1. EXECUTIVE SUMMARY:

With the increasing supermarket in Australia, Coles supermarket has got a firm competition with its peers like Woolworths, Aldi etc. In order to increase the sales and maintain the leadership value of the Coles a study have been conducted. Different methodologies have been used in the study namely Market basket analysis and Clustering. By both of these methods different patterns and outcomes have been generated which will help the business indeed. A precise study has been conducted by following a step by step procedure. Initially whole data is preprocessed by replacing the outliers and missing values using various techniques, once a healthy data is obtained methodologies like Market basket analysis and clustering has been applied to extract meaningful output from the data which can be used for increase of business. Different patterns of product purchase have been found out based on generating rules using apriori algorithm, even customer segmentation is done. Data analytics has been done on the preprocessed data, like finding out the most shopped item, from which area (post code) shoppers are more, which age group people are doing more shopping etc., In a single sentence it can be said as a Study to increase the business of the Coles.

2. INTRODUCTION:

Despite being one of the leader in supermarket industry, Coles is facing a stiff competition with other peer units. In order to maintain its brands value and increase its sales, this study is being conducted.

This study consists of analysis of a Coles supermarket data set, the end results of this study will help the business of the supermarket.

From this study, we will try to figure out following questions.

1. Which is the most purchased product and which is the least?
2. Which group of people do more shopping?
3. What is the average income of customers who do shopping in coles?
4. Which locality (Postcode) people will do more shopping?
5. What kind of products customers are purchasing? Especially what is the pattern of the purchase?
6. How many types of customers are classified?

3. DESCRIPTION OF DATA SET:

The Data set provided contains 58,100 rows (transactions) and 53 Columns (Variables). Out of 58,100*53 values, 9906 were missing values (NA's) which were handling accordingly as explained in the following description. Apart from this one more column is added in order to mention the cluster classification for different transactions. As a whole, Variables can be split into two classes. One is transactional variable, purchase variable. Transactional variable contains all the information about the customer like Receipt ID, Postcode, Income, Sex etc., whereas the purchase variable consists of all the products that are kept in selling. Find Table 7.1 for the description of variables.

Data Preprocessing and Data Analysis:

Data Preprocessing is done using different methods for different variables, Below is the description of methods for each variable.

Receipt ID:

Receipt ID has to be unique as it is the way to distinguish customers. However, 9 values were found to be duplicate out of 58,100. As it's a small number and this variable is not much used in the study, no action has been taken.

Value:

No missing values (NA's) are found in this variable. But there were 3 outliers (Extreme Values), which were replaced by mean (77.2861). Find the figures 7.1 and 7.2 for the description of Value variable before and after preprocessing.

Pmethod:

Pmethod is a categorical variable and values should be between 1~4 as per description, But few wrong entries were present. 97 observations were having values ranging from 5~101. All the observations were replaced by the mode value of the variable. Find the figures 7.3 and 7.4 for graph description of the data before and after preprocessing. Most of the payments are done by credit card (24817 transactions) which account to about 42.7% of total Payments.

Sex:

No missing values, error data or outliers were found in this variable. Female count is more than female count among customers. Female count is 34747, which accounts to 59.8% of total customers. Find the figure 7.5.

Homeown:

Homeown variable is a categorical data and its value should be between 1~3 according to description. But there are values ranging from 3 ~102 which are erroneous. These values are replaced by the mode count of this variable. Most of the customers are homeown, there are 42069 customers who are homeown, which account to about 72.4% of the total. Find the figures 7.6 and 7.7.

Income:

There was one missing value among the income variable, which was replace by the mean value ($74838). Most of the customers annual income is in the range of $50000~$100000. There are around 50,000 customers in this range. Find the figure 7.8.

Age:

There was one missing value in this variable which is replaced by mean of all the ages of customers (39.72). Middle aged people are more among the customers. 30~40 age group

stands on the top followed by 40~50. 30~40 age group comprises among the 30,000 transactions. Find the figure 7.9.

Postcode:

Postcode variable contains the most number of missing values, which is 9887. A decision of replacing this values with the mode has been considered. Thought of predicting the postcode based on previous values was not considered, as the predictive value may lead to a unexisting postcode value. Although extreme values are found in the postcode, they were not deleted as the products related to all the post code are considered for the study. Extreme values of post code less than 100 and greater than 9999 was replaced by the mode value. Most of the customers are from the postcode of “2122”. Find the figures 7.10,7.11.

nChildren:

There are 2 missing values in this variable, both of them has been replaced by the mode value. Extreme values, count greater than 5 has been replaced by mode value. Number of customers with only 1 child is most. They account to about 20,000 in count. Find the figures 7.12, 7.13, 7.14.

Products:

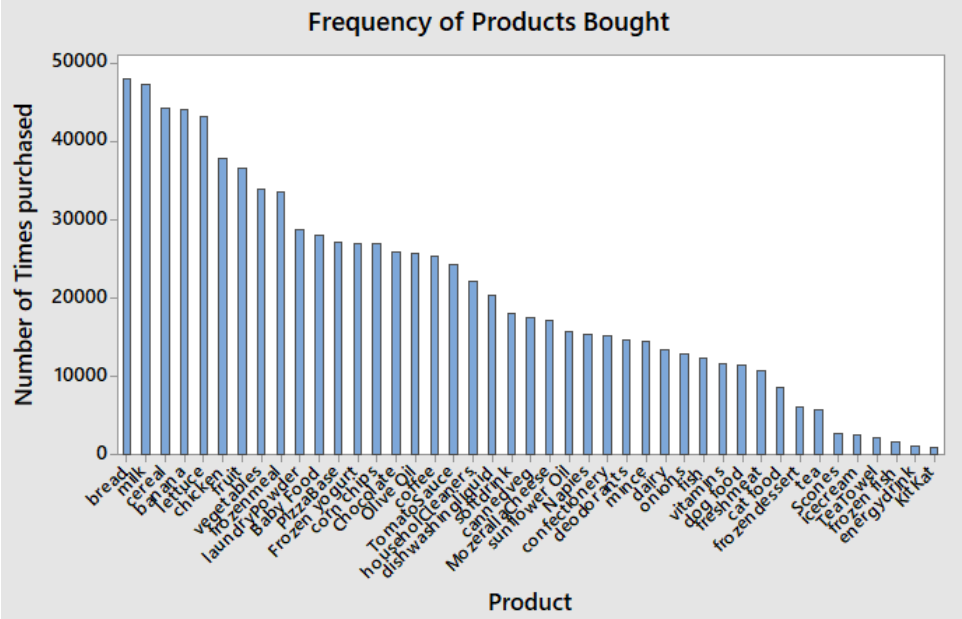
Fruits, Cannedveg, Cereal, Pizzabase, milk, confectionary:

These products contains some missing values which are replaced by 0.

Fruits, Fruitjuice:

These variables contain values other than 0 and 1. Those values are replaced by mode value.

Following is the summary of products with its transactions. Bread has the most number of transactions (48059) and kitkat has the least transaction (953). Bread accounts to 82.7% of transactions, which is the highest. Find the table, Table 7.2



4. METHODOLOGY:

Two types of methodologies are considered in the analysis of the dataset.

1. Market Basket Analysis
2. Clustering

Market Basket Analysis:

This analysis is basically performed to find out the patterns among the purchase of the products. It is a modelling technique based upon the theory that if you buy a certain group of items, you are more likely to buy another group of items.

Terminologies:

The rules in the market basket analysis can be written as ==> IF {A} THEN {B}.

The IF part of the rule is known as the antecedent and the THEN part of the rule is known as the consequent. The antecedent is the condition and the consequent is the result. The association rule has three measures that express the degree of confidence in the rule, Support, Confidence, and Lift.

Support:

Support is the number of transactions that include items in the {A} and {B} parts of the rule as a percentage of the total number of transactions. It is a measure of how frequently the collection of items occur together as a percentage of all transactions.

Confidence:

Confidence of the rule is the ratio of the number of transactions that include all items in {B} as well as the number of transactions that include all items in {A} to the number of transactions that include all items in {A}.

Lift:

Lift is the ratio of confidence to expected confidence. Expected confidence is the confidence divided by the frequency of B. The Lift tells us how much better a rule is at predicting the result than just assuming the result in the first place. Greater lift values indicate stronger associations.

For the Present Study, Support > 10% and Confidence > 80% is considered for obtaining the quality pattern out of the rules.

apriori algorithm is used in order to implement this methodology.

RCODE:

```
trans2 <- Simulated_Coles_Data[,c(10:53)]
trans2 <- as(trans2, "matrix")
trans2 <- as(trans2, "itemMatrix")
rules1 <- apriori(trans2)
rules1 <- sort(rules1, decreasing=TRUE, by="support") // Sort by Support,Lift,Confidence
inspect(rules1[1:100])
```

Cluster Analysis:

Based on the plot of sum of squares plot, the optimal number of clusters is two. Kmeans is used for this methodology. The variables used to describe a customer in this cluster analysis are transaction amount, income and age. No categorical variables were included as they were not appropriate for K-means clustering.

RCODE:

```
wss=(nrow(Simulated_Coles_Data)-1)*sum(apply(Simulated_Coles_Data,2,var))
for (i in 2:4) wss[i] = sum(kmeans(Simulated_Coles_Data,centers = i)$withinss)
plot(1:4,wss,type="b", xlab="Number of clusters",ylab="sum of squares", col = "blue")
cluster2 <- kmeans(Simulated_Coles_Data, 2, algorithm = "Lloyd")
Simulated_Coles_Data$cluster2 <- as.factor(cluster2$cluster)
cluster1_age=Simulated_Coles_Data$age[Simulated_Coles_Data$cluster2==1]
avg_age1= sum(cluster1_age)/length(cluster1_age) //Calculate avg values for two clusters
```

5. RESULTS:

Market Basket Analysis:

Once the rules are generated sorting is done by Support, Lift, Confidence.

Rules Based on Lift:

Following are the top 5 rules based on lift.

Table- Rule Table with highest lift

Antecedent	Consequent	Support	Confidence	Lift
{fish,vegetables,banana}	{householCleaners}	0.1013425	0.9182782	2.400970
{fish,vegetables}	{householCleaners}	0.1078313	0.8524969	2.228976
{cereal,fish,banana}	{householCleaners}	0.1017556	0.8379872	2.191038
{bread,fish,banana}	{householCleaners}	0.1085370	0.8328051	2.177489
{milk,fish,banana}	{householCleaners}	0.1049225	0.8183649	2.139733

Rules Based on Confidence:

Following are the top 5 rules based on Confidence.

Table- Rule Table with highest confidence

Antecedent	Consequent	Support	Confidence	Lift
{fruit,frozenmeal,TomatoSauce,vegetables}	{banana}	0.1045267	0.9955738	1.311094
{TomatoSauce,bread,vegetables,Olive Oil}	{banana}	0.1012909	0.9950964	1.310465
{fruit,TomatoSauce,vegetables,coffee}	{banana}	0.1019793	0.9947952	1.310068
{cereal,frozenmeal,TomatoSauce,bread,vegetables}	{banana}	0.1049914	0.9947815	1.310050
{fruit,cereal,TomatoSauce,bread,vegetables}	{banana}	0.1146299	0.9947722	1.310038

Rules Based on Support:

Following are the top 5 rules based on support.

Table- Rule Table with highest support

Antecedent	Consequent	Support	Confidence	Lift
{milk}	{bread}	0.6724441	0.8270497	0.9998457

{cereal}	{bread}	0.6370224	0.8358590	1.0104956
{banana}	{bread}	0.6359897	0.8375493	1.0125391
{cereal}	{milk}	0.6207057	0.8144493	1.0017042
{banana}	{milk}	0.6186059	0.8146561	1.0019586

Following are the rules which are ranked from all the three parameters

{fish,vegetables,banana} ==> {householCleaners}

This rule implies that around 10% of all Coles customers purchased fish, vegetables, banana and house hold cleaners in one transaction. About 91.8% who purchased fish, vegetables, banana also purchased house hold cleaners. A lift of 2.40 indicates that there is a very strong positive assosiation between this purchasing behaviours. This combination of products occured more than twice as often as expected.

{fruit,frozenmeal,TomatoSauce,vegetables} ==> {Banana}

This rule implies that around 10.4% of all Coles customers had fruits, frozenmeal, tomatosauce, vegetables, banana together in their basket. About 99.5% who purchased fruit, frozen meal, tomato sauce, vegetables also purchased banana. A lift of 1.31 indicates that this combination of products occured 31% more than expected.

{milk} ==> {Bread}

This rule implies that around 67.2% of all Coles customers purchased fmilk and bread in one transaction. About 82.7% who purchased milk also purchased bread.

Cluster Analysis:

Based on the cluster classification following outcomes are derived.

1. Average transaction amount of $76.70474 spent by customers who earn an average income of $67393.18 and of average age of 39.5 years.
2. Average transaction amount of $81.09686 spent by customers who earn an average income of $130724.8 and of average age of 41.3 years.

Find the figures 7.15 and 7.16

6. CONCLUSION:

Considerable amount of study has been conducted and answers to the raised questions are obtained. From the study it is clear that products like Banana, Milk, Bread are of highest sale. Whereas, products like kitkat, energy drink are of lowest sale. So, special promotions or offers

can be carried out in order to carry out even sales of the products. Moreover, steps need to be taken care so that all gender people can attract towards shopping.

Data used for the study was not clean. erroneous data, missing values were present. Prediction was done is order to clean the data. So, there might be minimal errors in the prediction. In mere future quality of the data can be improved in order to obtain accurate results.

Recommendations:

1. It is clear from the study that products like bread, milk, banana, fruits have the highest sale. So, special offers can be made combining highest sale products and lowest sale products so that people get exposure towards the lowest sale products.
2. Lowest sale products can be kept near the entrance of the supermarket, whereas highest sale products can be kept inside.
3. Considering the output market basket analysis, products location has to shifted so that customers can buy together those products.

By carefully taking consideration of all the points of study, steps should be taken in order to improve the business of Coles supermarket and thus, maintain its brand value.

7. APPENDIX:

Tables:

Table 7.1 Description of Variables

Variable Name	Description	Class Name	Type of Variable
Receipt ID	Transaction ID of Customer	Purchase	Discrete
Value	Value of the transaction	Purchase	Continuous
Pmethod	Payment method	Purchase	Categorical 1= Cash 2= Credit Card 3= Eftpos 4= Other
Sex	Gender of the customer	Purchase	Categorical 1= Male 2= Female
Homeown	Is the customer homeown	Purchase	Categorical 1= Yes 2= No 3= Unknown
Income	Annual Income of Customer	Purchase	Continuous

Age	Age of Customer	Purchase	Continuous
PostCode	Postcode of residence	Purchase	Discrete
nChildren	Number of Children	Purchase	Discrete
Products (Fruits, dairy, cereal etc)	Columns 10 to 53 contains all the product list available	Transactional	Categorical 1= Yes 2= No

Table 7.2 Product Frequency table

Product	Count
fruit	36646
freshmeat	10815
dairy	13424
MozerallaCheese	17133
cannedveg	17592
cereal	44279
frozenmeal	33541
frozendessert	6152
PizzaBase	27120
TomatoSauce	24404
frozen fish	1755
bread	48059
milk	47239
softdrink	18075
confectionery	15217
fish	12457
vegetables	33997
icecream	2595
energydrink	1112
tea	5733
coffee	25323
laundryPowder	28705
householCleaners	22221
corn chips	26940
Frozen yogurt	27083
Chocolate	25894
Olive Oil	25768
Baby Food	28055
Napies	15484
banana	44118
cat food	8566
dog food	11404

mince	14561
sunflower Oil	15768
chicken	37819
vitamins	11726
deodorants	14618
dishwashingliquid	20417
onions	12892
lettuce	43177
KitKat	953
TeaTowel	2156
Scones	2674

Figures:

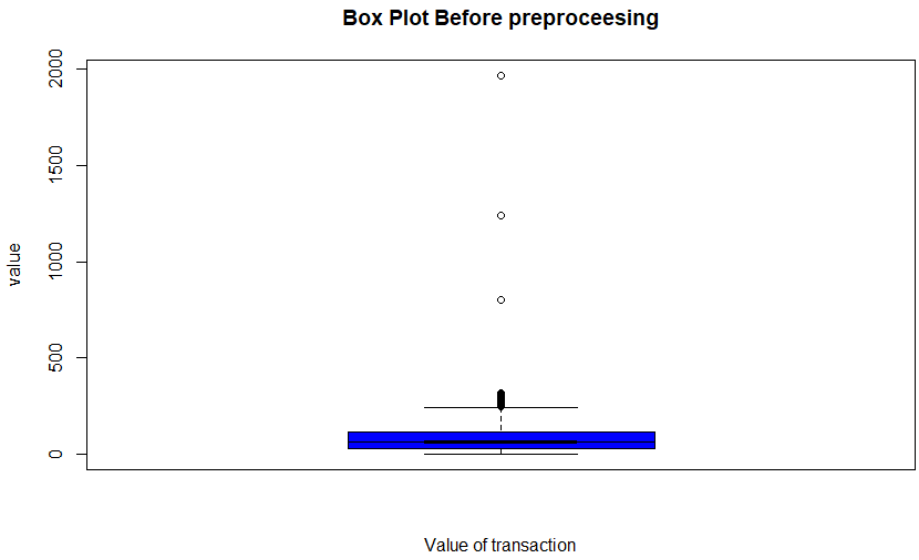


Figure 7.1 Box Plot of transactions before pre processing

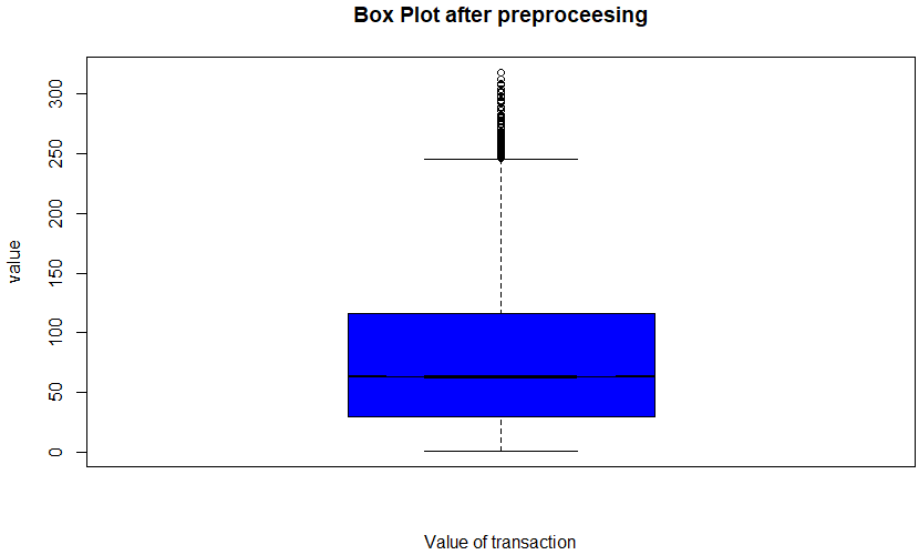
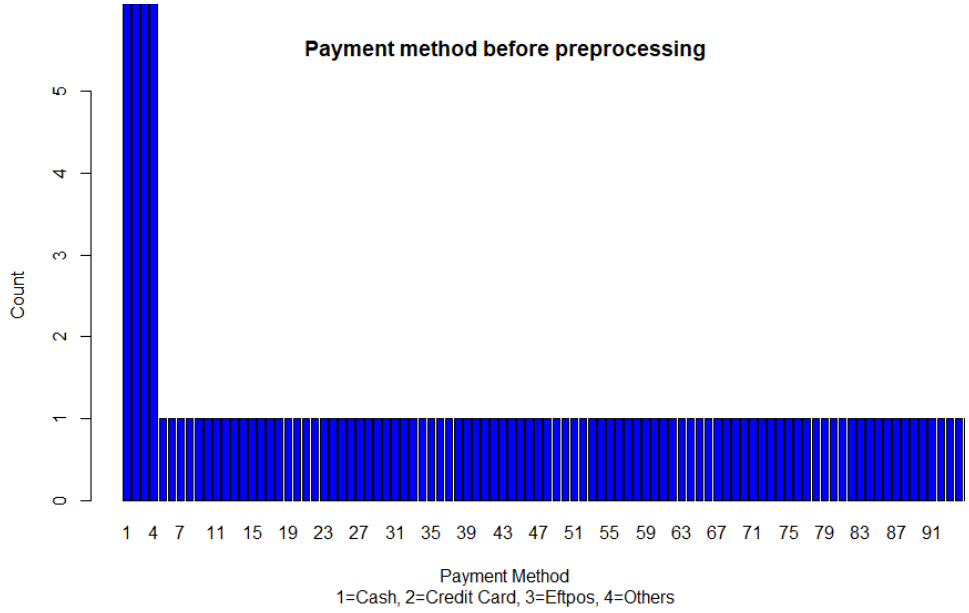


Figure 7.2 Box Plot of transactions after pre processing



7.3 Payment Method before preprocessing

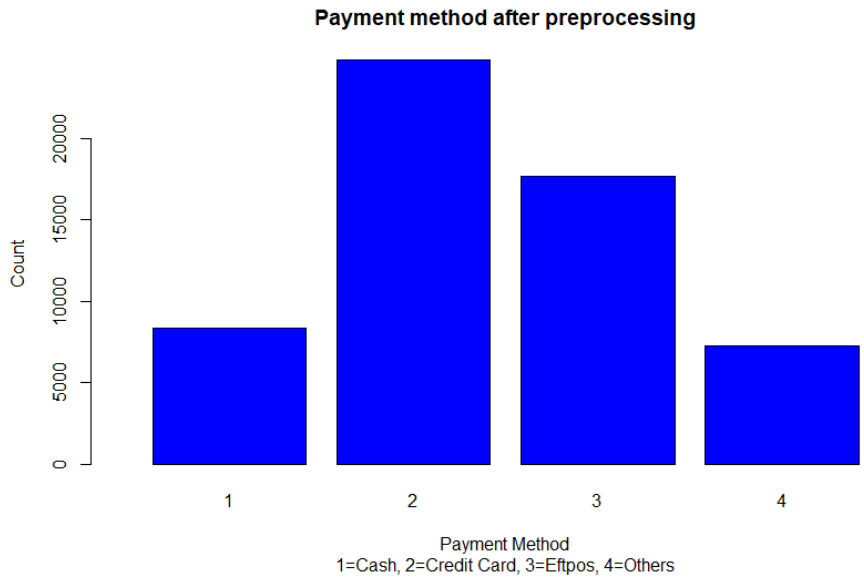


Figure 7.4 Payment Method after preprocessing

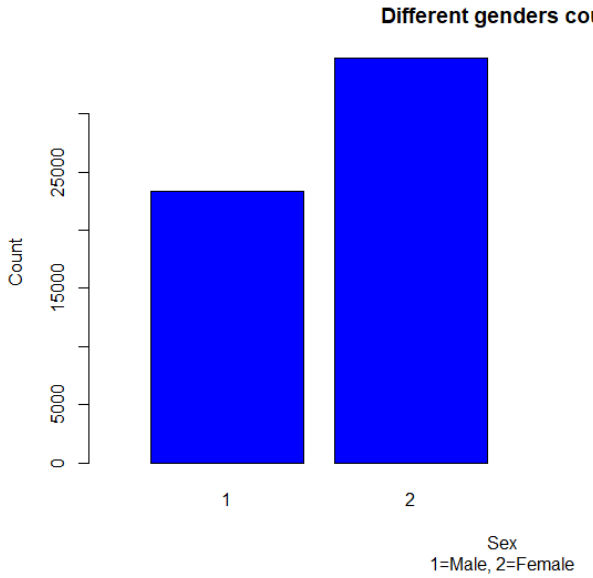


Figure 7.5 Different genders count

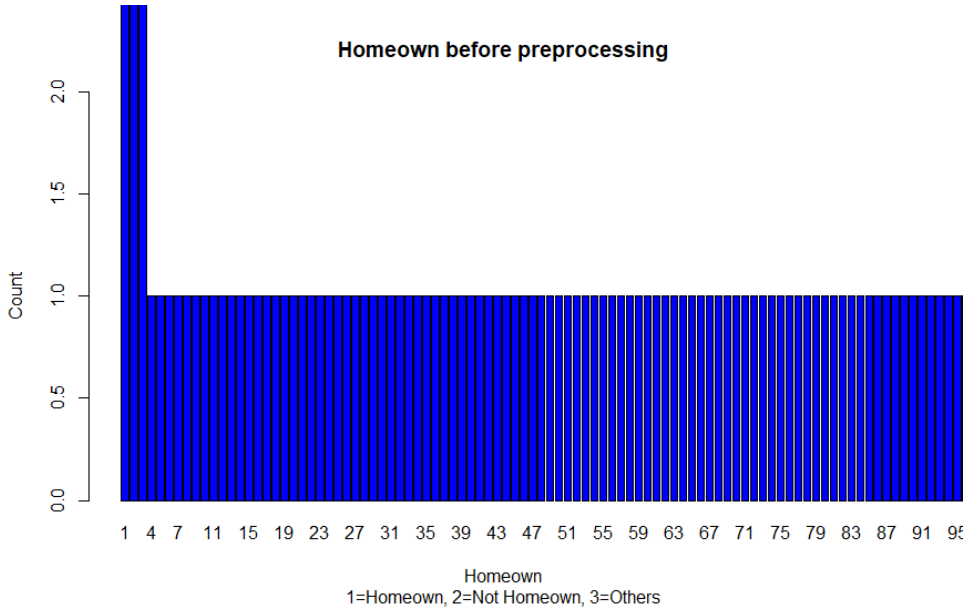


Figure 7.6 Homeown Count before preprocessing

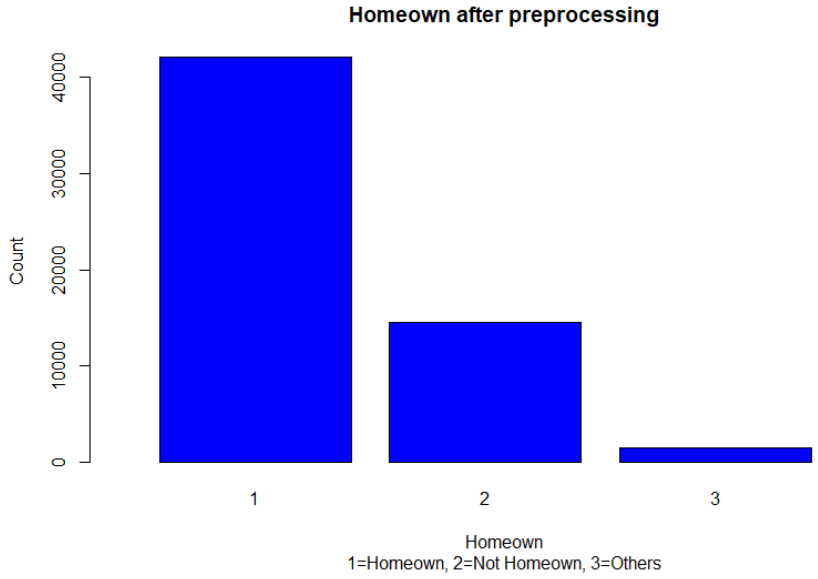


Figure 7.7 Homeown Count after preprocessing

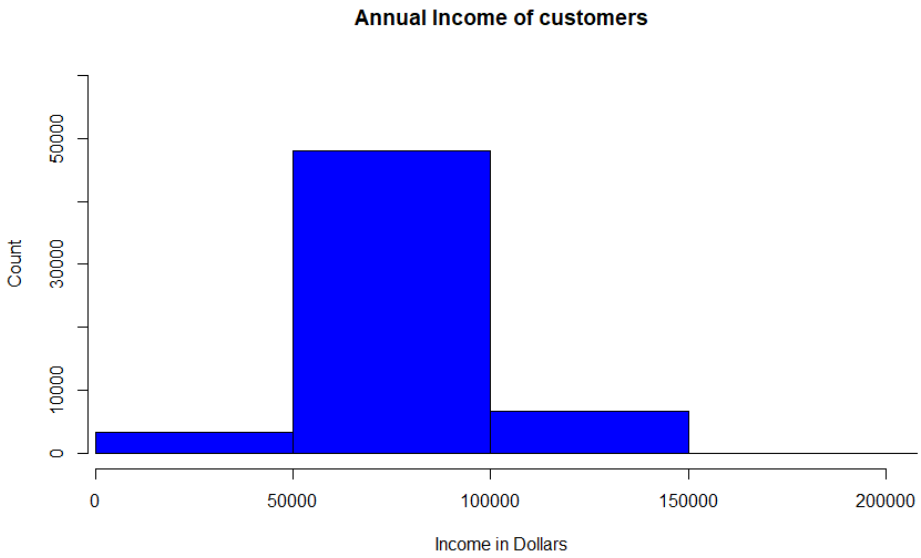


Figure 7.8 Annual income of customers

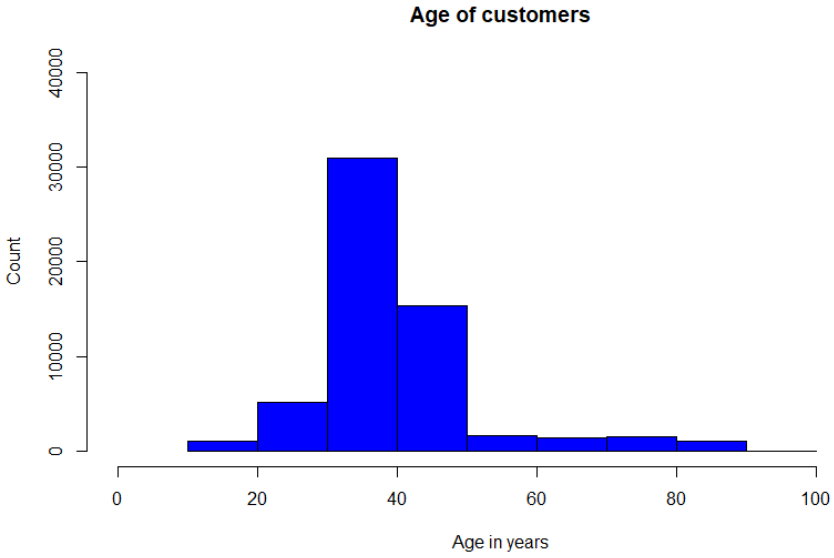


Figure 7.9 Age of customers

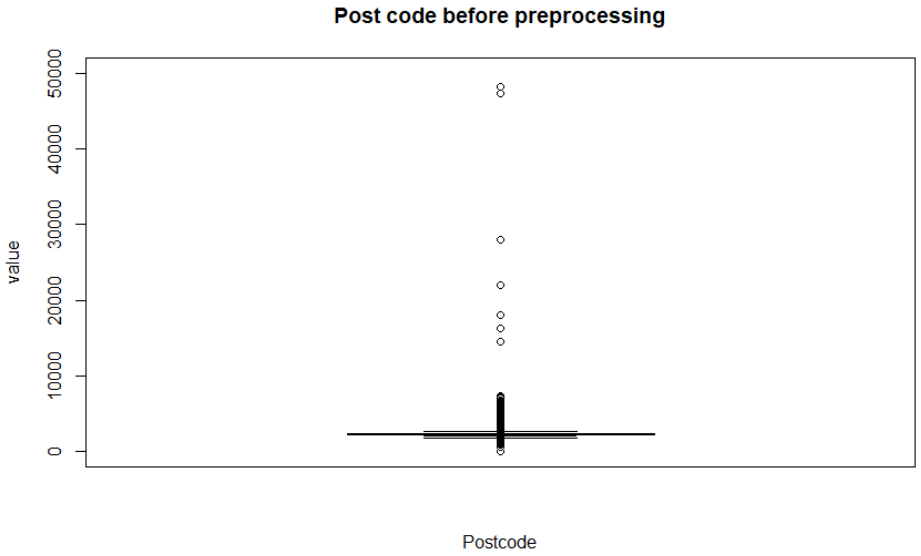


Figure 7.10 Postcode before preprocessing

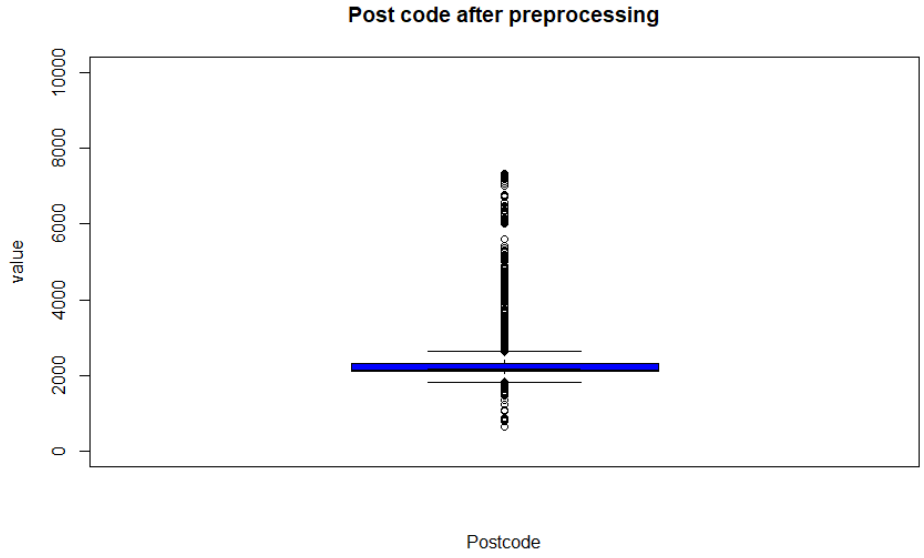


Figure 7.11 Postcode after preprocessing

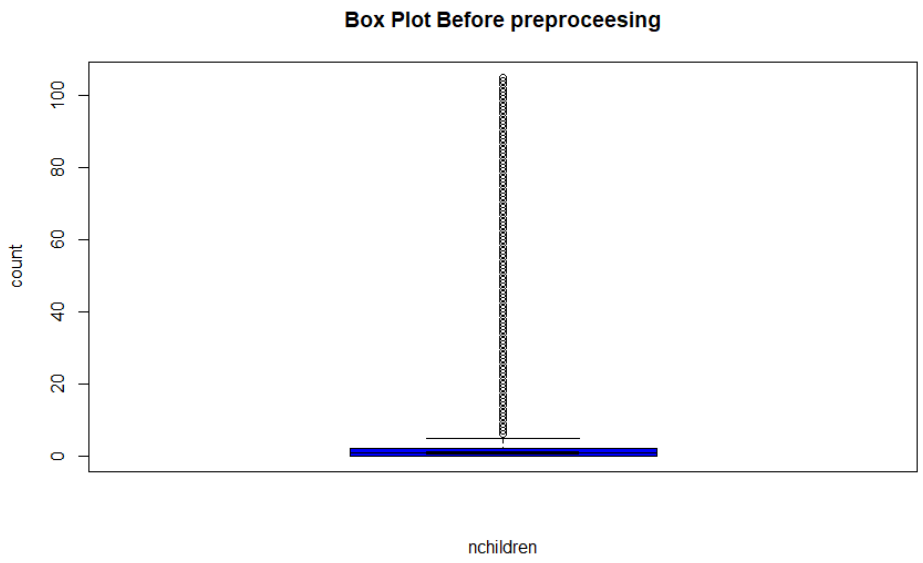


Figure 7.12 Box plot of number of children before preprocessing

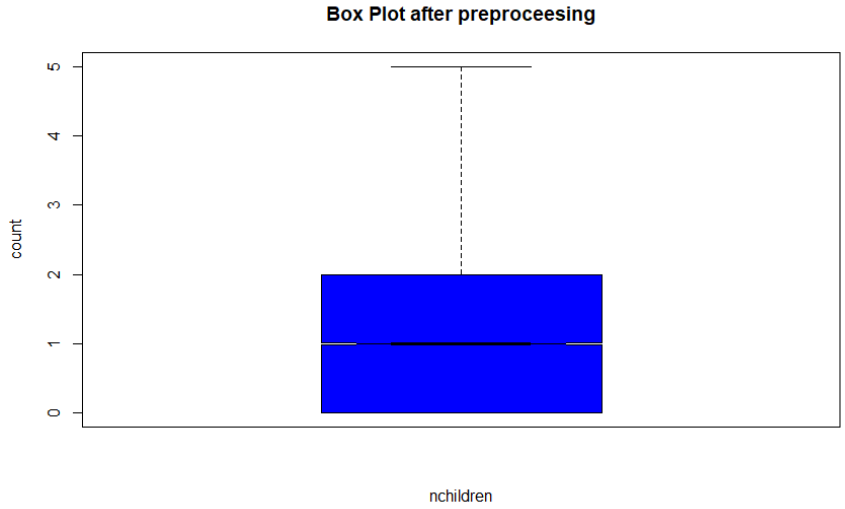


Figure 7.13 Box plot of number of children after preprocessing

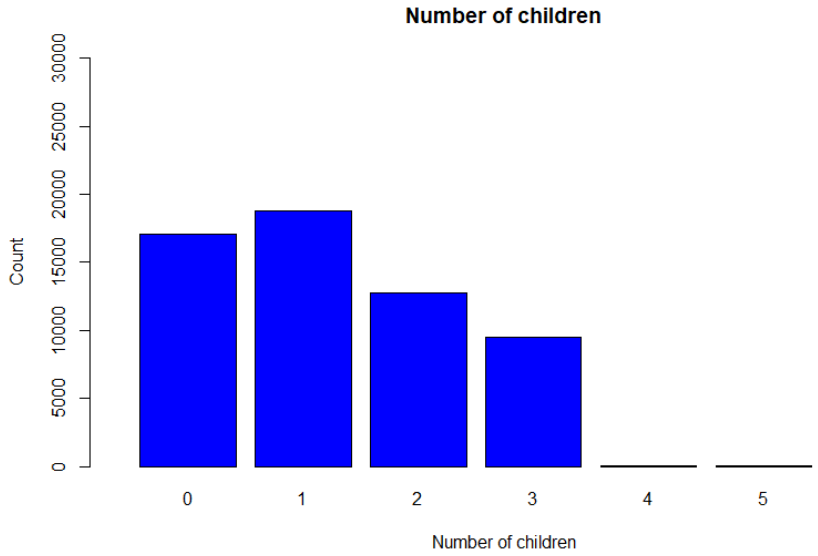


Figure 7.14 Number of children

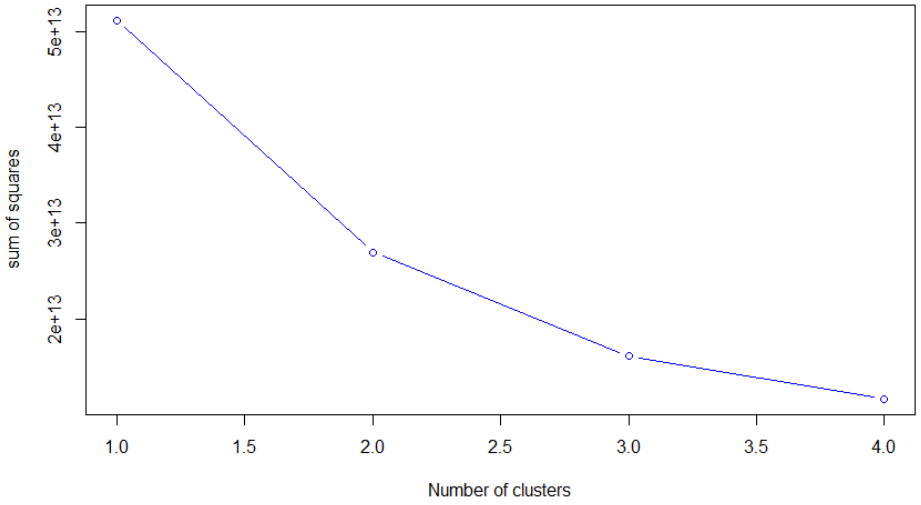


Figure 7.15 Number of clusters

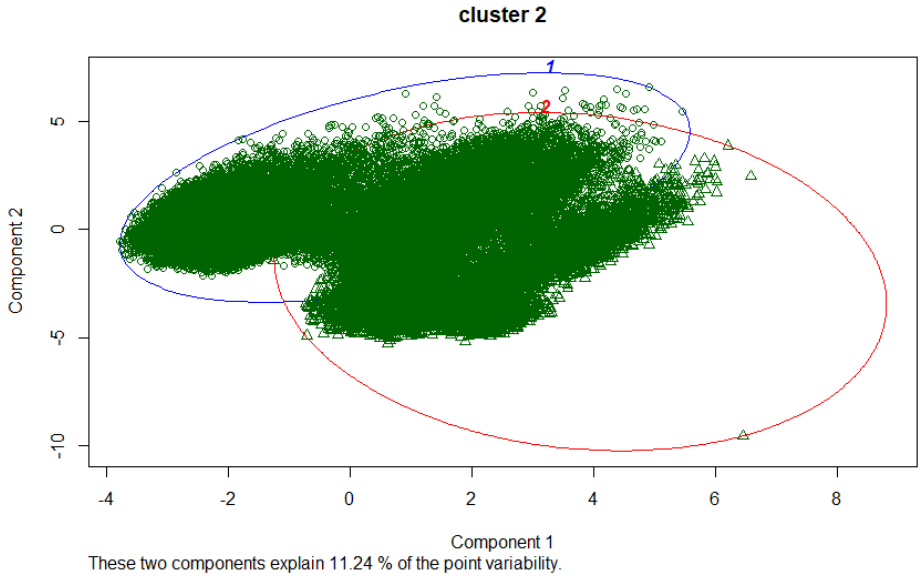


Figure 7.16 Two Clusters

FEEDBACK:

Feedback on the Project (Self Assessment)

1. Which part of this project you struggled most?

Clustering

2. Did you discuss your difficulties with your peers?

No, Searched in Web

3. What would have been helpful to overcome your difficulties identified in Q1?

More practice

4. Was the time given for this project sufficient?

Yes

5. Did the feedback you received for the lab exercises help you with this project?

Yes

6. Did the reading materials provided up till now help you with this project?

Yes

7. Reflect on one of the employability skills you gained by completing this project (no more than 250 words).

I got a kind of Real time working experience, which is very useful. Moreover, working on this kind of study which is end to end added on to my learning.