Market basket analysis
on simulated supermarket data

Executive summary:

Coles is one of the Australia's largest supermarket chains with 801 supermarkets, 883 liquor stores, 702 Coles express fuel stores and 89 hotels across the Australia and 107,000 team members. It has a large share of the market with about 21 million customer transactions each year. In recent year some European brands such as Aldi have come to Australia, there have become rivals for Australian brands like Coles. This research has been conducted to help analyse the different aspect of Coles customer transactions to found out some valuable information about them and think of better strategies to provide a better service to keep the current customers and try to attract new once. In this research method of market basket analysis and cluster analysis has been used to collect the main features of the data.

The association rules where found to see the what products has been purchased together to have a better understanding of patterns in transactions and found the most frequent items. Not surprisingly the most frequent items were bread, milk, cereal, banana which are all food item. In contrast the only non-food item that has a large frequency is household cleaners. Although, the cluster analysis results did identify two groups in customers. It did give us some overview of the customers and some of their attributes.

Introduction:

Nowadays keeping marketing has become so competitive, keeping customers and attract new ones is the main challenge of big companies. This is also the case for Coles, as it is facing more competitors and it want to continue its expanding trend it is necessary to have some insight of their customers and their need. In past decades by expansion of computer's power scientist have developed some method that could analysed the data in very fast pace and turn it into useful information. We would like to utilise these methods to answer two main questions:

Are there distinct group of Coles customers?

What are the product that are most frequently bought together?

Description of the data:

The data set has 50 variables and 58000 observations. Each row correspond to a transaction and first variable is the receipt ID the next eight variable give us some insight of the

customers about the age, gender and etc, the other next 41 variable corresponds to the product in the supermarket that customers have bought in their transactions. By taking a brief look at the data we could see the three of the nine variables related to customers have missing or invalid values. In addition, seven of the forty-one product variables have invalid or missing data too. In the following we will take a closer look to the variable in the data and trying to fix the problem using R.

Transaction value: This variable indicates the value of each transaction, it has the maximum of $1967.70 and minimum of $0.9. Base on the results it has the average amount of $77.2and median of 63.50$. The boxplot of this variable shows three some outliner. They wont cause problem since there is not too many observation, therefore we would not exclude them for further analysis. You can find the box plot of the value in figure G1. Also, boxplot of value of transaction base on the sex in figure G2 shows the value has the same variance on both sexes.

Pmethod: This variable shows the way the customers choose to pay for there purchases. It has four different levels cash, credit card, eftpos and other. In our data credit card has the highest frequencies with %42.61 followed by eftpos with %30.49 and cash with %14.36 and finally other with %12.51. it seems that more than one third of the coles customers choose to pay by credit card and about one of third choose to pay by eftops. The barplot based on paying methods is presented in G3.

Sex: This variable shows the gender of customers who but each transaction. The analysis of the data shows %40.21 of the customer are males and %59.97 are females. It seems that more than half of the customers are females. The bar plot of sex variable has could be find in figure G4.

Homeown: This variable shows if a customer owns a house or not. Base on the analysis %72.36 of the customers own a house, %25.13 don't own a house and we don't have any information about the rest %25. The barplot of owning a house in figure G5 shows the same results as it seems that majority of customers own a house. Moreover, the barplot of owning the house based on the sex in figure G6 illustrates that the number of females that own a house is larger than males who own a house.

Income: This variable illustrates the annual income of the customers. The analysis shows that the max income is $650235, and the minimum income is $6000. Furthermore, the average income is $74850, and the median income is $70169. The boxplot of the variable shows that the $650235 is an outlier. Although this income is probable we would exclude it from the data for cluster analysis. This variable has one missing data since the data is right-skewed we will substitute it by the median for the analysis. The boxplot of the income is in figure G7. In addition, the boxplot of the income based on the sex of the customers is in figure G8 it indicates that the variability of income is the same among both sexes.

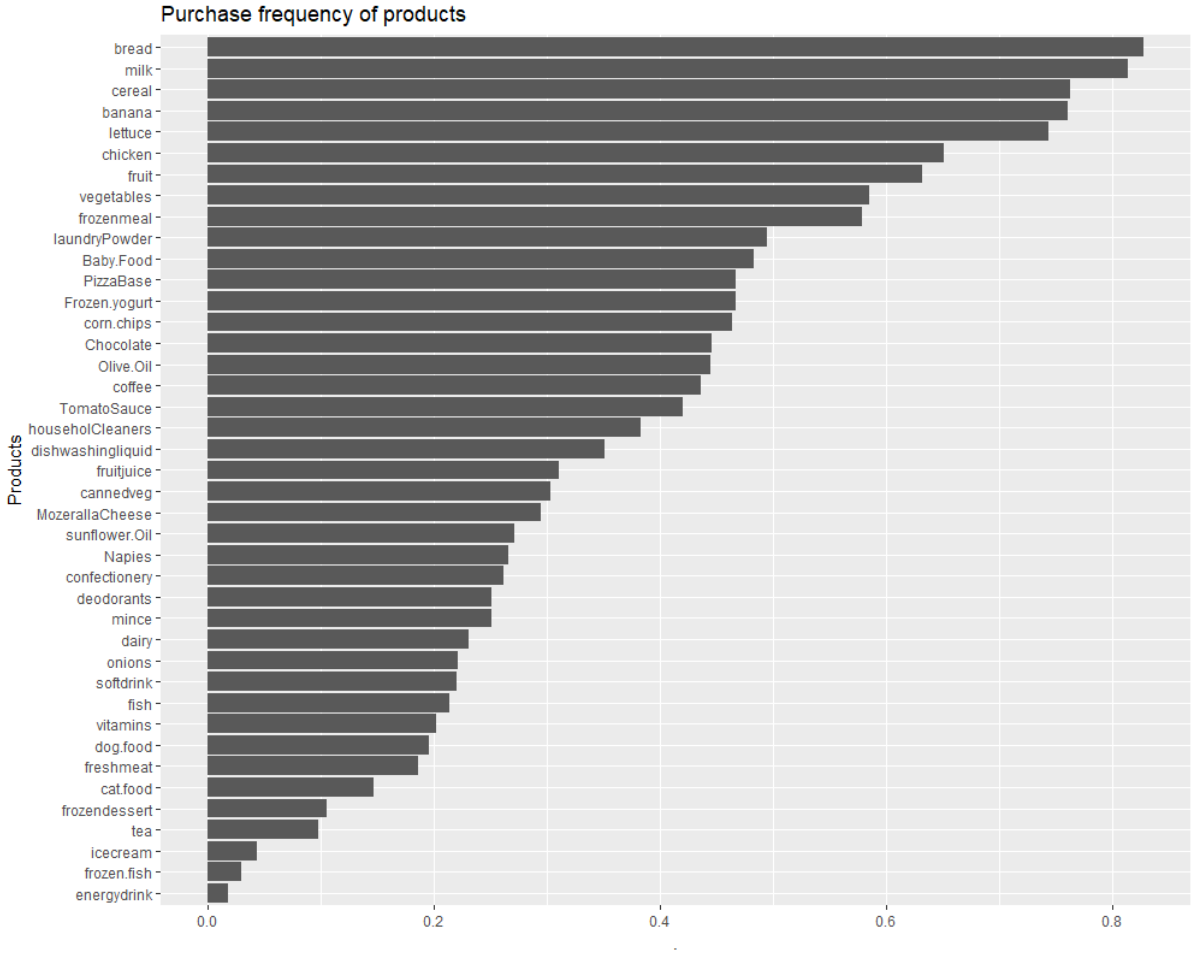
Age: This variable corresponds to the age of the customers. The analysis shows the maximum age of the customers is 95 and the minimum is 10. The average age is 39.7 and the median is 38. Majority of customers are between 34 and 39.7 years old. There is a missing value in this data. Since the data is slightly right-skewed we will substitute it by the median. The histogram of the age could be found in G9.

Number of children(nchildren): This variable shows the number of children each customer has. The maximum of the children a customer has is 11. Base on the analysis, %32.13 of the

customers have one child and % 29.47 has no children, %21.99 has two children and %16.31 has three children. Only one customer has eleven children which is probable. It seems that one of every two has at least a child. On average customers have 1.255 children and median is 1 child. This variable has 2 missing values and we will substitute them with median. Barplot of customers based the number of the children could be seen in figure G10.

Postcode: The postcode variable is not recorded properly and has about 10000 missing values. This could not be used in the analysis, so we have excluded it from the further analysis.

Products: These 40 variables show the purchase of the products in transactions. The graph below shows the frequency of purchase of each item. According to the graph bread milk cereal, banana are most frequent items. In contrast, energy drinks, frozen fish, ice cream, tea, frozen dessert are the least frequent item. The most frequent items are in more than %70 frequency in transactions while the least frequent items are only in less than %10 of the transactions. There are some missing values in the data but we consider them as not purchased.



Methodology:

Data has been gone through a prepossessing in R it has been carefully examined to find the missing values and outliers. Decisions has been made according to the parameters of the data

to either substitute them other value or remove them from the data. In addition, the data has been analysed using statistical and graphical tool to have a better overview of the data and detect unusual pattern and values in it.

Market basket Analysis: For conducting this analysis we will chose the apriori algorithm which will find us pattern among the product purchasing dataset. This algorithm will first look for the combinations of the items being purchased together and then will check how often has this combination occurred in the data set based on some parameter that we will chose. Basically, every association rule is like A→B, in which A is called antecedent and B is called consequent, it is good to mention A and B are different product. First parameter of the apriori is called support which is percentage of the transaction that a rule has been occurred. The next one is called confidence which the proportion of the transactions which the B is in them given that A is in them. And last but not the least is the lift which is the proportion of all transaction including A and B divided by the proportion of transaction including A and proportion of transactions including B. The lift is a measure for reliability of your rule. In R we use the arules package which include the apriori command we will set support parameter to 0.1 because a resendable pattern should be supported by at least %10 of the data. We want a confidence of %80 so we will set the parameter to 0.8 because we don't want to end up with some trivial rules we need more actionable rules, rules that could be used for marketing purposes. The will would sort the rules by maximum lift to find the best rules and will report them. Also, we would set the minimum length of a rules to 2 because a rule with the length of one is not something useful and will set the maximum length to 3 because any rules larger than that would not be interesting and actionable.

Cluster analysis:

For conducting a cluster analysis, k-means algorithm has been used to identify distinct groups in customers. This algorithm can only find clusters in continuous variables; therefore, we will only consider our continuous variable including Value, income, age, number of children. There are other types of clustering like hierarchical clustering which is not suitable here because the number of observation is to large. The graph sum squares against clusters could be found in figure G11. According to the graph the optimum number of clusters to use in our analysis is two.

Results:

The rules that we are going to explain below are sorted based on the lift and they have the length of three they are actionable rules that means they could be used in business and marketing purposes. However, here are rules by the length of 2 which are mainly trivial rules like {Nappies} → {Baby food} with support of %22.43 and confidence of $84.24 and lift of 1.74 as we mentioned above this rule is so trivial and would not give us new information about the transaction patterns and we could not use it for marketing purposes. Also, if we sort the rules of length 3 base on the confidence we might end up with some trivial rules like {tomato sauce, vegetable} → {banana} with support of %23.93 and confidence of %97.71 and lift of 1.28 this rules does not give us any insight cause the confidence is about %100 so it is considered as trivial rule. Furthermore, all these products are from the food section, so it is very likely that all them placed near each other.

Market basket analysis:

Rules1: {Fish, vegetable} → {Household Cleaners}

Fish and vegetable have purchased together in %10.79 percent of the transactions. It seems that %85.34 of the times when fish and vegetable have been purchased together household cleaners have been purchased with them too. The lift equal to 2.2308 suggest a strong association between the pattern of fish and vegetable together and purchasing household cleaners.

The fish and vegetable are in food category and house cleaners is in chemical products. It seems that customers most of the times purchase these items together. It suggests that it is better to offer promotion on these item at the same time to increase the confidence of this rule. Coles can take advantage of these rules to improve the purchase frequency of items like household cleaners which is among the less popular products.

Rules 2: {Fish, banana} → {Household Cleaners}

Coles customers have purchased fish and banana together in %12.92 of the all transactions. It shows that %81.73 of the times when fish and banana have been purchased household cleaners have been purchased too. This rule has the lift 2.13 which shows a strong positive association between the pattern fish and banana together and household cleaners. Again, like the previous rules although, household cleaners are a chemical product both fish and banana are food products. Banana is frequently purchased but fish is bought rarely, if we some how try to improve the frequency of the fish been bought it will help the purchasing rate of household cleaners.

Rules 3: {Tomato sauce, nappies} → {Household Cleaners}

Purchasing Tomato sauce and nappies has been observed in %10.26 of the transactions. It seems that %95.61 of the transactions when tomato sauce and nappies has been purchase baby food has been purchased too. In addition, this rule has the lift 1.978 which suggest a strong association between purchasing tomato sauce and nappies and baby food together.

Rules 4: {Olive oil, Nappies} → {Baby food}

Based on the analysis we can say that olive oil and nappies has been purchased to together in %10.63 of all the transactions. We can see that when olive oil and nappies has been purchased it has %94.02 chance for the baby food to be in the basket. The lift of this rules is 1.945 which means strong association between purchasing pattern of olive oil and nappies and baby food. This is an interesting rule because we might guess that nappies and baby food would be purchased together but the exciting of olive oil in the rules is surprising. It suggests that if we can have promotion on olive oil and nappies together it will somehow increase the purchase frequency of baby food.

Rules 5: {chocolate, Nappies} → {Baby food}

After the analysis we can say that chocolate and nappies have been purchase together in %10.63 of all the transactions. The analysis illustrates that %92.74 of the times when chocolate and nappies have been purchase together, baby food have been purchased too. This rule has the lift of 1.91 which suggest a strong association between the chocolate and nappies

pattern and baby food. For explaining this rule, we can imagine the scenario that one of the parent has came to supermarket with an older sibling of the baby and make he or she happy they will buy a chocolate for him or her. So, it could be recommended to put some chocolate near the nappies to improve the support the rule.

Cluster analysis:

For choosing the optimum number of clusters, we have plot the sum squares against the number of clusters and graph suggest that two is the best choice for number of clusters. Therefore, we have done our analysis and come up with these two clusters:

Cluster 1: Average transaction value ($80.55), Customers with average income ($131249.34), with average age of (41.16), with average number of children (1.34)

Cluster 2: Average transaction value ($76.71), Customers with average income ($67477.61), with average age of (39.50481), with average number of children (1.242)

As we can see the average transaction value of the clusters are close to each other. The same thing holds for the number of children and age. However, the average income is very different. The average income of the first cluster is twice as much as the second cluster. Although, it does not give us any actionable information for marketing. In the figure G12 you can find a graph of clusters.

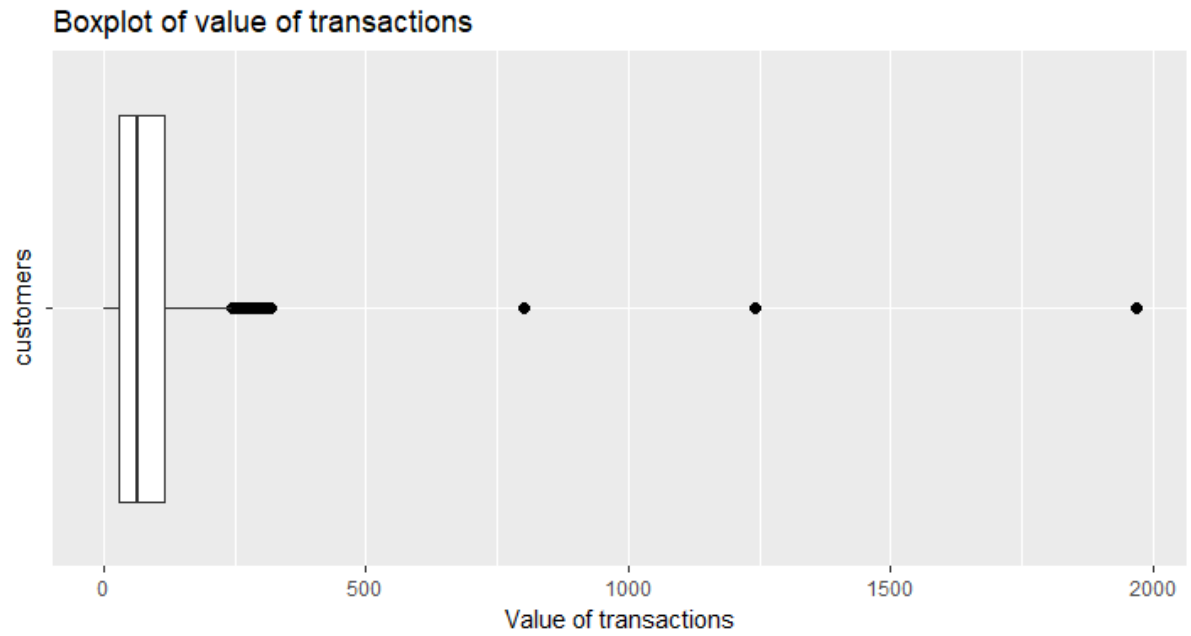
Conclusion:

In market basket analysis we have identify the most frequent items in the basket of customers which wear bread banana milk and cereal. As we can see they have also occurred in our top association rules too. Nappies and household cleaner were also very common in the association rules, though they don't seem to be frequently bought items items, so it is suggested that these items should somehow place the frequent items such as bread, banana and milk. In addition, Coles could put some promotion and sale bundles to improve the frequency of these items. According to our analysis a big proportion of Coles customers have at least one children so they very likely to by nappies. The Coles marketing department can consider these facts when they are planning for future.

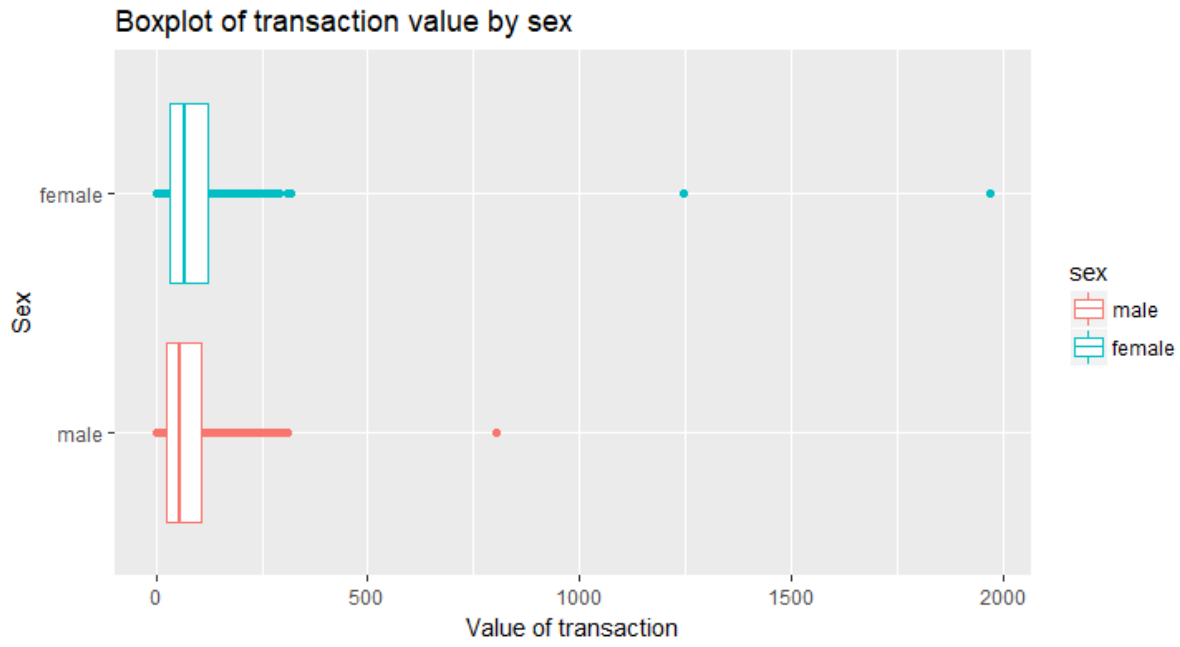
The cluster analysis results may not seem useful in the first sight. But the difference I the average income might be actionable. We know that a cluster of customers do have a medium income that suggests that Coles can put advertisement in the public transport area that these people are very likely to use. For the other group which represent a high-income group Coles could advertise in magazines in the plane cause these people are more likely too be businessman and they are very likely too travel by plane a lot.

Appendix:

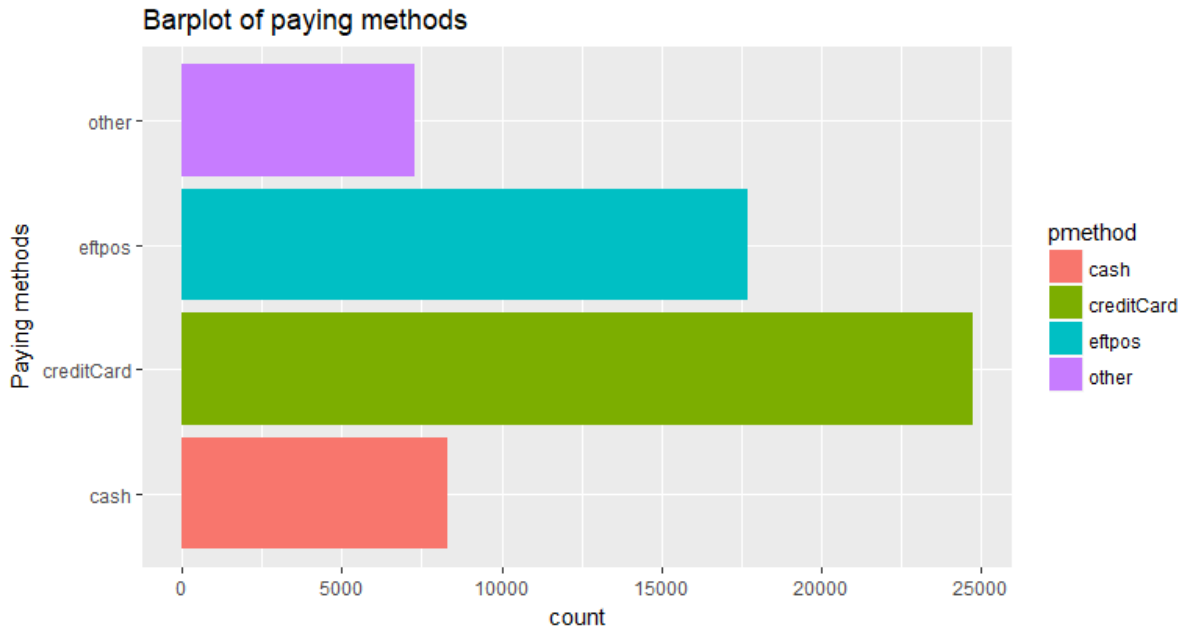
G1.


Boxplot of value of transactions
customers
0
500
1000
1500
2000
Value of transactions

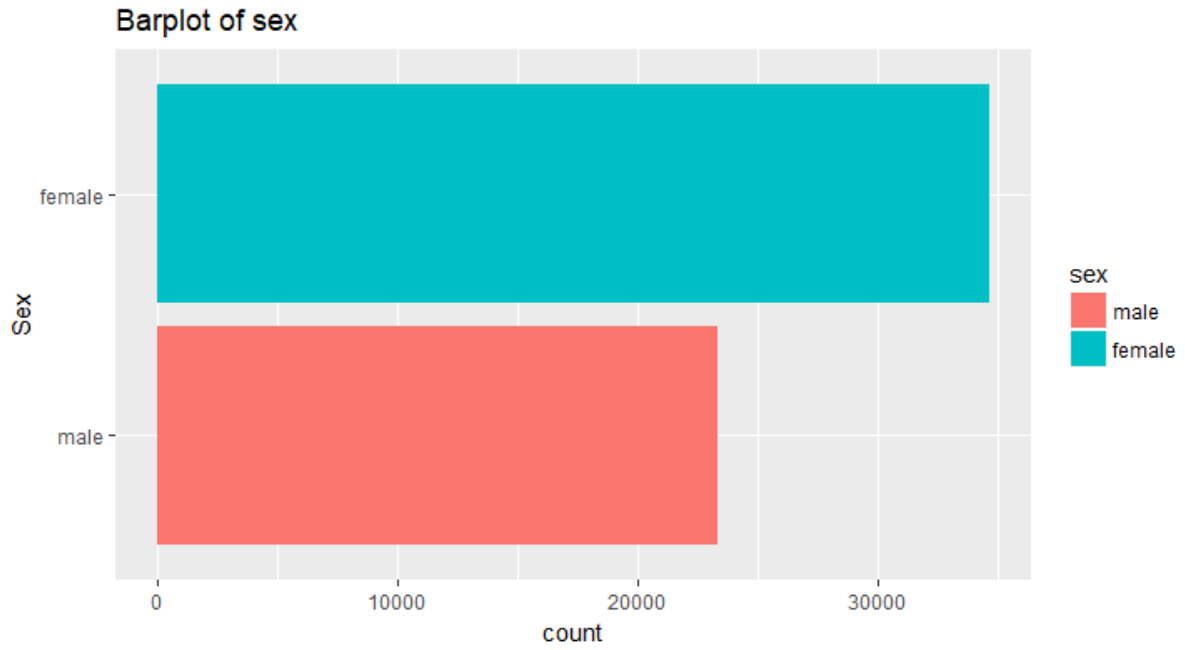
G2.


Boxplot of transaction value by sex
female
male
Sex
0
500
1000
1500
2000
Value of transaction
sex
male
female

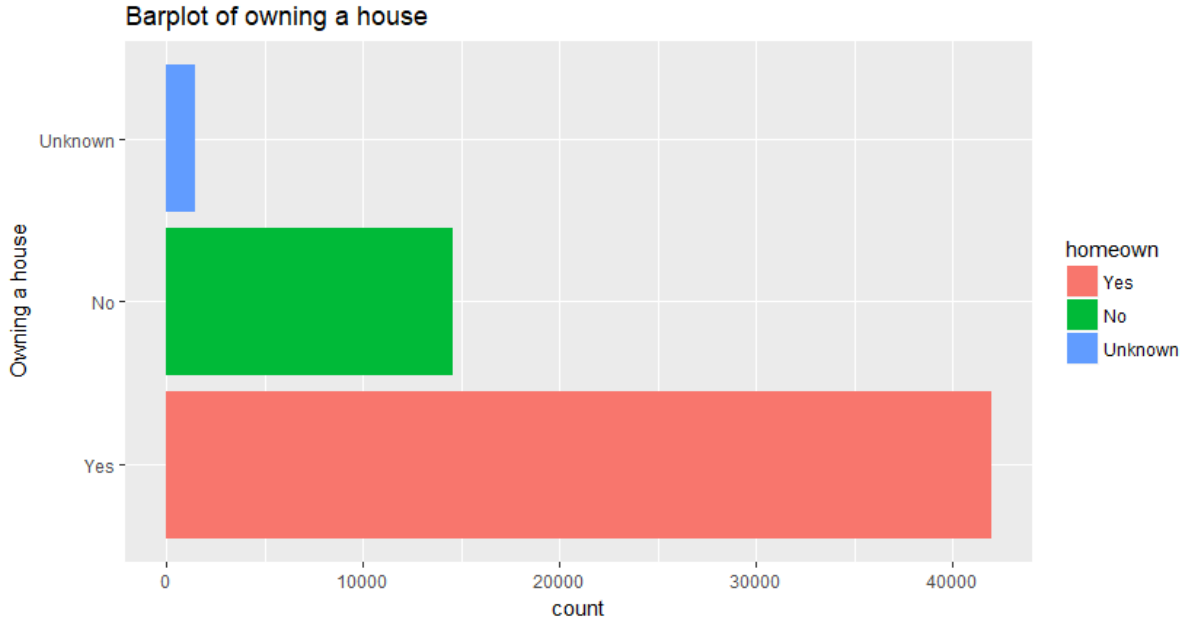
G3:


Barplot of paying methods
other
eftpos
creditCard
cash
Paying methods
0
5000
10000
15000
20000
25000
count
pmethod
cash
creditCard
eftpos
other

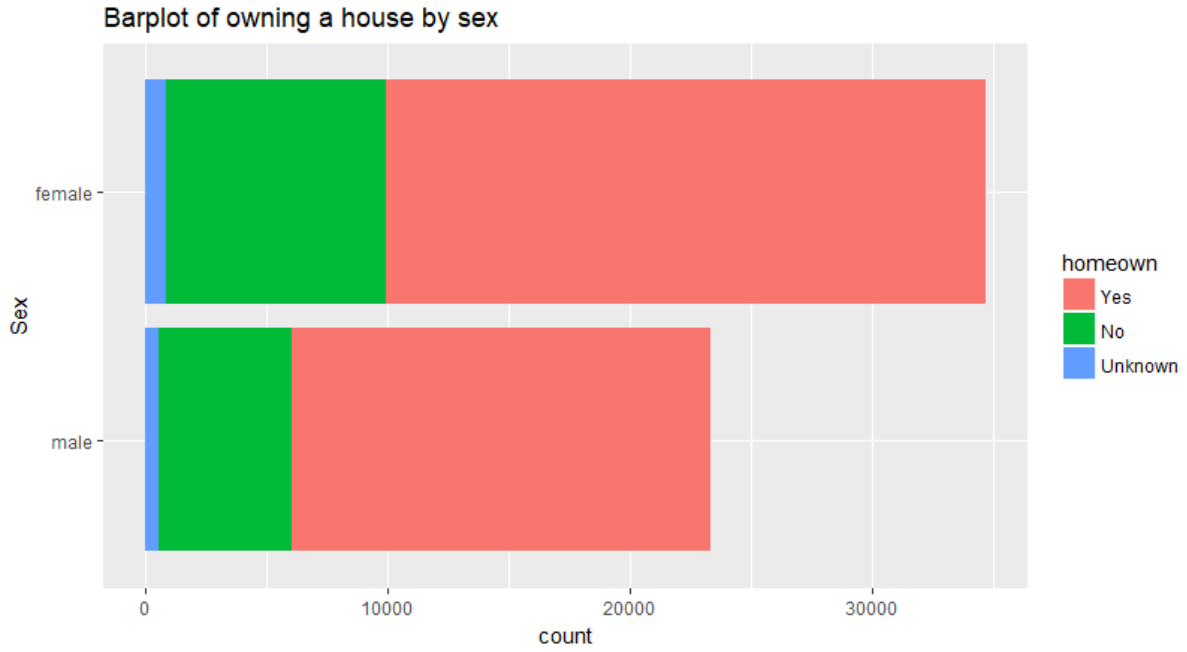
G4.


Barplot of sex
female
male
Sex
0
10000
20000
30000
count
sex
male
female

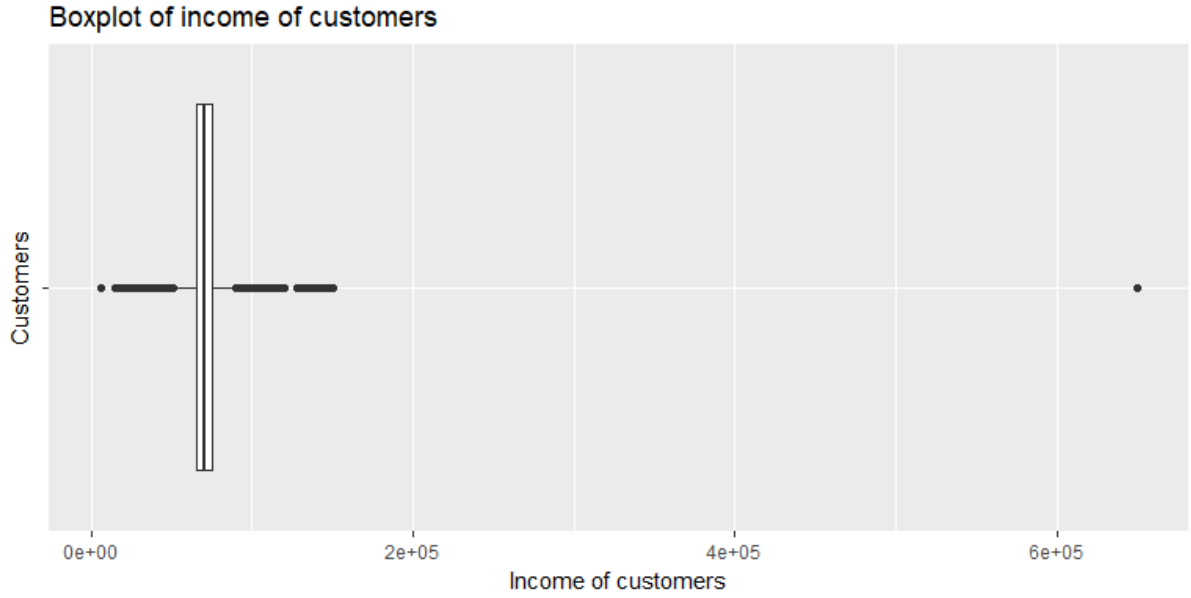
G5.


Barplot of owning a house
Owning a house
Unknown
No
Yes
0
10000
20000
30000
40000
count
homeown
Yes
No
Unknown

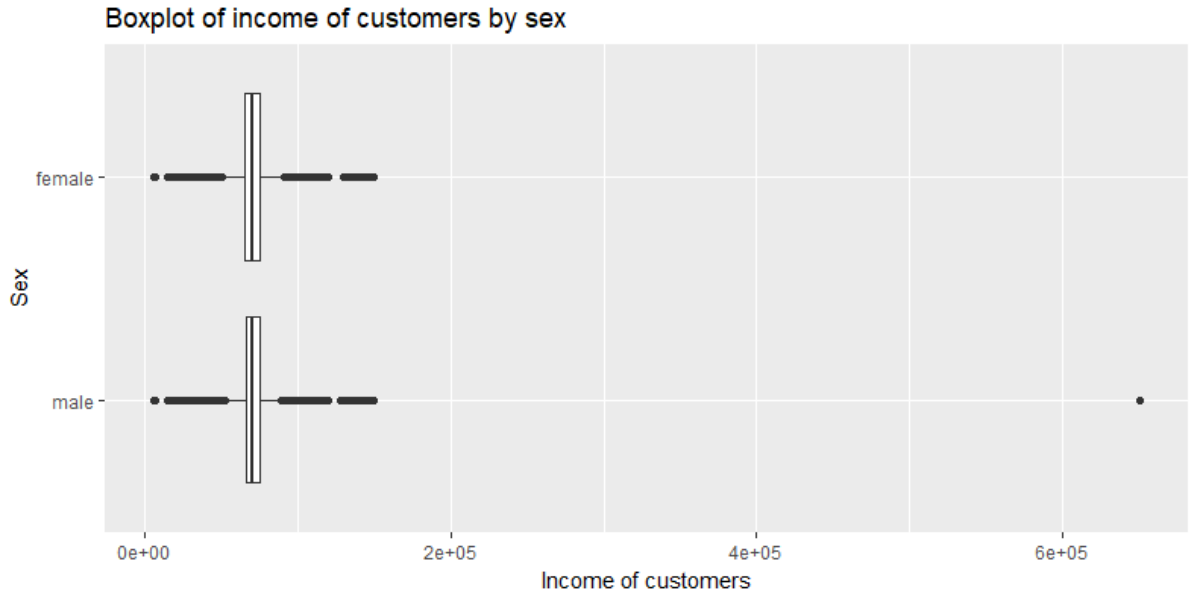
G6.


Barplot of owning a house by sex
Sex
female
male
0
10000
20000
30000
count
homeown
Yes
No
Unknown

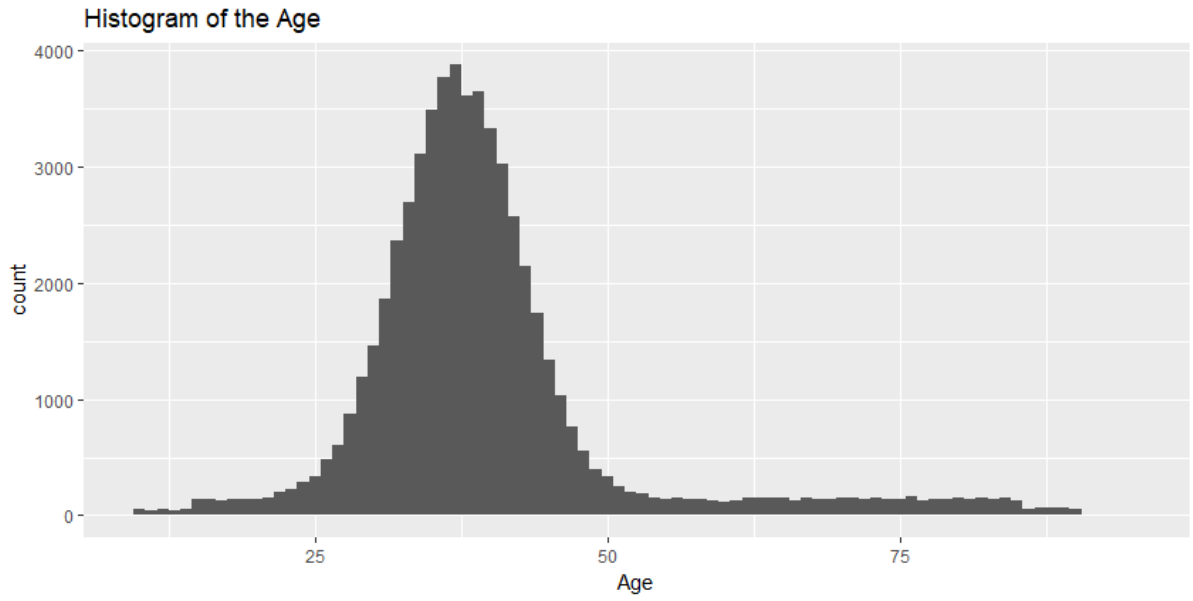
G7.


Boxplot of income of customers
Customers
0e+00
2e+05
4e+05
6e+05
Income of customers

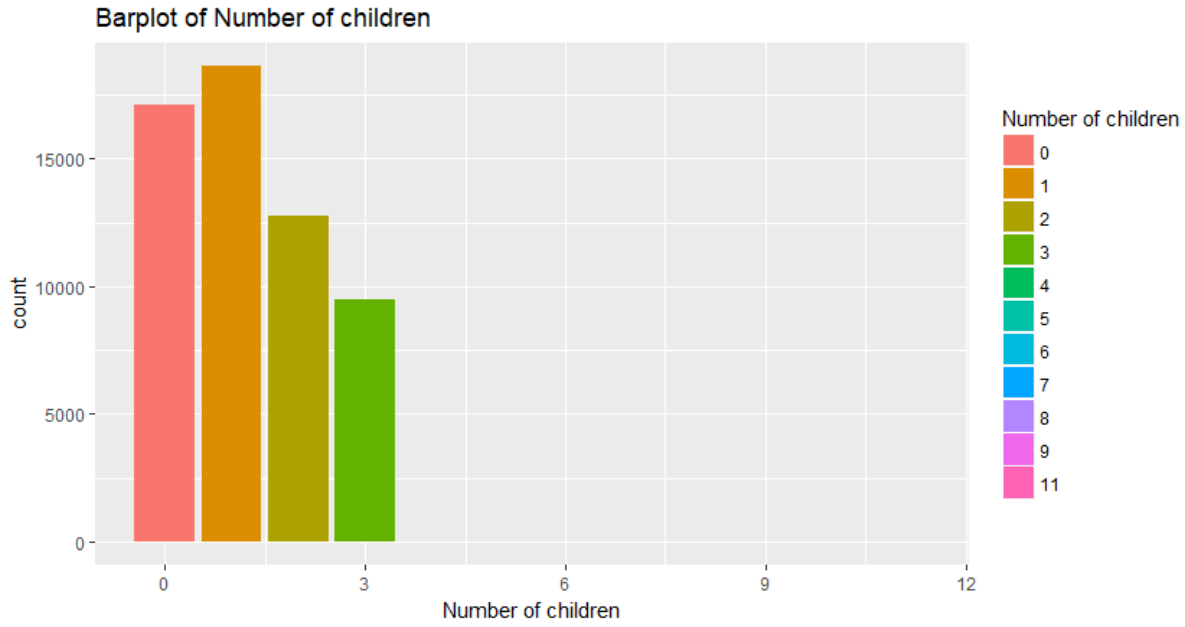
G8.


Boxplot of income of customers by sex
female
male
Sex
0e+00
2e+05
4e+05
6e+05
Income of customers

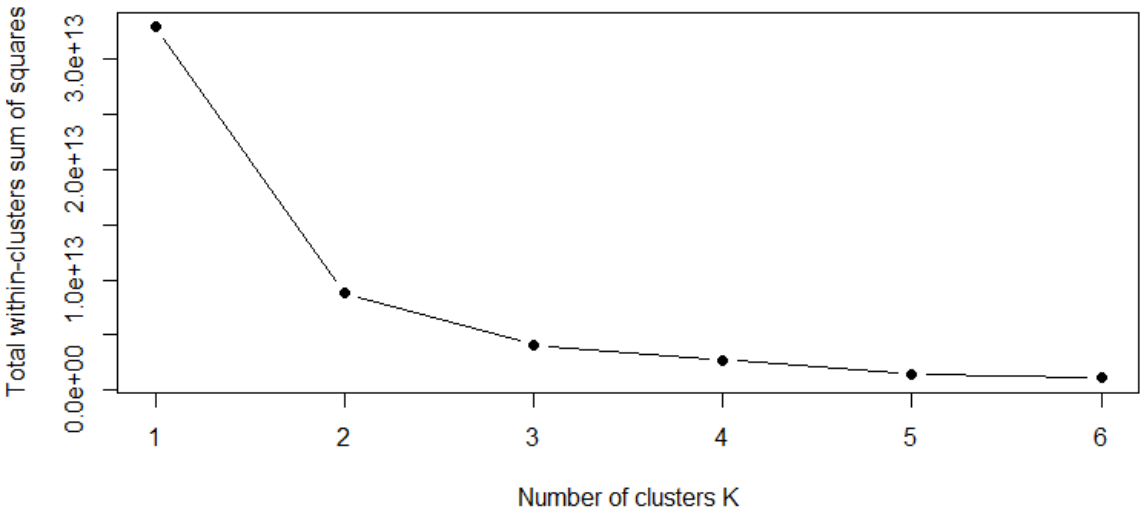
G9.


Histogram of the Age
4000
3000
2000
1000
0
count
25
50
75
Age

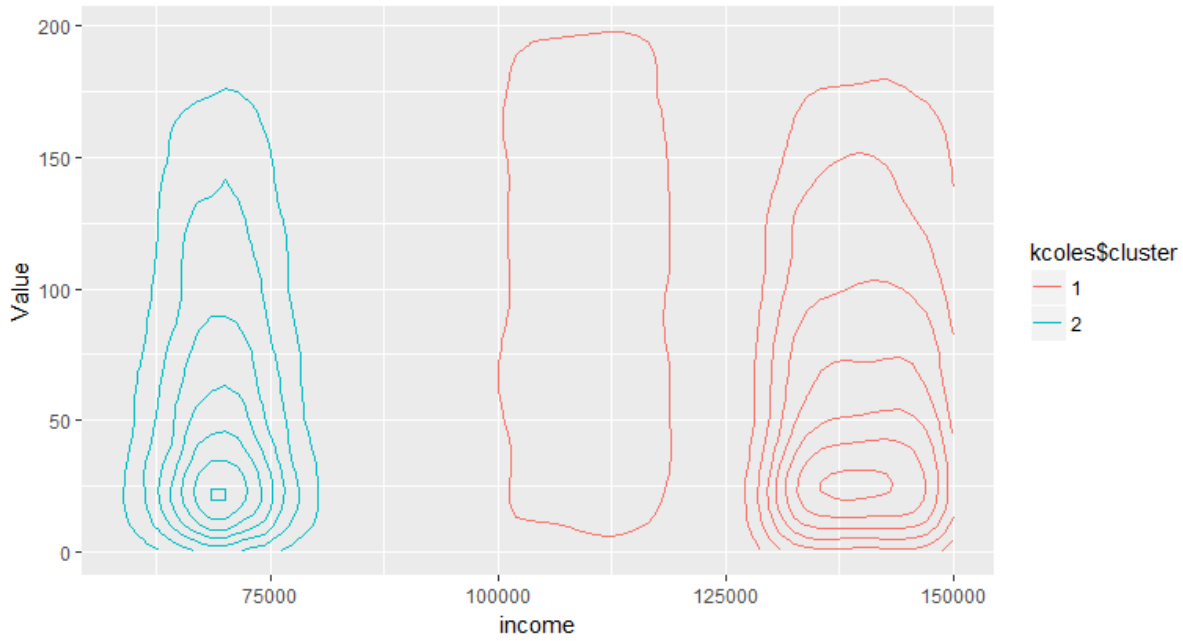
G10.


Barplot of Number of children
15000
10000
5000
0
count
0
3
6
9
12
Number of children
Number of children
0
1
2
3
4
5
6
7
8
9
11

G11.


Total within-clusters sum of squares
0.0e+00
1.0e+13
2.0e+13
3.0e+13
1
2
3
4
5
6
Number of clusters K

G12.


200
150
100
50
0
Value
75000
100000
125000
150000
income
kcoles$cluster
1
2