STAT828 DATA MINING – UNDIRECTED KNOWLEDGE DISCOVERY PROJECT
FRANCESCO PALERMO
Student ID: 45539669

1) Executive Summary

Strategy is essential to a successful organization and to make a successful strategy it's important to have as much information as possible. Coles' strategy is about "making life easier"[1] and through analysis of a random dataset of transactions from Coles' supermarkets valuable insight can be gained into the shopping habits of the supermarket's customers.

A market basket analysis has identified several different purchasing behaviours and this knowledge can be used to: improve store layout to encourage a higher spend per transaction, or target promotions and marketing to certain demographics and even day-parts. Association rules indicate that there are several items in the supermarkets that are often purchased together. Another approach, a clustering algorithm, has been applied to the dataset to organise the customers into groups. After applying these methods, Coles has the knowledge to be able to make life easier for their customers to encourage brand loyalty and stimulate sales growth.

For example: items frequently purchased together should be arranged close to each other in the supermarket, and items rarely purchased might be displayed near the cash register at a slightly discounted price to encourage an impulse buy.

2) Introduction

After applying several data mining techniques to the dataset, the following questions have been answered:

Which distinct customer groups can be found in the data set in terms of income, age and purchase value?

Which products are purchased the most and which products are purchased the least? Are there any patterns between purchases, for example, which products are often purchased together?

3) Description of the Data Set

a) Original Data

The simulated dataset, which was provided in an Excel format, was analysed using R. The dataset contains 58,088 customers (or rows) with 53 variables (or columns). Each customer is identified by their transaction ID (with a variable label: RecieptID) and using this variable an existing duplicates check has been applied to ensure that the data used does not contain multiple entries of the same transactions. There are 9 rows which contain the same ReceiptID, however in all other aspects they are different which means they refer to different customers. Therefore, it would not be beneficial to eliminate this data.

By changing the duplicates with values from ReceiptID 658101 to 658109 each customer now has a unique RecieptID.

The variables from the simulated Coles dataset can be divided into three categories.[2]

[1] "Our Strategy", Coles Group, https://www.colesgroup.com.au/about-us/?page=our-strategy

[2] The variable type shows the real nature of the variable. This may differ from the dataset type, for example; "pmethod" is a categorical variable, but in the dataset, it is an integer.

1. Customer demographics; these variables describe customer characteristics such as age, income, sex, etc.
2. Transaction data; these variables describe information regarding each purchase.
3. Product profile; these binary variables describe whether the customer buys the product or not.

Customer Demographics

Variable	Description	Type	Values
Sex	Customer Sex	Binary	1 = Male, 2 = Female
homeown	House Ownership	Categorical-nominal	1 = Yes, 2 = No, 3 = Unknown
income	Customer Income P/A	Numeric-continuous	Positive real number
age	Age in Years	Numeric-continuous	Positive integers
postcode	Postcode	Categorical-nominal	Four numeric characters
nchildren	Number of Children	Numeric-discrete	Positive integer

Transaction Data

Variable	Description	Type	Values
receiptID	Transaction identifier	Categorical-nominal	Unique ID
value	Value of transaction	Numeric-continuous	Positive real number
pmethod	Payment method	Categorical-nominal	1 = Cash, 2 = Credit card, 3 = Eftpos, 4 = Other

The other 44 flag variables are the product variables that display a binary “1” if the customer purchases that specific item, and “0” if they haven’t purchased.

b) Data Preprocessing

All the 53 variables have been analysed in order to find missing values, non-conformant data and outliers. Depending on the nature of the variable some values have been imputed or modified.

Customer demographics

“sex”: This value has been transformed from an integer to a character (0 for female, 1 for male). No missing values were found.

“homeown”: No missing values were found. These categorical variables could be separated into three categories of homeownership (1 = yes, 2 = no, and 3 = unknown). “Value 3” has not been treated as faulty data as it has been assumed that the customer purposefully did not share this information. The 99 values that exceed 4 have been substituted with the mode (1) which indicates homeownership.

"income": To begin with, the variable has been changed to numeric. One missing value was detected and replaced with the median of the distribution ($70169). There were many outliers found for both low and high income, however they are most likely true values as individual incomes can vary greatly and they were not altered (Appendix B).

"age": The variable was first changed to an integer – this way all decimal ages were rounded down to the nearest integer (23.2589 was rounded down to 23, for example). One missing value for the female customers was discovered and then imputed with the group median which was 38 years old. There were some extreme values in the group, but they were not so extreme as to be impossible, so nothing was done to them.

"PostCode": This variable had by far the most missing or incorrect data out of all the variables in the dataset. There were 10,365 missing values in total. Since Sydney's postcodes begin with 2000 and end with 2914[3], the values not in that range (around 1,000), have been replaced with N/A. Data is too expensive to justify cancelling a variable with 80% of its values being correct, so this variable was kept in the analysis.

"nChildren": There were no missing values. From the distribution frequency, the outcomes related to number of children decreased rapidly after 11 children. Any observations from 12 to 14 children were analysed to understand if it was appropriate data or not.

ReceiptId	Age	nChildren
658007	73	12
658008	54	13
658009	23	14

The first two rows in the table above can be realistic, so neither of them can be excluded, however the last row must be considered incorrect data. This value has been replaced with the median, as well as any other values greater than 14.

One final check was completed and in 167 instances, customers younger than 14 were recorded as having children. This was not realistic, so those values were substituted with zero.

Transaction data

"ReceiptID": This value has been briefly discussed in Section 3.1- no missing values were found. The only changes made were the replacement of the duplicates with the values from ReceiptID 658101 to 658109.

"Value": The first step was changing the character from factor to integer. No missing values were found, so the next step was addressing the 3 obvious outliers, values 802.1, 1243.0, 1967.7 (Appendix A). These values were not eliminated because they can be associated to true

[3] https://www.training.nsw.gov.au/about_us/postcodes_byregion.html

values (for example, large transactions could be purchased due to supplies for a large party or a last-minute function).

"Pmethod": No missing values have been found. "Value 4" hasn't been modified as that can be referred to as cheque. The 98 values that exceed 4 are considered non-conformant data and have been substituted with the mode of the distribution which is 2 (credit card).

Product profile

"Fruit": There were multiple examples of non-conformant data for this binary variable. Ten entries in the column had labels: 11, 3, 7, 6, 4, and 'o'. The value 'o' has been translated to '0' as it has been assumed that the 'o' was mistakenly input instead of the integer. In addition, any value greater than 1 has been substituted for 1 because another assumption has been made that the number of fruits purchased was input, rather than 1 or 0 (yes or no) as it should have been.

"fruitjuice": Only 10 examples of incorrect data were found to be labelled with 2. They were replaced with 1, for the same reason explained above.

c) Data Analysis

With a cleaner dataset, the question of "who are the customers?", can be answered with confidence. Graphs are usually the best method to display different characteristics in an intuitive way. First to be described is the customer's behaviour, followed by transaction behaviour. Finally, two tables will display the most and least common purchased products in the dataset.

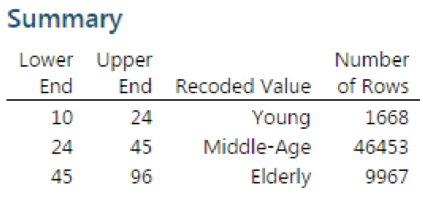
Customer Demographics

- Out of the total sample, 60 percent are female; as shown in Appendix C.
- In terms of homeownership, 2.5 percent of customers declined to share this information. Around 25 percent of customers stated they did not own homes, and the majority at 72 percent are homeowners (Appendix D).
- The majority of annual incomes in the dataset fall between $57000 and $80000. Half of the customers in the sample are earning, at most, an annual salary of $70169 (the median salary). According to the graph, it is possible to see two groups of customers: one group seems normally distributed around $70,000 and the other group is uniformly distributed around the mean $140,000 (Appendix E).
- The mean of customer age (which ranges from the youngest of 10 years to the oldest at 95 years old) is around 40 years old. The age distribution is skewed to the right, or older ages (Appendix F)
- Most of the customers in the dataset live in the postcode 2122 which includes the suburbs of Sydney, Eastwood and Marsfield. The region is known as Central and Northern Sydney.
- Customers had a number of children that varied from zero up to thirteen. Out of the customers who had children, the majority at 32.1 percent had only one child, followed by those who had two children at 21.9 percent and finally customers with three children at

16.2 percent. There were 29.7 percent of customers who stated they had no children at all (Appendix G)

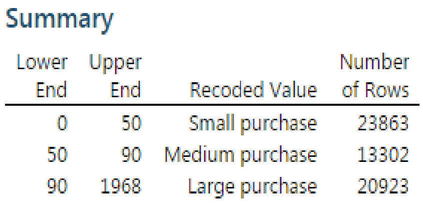
Transaction Data

For the purpose of this analysis, the following values have been chosen to represent the different categories.



Summary

Lower End	Upper End	Recoded Value	Number of Rows
10	24	Young	1668
24	45	Middle-Age	46453
45	96	Elderly	9967



Summary

Lower End	Upper End	Recoded Value	Number of Rows
0	50	Small purchase	23863
50	90	Medium purchase	13302
90	1968	Large purchase	20923

- The most common payments methods are Credit Card (42.71%), and Eftpos (30.45%) with together account for over 70 percent of payment methods. Customers paid with cash only 14 percent of the time (Appendix H).

- The graph illustrating transaction value by frequency has been created by excluding the outliers to get a better picture of the more common purchase values. Half of the customers spend $64 or less per transaction. Only 25 percent of customers spend more than $116 dollars. In addition, the value seems to be following a Fisher-Snedecor distribution (Appendix I)

- In general, female customers spent more than males. In fact, the median spend of female customers is $68 whereas the median spend of the males is $57 (Appendix J).

- The next graph shows the interesting phenomenon that the largest demographic of customers (the middle – aged) spend most of their money on small purchases of $50 or less. Elderly customers and young customers are more likely to spend their money on large purchases of $90 or more (Appendix K).

Product Profile

The two graphs below answer the question "Which products are most and least frequently purchased?" clearly.

Most purchased products

Item	Times purchased	Proportion Of basket
Bread	48051	82.7%
Milk	47228	81.3%
Cereal	44277	76.2%
Banana	44113	65.9%

Least purchased products

Item	Times purchased	Proportion Of basket
KitKat	952	1.64%
Energy drink	1112	1.9%
Frozen fish	1753	3.02%
Tea towel	2156	3.71%

The summary of product frequency is illustrated in Appendix L.

4) Methodology

a) Market basket analysis by association rule mining

Market basket analysis uses data about what customers have purchased to know more about who they are and why they purchase certain items. Market basket analysis demonstrates which products tend to be bought together and which are would be best for promotions.

The data mining technique used is Association rules. Association rules are used to discover frequent patterns in data sets. Association rules are useful because of the clarity and utility of the outcomes, which are in the form of rules about groups of products.

Market basket analysis has been applied on the dataset via SPSS Modeler with the apriori algorithm, which then selected the 44 binary variables; the product variables.

The threshold parameter for rule support was set at 10% because the number of transactions is large. This value guarantees that the itemsets included in the analysis occur at least 5808 times in the dataset.

Maximum rule confidence was set to 0.85%, in order to capture strong associations between antecedents and their consequents. The maximum number of antecedents has been set at 3.

b) Clustering for customer segmentation

In order to find different groups based on the customer demographics within Coles customers, cluster analysis was used. The method applied is k-means.

The k-means algorithm can be divided into 3 different steps. In the first step, the algorithm arbitrarily selects K data points to be the seeds. The next step gives each record to the closest seed. In other words, the algorithm measures the Euclidian distance of each row to each seed and chooses the minimum distance for this step. The final step computes the centroids of the clusters. The centroids will then be the seeds for the next iteration of the algorithm. Each step is repeated until the cluster boundaries stop changing.

This method sorts the items according to how alike and dislike they are to the other clusters. The characteristics of K-means are that it is only suitable with the use of continuous variables and the number of clusters needs to be decided at the beginning of the algorithm.

K-means was a more appropriate choice for this larger dataset than hierarchical clustering which only works with smaller datasets. The variables used were age, value and income since it was discovered that they were the strongest predictors (for example, number of children would have not added anything relevant to the clusters). The variables have been normalised in order to be in the same measurable scale. Finally, the number of clusters chosen to be 3 because it gave a good separation of customers in each cluster, something that 4 clusters did not do.

5) Results

a) Association Rules

The results obtained after running the apriori algorithm are summarised below. To begin with, 1159 rules were discovered. These rules have then been narrowed down to those with the highest rule support, confidence and lift (Appendixes M, N and O).

Rule 1.

{fruit + cereal} → {bread}

This rule indicates that approximately 41.7% of all Coles customers purchase both fruit, cereal and bread. Roughly 85% who bought fruit and cereal also bought bread. A lift of 1.02 shows the combination of products occurs no more than expected.

Rule 2.

{vegetables + lettuce} → {banana}

This rule specifies that approximately 39.5% of all Coles customers purchase both vegetables, lettuce and bananas. Around 9% (1-0.9) of customers who purchased vegetables and lettuce did not buy bananas – to reduce this number it would be beneficial to target online customers with a suggestion to purchase bananas after they buy vegetables and lettuce. With a lift of 1.1, this combination of products occurs 10% more than usually expected.

Rule 3.

{tomato sauce + chocolate + vegetables} → {banana}

Almost everyone (99%) who bought tomato sauce, chocolate and vegetables also purchased bananas. Approximately 11.7% of all Coles shoppers who purchased tomato sauce, chocolate and vegetables, also purchased bananas. This combination of items happened 30% more than expected.

Rule 4.

{fish + laundry powder} → {household cleaners}

This rule implies that this combination of products occurred more than twice as often as expected (with a lift of 2.2). Approximately 14% (1-0.86) of customers who purchased both fish and laundry powder neglected to buy household cleaners. Since these items are not usually near each other, proximity would increase the likelihood of purchase.

Rule 5.

{nappies + tomato sauce} → {baby food}

This rule indicates that 10.2% of all Coles customers purchase nappies, tomato sauce and baby food. Around 95% who bought nappies and tomato sauce also bought baby food. A lift of 1.9 means this combination of products occurred 90% more than expected.

b) Cluster Analysis Results

Income and age have been the biggest predictors of importance in the k-means algorithm as illustrated in Appendix P. The size of the larger cluster contains 47537 customers, which is 81.8% of the whole. The smallest cluster contains 4634 customers which is only 8 percent of the whole (Appendix Q). The clusters meaning and description are now going to be analysed.

Cluster 1; Middle – aged wealthy professionals

This 10.2% of customers refers to the people who earn the highest incomes. This group does not spend the most nor the least on grocery shopping – their purchase habits lay somewhat in the middle of Coles customers purchase behaviour.

Cluster 2; Elderly retirees

This cluster refers to the elderly customers in the dataset, who spend on average, more than any other customer on shopping. Their income, which most likely is a pension, is lower than the mean.

Cluster 3; Middle – aged working class

This 81% refers to middle-aged customers whose salaries are slightly below the population average. Their shopping habits fall in the middle of Coles customers average purchase per transaction.

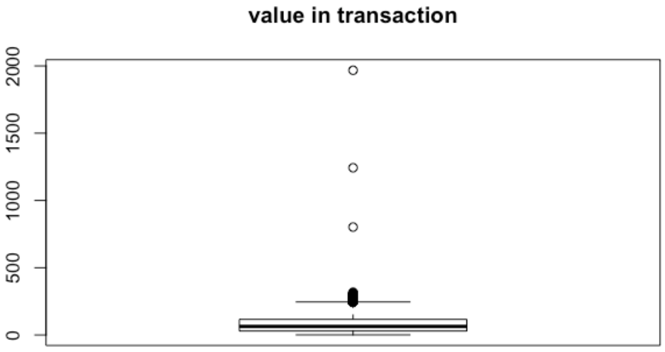
Cluster summaries and a comparative bar chart are listed below in Appendix R and S.

6. Conclusion

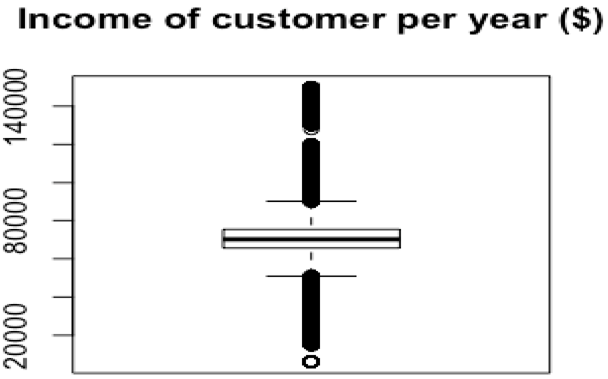
After cleansing and analysing the dataset, the following suggestions might be implemented to drive sales growth and improve customer service (and therefore loyalty). Ensuring the store layout is optimised to capture more sales is essential. Placing the least purchased items under a discounted price before the register might encourage impulsive shoppers. Items purchased together often should be kept near each other for convenience and having targeted promotions would be a good tactic as well. The key takeaway is that the store must be strategically organised and knowing which items get purchased the most, or together, as has been laid out in this report, will help organise that strategy. It is also clear that the largest group of customers (middle aged) are purchasing less at a time, so it will be important to ensure stores suit these preferences by having enough self-checkout registers for the large number of customers making smaller purchases of $50 or less. Convenience is very important so the easier and quicker it is to pay the better.

With the dataset being simulated it is difficult to be sure the results are realistic. Another obstacle was the incomplete postcode variable which could provide valuable insight into purchasing patterns in the various suburbs – the CBD versus regional suburbs for example. It was difficult to organise the customers into specific demographics other than age, income, etc, so a broader dataset with more variables might yield more detailed results. More variables would capture more hidden rules and purchasing behaviours and lead to stronger strategies to lead Coles to success.

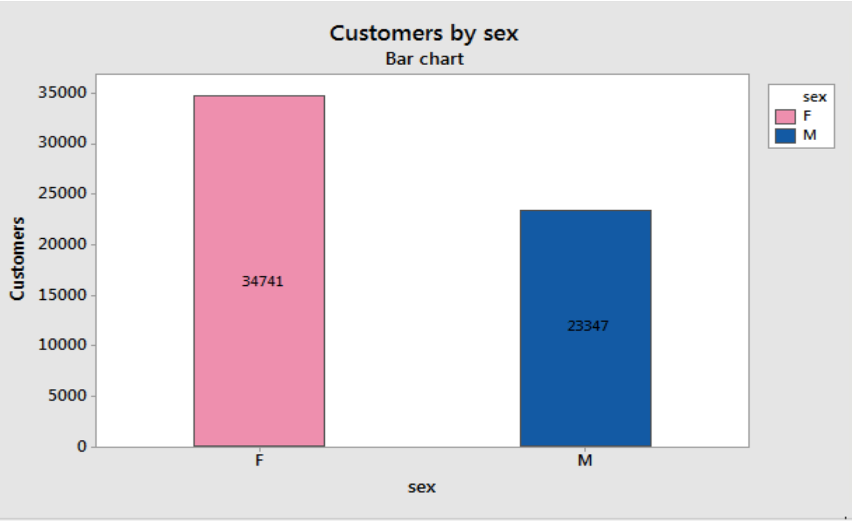
7) Appendix



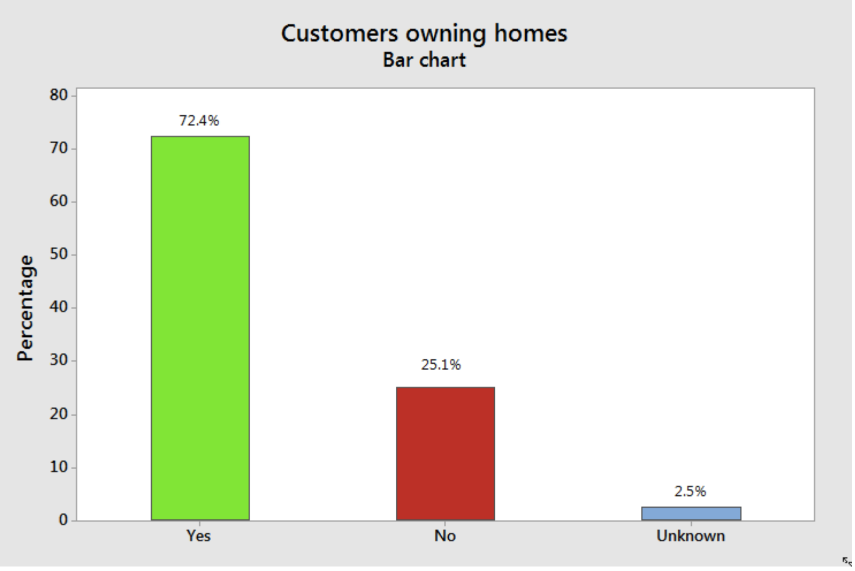
A.



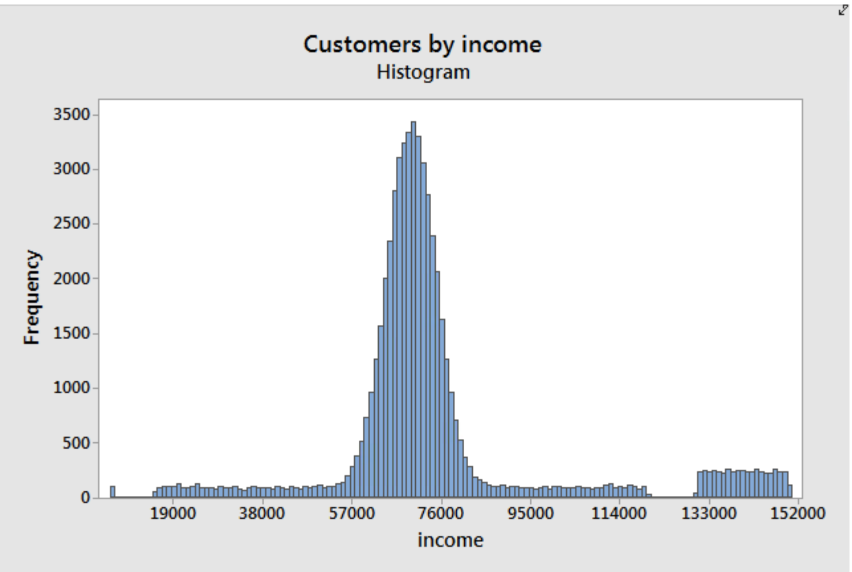
B.


Customers by sex
Bar chart
Customers
35000
30000
25000
20000
15000
10000
5000
0
34741
23347
F
M
sex
sex
F
M

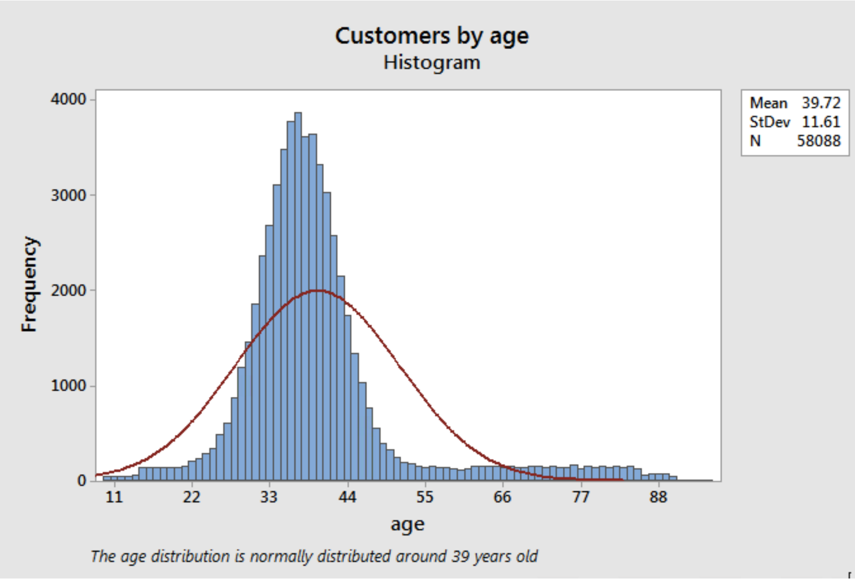
C.


Customers owning homes
Bar chart
Percentage
80
70
60
50
40
30
20
10
0
72.4%
25.1%
2.5%
Yes
No
Unknown

D.

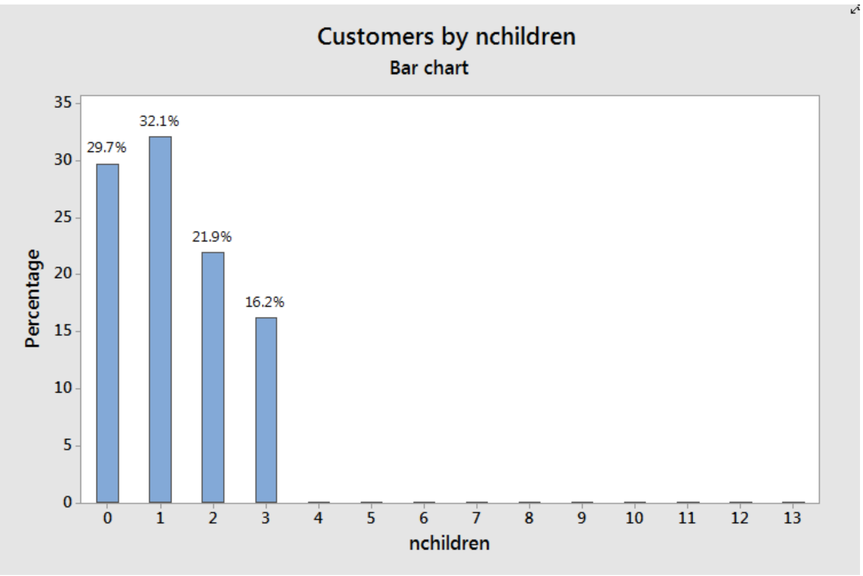


E.

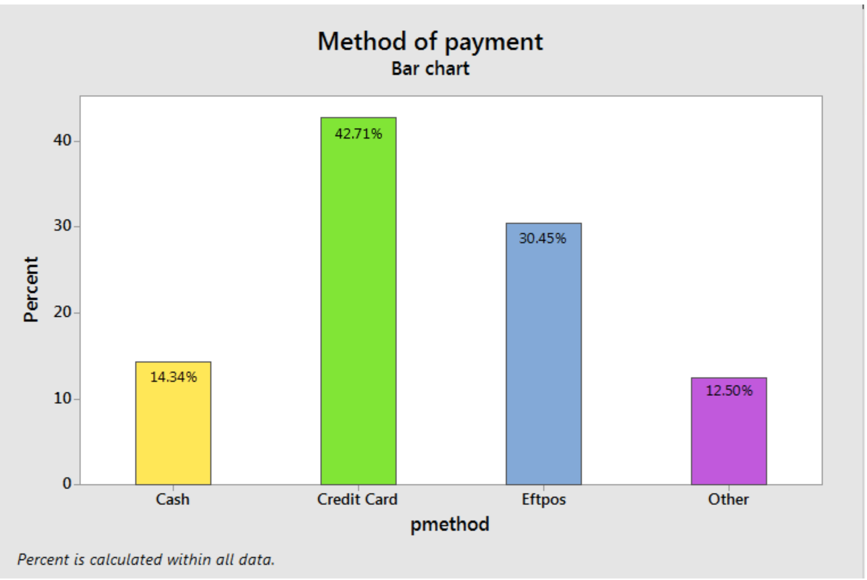


The age distribution is normally distributed around 39 years old

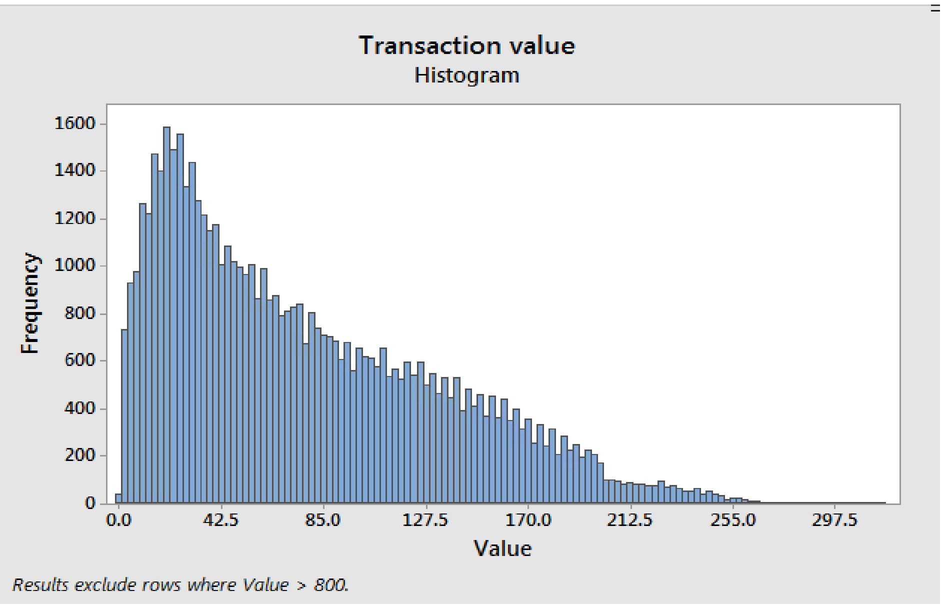
F.


Customers by nchildren
Bar chart
Percentage
35
30
25
20
15
10
5
0
29.7%
32.1%
21.9%
16.2%
0
1
2
3
4
5
6
7
8
9
10
11
12
13
nchildren

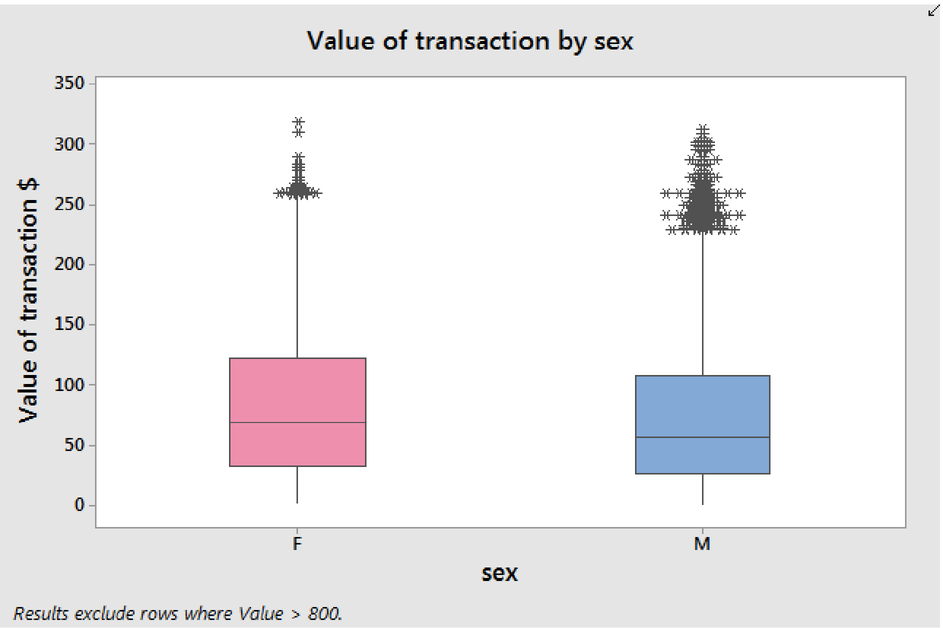
G.


Method of payment
Bar chart
Percent
40
30
20
10
0
14.34%
42.71%
30.45%
12.50%
Cash
Credit Card
Eftpos
Other
pmethod
Percent is calculated within all data.

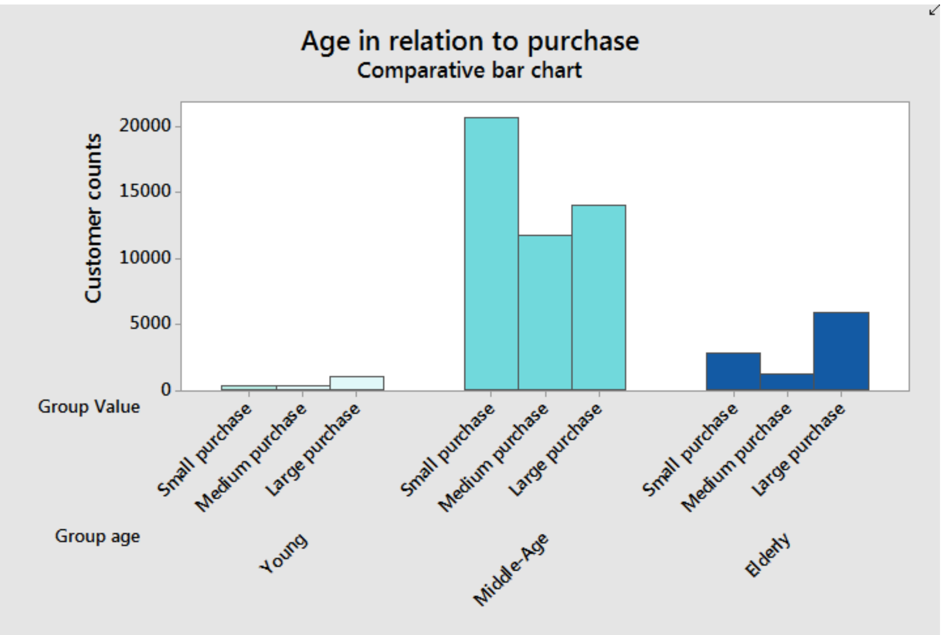
H.


Transaction value
Histogram
Frequency
1600
1400
1200
1000
800
600
400
200
0
0.0
42.5
85.0
127.5
170.0
212.5
255.0
297.5
Value
Results exclude rows where Value > 800.

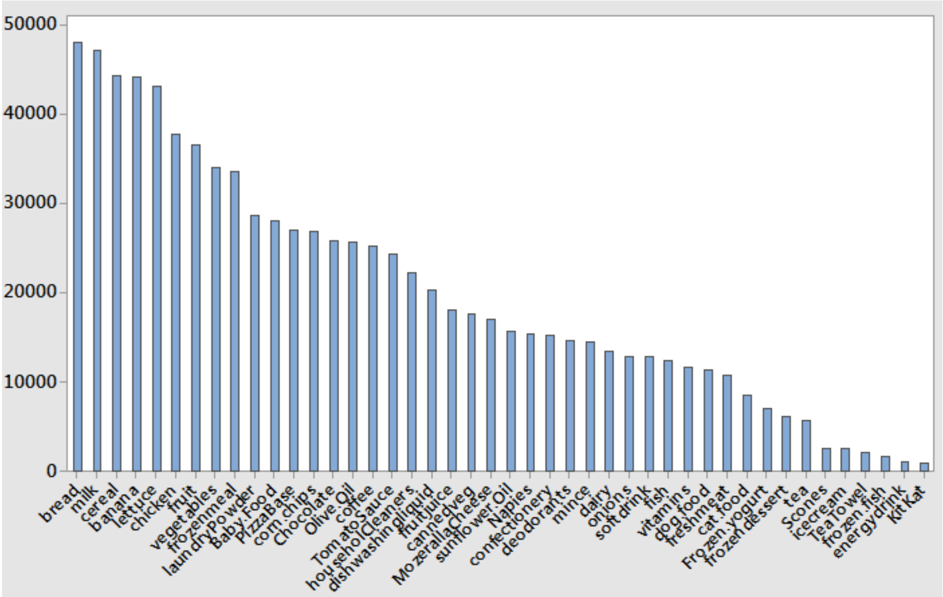
I.


Value of transaction by sex
Value of transaction $
350
300
250
200
150
100
50
0
F
M
sex
Results exclude rows where Value > 800.

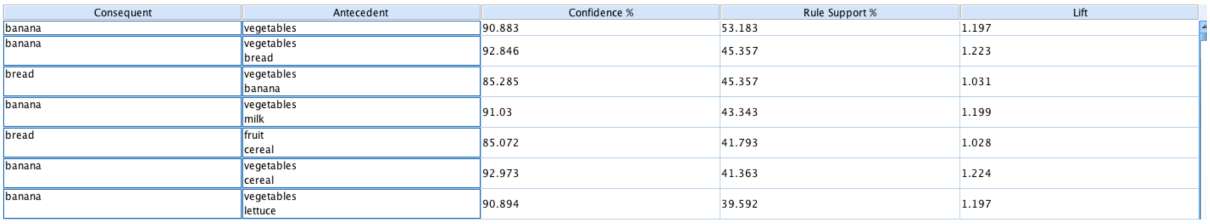
J.



K.

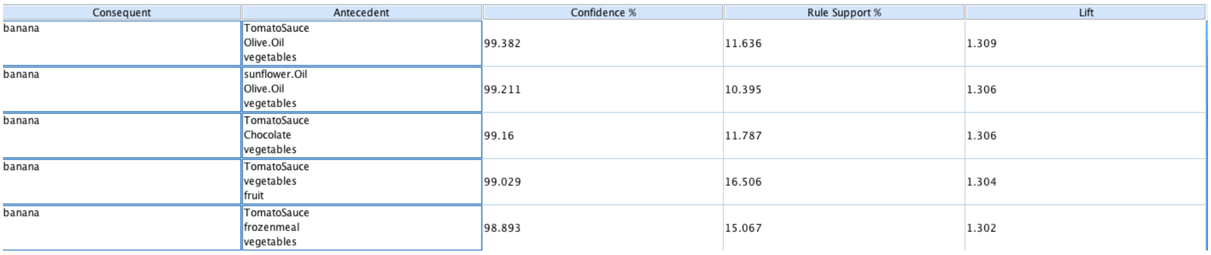


L.



Consequent	Antecedent	Confidence %	Rule Support %	Lift
banana	vegetables	90.883	53.183	1.197
banana	vegetables bread	92.846	45.357	1.223
bread	vegetables banana	85.285	45.357	1.031
banana	vegetables milk	91.03	43.343	1.199
bread	fruit cereal	85.072	41.793	1.028
banana	vegetables cereal	92.973	41.363	1.224
banana	vegetables lettuce	90.894	39.592	1.197

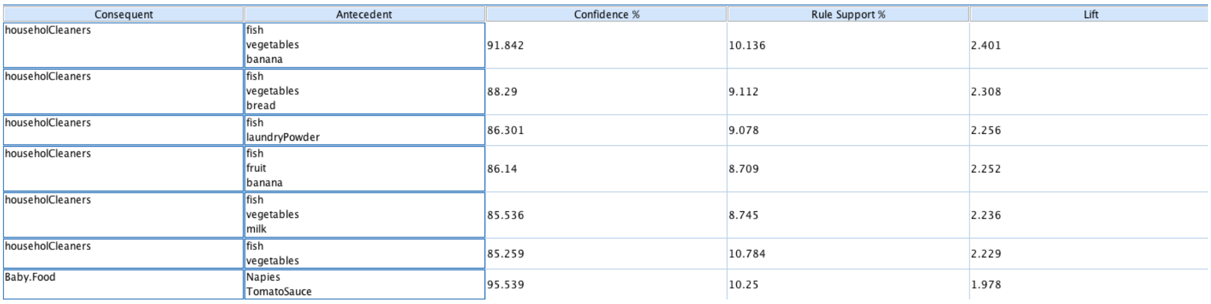
M.



Consequent	Antecedent	Confidence %	Rule Support %	Lift
banana	TomatoSauce Olive.Oil vegetables	99.382	11.636	1.309
banana	sunflower.Oil Olive.Oil vegetables	99.211	10.395	1.306
banana	TomatoSauce Chocolate vegetables	99.16	11.787	1.306
banana	TomatoSauce vegetables fruit	99.029	16.506	1.304
banana	TomatoSauce frozenmeal vegetables	98.893	15.067	1.302

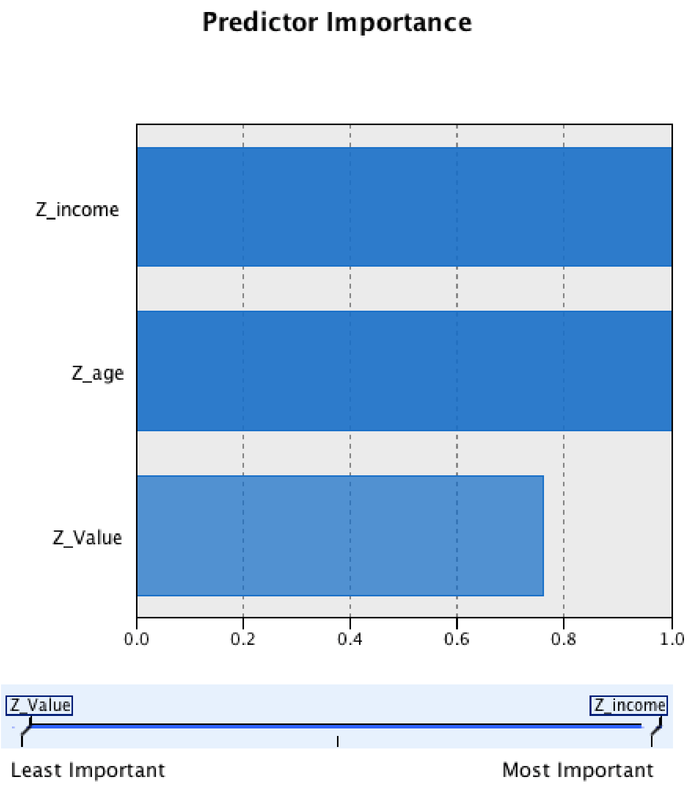
N.

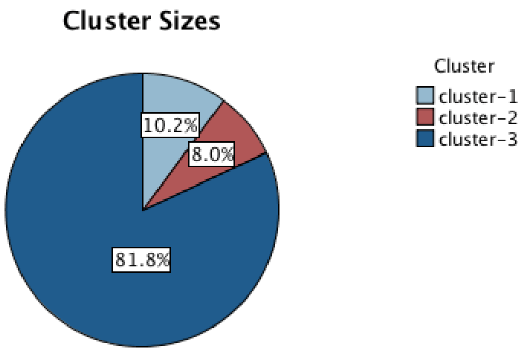
O.

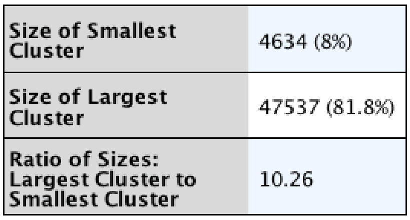


Consequent	Antecedent	Confidence %	Rule Support %	Lift
householCleaners	fish vegetables banana	91.842	10.136	2.401
householCleaners	fish vegetables bread	88.29	9.112	2.308
householCleaners	fish laundryPowder	86.301	9.078	2.256
householCleaners	fish fruit banana	86.14	8.709	2.252
householCleaners	fish vegetables milk	85.536	8.745	2.236
householCleaners	fish vegetables	85.259	10.784	2.229
Baby.Food	Napies TomatoSauce	95.539	10.25	1.978

P.

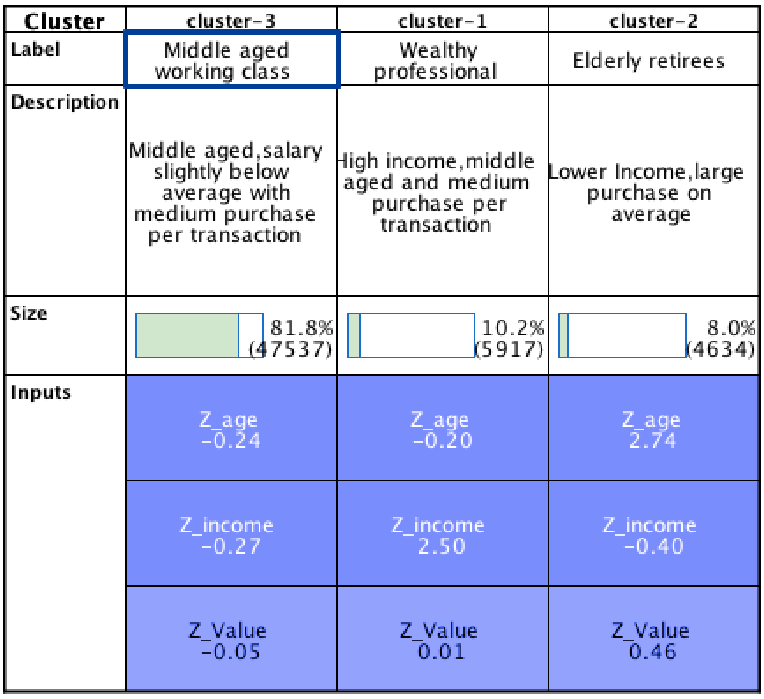






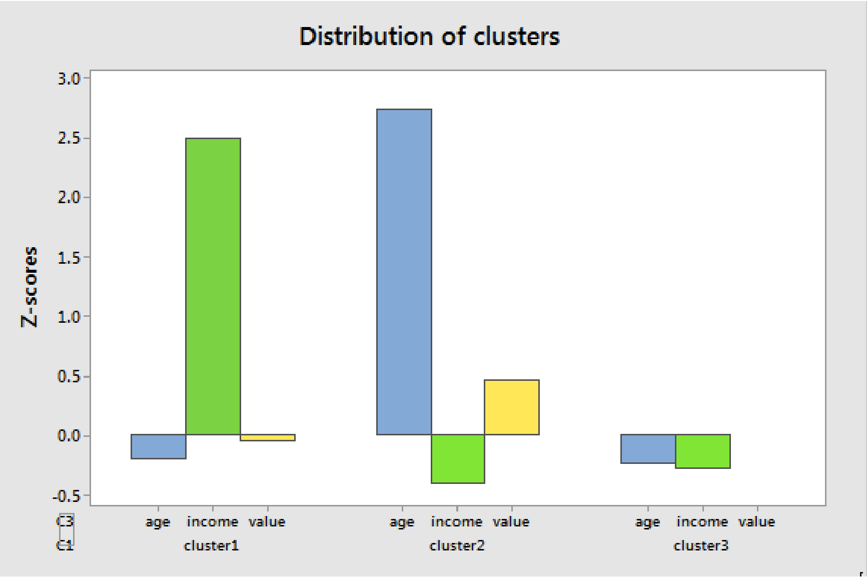
Size of Smallest Cluster	4634 (8%)
Size of Largest Cluster	47537 (81.8%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	10.26

Q.



Cluster	cluster-3	cluster-1	cluster-2
Label	Middle aged working class	Wealthy professional	Elderly retirees
Description	Middle aged,salary slightly below average with medium purchase per transaction	High income,middle aged and medium purchase per transaction	Lower Income,large purchase on average
Size	81.8% (47537)	10.2% (5917)	8.0% (4634)
Inputs	Z_age -0.24	Z_age -0.20	Z_age 2.74
	Z_income -0.27	Z_income 2.50	Z_income -0.40
	Z_Value -0.05	Z_Value 0.01	Z_Value 0.46

R.



S.

References

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

Feedback

I struggled the most with the description of the dataset. I didn't discuss my difficulties with my peers at school. A better knowledge of R would have helped me with this project. Yes, although the project took me a long time, I felt I had enough time to complete the project. The reading materials have really helped with this project. The theoretical lessons I have learned throughout the course helped me as well. R and SPSS are some of the most commonly used software. My skills in both were limited at the beginning of this subject, now I feel much more confident about them. This was the first time I have learned about Market Basket Analysis and there are any ways I can apply this knowledge in the future.