MARKET BASKET ANALYSIS
SIMULATED COLES DATA

Table of Contents

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**, **All age & Broke** and **Rich & Wise**. 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 in the spreadsheet(.xlsx) format. The spreadsheet consists of:

- 58,100 observations(rows)
- 53 variables(columns)
  - 3 Transactional 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 the 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 outliers can cause problems in the 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 that 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 leveled. This indicates that whatever caused these invalid values for *nchildren* also caused invalid entries for pmethod and homeown variable. This raises the question of 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 a 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 a lot of young customers(below 16 years) were listed as house owners, having multiple children and earning high salaries which do not make any sense so I decided to remove all the entries where the 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 the analysis. There are plenty of invalid entries(alphanumeric values) for *PostCodes*.

Basket Items: - *PizzaBase*, *milk*, *confectionery*, 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 parents.

NOTE: - I have created a new variable for I have created new variables for categorical data(*sex*, *pmethod*, *nchildren*) just for visualization purposes 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 a 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 customers 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:

1. Bread, 82.85% of the times
2. Milk, 81.34% of the times
3. Cereal, 76.35% of the times
4. Banana, 76.18% of the times
5. Lettuce, 74.31% of the times

The **LEAST** purchased items were

1. KitKat, 1.64% of the times
2. energydrink, 1.85% of the times
3. frozen fish, 2.94% of the times
4. TeaTowel, 3.70% 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 the ***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, catalog 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 the 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]*

Clustering

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). It is an unsupervised classification technique, which means that classes are not predefined. I intend to uncover the customer segmentation based on their features like spending habits, income, and age that's why I chose clustering to analyze data.

There are many clustering techniques: Hierarchical Clustering which decomposes the data using either divisive or agglomerative strategies. Generally, they are preferred as they don't require the number of clusters o be specified in advance. However, our data set is too big for such an algorithm and it would take a significant amount of time. So, I chose KMeans for clustering as it would be much more efficient. There are also many ways to carry out KMeans depending upon which distance metric we use to create clusters but here I'll be using Euclidean distance.

The KMeans clustering method is a partitioning algorithm. Once we specify the number of clusters in the data set, the algorithm will randomly select a centroid for each cluster(generally points from data set), find the closest data points and assign them to the centroid, recalculate the centroid(by averaging all the data points in the cluster). The above steps are repeated until there is no change in the centroids. There are many limitations of this KMeans, one of them is this can only be applied to the continuous variable. I wanted to group customers based on age, spending, and income and all the variables are continuous so we can perform KMeans easily. Another problem with Kmeans is that the centroids are sensitive to the magnitude of the values as they are calculated by taking a mean of the data points. So before I apply Kmeans I will standardize variables(z-scores) as income, age and Value are not on equal footing and this can cause inefficient results sometimes. For example for 4 clusters, variable income has a larger value, it is going to dominate the cluster results and clusters are going to form based on different values of income***[Appendix XVII]***. As we can see there are 4 visible groups in the income, because income has larger values compared to other variables.

How many clusters?: - We can determine the optimal number of clusters by plotting the graph of a number of clusters versus the withing sum of squares. Lesser the withing sum of square better the clusters are.

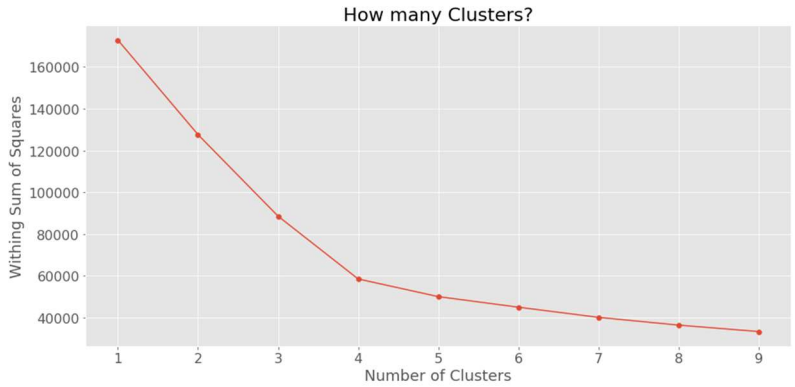


Figure 1 – How many clusters?

Based on the graph on the left side, there is an obvious elbow at 4 clusters. Here we can't choose 9 cl usters even though the within sum of square is the least for it. We have to choose an optimal number of clusters, choosing 9 clusters would make interpretation of the clusters difficult. So, I am choosing 4 clusters as it has balanced within sum of squares.

Results

Market Basket Analysis(MBA)

There were a plethora of association rules generated by the Apriori algorithm. I have divided them into 3 groups: Rules with high support, high confidence and high lift.

HIGH SUPPORT

Rule	Antecedent	Consequent	Support ↑	Confidence	Lift
S1	bread	cereal	63.90	77.13	1.01
S2	banana	bread	63.87	83.85	1.01
S3	cereal	milk	62.22	81.49	1.00
S4	milk	banana	62.08	76.32	1.00

Table 2 Rules With High Support

Bread & Cereal is the most popular combination of products being bought 63.9% of the time. The second and third most popular combinations of the products are Bread & Banana(63.8%) and Milk and Cereal(62.2%). These rules are also called the most obvious rules as they are the most common items bought by the customers. These rules don't have much value to us.

NOTE: - *Due to the high frequency of some products like bread, milk, cereal, banana, and lettuce, they were appearing in almost all the rules. I have decided to drop these 5 items so that I can discover the relationships between products that are not obvious.*

HIGH CONFIDENCE

Rule	Antecedent	Consequent	Support	Confidence↑	Lift
C1	(Napies, TomatoSauce)	(Baby Food)	10.33	95.81	1.98
C2	(Olive Oil, Napies)	(Baby Food)	10.72	94.26	1.95
C3	(coffee, householCleaners, frozenmeal)	(vegetables)	13.75	93.98	1.61
C4	(coffee, TomatoSauce, householCleaners)	(Vegetables)	62.08	76.32	1.00

Table 3 Rules With High Confidence

The probabilities that Customers will buy Baby Food are 95.8% and 94.2% when they buy (Napies & TomatoSauce) and (Napies and Olive Oil). Similarly, probabilities that customers will buy Vegetables are 93.9% and 93.5% when they buy (householCleaners, frozenmeal & coffee) and (householCleaners, TomatoSauce & coffee). However, these combinations of products are less likely to occur(10% - 14%).

Rules C1 and C2 are rather strong - they both have high lift and confidence indicating that customers who buy antecedent products are 1.9 times more likely to buy the consequent products. Rules C3 and C4 are less obvious relationships. Although the lift for these rules is low, coles could probably increase sales by advertising/marketing householCleaners, coffee, frozenmeal, TomatoSauce with Vegetables(if possible stock them near).

HIGH LIFT

Rule	Antecedent	Consequent	Support	Confidence	Lift↑
L1	(fish, vegetables)	(householCleaners)	10.82	85.77	2.24
L2	(fruit, fish)	(householCleaners)	10.32	79.06	2.07
L3	(Napies, TomatoSauce)	(Baby Food)	10.33	95.82	1.98
L4	(Napies, Olive Oil)	(Baby Food)	10.72	94.23	1.95
L5	(Napies, Chocolate)	(Baby Food)	10.69	92.94	1.92
L6	(Napies, fruit)	(Baby Food)	14.94	92.85	1.92

Table 4 Rules With High Lift

Rule L1 - As we can see, customers who bought Vegetables & Fish are 2.2 times more likely to buy householCleaner. These items were bought together for like 10% of the time but 85% of the customers who bought Vegetables & Fish also bought householCleaners. Customers who bought fish and fruit together are 2 times more likely to but householCleaners with low support and high confidence. These products need to be marketed together or perhaps give a discount on householCleaner to the customer who buys vegetables, Fish or Fruit.

Obvious rules – L5 & L6

Parents often buy fruit and chocolates for their young ones. Rules L3 & L4 are not as obvious as L5 & L6 but Coles can still use them. By looking at the rules with Baby Food in consequent(with low support), We can give discounts on Baby Food if customers by (Napies & fruit), (Napies, Chocolate), (Napies, Olive Oil) or (Napies & TomatoSauce) to increase the support of these transactions[If it is not possible to carry out marketing for all the combinations, what Coles can do is it should make sure that these products are as close as possible to Baby Food].

Clustering

After running KMeans on standardized variables z_income, z_value and z_age we get the following clusters.

Cluster	Size	Median Income	Median Value	Median Age	Homeowners (Yes / No)*	Sex (F / M)
1	7.7%	$66,706	$103	72	43% / 56%	70% / 30%
2	27.9%	$69,348	$134	35	72% / 25%	63% / 37%
3	54.8%	$69,434	$37	38	76% / 22%	57% / 43%
4	9.5%	$138,546	$61	37	80% / 18%	59% / 41%

Table 5 Cluster Features

[- rest of % is for category "Unknown"]*

[Values for variables are rounded to near integer]

CLUSTER EXPLANATION

I. Cluster 1 - Elderly High Spenders

1. Customers in cluster 1 are **older** than customers in other clusters having a median age of **72 years** and have **low income**.
2. Despite their low income, Customers in this cluster tend to **spend more** at the grocery store.
3. This cluster has the **highest % of card payments** which is around **53%**.
4. The cluster is **small in size** comprised of 7.7% of the total customers, which might be justified by saying that **old people often avoid** to **leave their house** and drive to the grocery stores. This nature of old people also **justifies their high spending**, because they **tend to buy more groceries at a time** to avoid making more trips to the store which **in turn causes fewer old people**(small size of cluster) visiting the store.
5. Customers in cluster one have **more high number of children(>3)** compare to other clusters, which can be easily linked to their old age. In the past, people tend to have more children compared to now.
6. The majority of the people in this cluster **do not own a house**.
7. This cluster has the **highest % of the female customers** which is **70%**.
8. This cluster has the **highest % of the customer owning a cat as a pet**, which is **45%**.

Basically, this cluster can be seen as **less number of people(old woman) earning high and spending more money**.

II. Cluster 2 - High Spending Youth

1. Customers in cluster 2 are **younger** than customers in other clusters having a median age of **35 years** and have **low income**.
2. Despite their low income, Customers in this cluster tend to **spend more** at the grocery store.
3. The cluster is comprised of **28% of the total customers**.
4. Customers in cluster 2 have **more high number of children(>3 and <9)** compare to other clusters, which can be **easily linked to their high spending**(*more children more spending*).
5. High spending of these customers can be also justified by saying that since customers in this cluster are generally young, they might be **working professionals** and given that they are working they might **not have time to cook** things by themselves so in turn, they **buy more grocery** which probably includes **already cooked meals** which in turn **increases their**

spending at the store. As they are working professional they might **not visit the store more often** and they might **buy groceries for a longer period of time** which is also one of the reasons for their high spending.

6. **The majority** of the people in this cluster **own a house**.
7. **Most** of the customers in this cluster are **female**.

Basically, this cluster can be seen as a **moderate number of people earning low and spending more money**.

III. Cluster 3 – All age & Broke

1. Customers in cluster 3 have a median age of 38 years, the age range of 16-59 years, and have **low income**.
2. Customers in cluster 3 as their income permits they spend a **low amount of money** at the store.
3. This cluster is the **largest cluster** among all the clusters, comprised of 55% of the total customers.
4. Their **low spending** and **large size of clusters** suggest that they **often visit the store** and buy **groceries** and other stuff for a **shorter period of time**.
5. Customers in this cluster have **3 or fewer children** which also reflects their **low spending**. **Most** of the customers in this cluster have **1 child**.
6. Customers in this cluster have the **lowest % of female customers**, which is **56%**.
7. **Most** of the customers are **female** and they **own a house**.

Basically, this cluster can be seen as a **High number of people(of all age) earning less and spending less money**.

IV. Cluster 4 - Rich and Wise

1. **The median** age of the customers in this cluster is **37 years**. But Age range is **16-75 years** with 7 years of standard deviation.
2. Customers in this cluster have **high incomes** compared to customers in the other clusters. **The median income** of this cluster is the **highest** among all the clusters.
3. Even though the customers in this cluster **earn more**, they tend to **spend the moderate**(wise) amount of money at the store.
4. This cluster is quite **small** and comprises of 10% of the customers.
5. **Most** of the customers in this cluster **don't have a child**, which **reflects on their spending** at the store(*fewer Children - less grocery - less spending*).
6. **Not** having **children** can be justified by saying that as they are high-income earners(more career-driven) they are working at higher level jobs(CEOs, CTOs, CFOs and so on) because of high earning.
7. This cluster has the **highest % of cash payments** among all the clusters which is around **15%**.

Basically, this cluster can be seen as a **Low number of people(of all age) earning more and spending a moderate amount of money**.

[Find all the cluster related graphs in Appendix]

Conclusion

- Data Quality – Coles data set raised concerns about data quality – the age, Postcode variable and the sequential invalid entries in other variables like pmethod, nchildren, homeown.

- Despite the data quality issues, it has provided great insights. The key results and recommendations are:

    - Coles should put newly marketed products near the shelf of bread, milk, cereal, and banana So that customers are exposed to the new products. Coles should put these products to the end of the aisle So that customers have to walk more in the stores, in turn, they will be tempted to buy other products they see on their way to the most purchased product aisle.
    - Baby Food should be displayed near the shelves of Napies, Fruit, Chocolate and Olive Oil.
    - Coles could probably increase sales by advertising/marketing householCleaners, coffee, frozenmeal, TomatoSauce with Vegetables(if possible stock them near).
    - Customers who bought Vegetables & Fish and Fruit and Fish are 2.2 and 2.0 times more likely to buy householCleaner respectively. Coles should spatially separate fish, vegetables, fruit, and householCleaners for greater travel distance so that customers will be encouraged to purchase other products.
    - Coles should start marketing campaigns for family or house related products towards young customers(from cluster 3) as the majority of them are home owners and parents. These groups of customers are on a tighter budget and they will respond better to deals and discounts on the branded products that coles has to offer.
    - On the other hand, luxury products should be targeted to customers(from cluster 1) who do not own a house and are females independent of their income and age.

Future Analysis

- Recording customer's postcode adequately would help identify patterns by states or suburbs.
- Recording data of as many customers and as many transactions as coles can by using various methods like using a points card like Flybuys. More data and correct data would help coles get meaningful insights from the data.
- Quantity of items purchased alongside the value would be more helpful in assessing customer's preferences and spending power.
- MBA can be carried out on more generalized product categories for less-frequently purchased products, for example, put all the frozen items in one category that would increase the support of the products.

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]

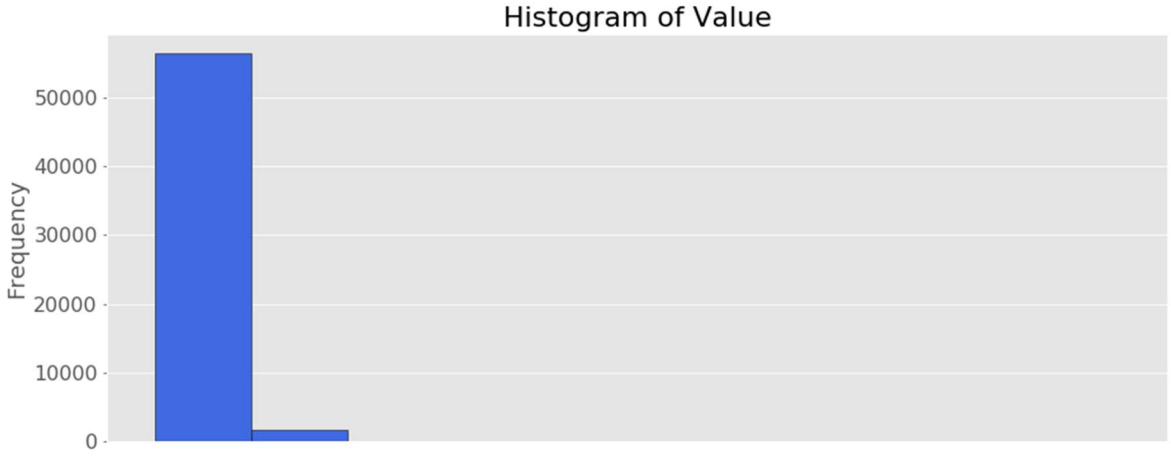
```
data[data.duplicated("ReceiptID", keep = False)].sort_values("ReceiptID")
```

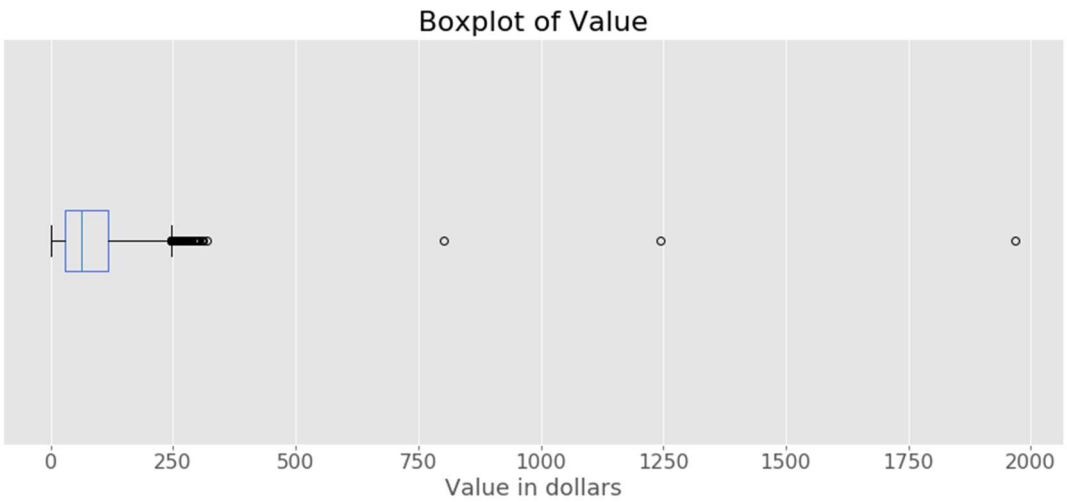
	ReceiptID	Value	pmethod	sex	homeown	income	age	PostCode	nchildren	fruit	...	sunflower Oil	chicken	vitamins	deodorants	d
79	600090	142.000000	3	2	1	50226.000000	81.000000	2250	2.0	1	...	0	1	0	0	
89	600090	96.000000	2	2	1	95560.000000	12.000000	2422	3.0	1	...	1	1	1	0	
82	600093	45.000000	2	2	1	17984.000000	30.000000	2445	1.0	1	...	0	1	1	0	
92	600093	152.000000	3	1	2	117353.000000	55.000000	2460	3.0	1	...	1	0	0	0	
87	600099	154.000000	2	2	3	80687.000000	47.000000	2111	0.0	0	...	0	1	0	0	
98	600099	17.000000	2	2	1	64145.000000	42.000000	2285	0.0	0	...	1	1	0	0	
12294	612295	16.000000	4	1	2	67443.834744	38.810297	NaN	3.0	1	...	1	1	0	1	
12296	612295	28.740141	2	2	1	79696.212416	39.079175	NaN	2.0	0	...	0	1	0	0	
12306	612305	14.506269	4	2	1	72757.193087	31.349241	NaN	3.0	0	...	0	0	1	0	
12304	612305	22.287095	2	2	1	72118.969656	28.102349	NaN	1.0	1	...	0	1	0	0	
12314	612315	11.000000	4	1	2	69802.218682	43.819375	NaN	2.0	1	...	0	1	0	0	
12316	612315	118.441447	3	2	1	71475.078687	35.108421	NaN	2.0	1	...	0	0	0	0	
12344	612345	79.635933	3	2	1	74571.622479	29.822062	NaN	1.0	1	...	1	0	0	0	
12346	612345	4.872894	3	1	1	68899.046618	41.956245	NaN	1.0	1	...	1	1	1	1	
12349	612350	40.061788	3	1	1	59029.603313	40.584459	NaN	2.0	1	...	0	1	0	1	
12369	612350	11.000000	1	1	1	67110.115945	40.696993	NaN	1.0	1	...	0	0	0	1	
12394	612395	28.635672	1	2	1	70889.639920	35.552522	NaN	1.0	1	...	1	1	0	0	
12396	612395	33.817973	2	1	1	73544.316849	40.745527	NaN	1.0	1	...	1	0	1	0	

18 rows × 53 columns

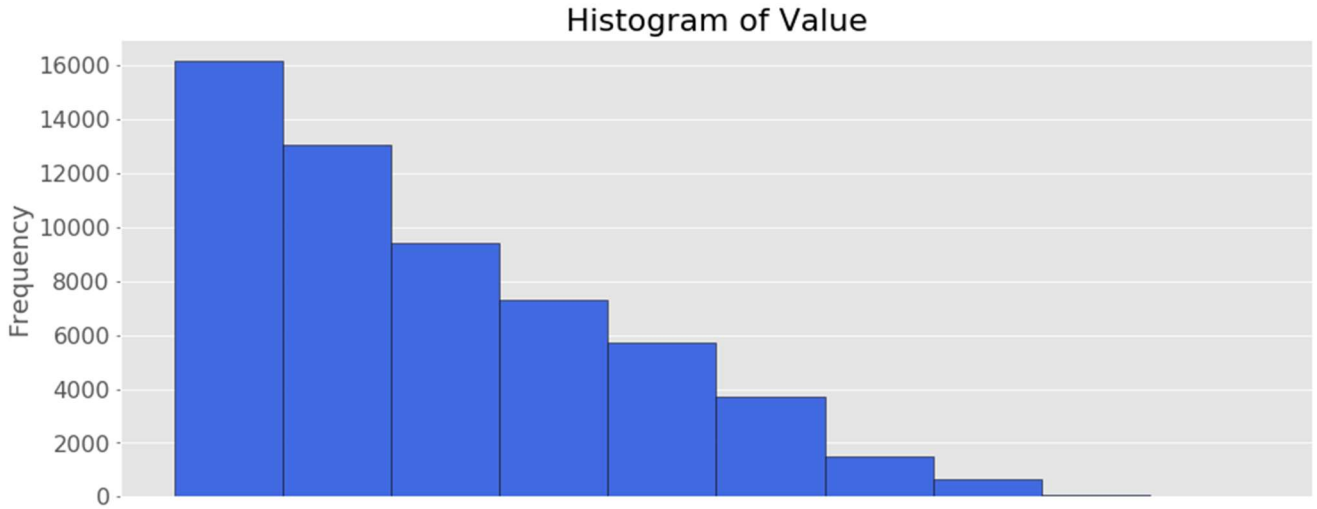
Appendix V – Value

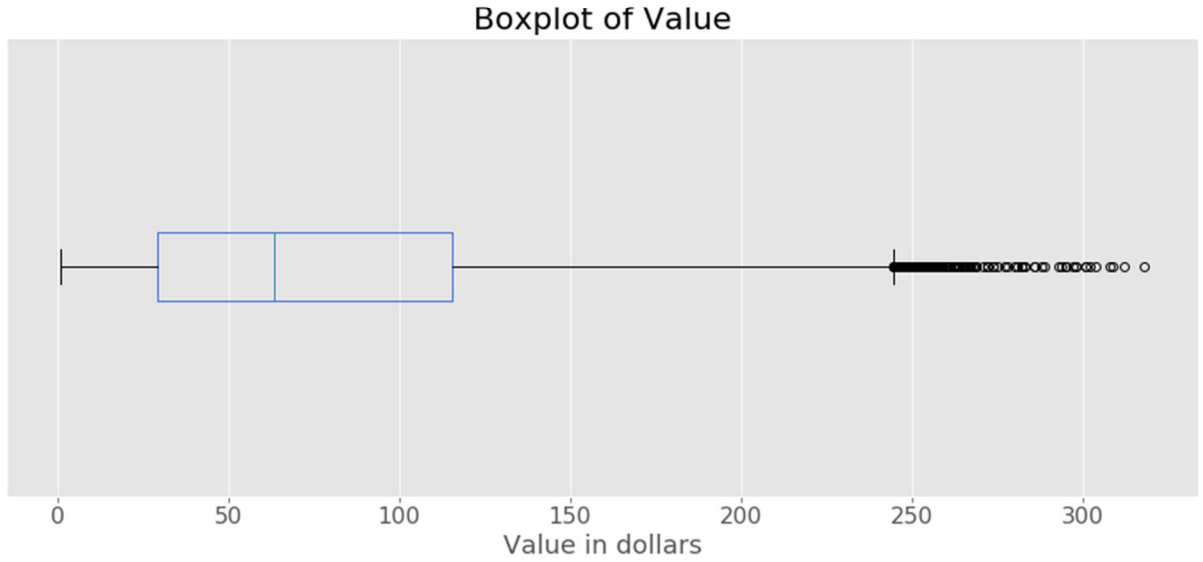
Before cleaning





After cleaning





Summary of Value

```
count    57622.000000
mean        76.981598
std         56.645618
min          0.929691
25%         29.426347
50%         63.287992
75%        115.540338
max        318.000000
Name: Value, dtype: float64
```

Male VS Female Value

```
Male :
 count    23223.000000
mean        71.248935
std         54.305460
min          0.929691
25%         26.377209
50%         56.722715
75%        106.617665
max        312.000000
Name: Value, dtype: float64
Female :
 count    34399.000000
mean        80.851759
std         57.852728
min          1.305639
25%         31.854123
50%         68.170163
75%        121.176568
max        318.000000
Name: Value, dtype: float64
```

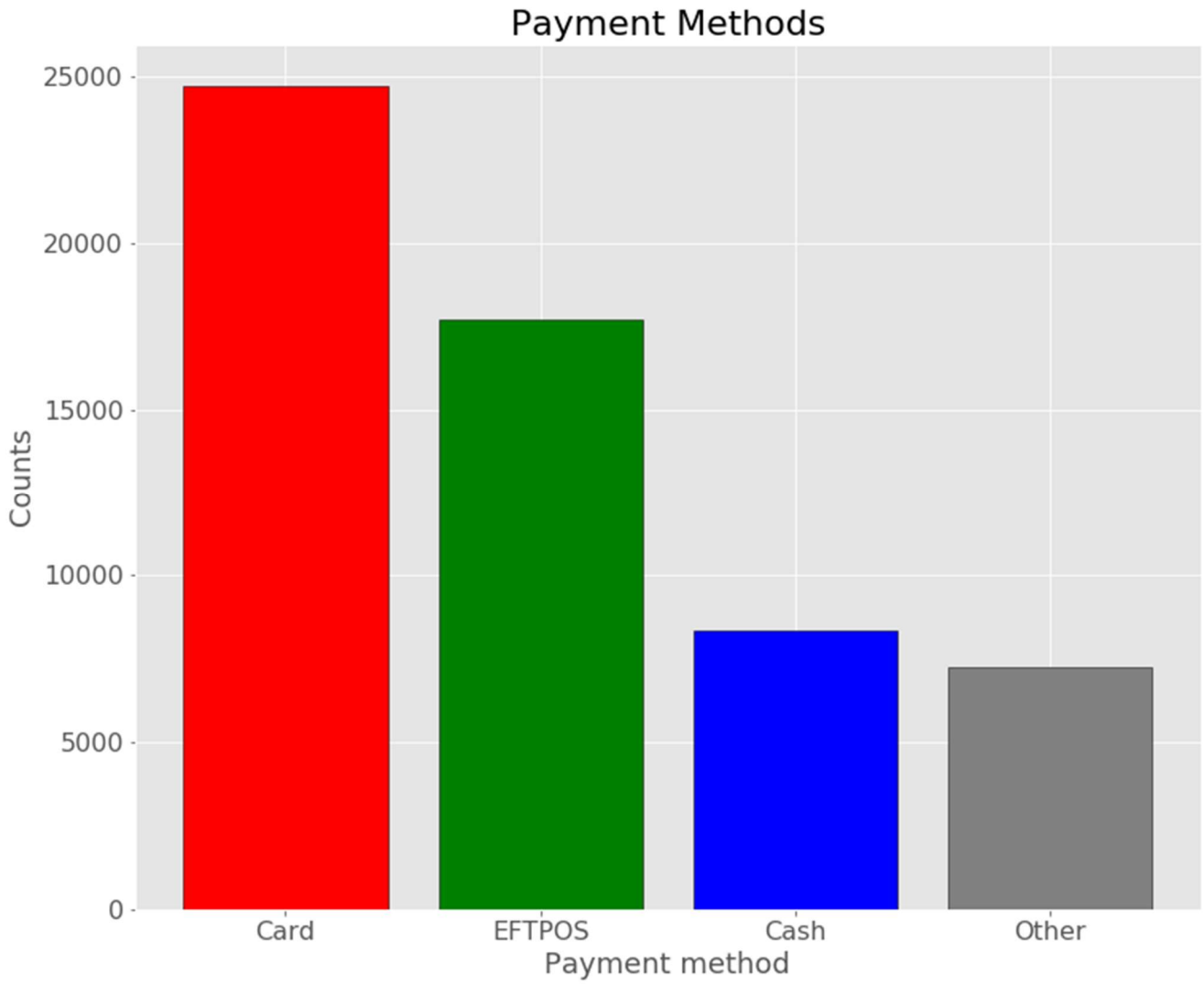
Before cleaning

```
# Checking for invalid entries
data[data["pmethod"] > 4]
```

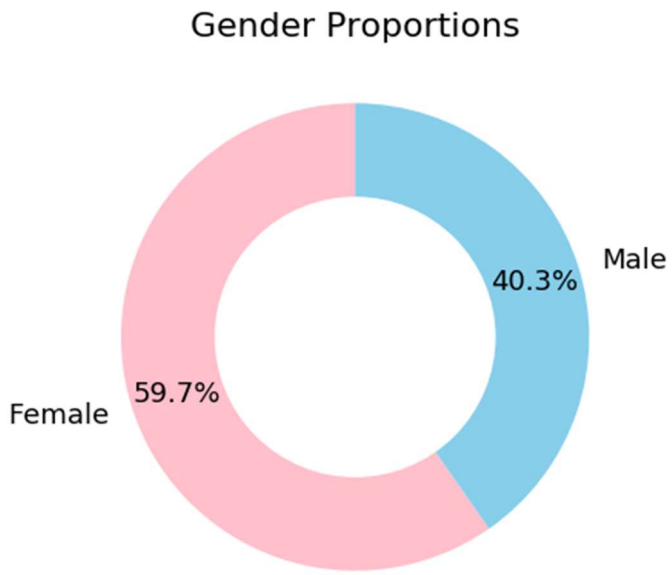
	ReceiptID	Value	pmethod	sex	homeown	income	age	PostCode	nchildren	fruit	...	sunflower Oil	chicken	vitamins	deodorants	dishwashingliquid
58003	658004	124.0	5	2	6	90915.0	11.0	2304	9.0	0	...	0	1	1	1	1
58004	658005	178.0	6	2	7	105121.0	23.0	2231	10.0	1	...	0	1	0	0	0
58005	658006	142.0	7	1	8	102043.0	90.0	2192	11.0	1	...	1	0	0	1	0
58006	658007	95.0	8	2	9	111302.0	73.0	2271	12.0	0	...	1	1	1	0	0
58007	658008	196.0	9	1	10	25396.0	54.0	2130	13.0	0	...	0	0	1	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
58095	658096	184.0	97	2	98	90021.0	73.0	2391	101.0	0	...	0	1	0	1	0
58096	658097	163.0	98	2	99	69559.0	90.0	2416	102.0	0	...	0	1	0	1	1
58097	658098	20.0	99	2	100	119925.0	17.0	2351	103.0	0	...	1	1	1	1	1
58098	658099	106.0	100	2	101	54001.0	55.0	2446	104.0	1	...	1	1	0	0	1
58099	658100	59.0	101	2	102	72683.0	26.0	2126	105.0	0	...	0	1	0	0	0

97 rows × 53 columns

After Cleaning



Appendix VII – sex



Appendix VIII – homeown

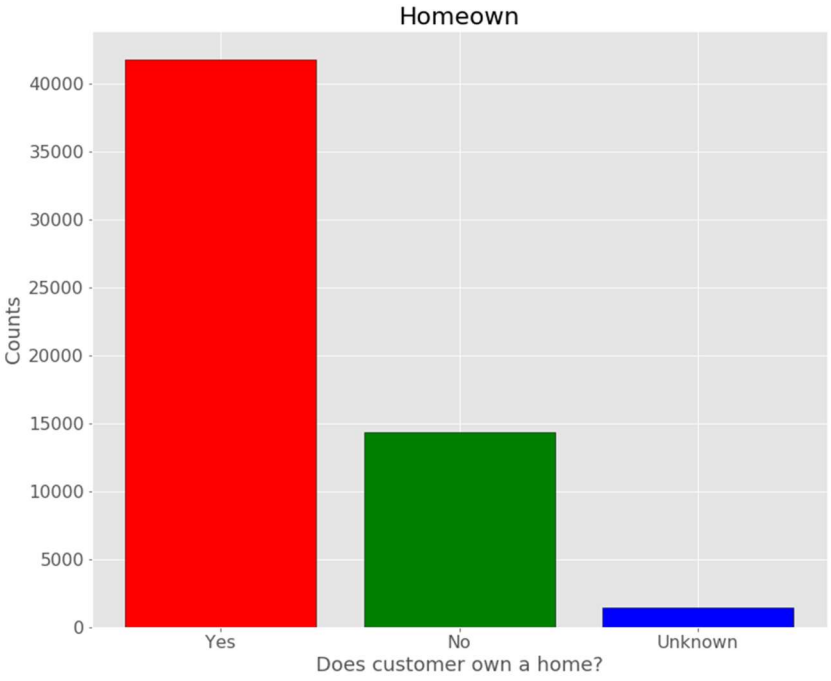
Before cleaning

```
# Filtering the data
data[data["homeown"] > 3]
```

	ReceiptID	Value	pmethod	sex	homeown	income	age	PostCode	nchildren	fruit	...	vitamins	deodorants	dishwashingliquid	onions	lettuce	K
58001	658002	48.0	3	2	4	60985.0	46.0	2143	7.0	0	...	0	1	1	0	0	
58002	658003	68.0	4	1	5	101239.0	73.0	2193	8.0	0	...	0	1	1	0	1	

2 rows × 55 columns

After cleaning

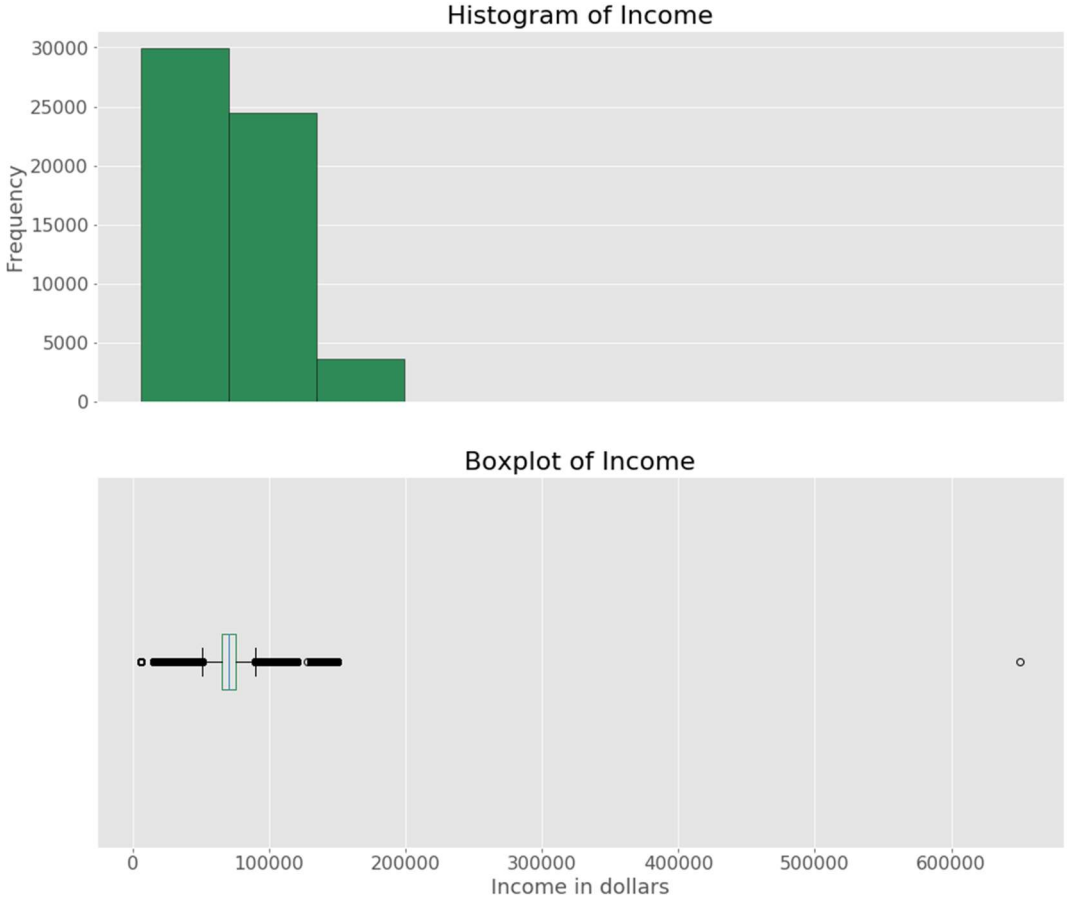


Value for Homeowners VS Not Homeowners

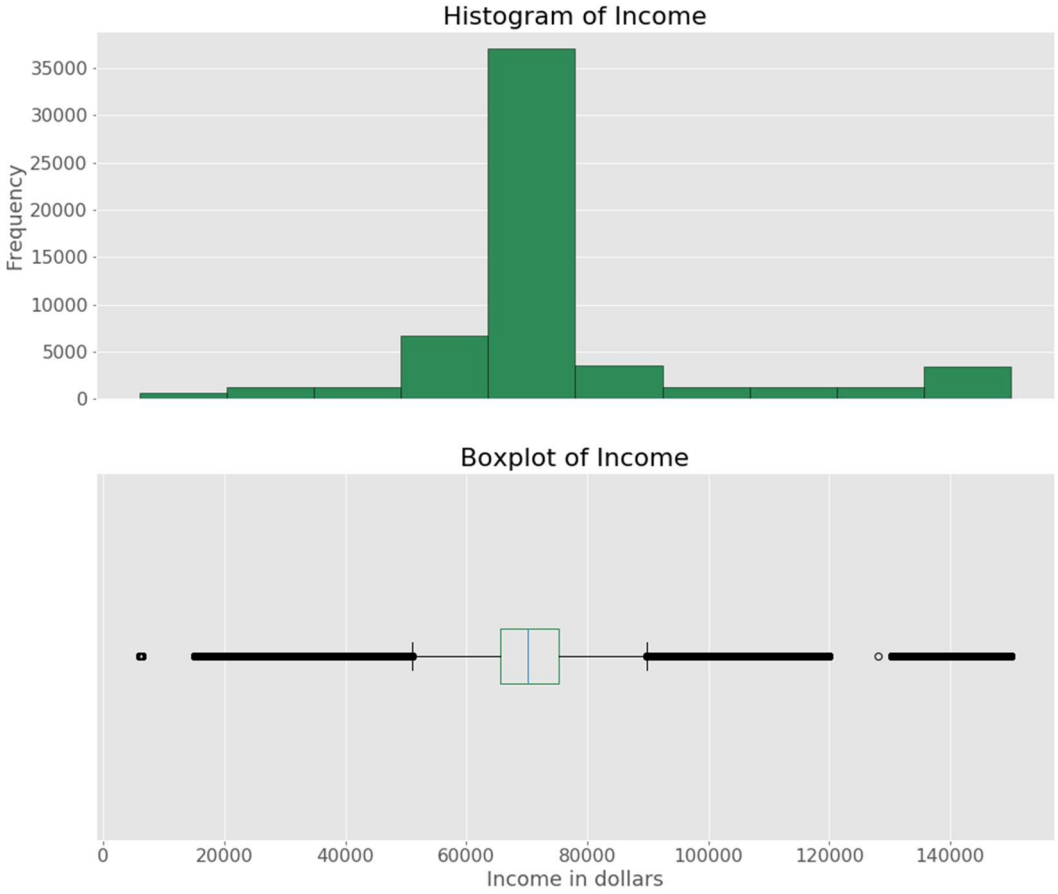
```
HomeOwners :
 count    41787.000000
mean        74.941837
std         56.233129
min          0.936863
25%         28.515251
50%         60.867316
75%        111.666179
max        318.000000
Name: Value, dtype: float64
NOT HomeOwners :
 count    14391.000000
mean        83.279046
std         57.508268
min          0.929691
25%         33.000000
50%         72.241693
75%        126.000000
max        295.000000
Name: Value, dtype: float64
```

Appendix IX – income

Before cleaning

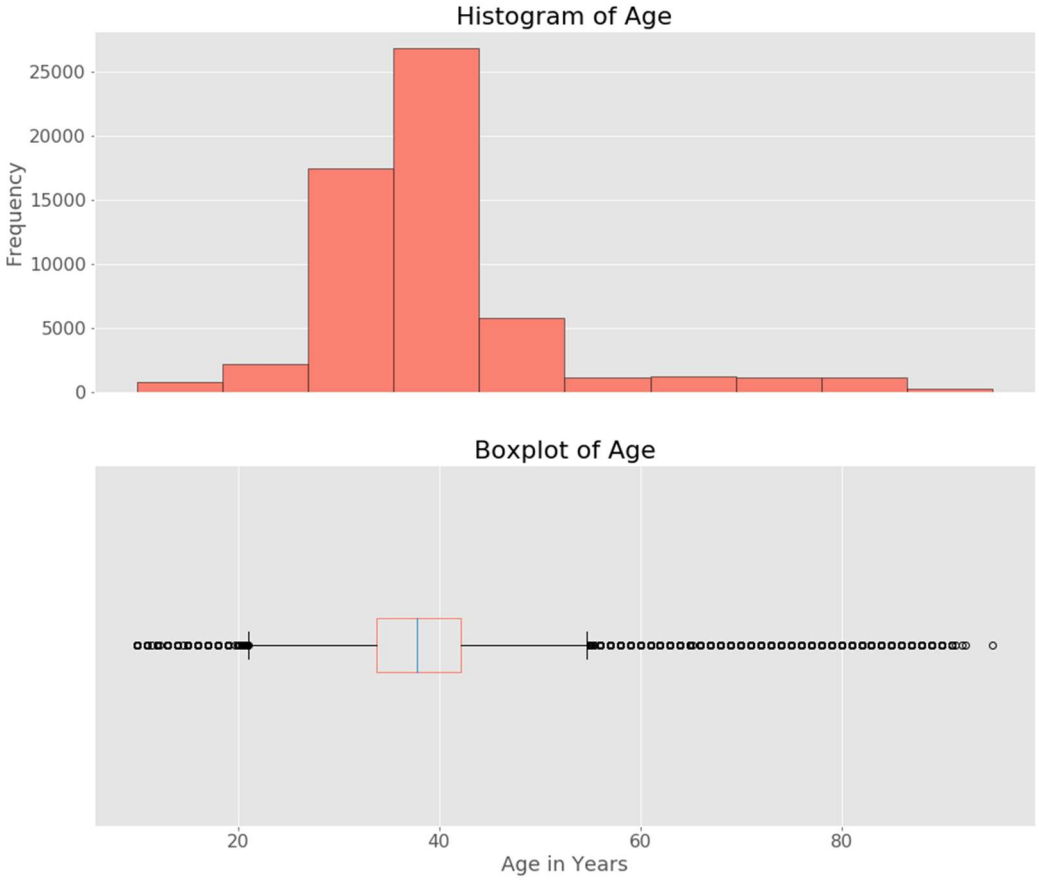


After cleaning

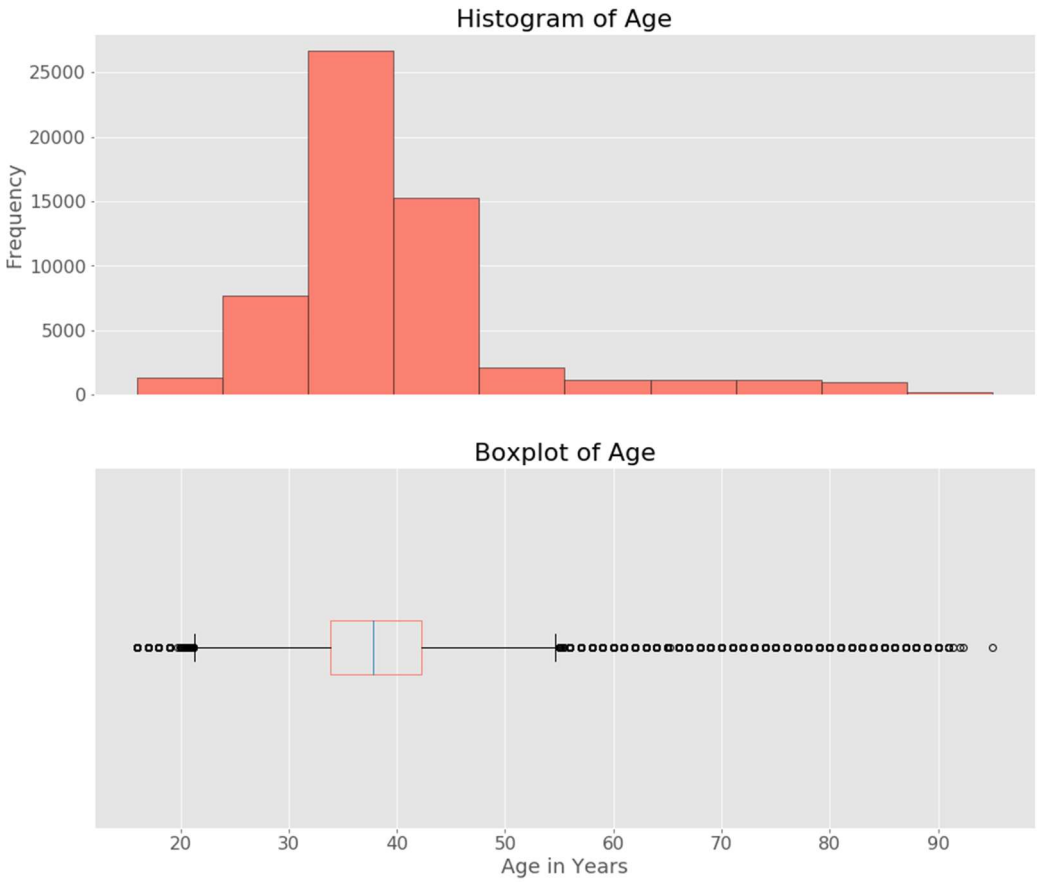


Appendix X – age

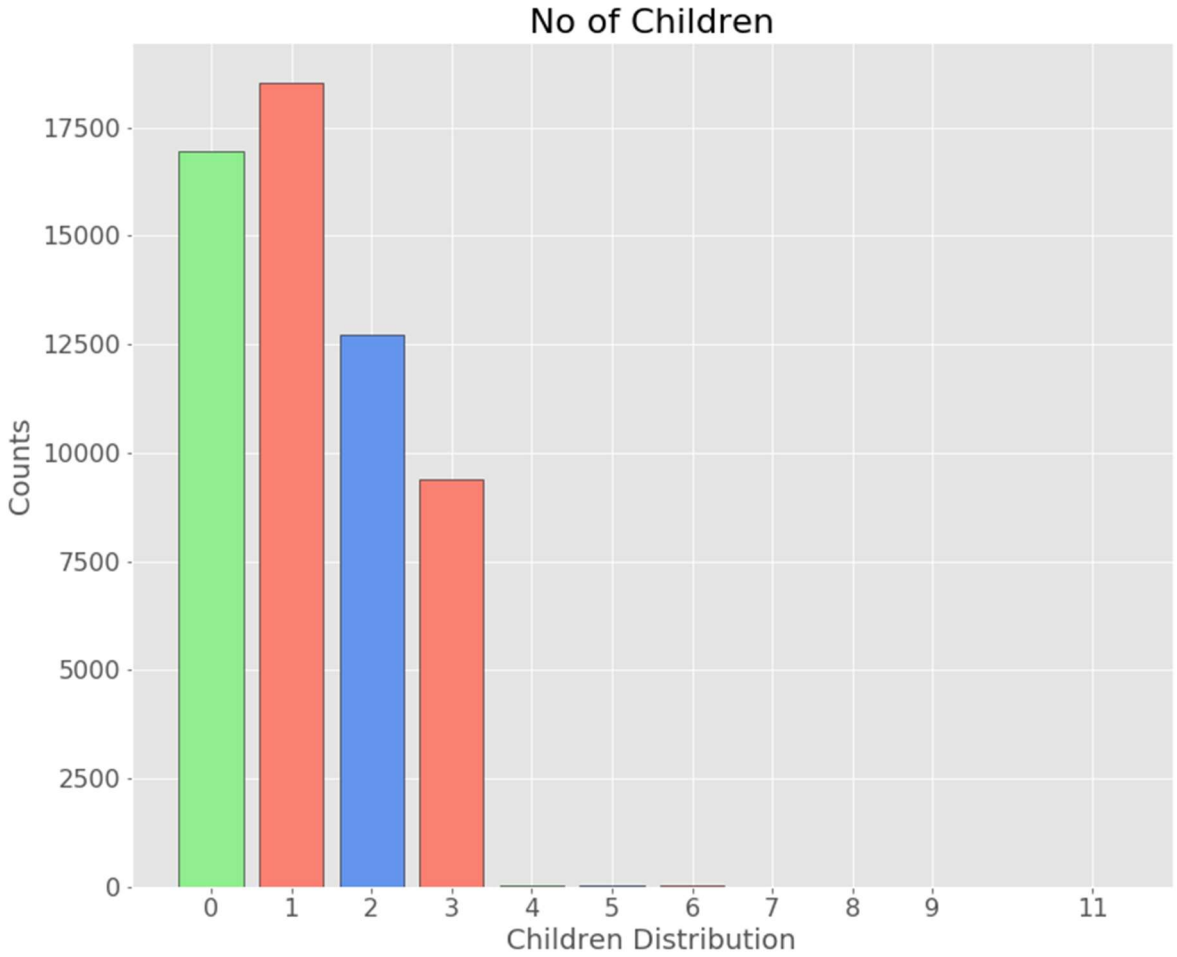
Before cleaning



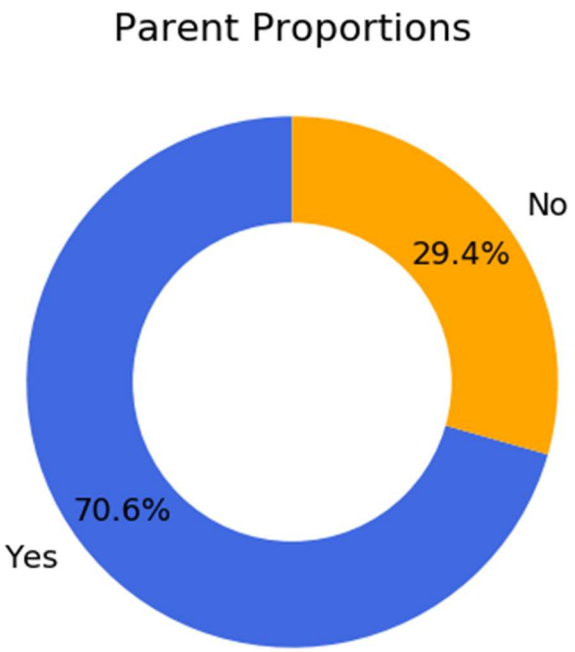
After cleaning



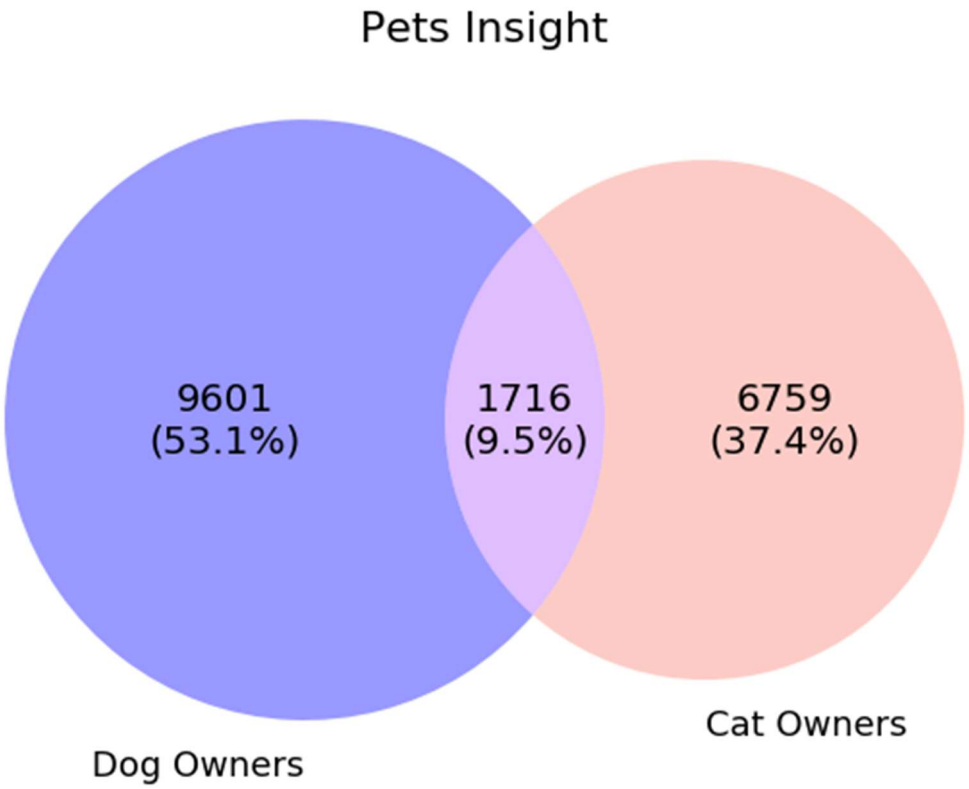
Appendix XI – nchildren



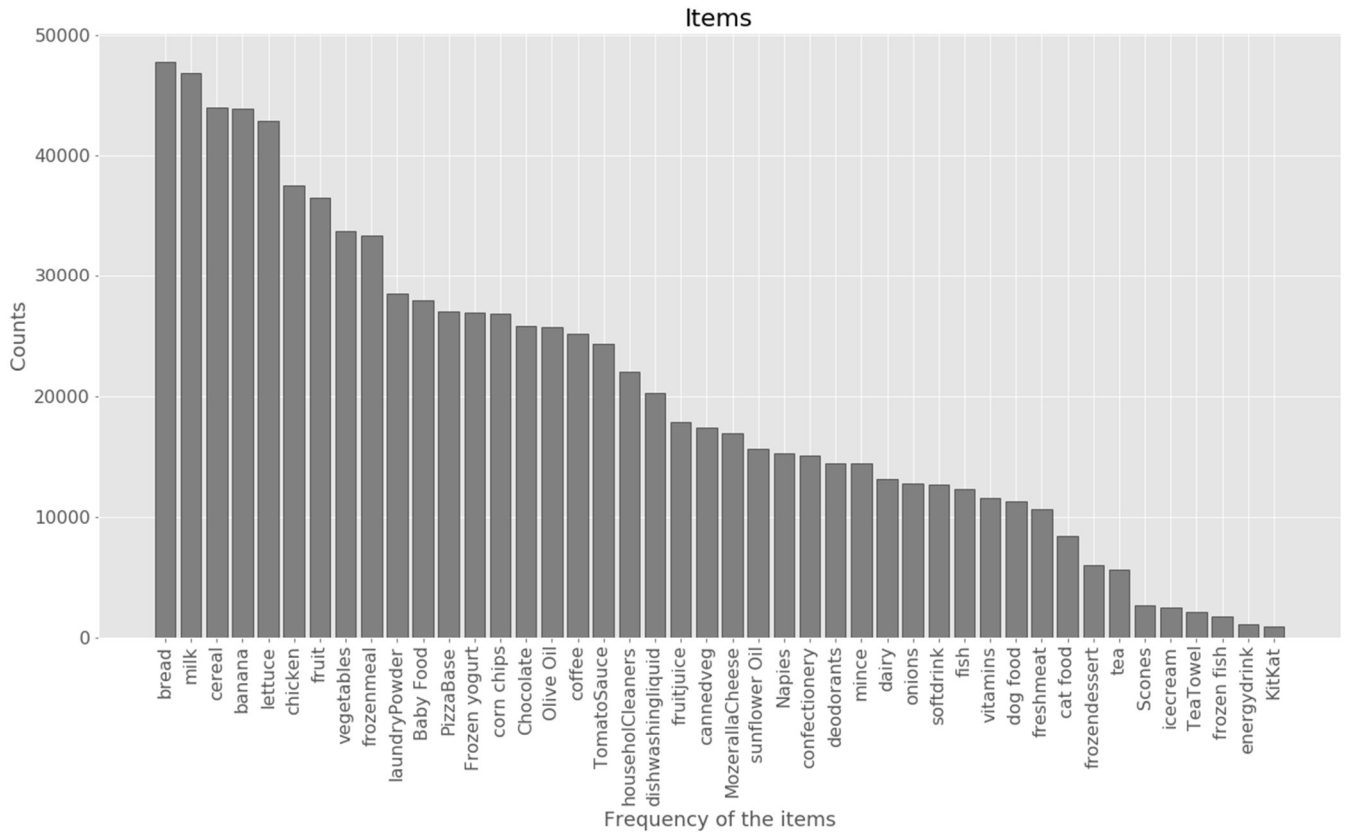
Appendix XII – parent



Appendix XIII – Pet Owner



Appendix XIV – Basket items



Appendix XV – Summary for Numerical Variable

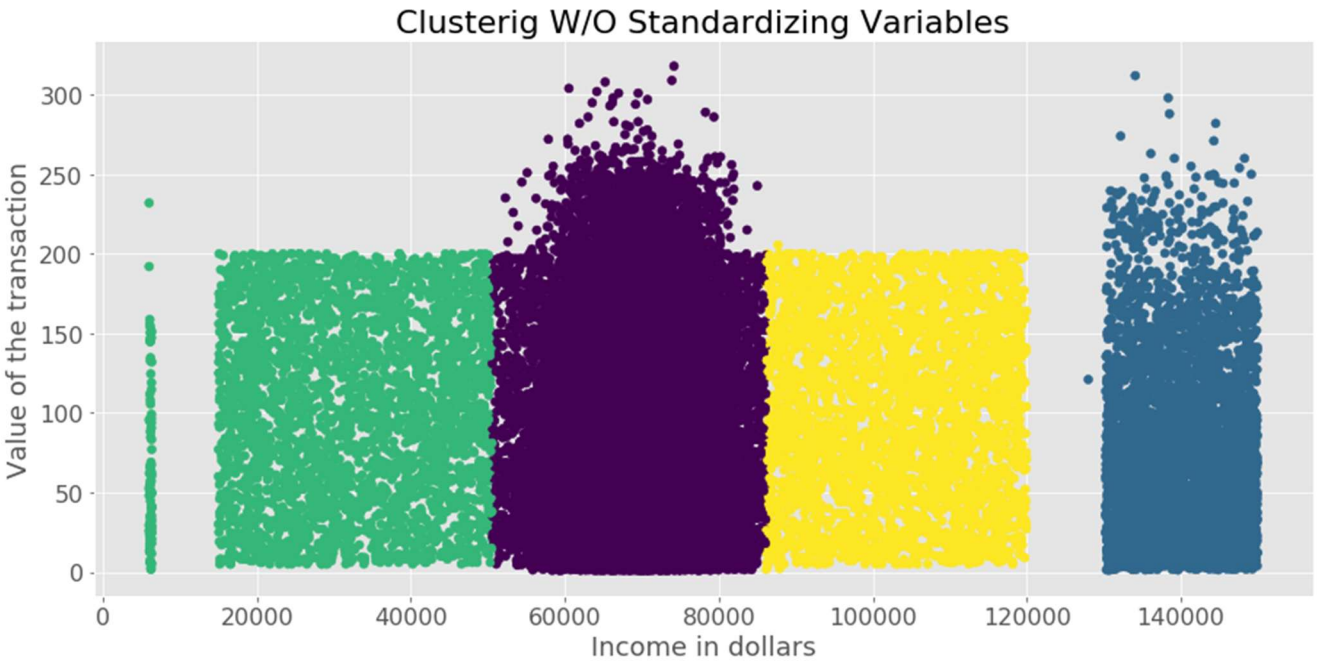
Variable	Min	25%	50%	75%	Max	Mean	Std.Dev
Value(in $)	0.92	29.43	63.29	115.54	318.00	76.98	56.64
income(in $)	6000.23	65,623.48	70,170.42	75324.32	149981.00	74884.04	23761.12
age(in years)	16	33.84	37.83	42.24	95	39.87	11.40

Appendix XVI – Summary for Categorical Variable

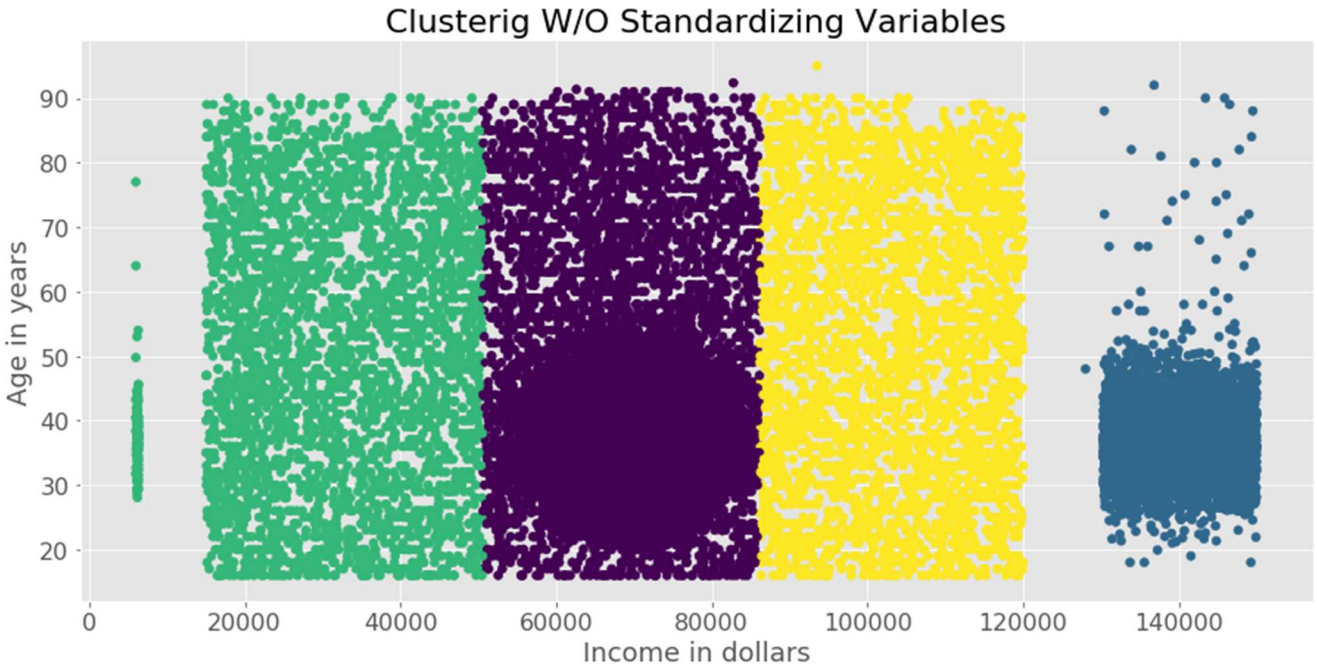
Variable	Frequency – Percentage
pmethod	1 = Cash - 24,488 – 42.5% 3 = Eftpos – 17,575 – 30.5% 2= Card – 8,310 – 14.4% 4 = Other – 7,249- 12.6%
sex	2 = Female – 34,399 – 59.7% 1 = Male – 23,223 – 40.3%
homeown	1 = Yes – 41,787 – 72.5% 2 = No – 14,391 – 25% 3 = Unknown – 1,444 – 2.5%
parent	Yes – 40,683 – 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, 5, 6, 7, 8, 9, 11 = 49 ~= 1.1%

Appendix XVII – Income dominating the cluster results

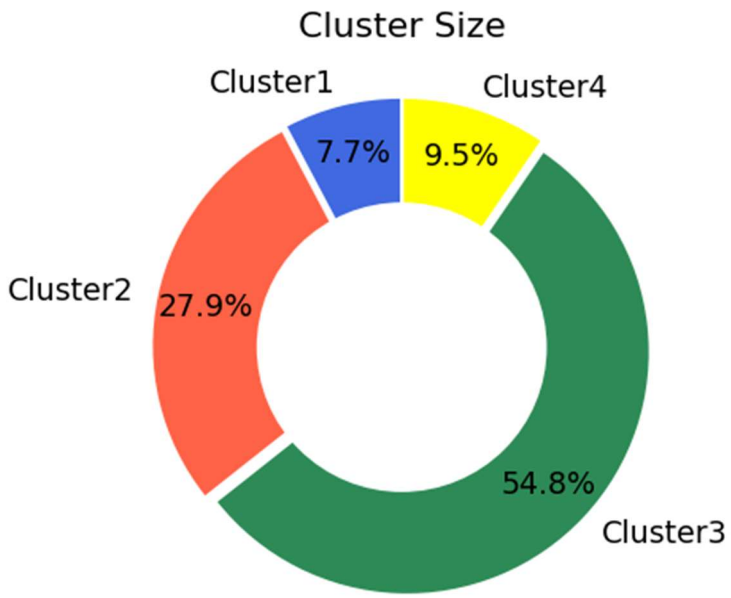
1. Income VS Value



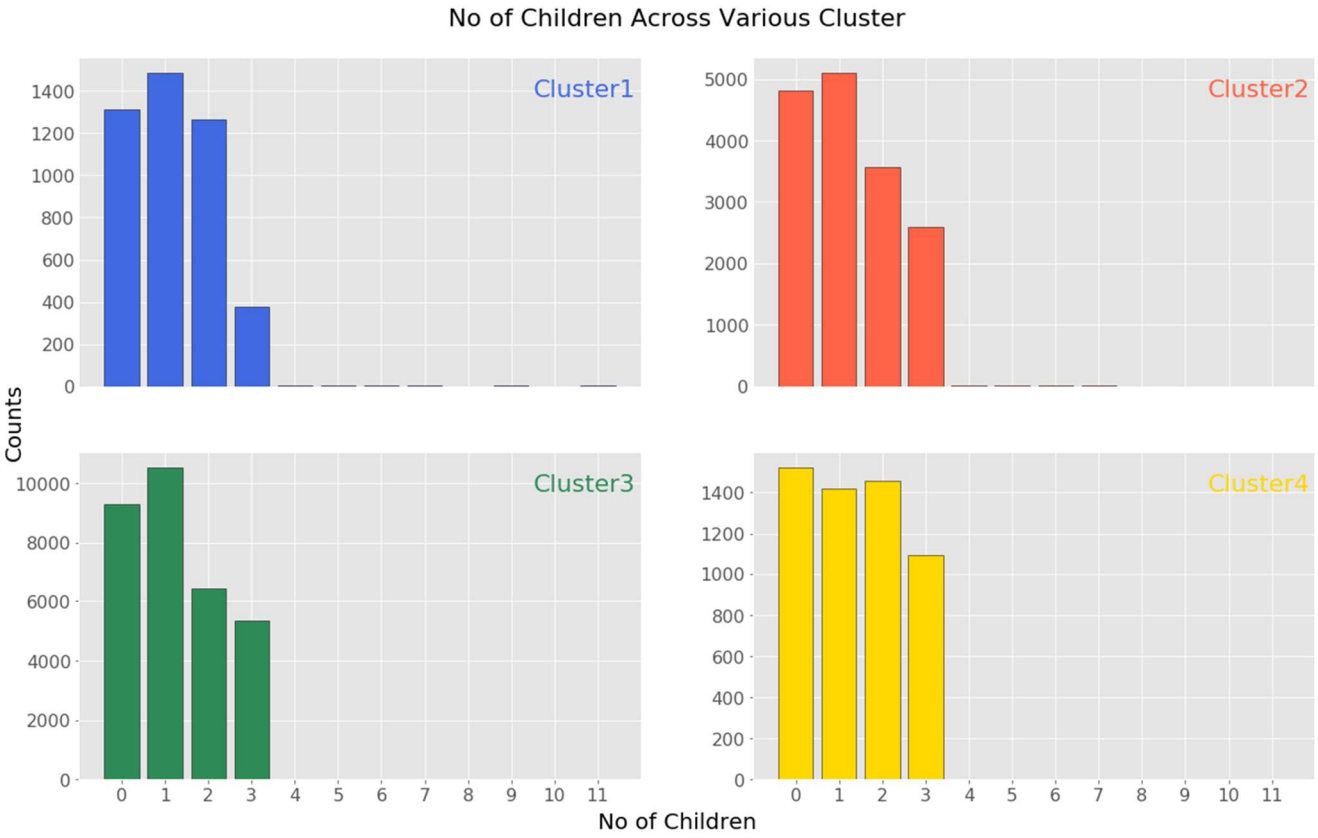
2. Income VS Age



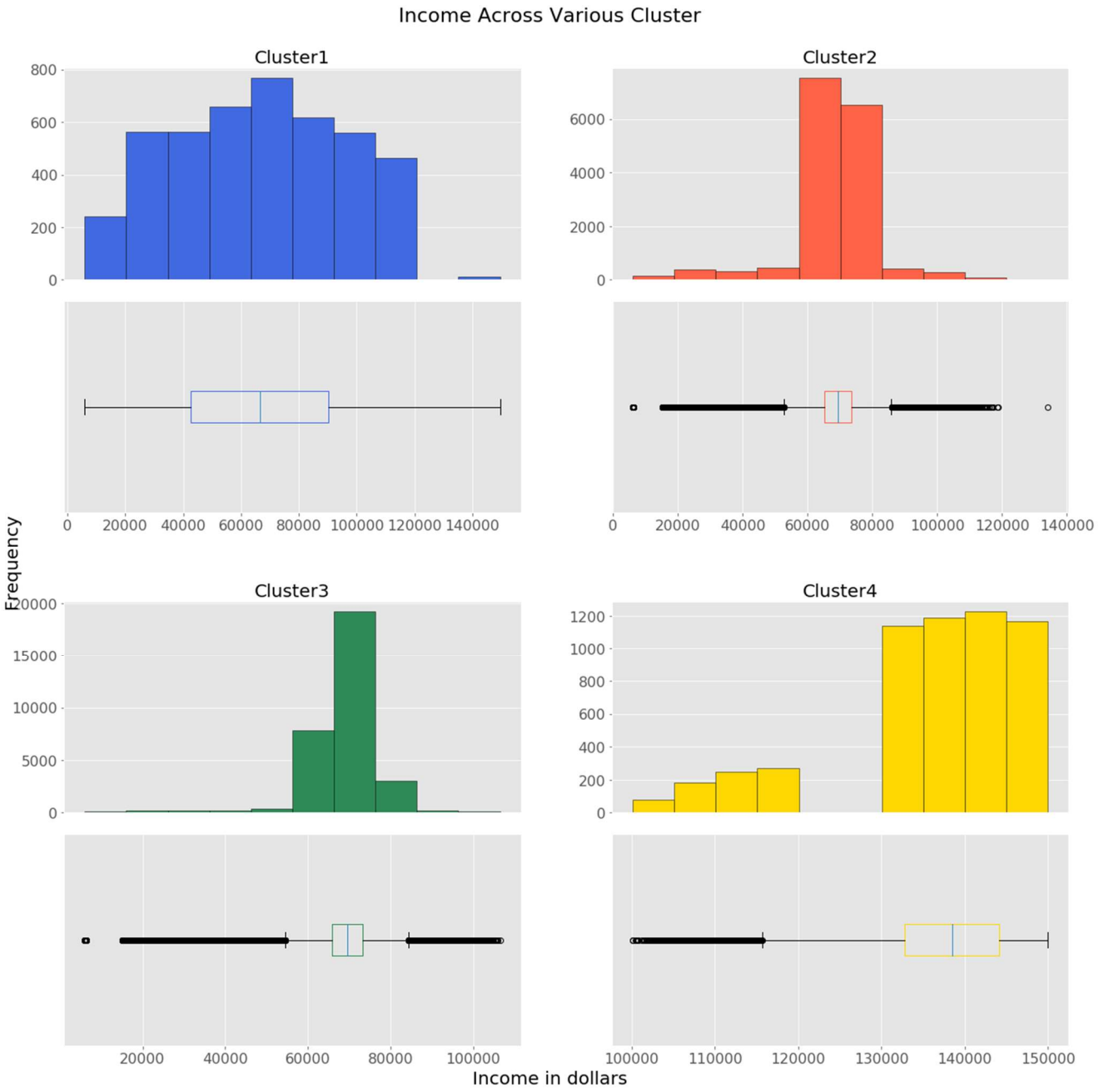
Appendix XVIII – Cluster Size



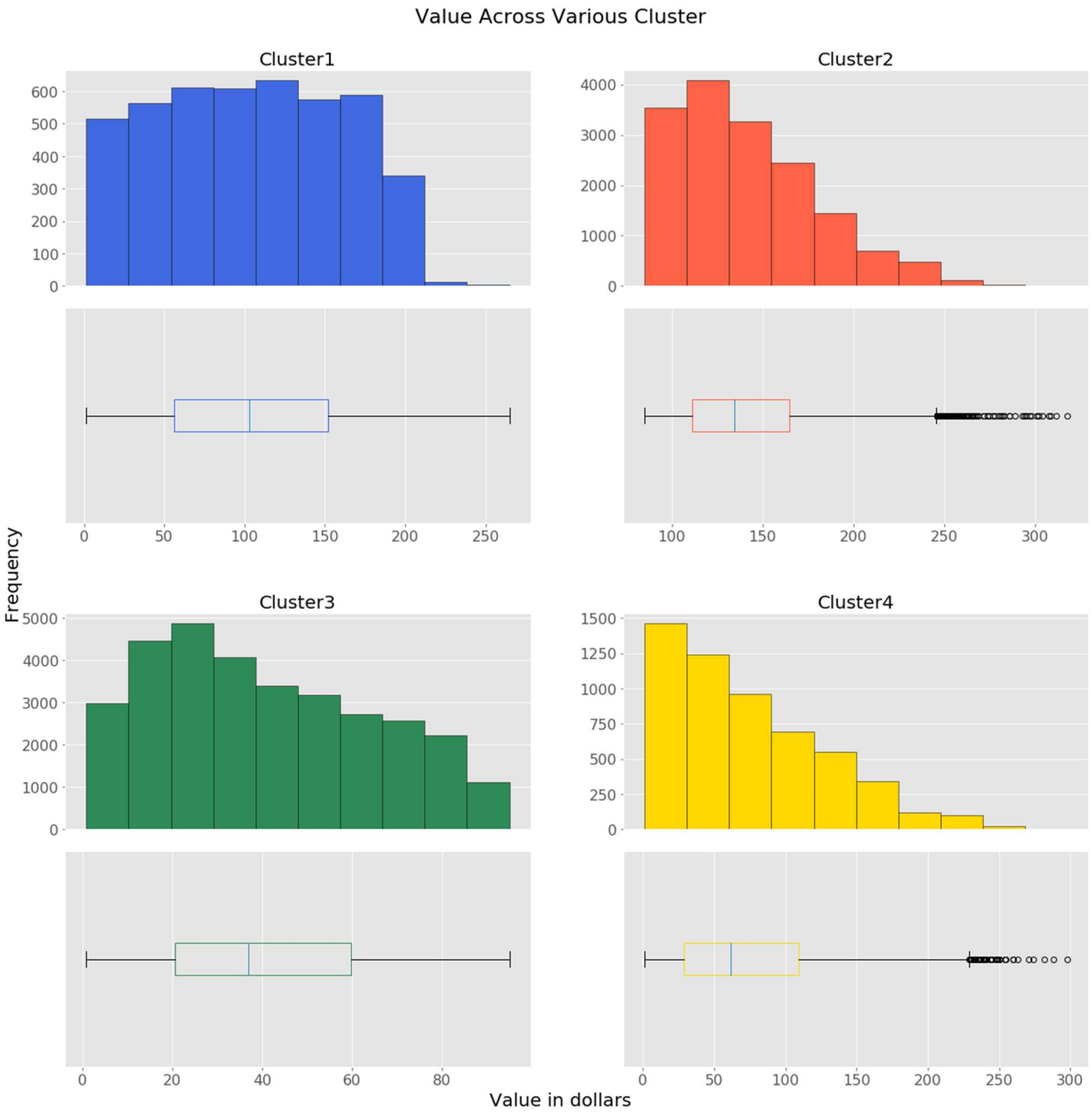
Appendix XIX – Number of Children across clusters



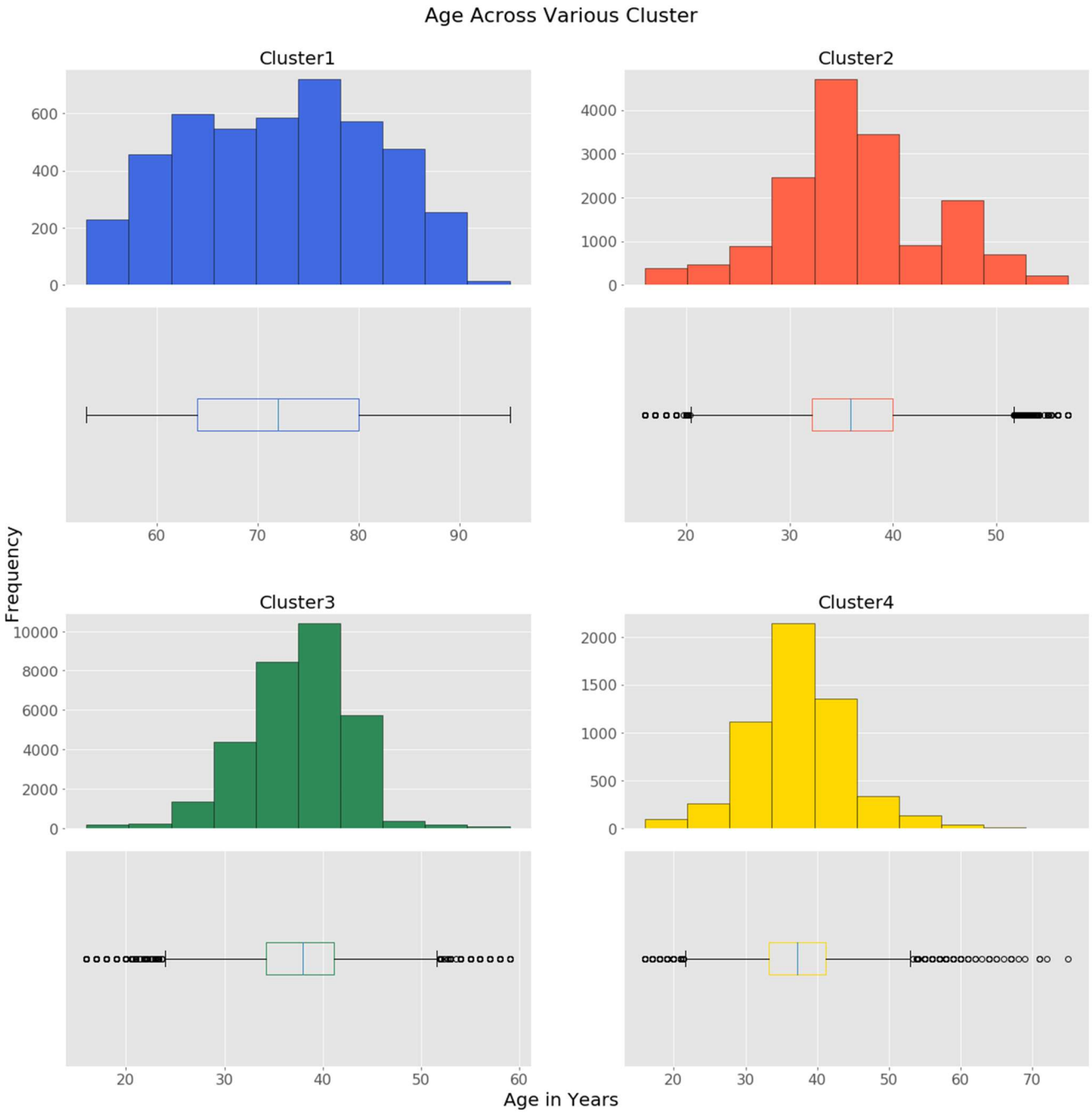
Appendix XX – Income across clusters



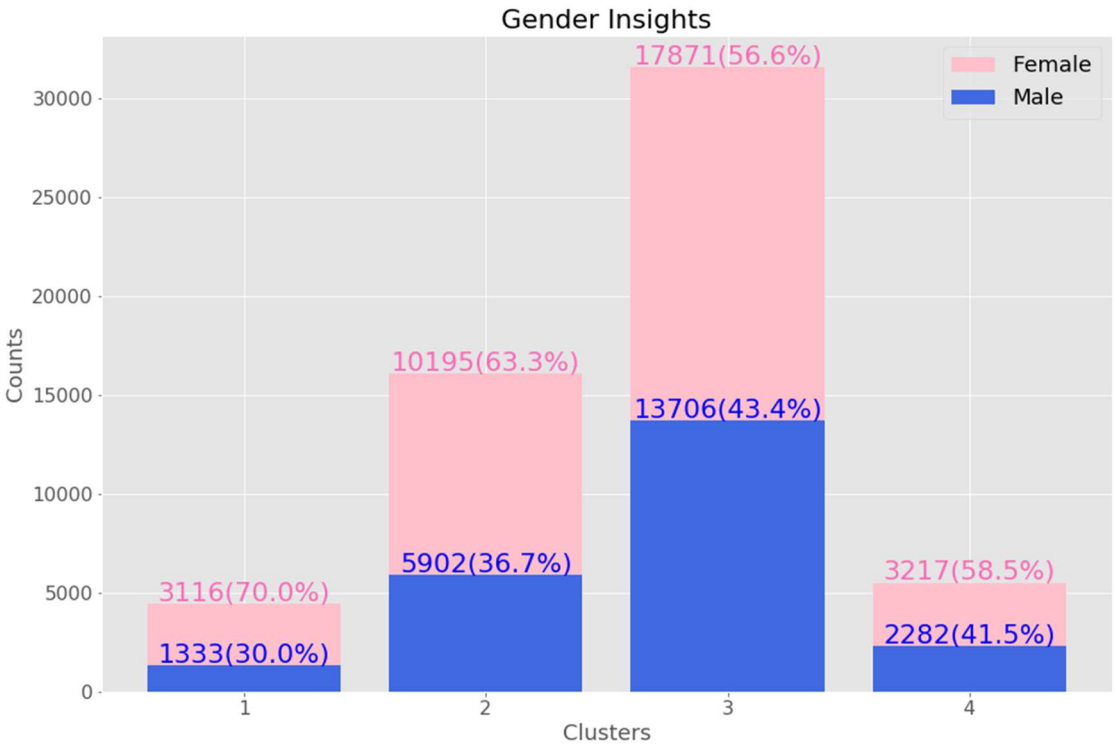
Appendix XXI – Value across clusters



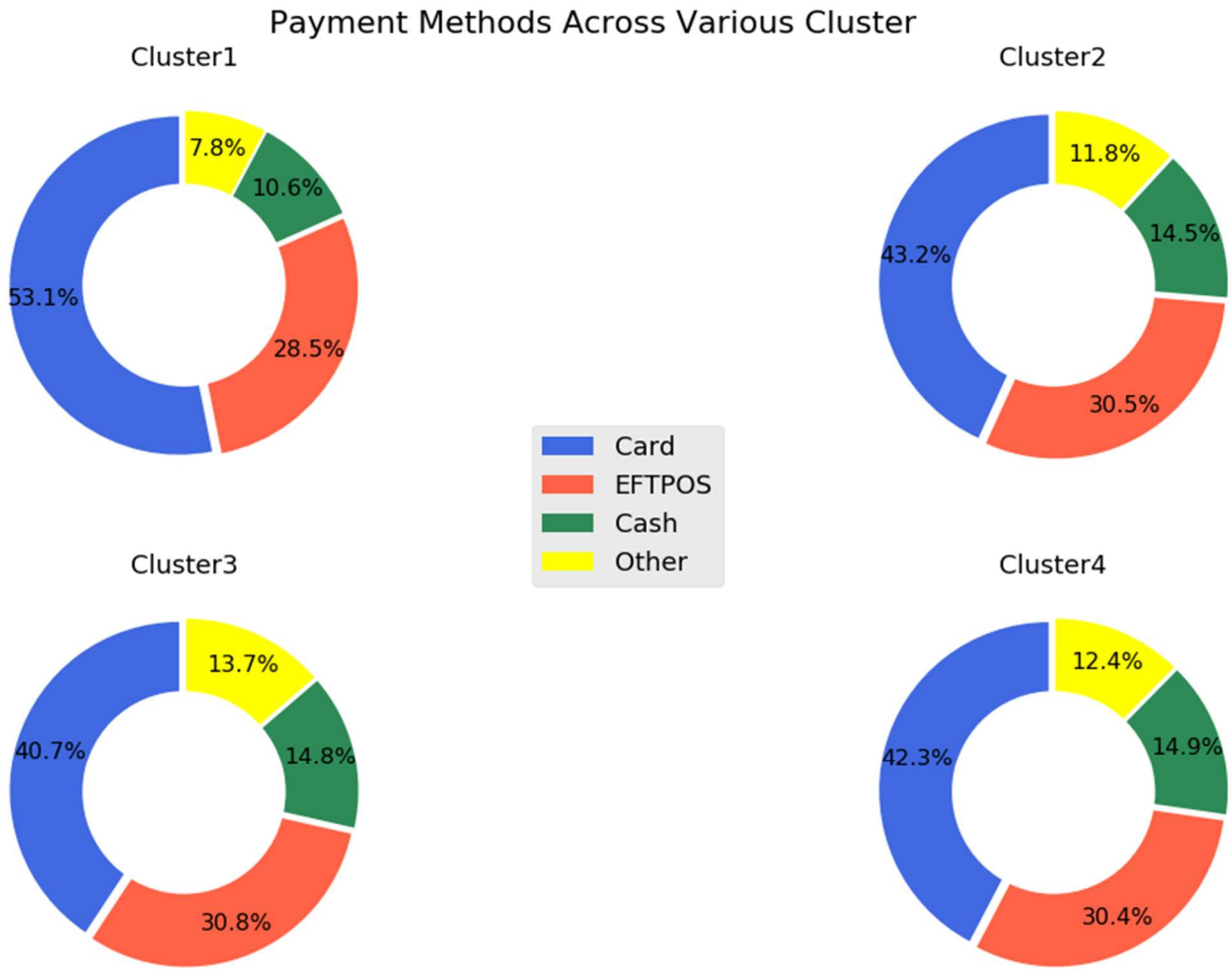
Appendix XXII – Age across clusters



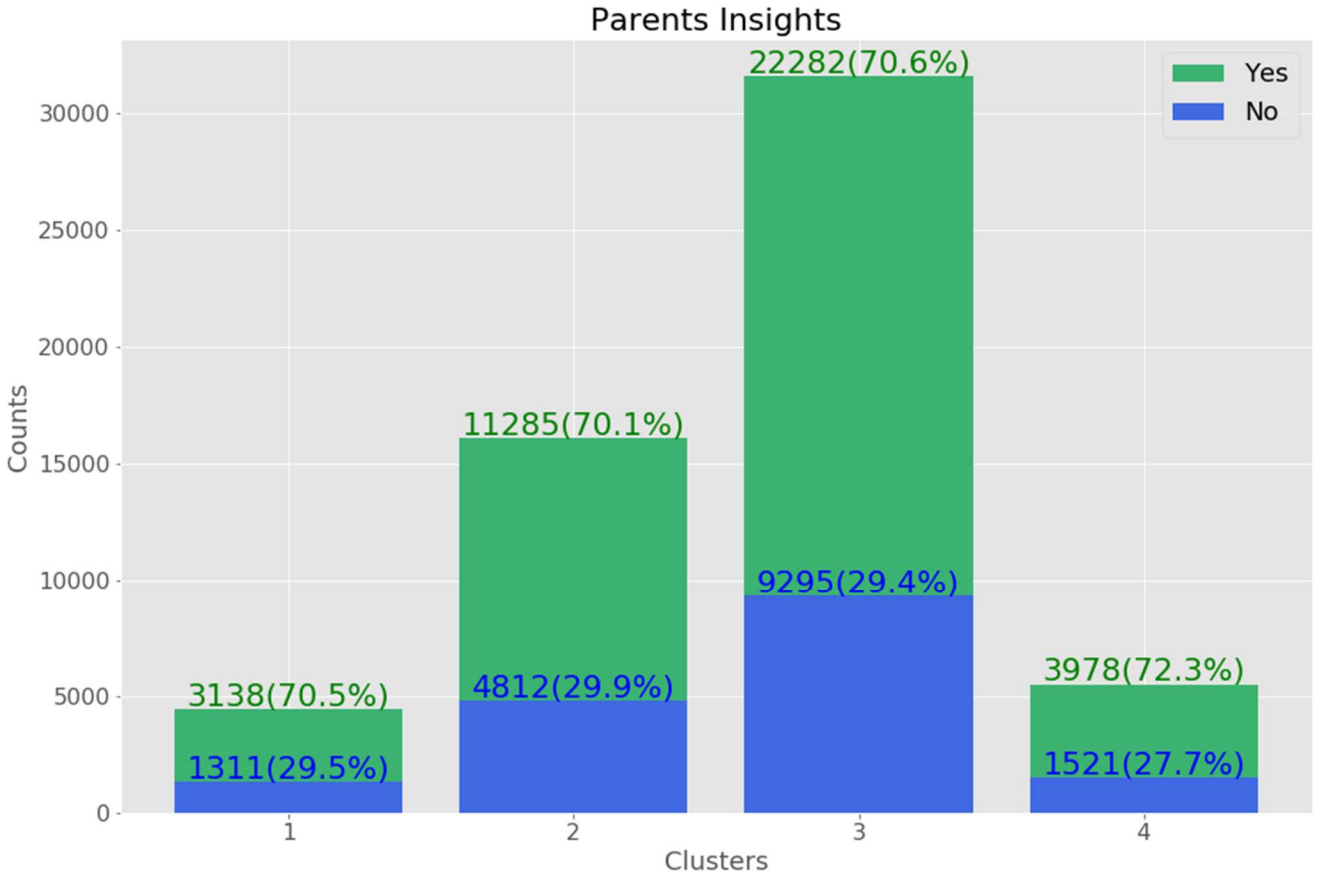
Appendix XXIII – Gender Insights across clusters



Appendix XXIV – Payment Methods across clusters

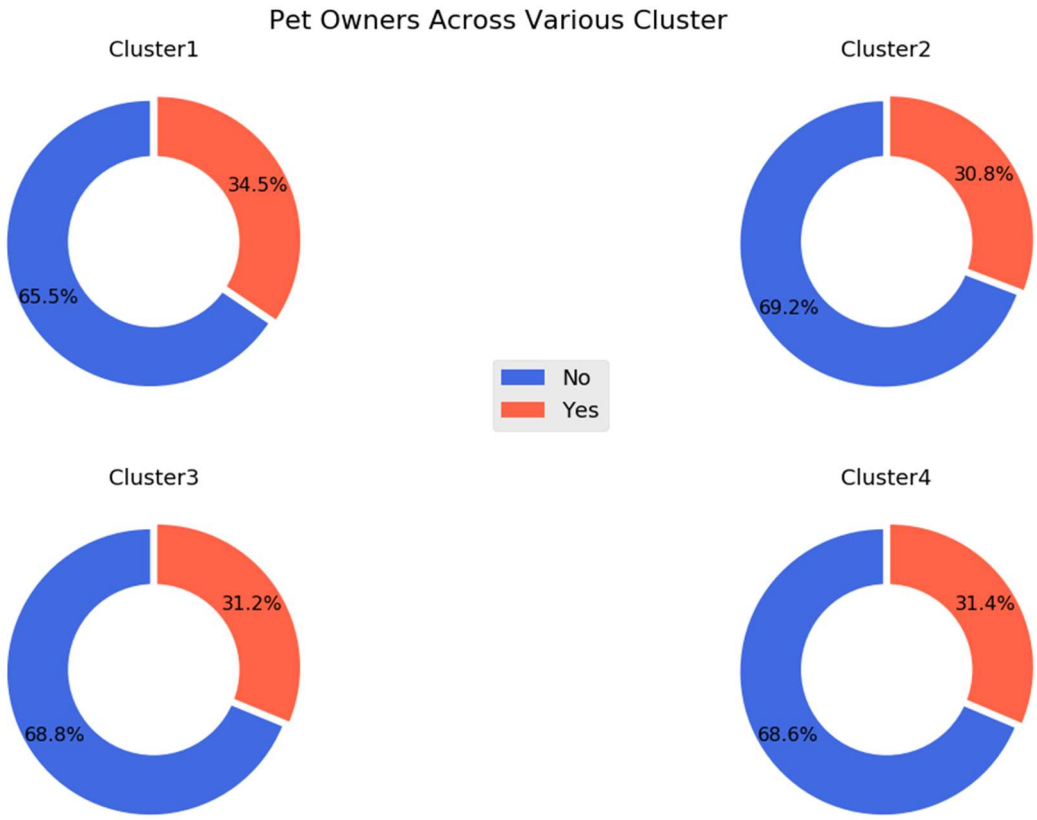


Appendix XXV – Parents Insights across clusters

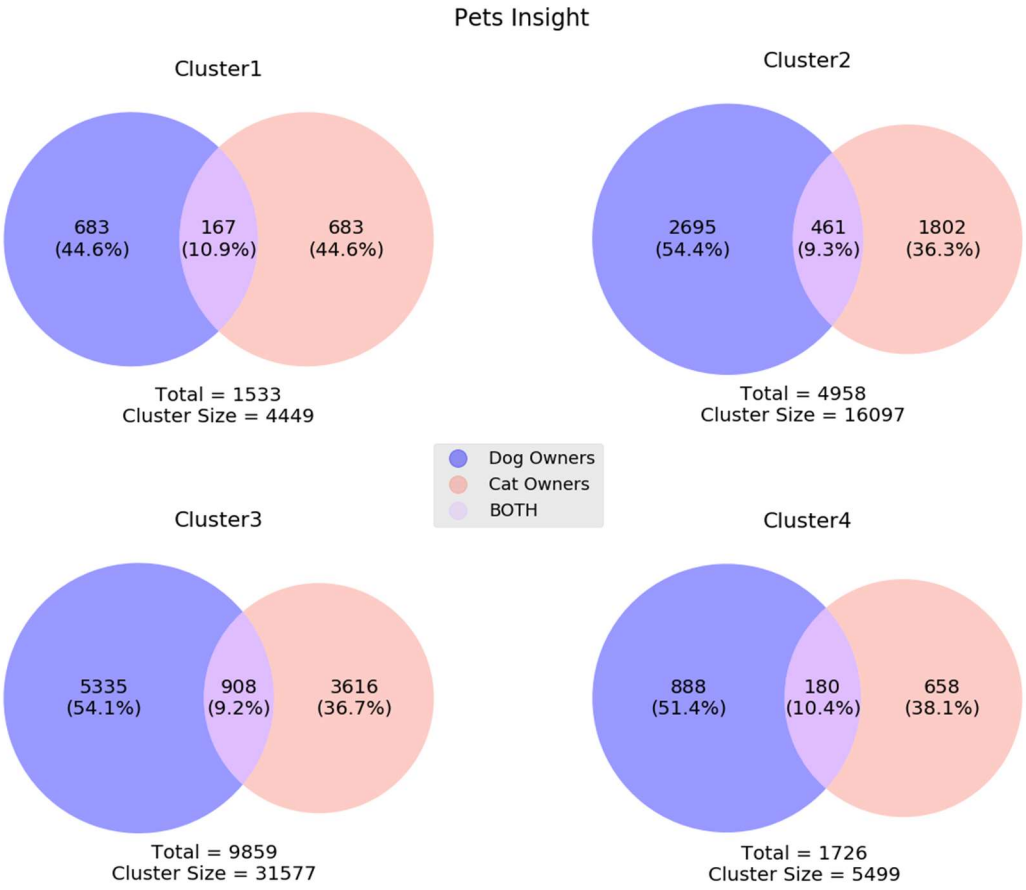


Appendix XXVI – Homeown across clusters

Appendix XXVII – Pet Owner across clusters



Appendix XXVIII – Pet Insight across clusters



Appendix XXIX – 3D(Income, age & Value) Scatter plot