**MARKET BASKET ANALYSIS**

**SIMULATED COLES DATA**

**Table of Contents**

[Executive Summary 1](#_Toc32961638)

[Introduction 1](#_Toc32961639)

[Description of the dataset 2](#_Toc32961640)

[Original Data 2](#_Toc32961641)

[Data Preprocessing 2](#_Toc32961642)

[Consumer Profile 4](#_Toc32961643)

[Methodology 4](#_Toc32961644)

[Market Basket Analysis(MBA) 4](#_Toc32961645)

[Appendix 6](#_Toc32961646)

[Appendix I – Transactional variables 6](#_Toc32961647)

[Appendix II –Demographical and Socio-Economical Features 6](#_Toc32961648)

[Appendix III – Basket Items 7](#_Toc32961649)

[Appendix IV – ReceiptID[Duplicates] 9](#_Toc32961650)

[Appendix V – Value 9](#_Toc32961651)

[Appendix VI – pmethod 11](#_Toc32961652)

[Appendix VII – sex 12](#_Toc32961653)

[Appendix VIII – homeown 12](#_Toc32961654)

[Appendix IX – income 13](#_Toc32961655)

[Appendix X – age 14](#_Toc32961656)

[Appendix XI – nchildren 16](#_Toc32961657)

[Appendix XII – parent 16](#_Toc32961658)

[Appendix XIII – Pet Owner 17](#_Toc32961659)

[Appendix XIV – Basket items 17](#_Toc32961660)

[Appendix XV – Summary for Numerical Variable 18](#_Toc32961661)

[Appendix XVI – Summary for Categorical Variable 18](#_Toc32961662)

# **Executive Summary**

This report provides an analysis of the Simulated Coles data which consists of 58,100 observations and 53 variables. The aim of the analysis is to explore questions about which products customers are buying and what kind of customers are shopping at Coles. The answers to these questions can be used to enhance customer experience and/or to maximize profits which can be achieved by improving store layouts, offering discounts or marketing different products together, and determining the best place to shelf the specific products.

After cleaning and preprocessing the data, Clustering analysis and Market Basket analysis have been applied to the data. Clustering analysis results revealed that there 4 groups of customers based on income, age and spending value. These 4 groups are named **Elderly High Spenders, High Spending Youth, Young & Broke and Rich & Wise Youth. These groups are named based on their characteristics of the customers in the groups. Market Basket Analysis results uncovered that there is a high correlation between products like bread, milk & banana and nappies & baby food. Some less obvious product combinations were fish, vegetables & household cleaners and coffee, bread, vegetables, frozen meals & household cleaners.**

**This report also offers a detailed explanation of data cleaning and analytical processes, data mining techniques, analysis outcome, and suggestions for future analysis.**

# **Introduction**

Coles is one of the leading supermarkets in Australia, generally dealing with the provision of everyday goods. The emergence of competitors in the market has meant that the competition has become much fiercer and It is crucial for Coles to take strategic actions to beat the competition. One way to do that is to use the data Coles has been collecting from its customers and analyze data to gain a competitive edge. This report aims to use unsupervised knowledge discovery to find hidden patterns in customers’ transactions and to form customer segments. We’ll be seeking a better understanding of two aspects:

* **Customer Segments**: - What are the different segments or groups and their demographic as well as socio-economic features.
* **Products Patterns**: - Popularity of the products, what kind of products are being bought together, the likelihood of two or more products being purchased together.

By better understanding these two aspects and finding answers to the questions related to these two aspects, Coles would be able to form sophisticated strategies that would help Coles to lead the competition. A better understanding of the customer segments would help Coles create more effective and successful market campaigns and design eye-catchy catalogs of the products whereas better knowledge of product patterns would allow coles to improve in-store layouts and product placements.

# **Description of the dataset**

## Original Data

The original dataset used in this analysis was provided into spreadsheet(.xlsx) format. The spreadsheet consists of:

* 58,100 observations(rows)
* 53 variables(columns)
  + 3 Transcational Variables. *[Appendix I]*
  + 6 Variables representing Demographics and Socio-Economical features of customers. *[Appendix II]*
  + 44 Basket Items. *[Appendix III]*

The table below offers a brief description of each variable(For detailed table go to Appendix)

|  |  |  |
| --- | --- | --- |
| **Variable** | **Type** | **Description** |
| ***ReceiptID*** | Ordinal | Transaction identification number |
| ***Value*** | Continuous | Dollar amount of the transaction |
| ***pmethod*** | Nominal | Payment method: 1 = Cash, 2 = Card, 3 = Eftpos, 4 = Other |
| ***sex*** | Binary | 1 = male, 2 = female |
| ***homeown*** | Nominal | If they own a house: 1 = Yes, 2 = No, 3 = Unknown |
| ***income*** | Continuous | Customer’s income per annum in dollars |
| ***age*** | Discrete | Age last birthday |
| ***PostCode*** | Nominal | Post code of customer’s current address |
| ***nchildren*** | Ordinal/Discrete | Number of children of the customer |
| **Basket Items** | Binary | Supermarket items: 0 = not purchased, 1 = purchased |

***Table 1 Variable Description of Original Data***

## Data Preprocessing

**ReceiptID**:- There are 9 duplicate values of *ReceiptID*. These duplicate values were seen in only ReceiptID and not in other variables. Since this variable is supposed to be a primary key for this data, we will exclude this field from the analysis. ***[Appendix IV]***

**Value**: - There are no missing values in this column. But extreme outliers have been observed which are *$802.0592, $1242.9862 and $1967.6968*. These observations could be explained by saying that these transactions could be the orders for a big party or an occasion. Extreme outlier can cause problems in clustering process, So these values were replaced by the median value($63.58) to retain the original distribution of the variable. ***[Appendix V]***

**nchildren**: - There were 2 missing values initially. The range for this variable is 0-105. The shocking range led me to dig deeper into the data and I found that transactions which had value from 11-105 were also invalid for *pmethod* and *homeown* variable. After sorting the data using either of the variables I have come to know that the increase in these variables is at the same rate and these values are almost levelled. This indicates that whatever caused these invalid values for *nchildren* also cause invalid entried for pmethod and homeown variable. This raises a question of the data quality. ***[Appendix VI]***

**homeown**: - 99 invalid entries(values greater than 3), no missing values were found. ***[Appendix VII]***

**pmethod**: - 97 invalid entried(values greater than 4), no missing values were found. ***[Appendix VIII]***

The same course of action was taken for variables *nchildren*, *homeown* and *pmethod*, which was to remove all the transactions with the invalid entries as these transactions represent only 0.17% of the total transactions. After removing the invalid transactions, there were still 2 missing values in *nchildren* which were replaced by the median value which is 1.

**income**: - One missing value and one extreme outlier($650235.25), both values were replaced by the median income($70169.026563) to retain the distribution. There was break in income from $120,000 to $130,000 and then there were many observations which income greater than $130,000 which is not usual but all the other entries are valid so I let it be. ***[Appendix IX]***

**age**: - has one missing value which was replaced with the median age of 37 years. There are several records having float values(with decimal places), since I didn’t know how the data was collected I left the values as they were. The range for age is from 10-95 years. On closely exploring the age variable I have found out that lot of young customers(below 16 years) were listed as house owerns, having multiple children and earning high salaries which does not make any sense so I decided to remove all the entries where age of the customer is less than 16. Due to what I have found out about the variable, I warn the readers that the analysis result could be biased or inaccurate given the bad data quality. ***[Appendix X]***

**PostCode**: - 9792 missing values which represent 17% of the data because of that I decided to keep the records as I am not going to exclude *PostCode* from analysis. There are plenty of invalid entries (alphanumeric values) for *PostCodes*.

**Basket Items**: - *PizzaBase*, *milk*, *confectionary*, and *cannedveg* have 1 missing value, *cereal* has 9 and *Fruitjuice* has 10 missing values. *fruit* has 10 invalid entries, two of which were “o” and the others were values greater than 1. The “o” were probably intended to be zeros, so I replaced the values by 0. The rest of the values(>1) were replaced by 1 as it seemed that the error was due to customers purchasing more than 1 quantity of the item. All other missing values were replaced by 0(the most frequent category). ***[Appendix XIV]***

**Variable Created**

**Pet Owner**: - It was created using variables ‘*dog* *food’* and ‘*cat* *food’* to check how many customers own a pet(s) or not.

**parent**: - This variable was created using variable *nchildren* to discover how many customers are parent.

**NOTE**: - I have created new variable for I have created new variables for categorical data(*sex*, *pmethod*, *nchildren*) just for visualization purpose named *sex*\_*processed*, *pmethod*\_*processed* and *nchildren*\_*processed*. So, after all the data preprocessing there are 57,622 rows and 59 columns in the data. ***[Find visualizations for all necessary variables in the appendix]***

## Consumer Profile

Coles customers’ age range from 16 to 95 years averaging around 39 years. The median annual income is $70,170 with 50% of the customers earning between $65,623 and $75,324. There seems to be a gap between income $120,000 and $130,000 with no apparent reason that justifies it.

At least 60% of the customers are female and they spend median amount of $68.17 per transaction, where males spend $56.72 which indicates that generally women take care of the grocery. The median amount spent by all the customers on is around $63.28.

At least 73% of the customers own a house and they spend 16% less median value than custoemers who own a house and 70% of the customers are parents. Generally customers have 1, 2 or 3 children. There are very few occurrences where customers have more than 3 children.

The payment preferences of the customers are(in descending order): Card 42.49%, EFTPOS 30.5%, Cash 14.4%.

**FUN FACT**: - Around 31%(18,076) of the total customers own a pet(s). Out of those 18,076 customers 53% own a dog(s), 37% own a cat(s) and 10% own both a dog(s) and a cat(s).

The **MOST** purchased items were: The **LEAST** purchased items were

1. Bread, 82.85% of the times 1. KitKat, 1.64% of the times

2. Milk, 81.34% of the times 2. energydrink, 1.85% of the times

3. Cereal, 76.35% of the times 3. frozen fish, 2.94% of the times

4. Banana, 76.18% of the times 4. TeaTowel, 3.70% of the times

5. Lettuce, 74.31% of the times 5. Icecream, 4.36% of the times

Find visualization related to basket items in *Appendix XIV*. All other visualizations and summary statistics can be found in ***Appendix***.

# **Methodology**

## Market Basket Analysis(MBA)

The purpose of MBA is to discover patterns in customer purchases that could be useful to the retailer and help determine the correct stock levels, product placements on shelves, catalogue design and strategy for the next marketing campaign and target audience. The most useful weapon in the data mining arsenal is generating association rules using MBA when it comes to Transactional data.

Apriori algorithm is widely used to generate association rules. These rules are generated on the basis of how frequent a product is in the data. I used package *mlxtend* in Python to generate rules. The algorithm uses following metrics and terminology:

* **Rules**: - Given a rule "A -> C", A stands for antecedent(Item A) and C(Item C) stands for consequent.
* **Antecedent Support**: - It computes the proportion of transactions that contain the antecedent A.
* **Consequent Support**: - It computes the support for the itemset of the consequent C.
* **Support**: - Support is used to measure the frequency (often interpreted as significance or importance) of an itemset in a database(all the transactions here).
* support(A -> C) = support(A ∪ C)
* The 'support' metric then computes the support of the combined itemset A ∪ C -- '**'support' depends on 'antecedent support' and 'consequent support' via min('antecedent support', 'consequent support')**
* **Frequent Itemset**: - We refer to an itemset as a "frequent itemset" if you support is larger than a specified minimum-support threshold. Due to the downward closure property, all subsets of a frequent itemset are also frequent.
* **Confidence**: - The confidence of a rule A -> C is the probability of seeing the consequent in a transaction given that it also contains the antecedent.
* confidence(A -> C) = support(A -> C) / support(A)
* **This metric is not symmetric or directed; for instance, the confidence for A -> C is different than the confidence for C -> A.** The confidence is 1 (maximal) for a rule A->C if the consequent and antecedent always occur together.
* **Lift**: - The lift metric is commonly used to measure how much more often the antecedent and consequent of a rule A -> C occur together than we would expect if they were statistically independent.
* lift(A -> C) = confidence(A -> C) / support(C)

Before running the algorithm, we need to set threshold parameters. We set minimum antecedent support at 10%, which means that the antecedent product(or product set) must have occurred in the data at least 10% of the times. 10% threshold support seems rather low but since the data set is large (57,622 transactions), the probability of occurrence is relatively small. *[Threshold for lift has been set to 1]*

# **Appendix**

**[Red cell color suggests that data is dirty or missing in that column]**

**[Numbers in Quality of Data for one variable is independent of other variables]**

## Appendix I – Transactional variables

|  |  |  |  |
| --- | --- | --- | --- |
| **Transactional Data** | | | |
| **Variable** | **Type** | **Description** | **Quality of Data** |
| ReceiptID | Numeric - Unique Key | Unique transaction ID | 9 duplicate values |
| Value | Numeric - Continuous | Value of the transaction | No missing values, outliers detected |
| pmethod | Numeric - Categorical | Payment Method  (1 = Cash, 2 = Credit card, 3 = EftPOS, 4 = Other) | 97 erroneous entries [≈0.17%] |

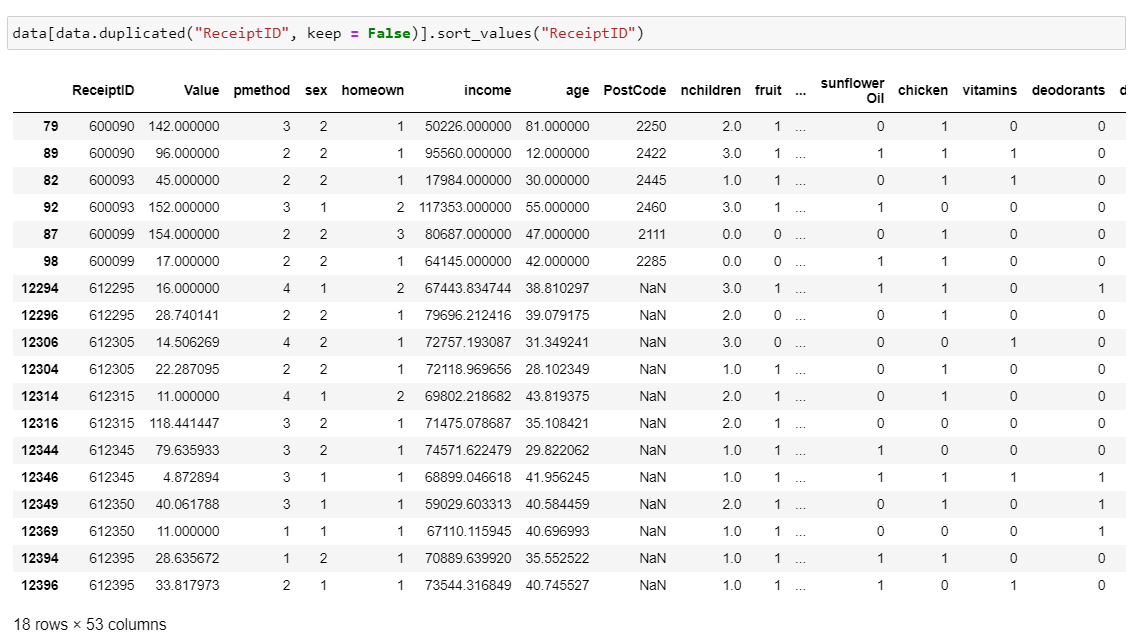
## Appendix II –Demographical and Socio-Economical Features

|  |  |  |  |
| --- | --- | --- | --- |
| **Demographical & Socio-Economical Data** | | | |
| **Variable** | **Type** | **Description** | **Quality of Data** |
| sex | Numeric - Categorical- Binary | Customer’s Gender  (1 = Male, 2 = Female) | No missing values or outliers |
| homeown | Numeric - Categorical | House Ownership  (1 = Yes, 2 = No, 3 = Unknown) | 99 erroneous entries [≈0.17%] |
| income | Numeric - Continuous | Customer’s Income in dollars (Per Annum) | 1 missing value [≈ 0.0017%], Outliers detected |
| age | Numeric - Continuous | Customer’s Age | 1 missing value [≈ 0.0017%], Outliers detected |
| PostCode | String - Categorical | Customer’s Postal Code | 9792 missing values + Erroneous entries [≈17%] |
| nchildren | Numeric - Discrete | No of Children that a customer has | 2 missing value + Erroneous entries detected [≈0.19%] |

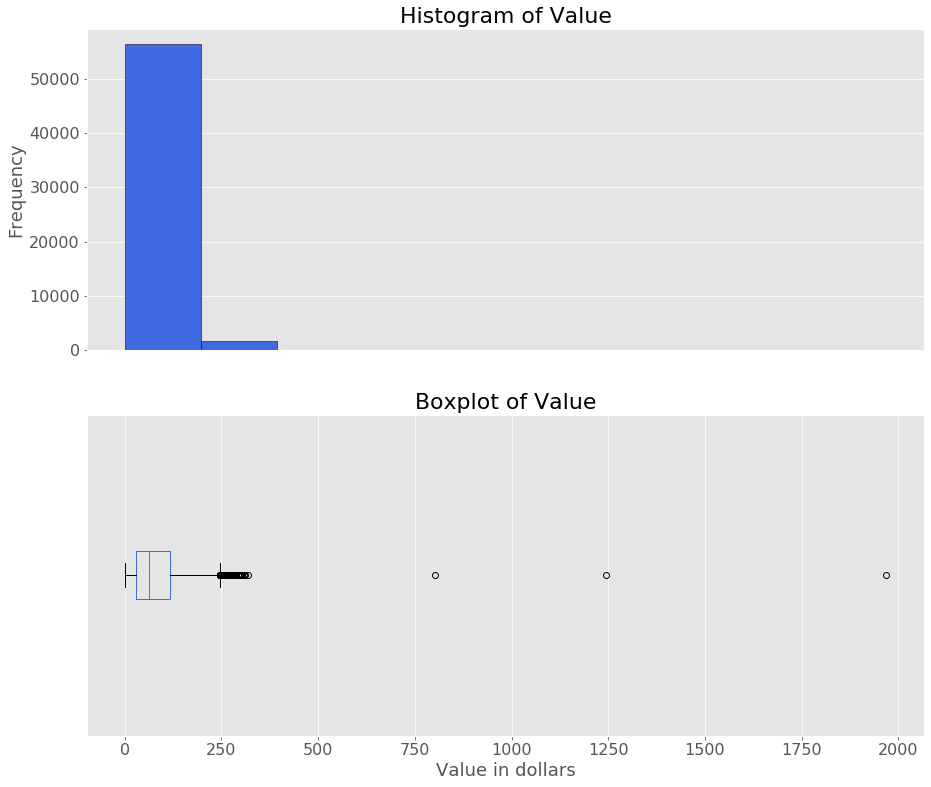
## Appendix III – Basket Items

|  |  |  |
| --- | --- | --- |
| **Basket Items** | | |
| **Variable** | **Type** | **Quality of Data** |
| fruit | String | 10 erroneous entries [≈0.017%] |
| freshmeat | Binary | No missing values or outliers |
| dairy | Binary | No missing values or outliers |
| MozerallaCheese | Binary | No missing values or outliers |
| cannedveg | Binary | 1 missing value [≈ 0.0017%] |
| cereal | Binary | 9 missing values [≈0.015%] |
| frozenmeal | Binary | No missing values or outliers |
| frozendessert | Binary | No missing values or outliers |
| pizzabase | Binary | 1 missing value [≈ 0.0017%] |
| TomatoSauce | Binary | No missing values or outliers |
| frozen fish | Binary | No missing values or outliers |
| bread | Binary | No missing values or outliers |
| milk | Binary | 1 missing value [≈0.0017%] |
| softdrink | Binary | No missing values or outliers |
| fruitjuice | Binary | 10 erroneous entries [≈0.017%] |
| confectionary | Binary | 1 missing value [≈ 0.0017%] |
| fish | Binary | No missing values or outliers |
| vegetable | Binary | No missing values or outliers |
| energydrink | Binary | No missing values or outliers |
| tea | Binary | No missing values or outliers |
| coffee | Binary | No missing values or outliers |
| laundrypowder | Binary | No missing values or outliers |
| householcleaners | Binary | No missing values or outliers |
| corn chips | Binary | No missing values or outliers |
| Frozen yogurt | Binary | No missing values or outliers |
| Chocolate | Binary | No missing values or outliers |
| Olive Oil | Binary | No missing values or outliers |
| Baby Food | Binary | No missing values or outliers |
| Napies | Binary | No missing values or outliers |
| banana | Binary | No missing values or outliers |
| cat food | Binary | No missing values or outliers |
| dog food | Binary | No missing values or outliers |
| mince | Binary | No missing values or outliers |
| Sunflower Oil | Binary | No missing values or outliers |
| chicken | Binary | No missing values or outliers |
| vitamins | Binary | No missing values or outliers |
| deodorants | Binary | No missing values or outliers |
| dishwashingliquid | Binary | No missing values or outliers |
| onions | Binary | No missing values or outliers |
| lettuce | Binary | No missing values or outliers |
| KitKat | Binary | No missing values or outliers |
| TeaTowel | Binary | No missing values or outliers |
| Scones | Binary | No missing values or outliers |

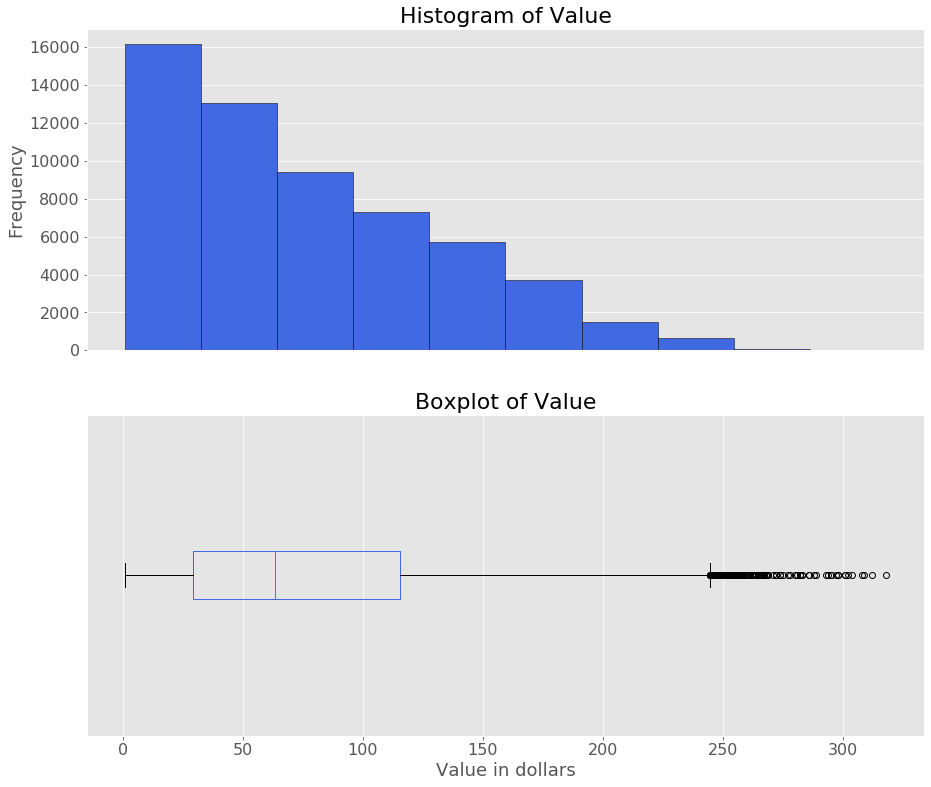
## Appendix IV – ReceiptID[Duplicates]



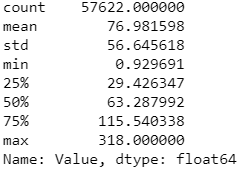
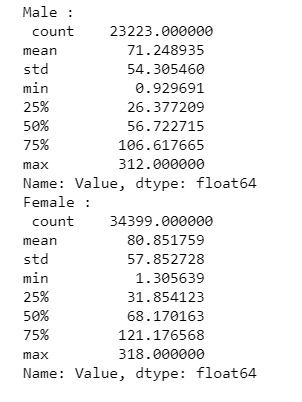
## Appendix V – Value

**Before cleaning**

**After cleaning**

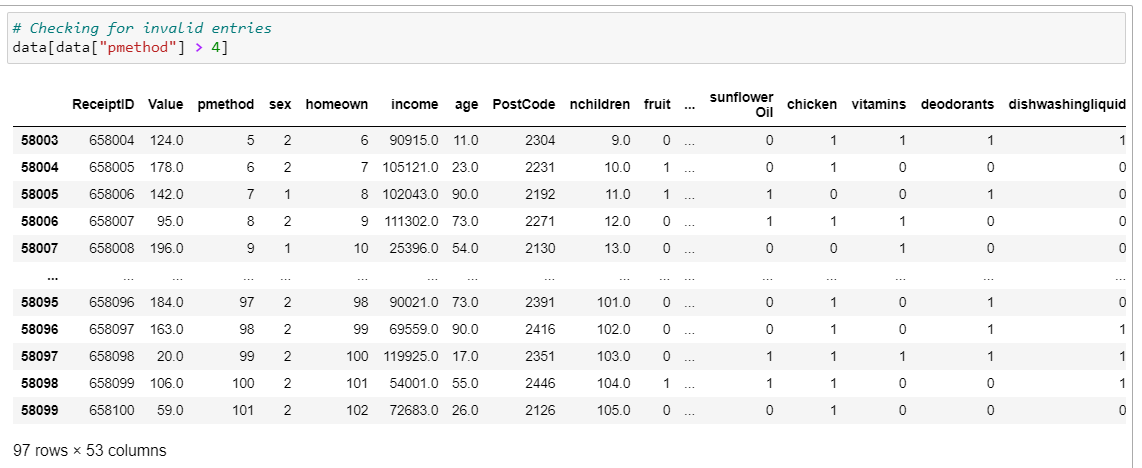
****

**Summary of Value Male VS Female Value**

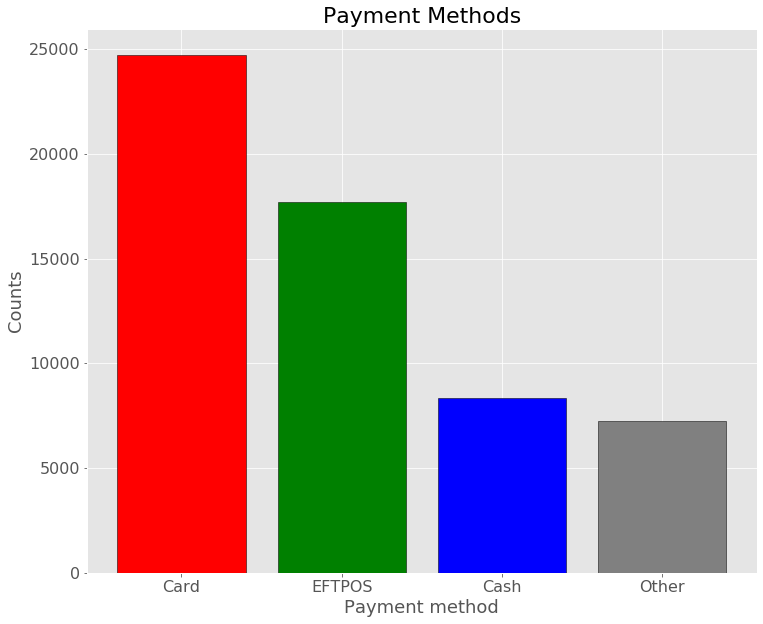
** **

## Appendix VI – pmethod

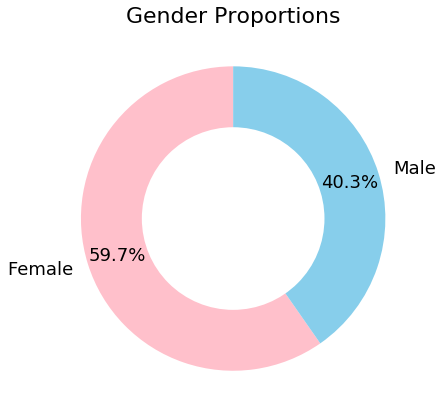
**Before cleaning**

****

**After Cleaning**

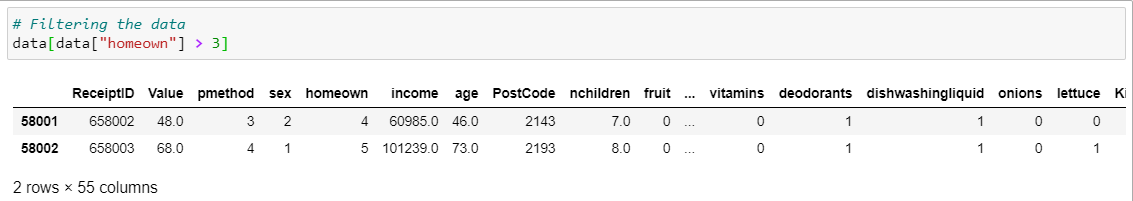
****

## Appendix VII – sex

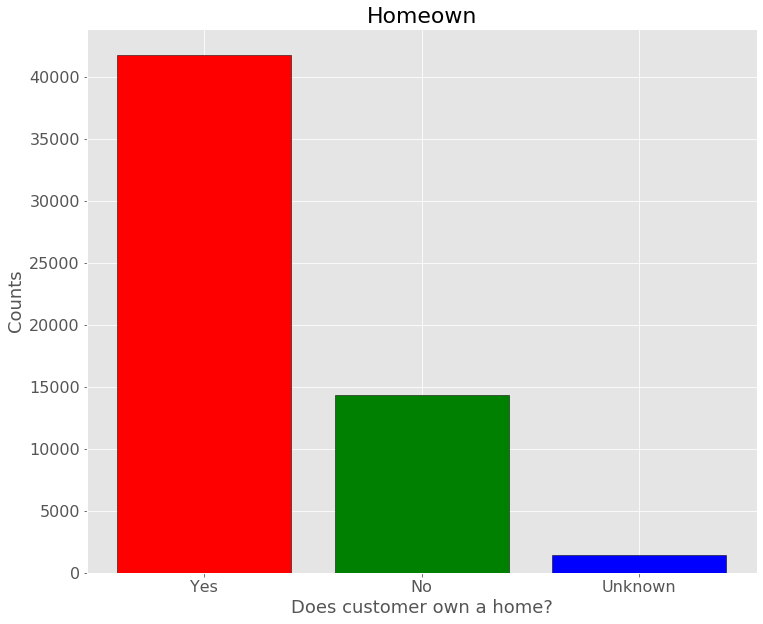
****

## Appendix VIII – homeown

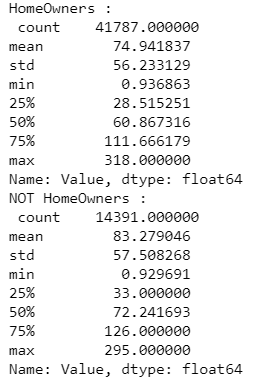
**Before cleaning**

****

**After cleaning**

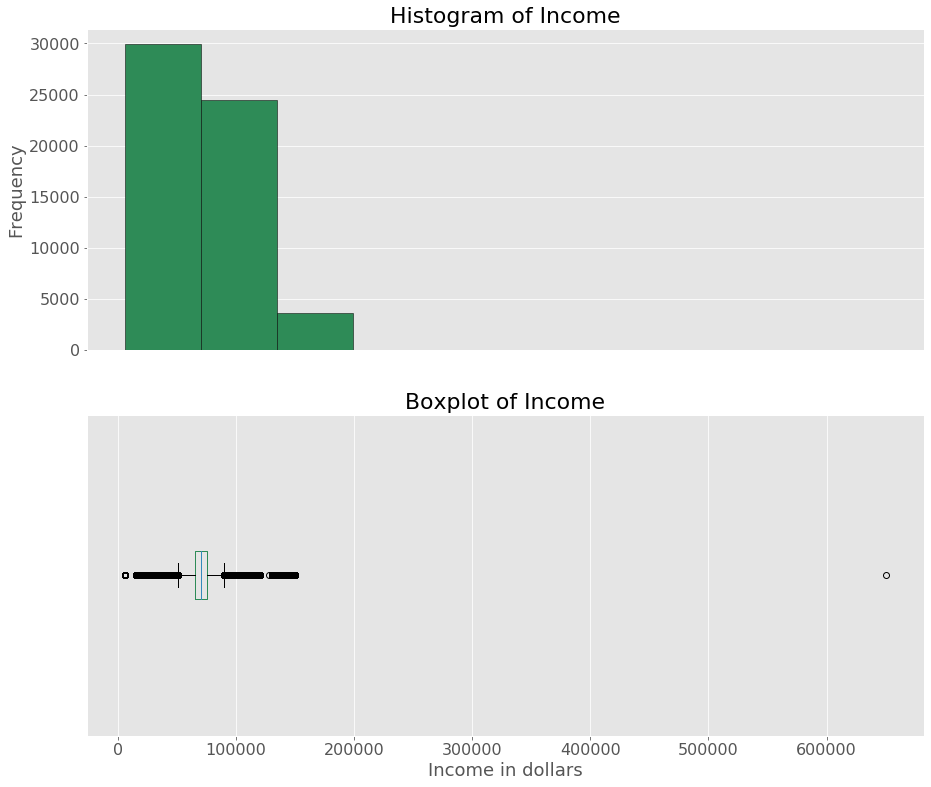
****

**Value for Homeowners VS Not Homeowners**

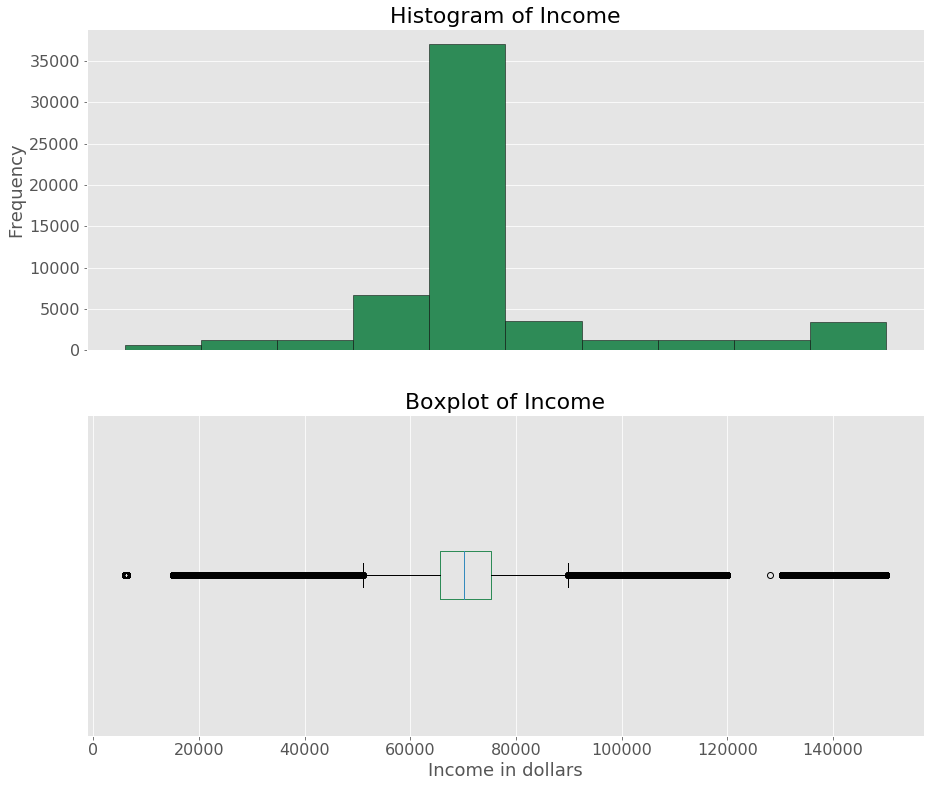
****

## Appendix IX – income

**Before cleaning**

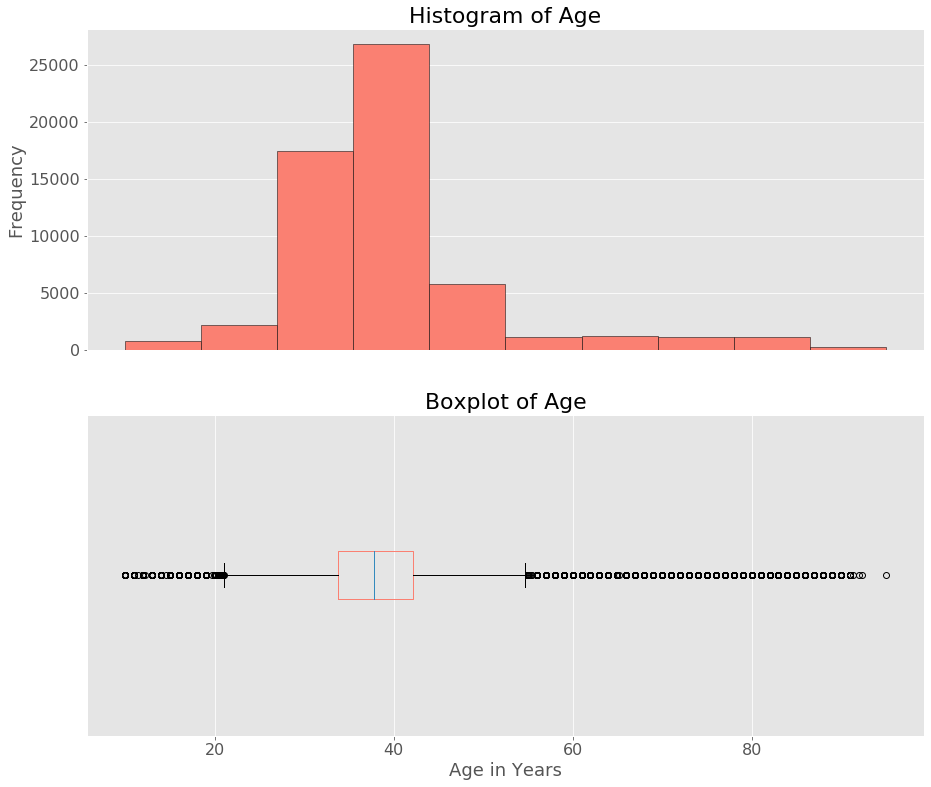
****

**After cleaning**

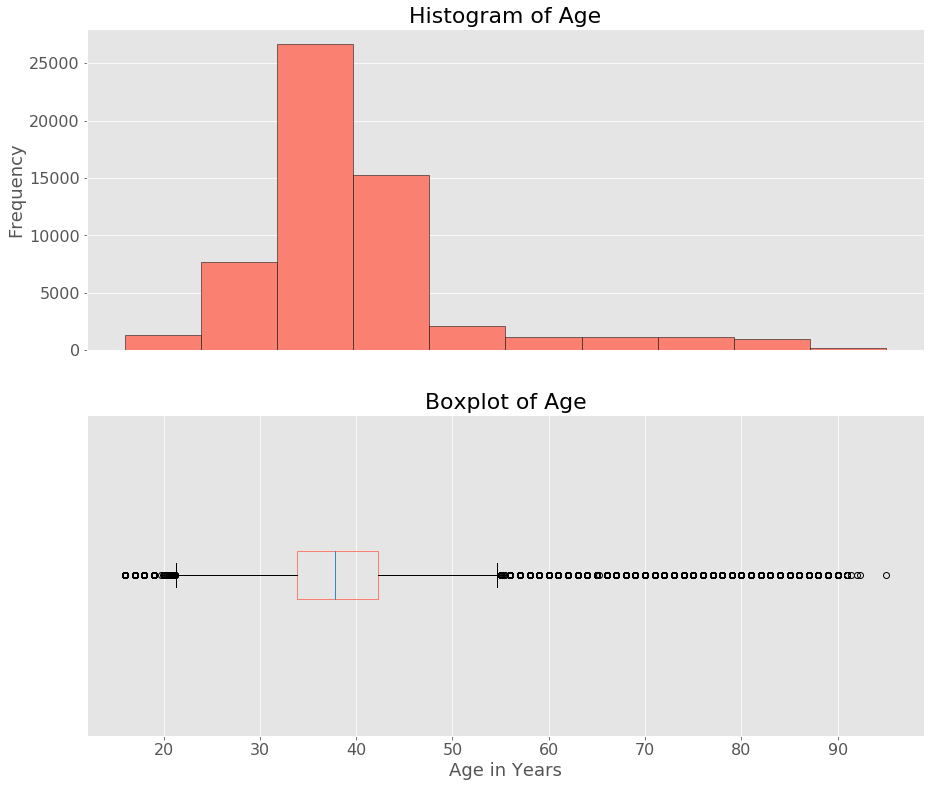
****

## Appendix X – age

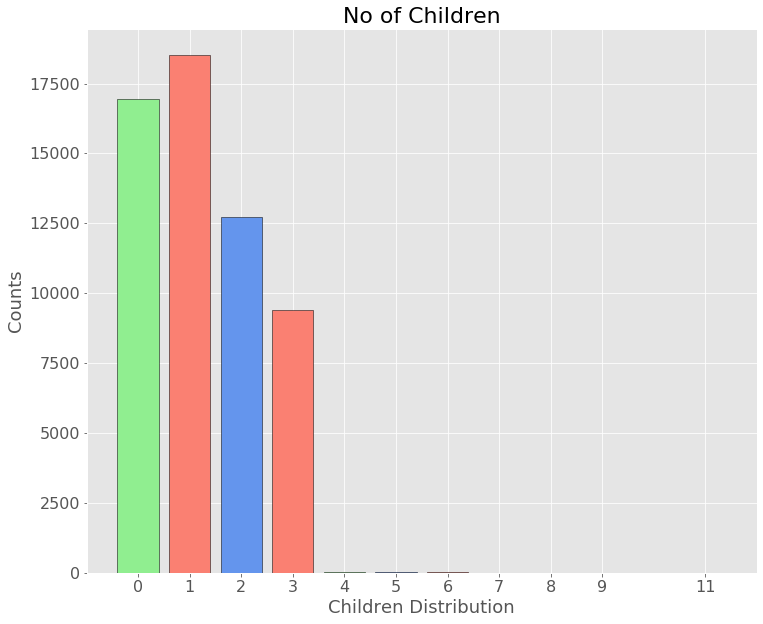
**Before cleaning**

****

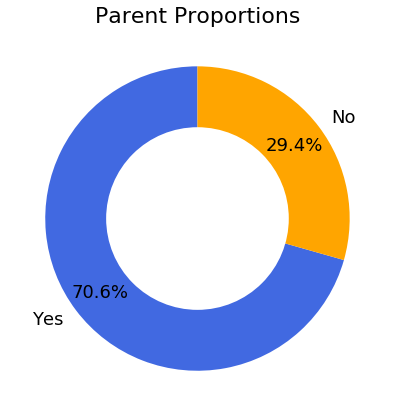
**After cleaning**

****

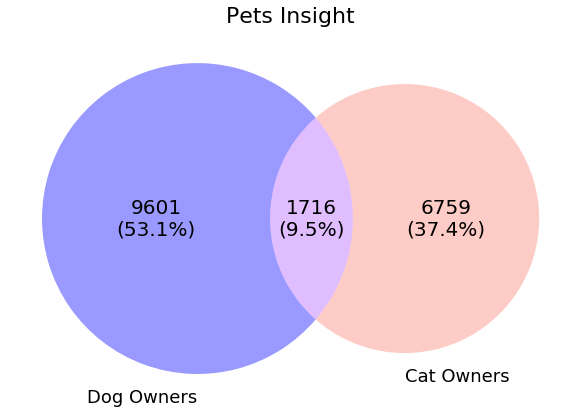
## Appendix XI – nchildren

****

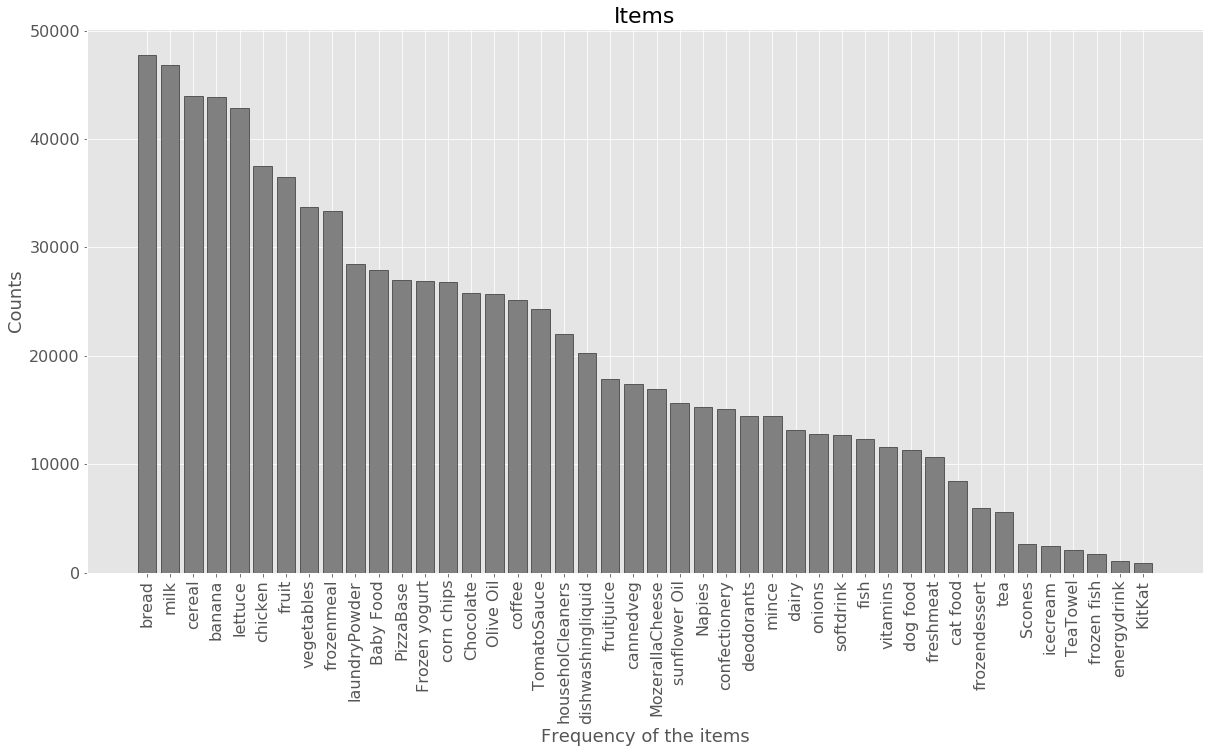
## Appendix XII – parent

****

## Appendix XIII – Pet Owner



## Appendix XIV – Basket items



## Appendix XV – Summary for Numerical Variable

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Min** | **25%** | **50%** | **75%** | **Max** | **Mean** | **Std.Dev** |
| ***Value(in $)*** | 0.92 | 29.43 | 63.29 | 115.54 | 318.00 | 76.98 | 56.64 |
| ***income(in $)*** | 6000.23 | 65,623.48 | 70,170.42 | 75324.32 | 149981.00 | 74884.04 | 23761.12 |
| ***age(in years)*** | 16 | 33.84 | 37.83 | 42.24 | 95 | 39.87 | 11.40 |

## Appendix XVI – Summary for Categorical Variable

|  |  |
| --- | --- |
| **Variable** | **Frequency – Percentage** |
| ***pmethod*** | 1 = Cash - 24,488 – 42.5%  3 = Eftpos – 17,575 – 30.5%  2= Card – 8,310 – 14.4%  4 = Other – 7,249- 12.6% |
| ***sex*** | 2 = Female – 34,399 – 59.7%  1 = Male – 23,223 – 40.3% |
| ***homeown*** | 1 = Yes – 41,787 – 72.5%  2 = No – 14,391 – 25%  3 = Unknown – 1,444 – 2.5% |
| ***parent*** | Yes – 40,683 – 70.6%  No – 16,939 – 29.4% |
| ***Pet Owner*** | No – 39,546 – 68.6%  Yes – 18,076 – 31.4% |
| ***nchildren*** | 1 = 18,519 – 32%  0 = 16,939 – 29%  2 = 12,712 – 22%  3 = 9,403 – 16%  4 = 16 – 0.0%  5 = 14 – 0.0%  6 = 12 – 0.0%  7 = 4 – 0.0%  8 = 1 – 0.0%  9 = 1 – 0.0%  11 = 1 – 0.0% |