.au/finance/business/retail/unwanted-supermarket-chain-coles-faces-uncertain-future/news-story/045dfc7167bc6063b5f6f51a0ab6ad15