Final Lab on Unsupervised learning using Customer Personality analysis dataset

Main Objective: In companies usually during marketing campaigns, all customers are not targeted with the same specific offers and coupons. Usually before the start of the marketing campaign, companies look into their historically stored customer data to identify group of customers where they are aware of the behavior of the customer, aware of their likes and dislikes and based on that they are targeted with certain products and coupons.

For our use case as part of the final lab in clustering techniques we are also going to delve into a similar customer data, where we have information about 2,240 customers like:

* Personal information about the customer like what is their highest level of education, family size, income

* Customer profile information like How recently the customer has purchased, how many times customer has complained about a product, likes and dislikes of the customer based on what and how much of certain products the customer has bought

* Customers interaction with previous campaigns, whether the customer was converted using first touch second touch or nth touch

* Place of purchase for the customer, like if the customer made purchases from web, catalog, store etc.

This information is really helpful in identifying different groups of customers, and helps companies target specific customers with specific offers,

Like someone who purchases a lot of wine can be offered discount in wine, or maybe given offers to be able to purchase meat and fruits with wine at discounted prices. This might be a lucrative offer for some customers but may not work with certain other customers who might have a lot of kinds at home and may prefer sweets.

In this project we will be first going over the data set and do some simple EDA followed by different ways of feature engineering and dimension reduction techniques.

We will be trying to identify at least 3 to 4 cluster of customers as well as try and learn about their preferences using customer profiling.

Description about the data: Customer Personality Analysis is a detailed analysis of a company's ideal customers. It helps a business to better understand its customers and makes it easier for them to modify products according to the specific needs, behaviors and concerns of different types of customers.

Customer personality analysis helps a business to modify its product based on its target customers from different types of customer segments. For example, instead of spending money to market a new product to every customer in the company's database, a company can analyze which customer segment is most likely to buy the product and then market the product only on that particular segment.

People:

- ID: Customer's unique identifier
- Year_Birth: Customer's birth year
- Education: Customer's education level
- Marital_Status: Customer's marital status
- Income: Customer's yearly household income
- Kidhome: Number of children in customer's household
- Teenhome: Number of teenagers in customer's household
- Dt_Customer: Date of customer's enrollment with the company
- Recency: Number of days since customer's last purchase
- Complain: 1 if the customer complained in the last 2 years, 0 otherwise

Products:

- MntWines: Amount spent on wine in last 2 years
- MntFruits: Amount spent on fruits in last 2 years
- MntMeatProducts: Amount spent on meat in last 2 years
- MntFishProducts: Amount spent on fish in last 2 years
- MntSweetProducts: Amount spent on sweets in last 2 years
- MntGoldProds: Amount spent on gold in last 2 years

Promotion:

- NumDealsPurchases: Number of purchases made with a discount
- AcceptedCmp1: 1 if customer accepted the offer in the 1st campaign, 0 otherwise
- AcceptedCmp2: 1 if customer accepted the offer in the 2nd campaign, 0 otherwise
- AcceptedCmp3: 1 if customer accepted the offer in the 3rd campaign, 0 otherwise
- AcceptedCmp4: 1 if customer accepted the offer in the 4th campaign, 0 otherwise
- AcceptedCmp5: 1 if customer accepted the offer in the 5th campaign, 0 otherwise
- Response: 1 if customer accepted the offer in the last campaign, 0 otherwise

Place of Purchase:

- NumWebPurchases: Number of purchases made through the company's website
- NumCatalogPurchases: Number of purchases made using a catalogue
- NumStorePurchases: Number of purchases made directly in stores
- NumWebVisitsMonth: Number of visits to company's website in the last month

Data Exploration Summary:

In the customer data there were around 3 categorical variables namely:

- Education
- Marital Status
- Dt_Customer

Income was the only float type variable and rest all were numerical variables

Income variable had around 24 null values. Because of the limited number of null values, I removed them from the dataset leaving me with 2,216 rows of data.

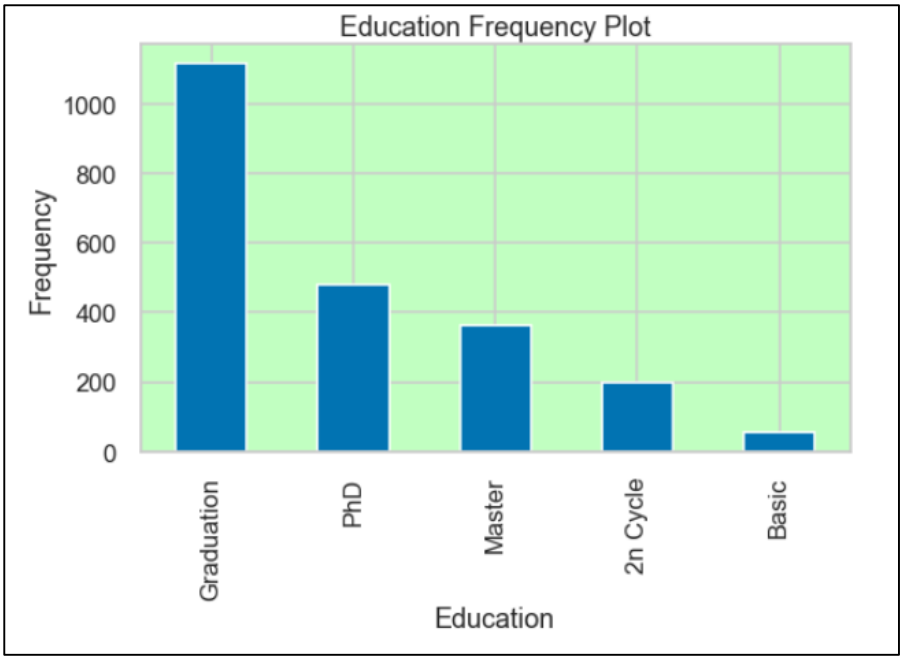
Next I looked into the distribution of the dataset and identified some more information from the data

- Year_Birth – Ranging from 1890 to late 1900's, there seems to be a few outliers in the data
- Income – Income is in the range of few hundreds to 700 K, most of the income lies within 0 to 100K with some outliers
- Numerical Categorical variables – There were some ordinal variables like Kid home, Teen home, Recency, Number of Deal/Web/catalog/store purchases, Number of Web visits per month, Accepted Campaign Offers variables, Complain, Response

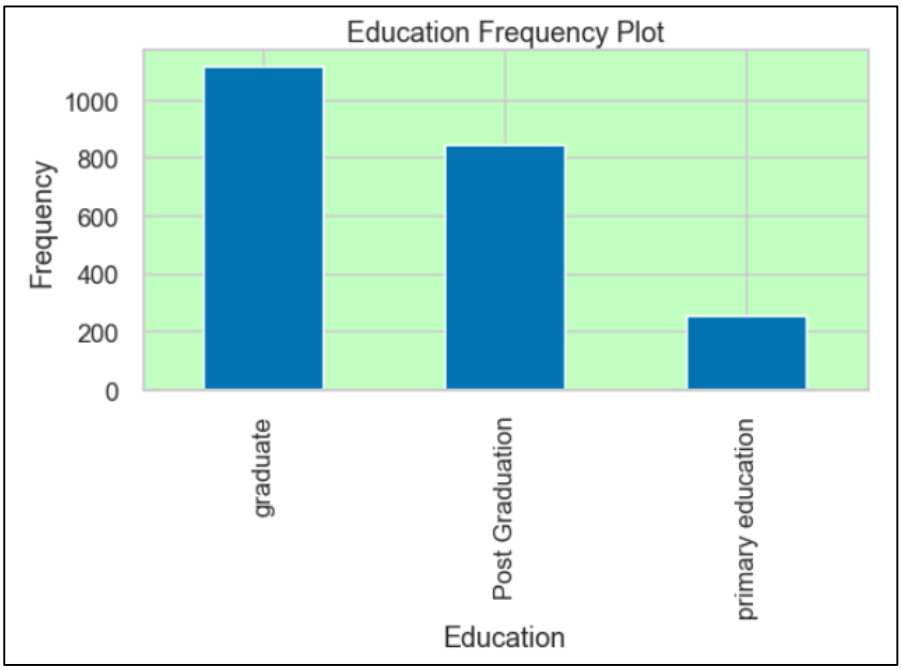
Two variables Z_CostContact and Z_Revenue only had one single value throughout, therefore carrying no information in them, so I dropped these variables

Feature Engineered Variables:

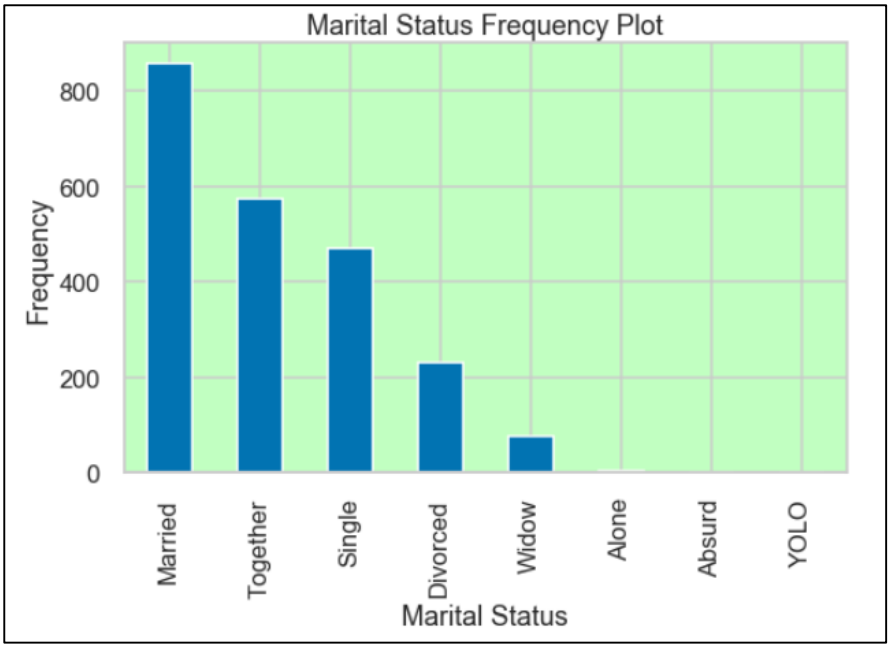
- Number of days as customer – Utilizing the field Dt_Customer i.e. the date at which the person became a customer and current date I identified number of days as a customer
- Age – Utilizing Year_Birth column and current date I identified the age of each of the customer
- Education – Variable had multiple categories of level of education and I wanted to reduce them into lesser categories
  Before



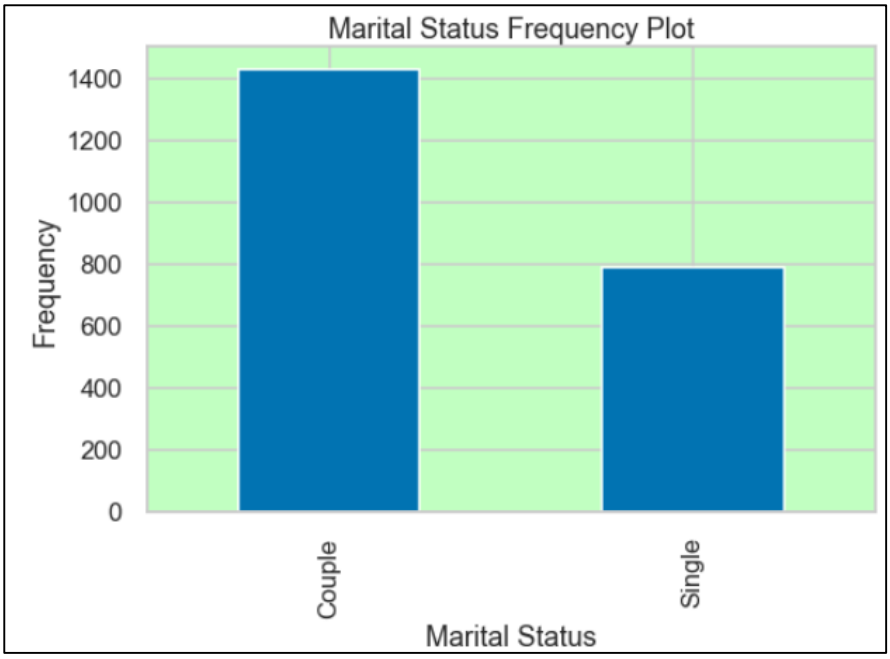
After



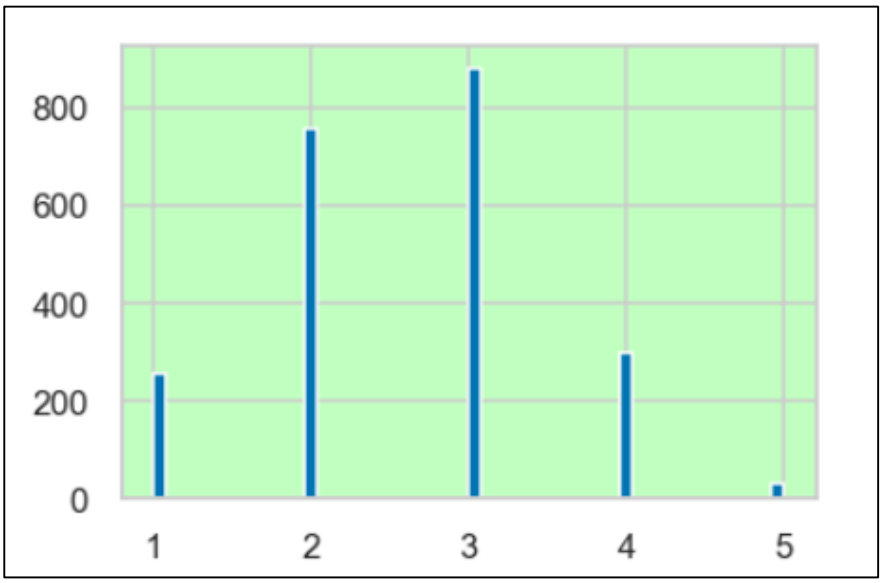
- Marital_Status – Similar to education marital status also had multiple categories and I reduced them to simpler categories
  Before


Marital Status Frequency Plot
Frequency
800
600
400
200
0
Married
Together
Single
Divorced
Widow
Alone
Absurd
YOLO
Marital Status

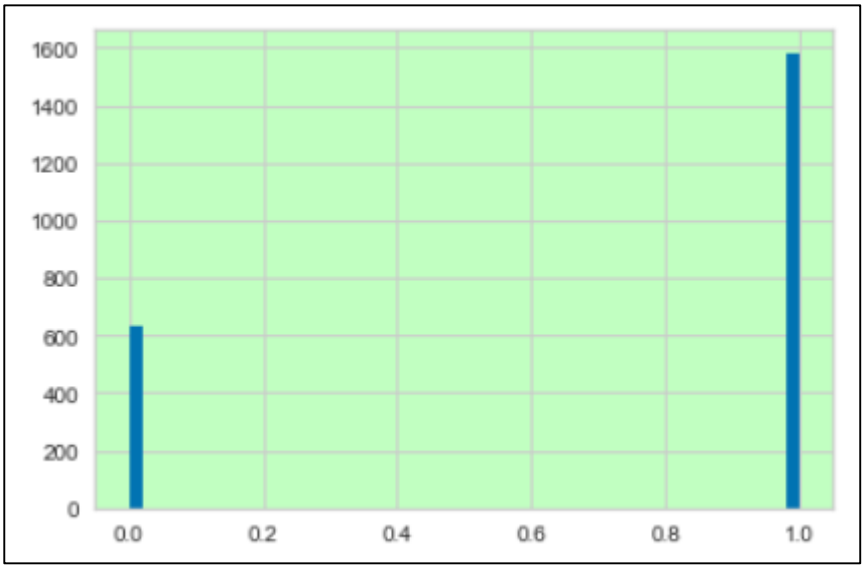
After


Marital Status Frequency Plot
Frequency
1400
1200
1000
800
600
400
200
0
Couple
Single
Marital Status

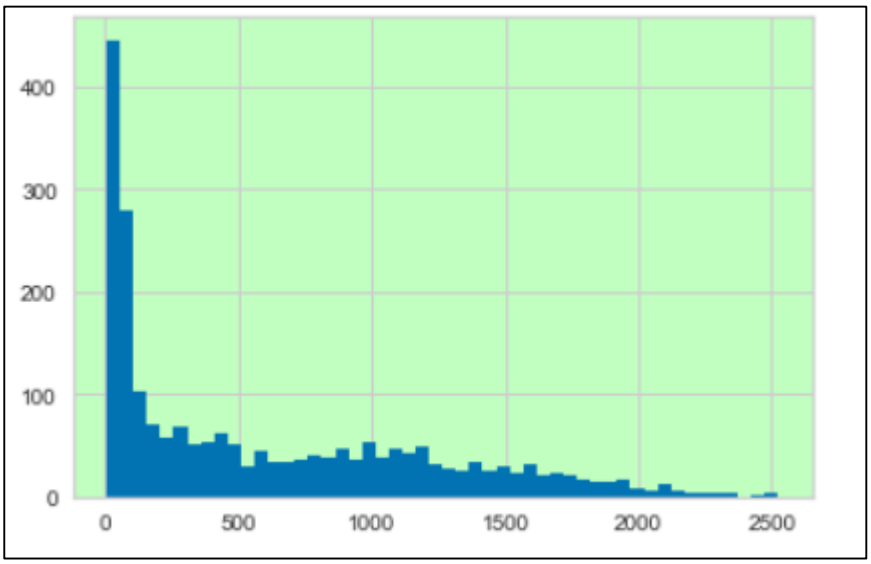
- Children – Created available children from the variables Kid home and Teen home to identify number of children in the house



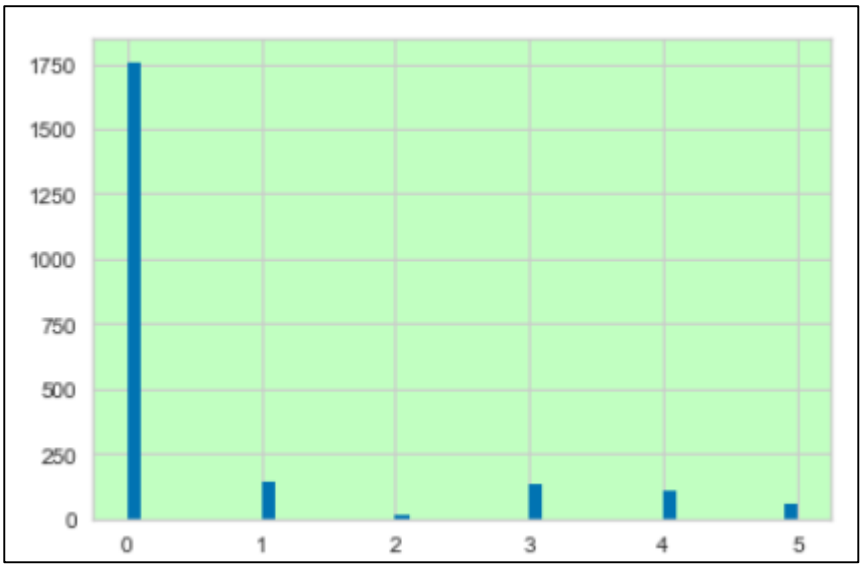
- Parent – Also created a variable Parent to identify if a customer is a parent



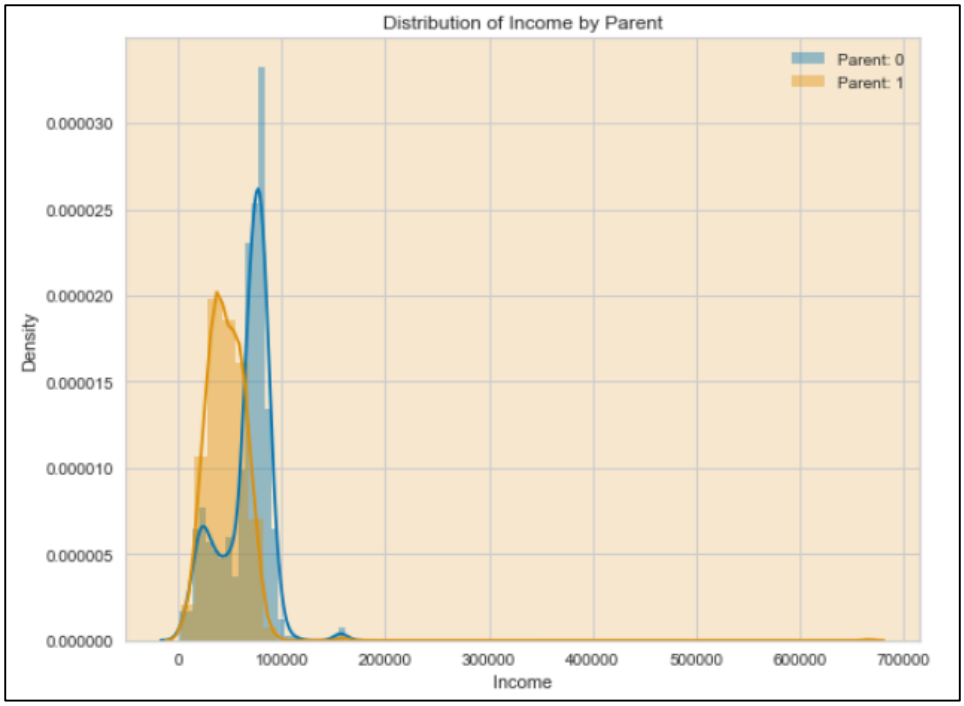
- Spent – Created a variable spent calculating total amount spent by each customer on all the available items like Wine, meat, fruits etc.

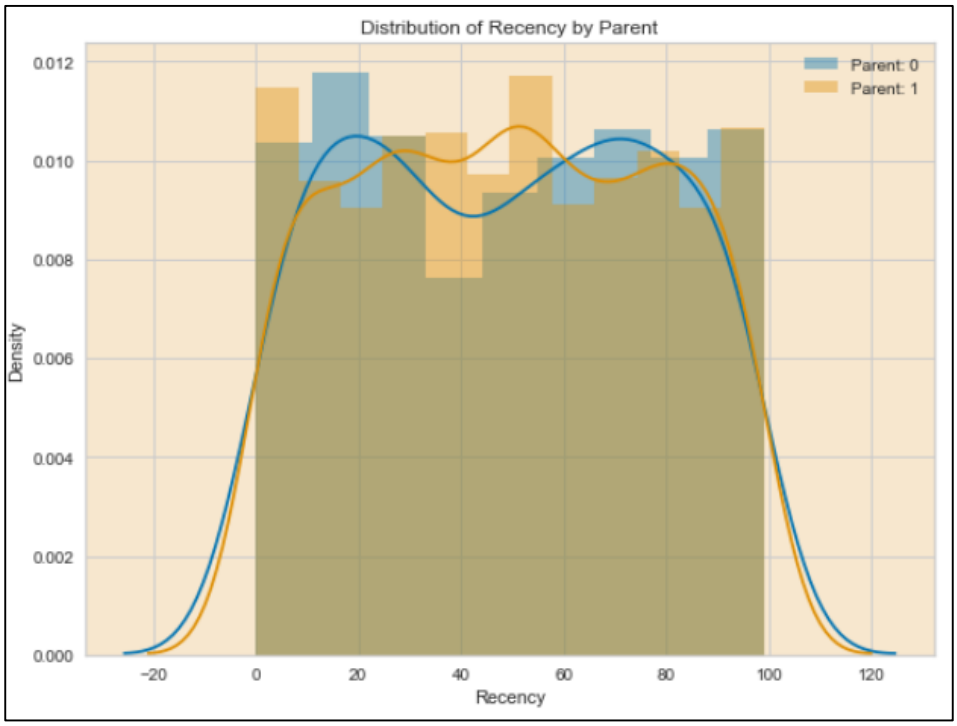


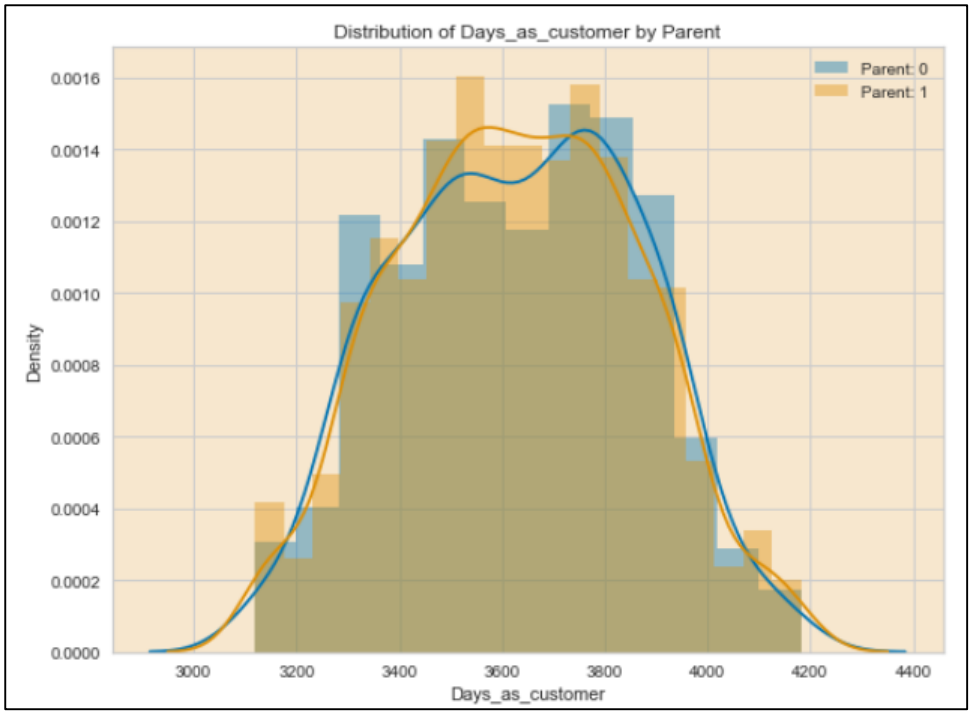
- Number of times contacted – Utilizing the response to campaign contact variables, I created a variable to tell us how many campaigns has a customer been contacted for and at what nth time contact did the customer responded, if in case a customer responded to the 3rd campaign as well 5th campaign then I took the min of the two, in this case the 3rd campaign. This variable would help us in identifying which customers are more responsive to campaign and which are not

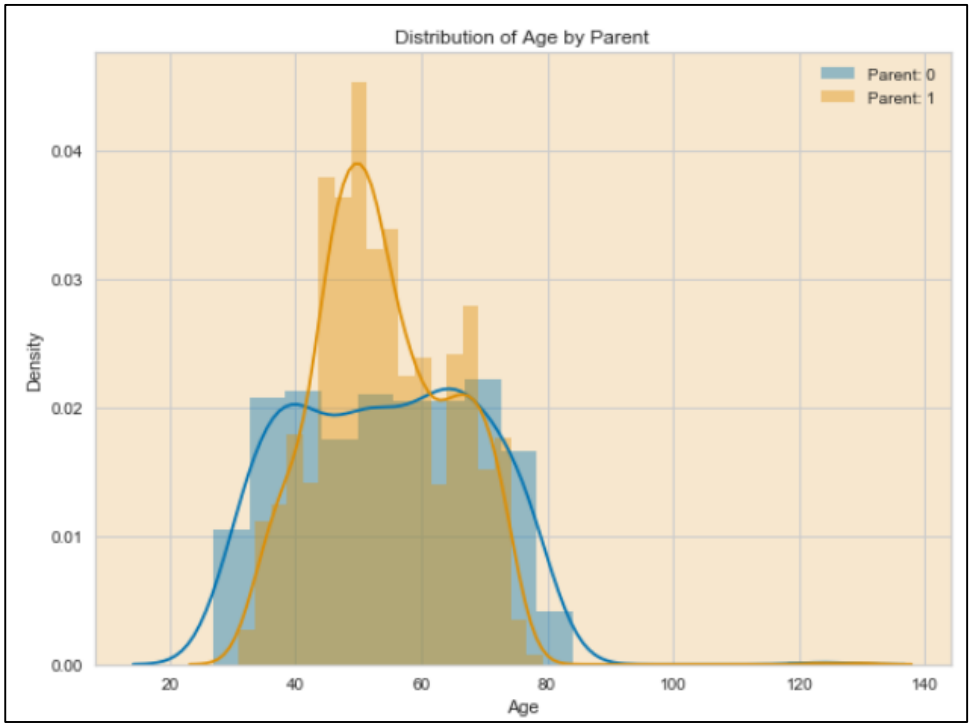


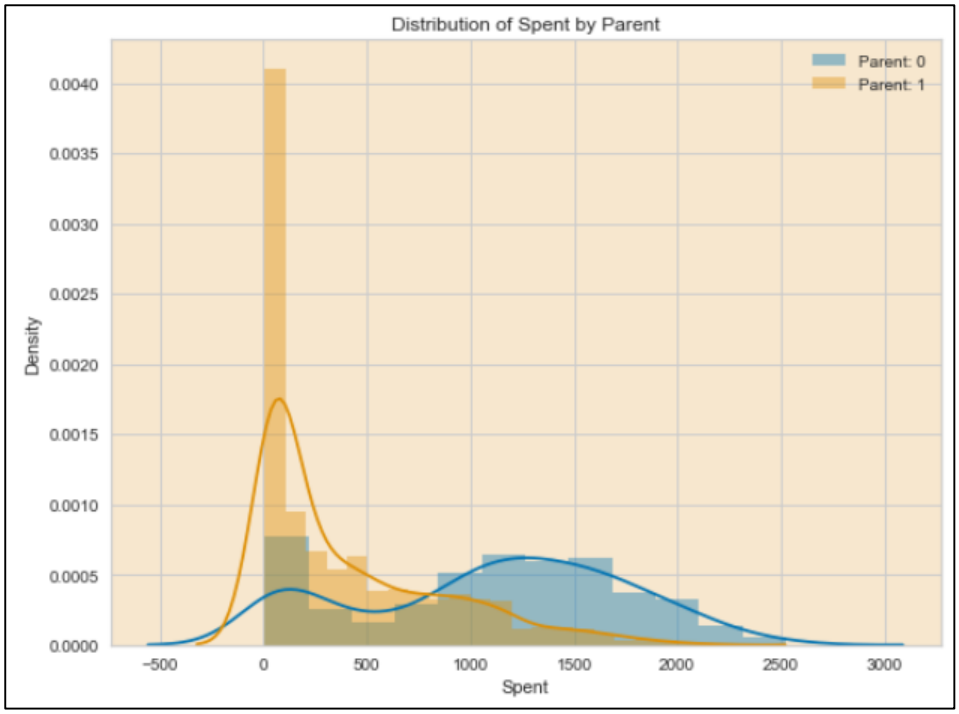
Next I looked at the distribution of some the numerical variables for customers who are parent vast no parent


Distribution of Income by Parent
Parent: 0
Parent: 1
Density
0.000030
0.000025
0.000020
0.000015
0.000010
0.000005
0.000000
0
100000
200000
300000
400000
500000
600000
700000
Income


Distribution of Recency by Parent
Parent: 0
Parent: 1
Density
0.012
0.010
0.008
0.006
0.004
0.002
0.000
−20
0
20
40
60
80
100
120
Recency


Distribution of Days_as_customer by Parent
Parent: 0
Parent: 1
Density
0.0016
0.0014
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
3000
3200
3400
3600
3800
4000
4200
4400
Days_as_customer

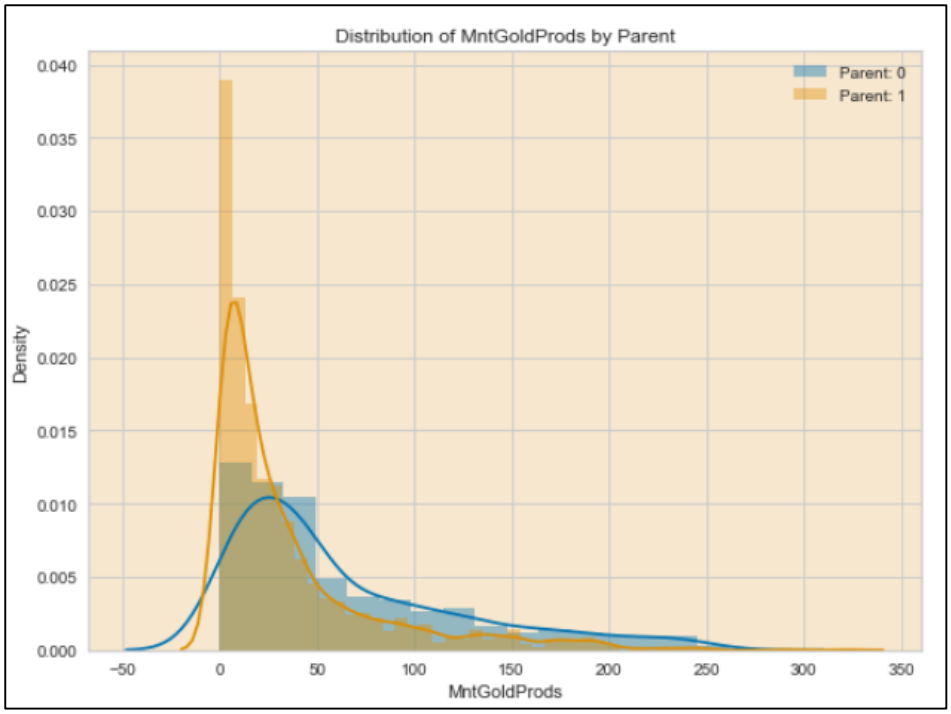

Distribution of Age by Parent
Parent: 0
Parent: 1
Density
0.04
0.03
0.02
0.01
0.00
20
40
60
80
100
120
140
Age



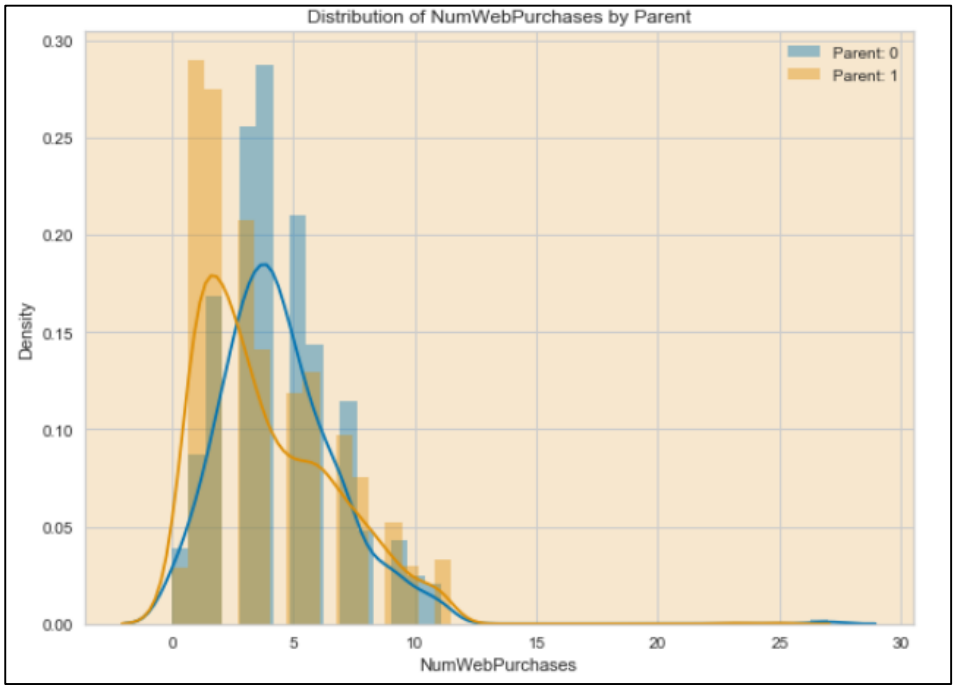
Few things observed from the above Distributions –

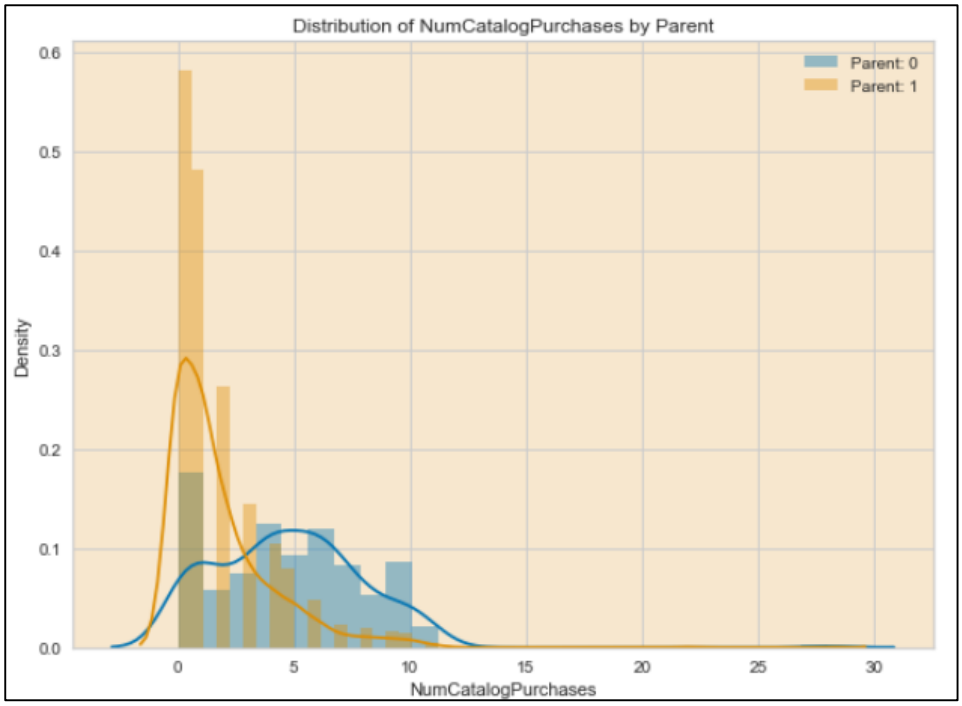
- Income – Income of customers as parents are lower compared to customers who are not parent
- Recency, Days as a customer – We do not see much of a difference
- Age – We see there are more customers in the age of 40 to 60 who are parents' vs people who are not parents have more of a uniform distribution
- Spent – Customers tent to spend much less when they are parent and the distributions is right skewed, whereas when they are not parent the peak is around 1000 to 1500, but the peak is not very long

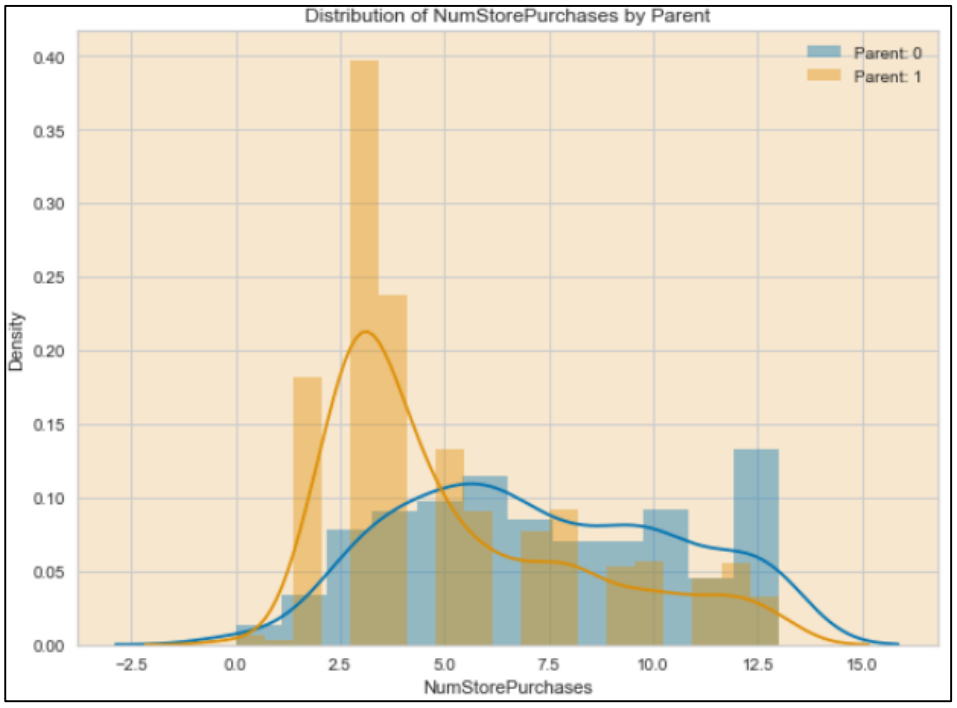
Next I also looked into the products that a parent vs not a parent tend to spend on but did not see much of a difference, except the fact that non parent tend to spend more on gold compared to parents

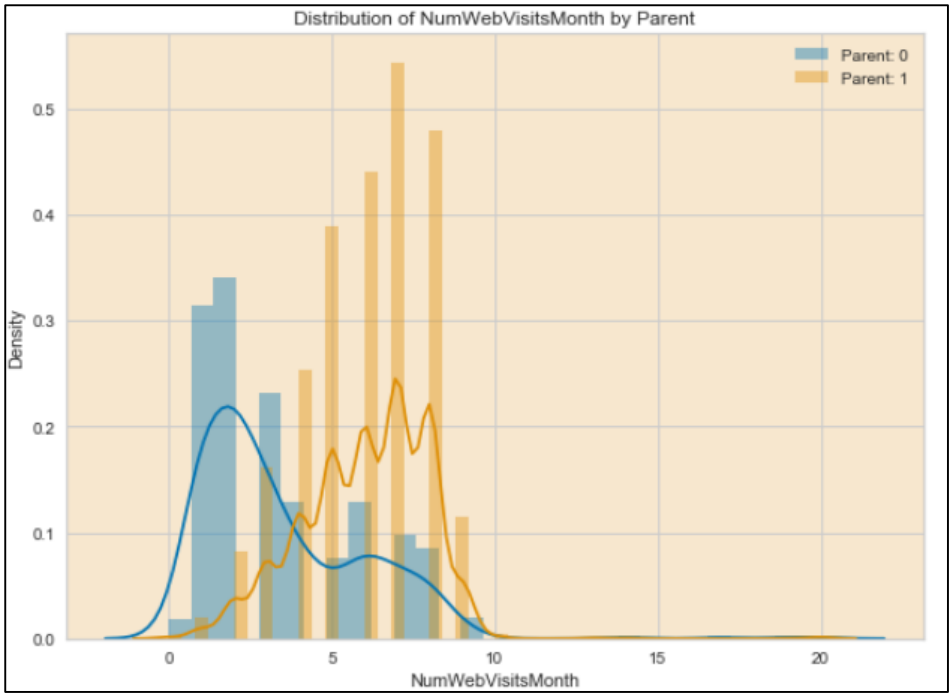


Also looked at the medium of purchase between parents and non-parents and these are some of the distributions




Distribution of NumCatalogPurchases by Parent
Parent: 0
Parent: 1
Density
NumCatalogPurchases
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
5
10
15
20
25
30


Distribution of NumStorePurchases by Parent
Parent: 0
Parent: 1
Density
NumStorePurchases
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
−2.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0



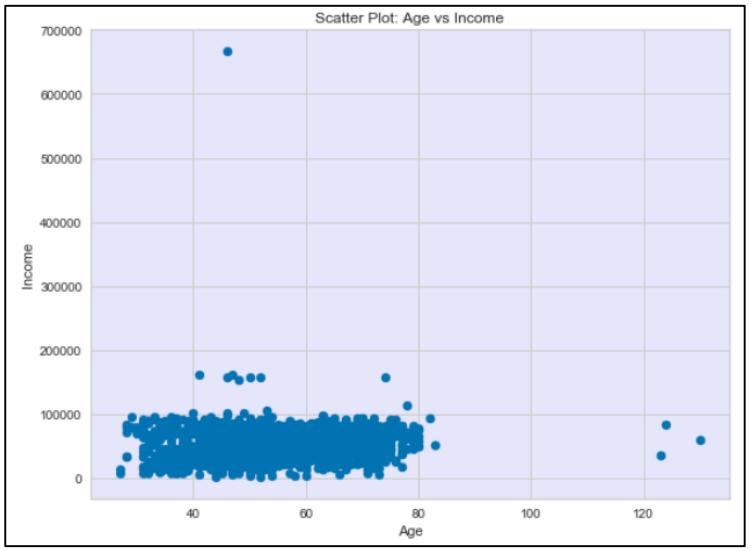
Overall from the above distribution we tend to learn that –

- Parents tend to make less web purchases compared to non-parents
- Non-parents tend to make more catalog purchases
- Both parents and non-parents make quite a bit of in store purchases
- Web visits for parents is more than Non-parents, this is bit counter intuitive given non-parents make more web purchases. This might mean that parents look into the web to see what products are available but go to the market to buy them after looking at their quality

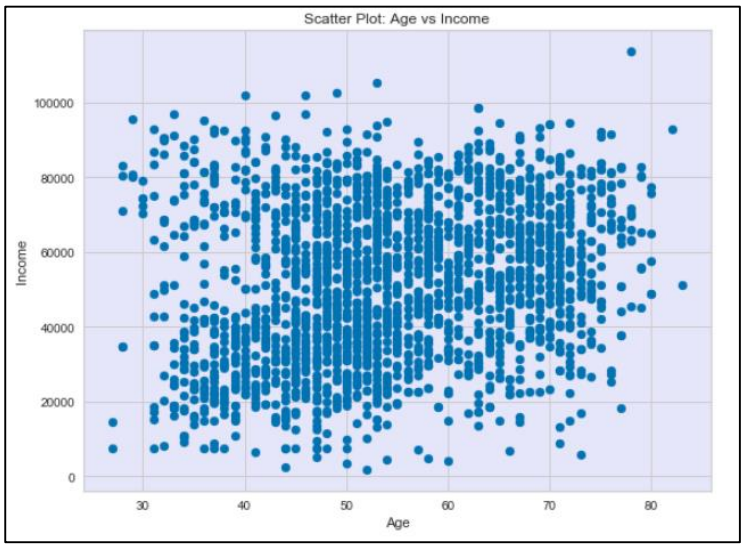
Removing Outliers:

As I already mentioned there were some outliers in age and income and we treated them by removing them from the dataset

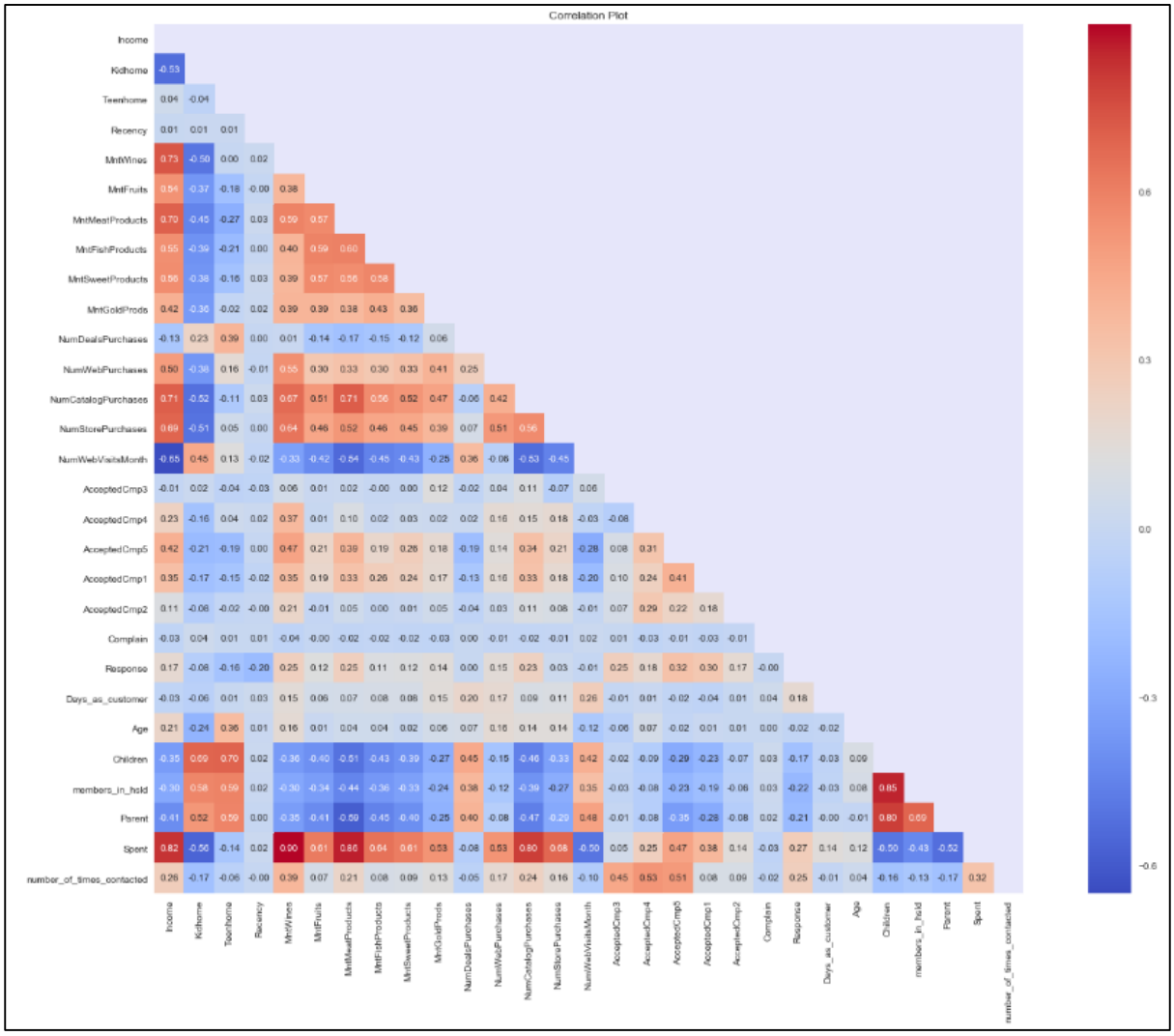
Before



After



Next we looked into the correlation plot in the variables and removed some of the features that are highly correlated or utilized to create new variables

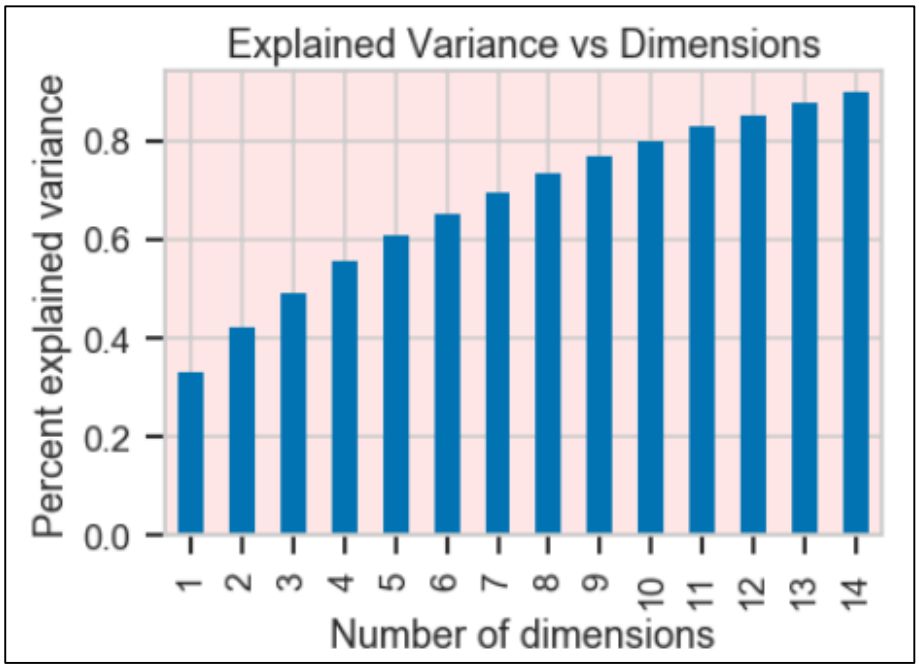


Label Encoding – As part of the feature engineering process there were some categorical variables like marital status and education that we needed to label encode.

Dimensionality Reduction using PCA –

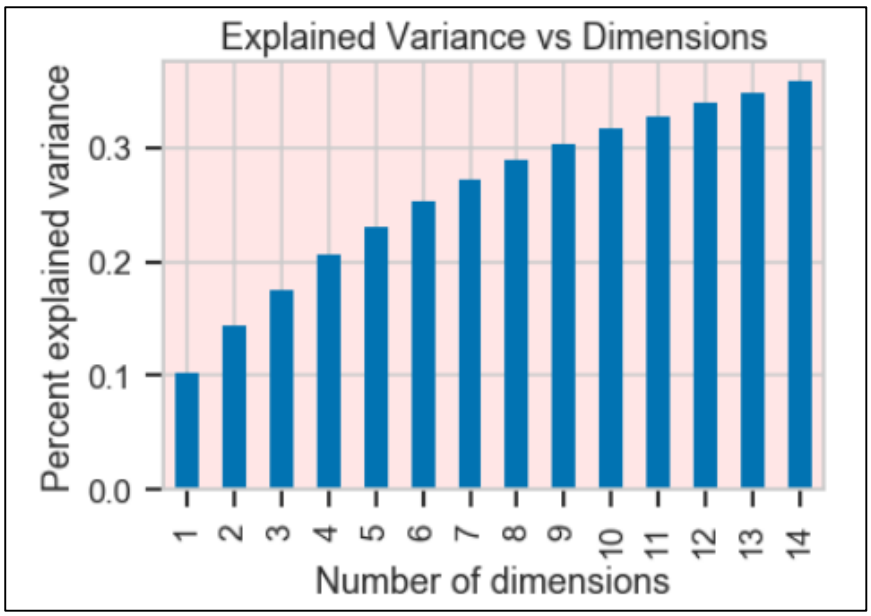
For directionality reduction I looked into both PCA, Kernel PCA since the reduced variables were not explaining much of the variance in the data

PCA –

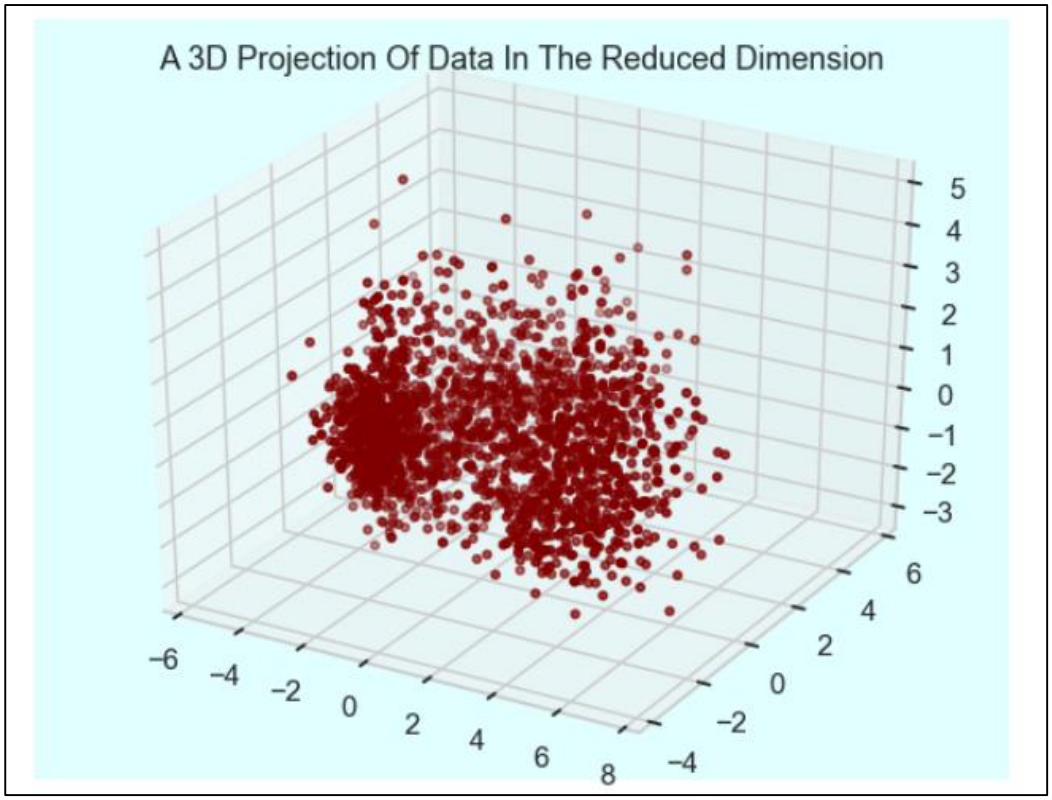


Only around 80 percent of the total variation in the data was getting explained with 10 principal components and top 3 principal components were only explaining 50 % of the total variance

Kernel PCA –

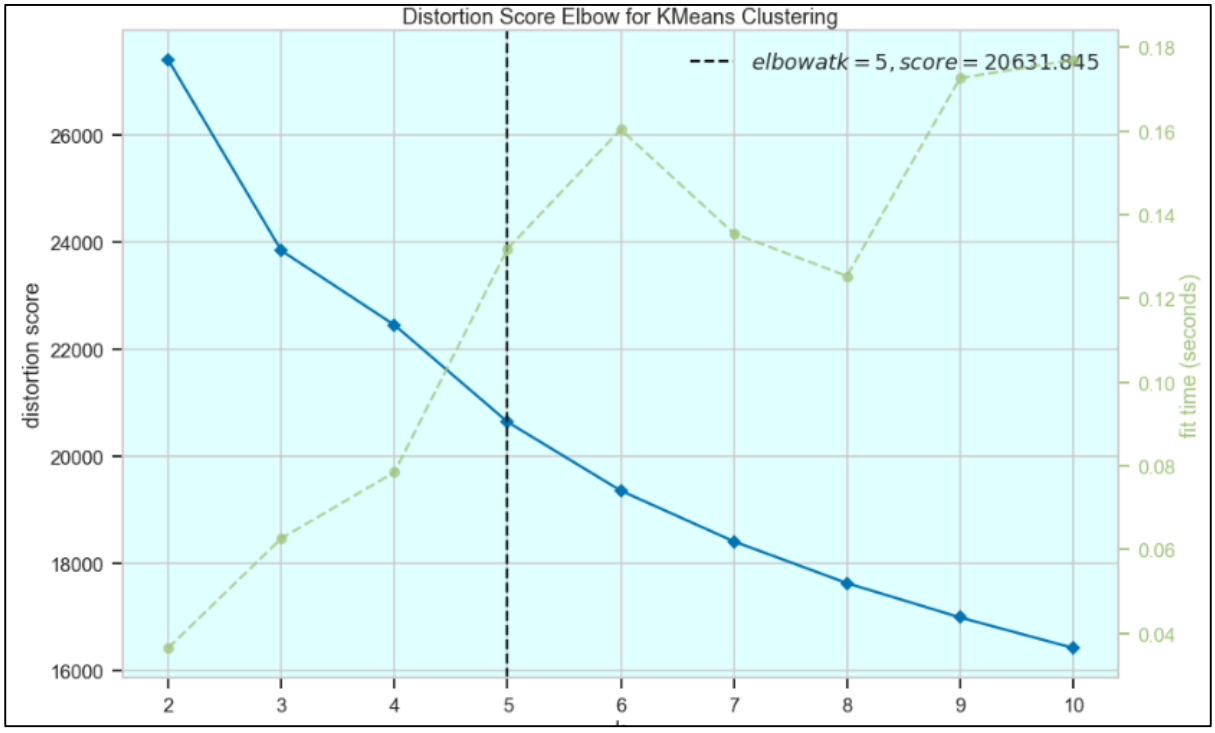


As it can be observed from the above plot that kernel PCA was performing worse than traditional PCA so I decided to go along with the top 10 components I got from the traditional PCA



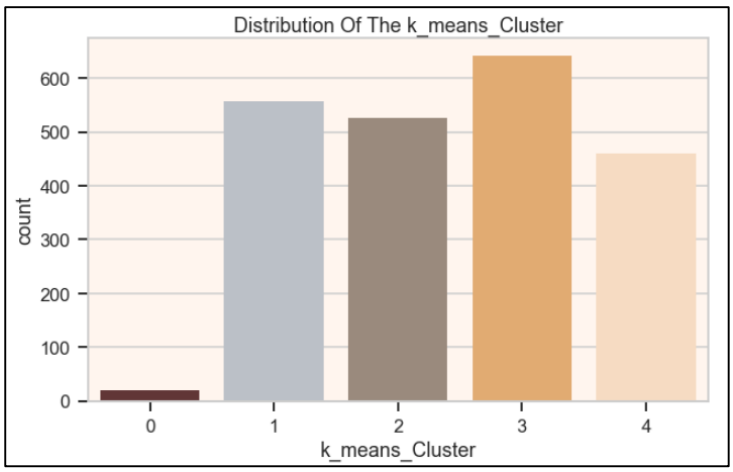
Clustering the data –

Using distortion score for elbow I looked into what should be the optimum number of clusters in the data and found that with K=5 would be the best value for clusters

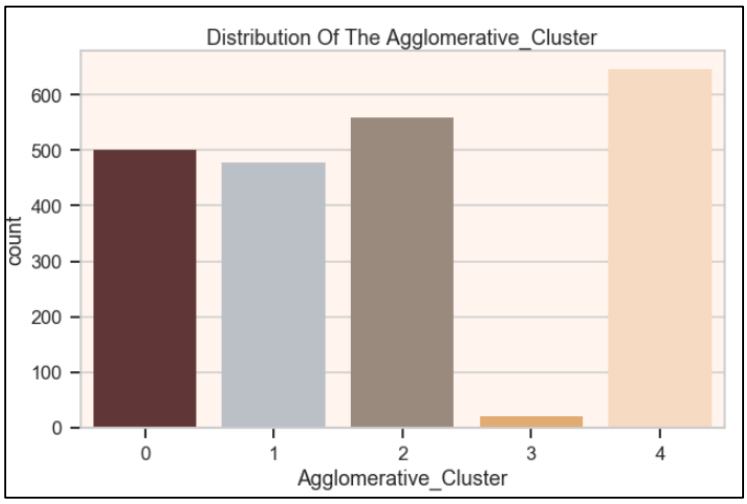


Next using K =5 I tried using 3 different clustering techniques to find which one gives me the best result

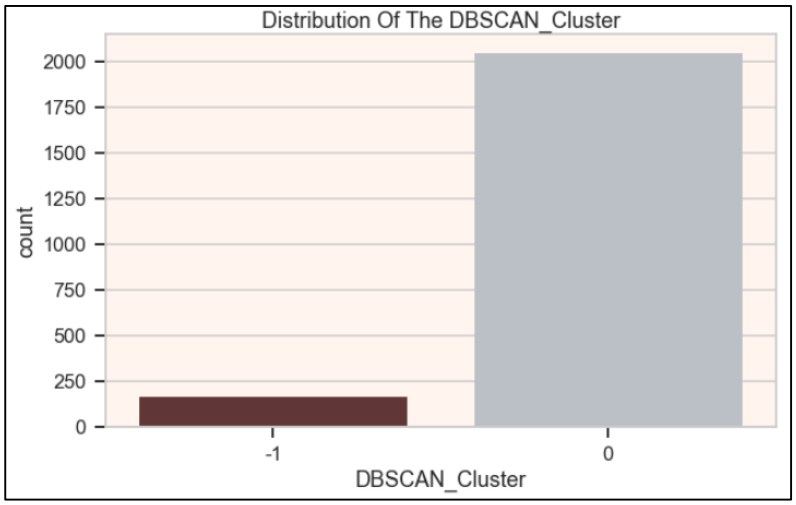
Kmeans – Using K =5 I was able to get a good divide of customers into 4 groups and the 5th group had around 20 customers, I also tried with multiple different values of K less and more than 5 as well but it was still giving me those 20 customers separately. It seems like the algorithm was able to find something in the data for those 20 customers



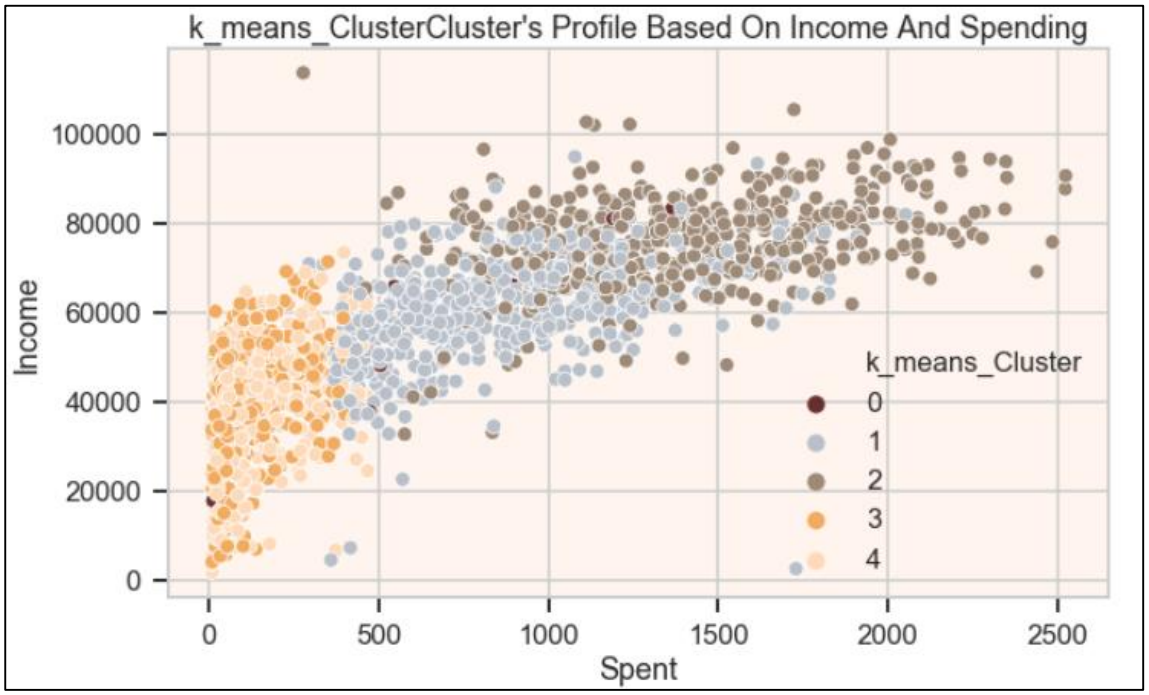
Agglomerative Clustering – Similar to K means clustering I tried clustering the data with multiple values of K but the most stable clusters came with K =5. Also for agglomerative clustering as well I was getting a 5th cluster with only 20 customers.

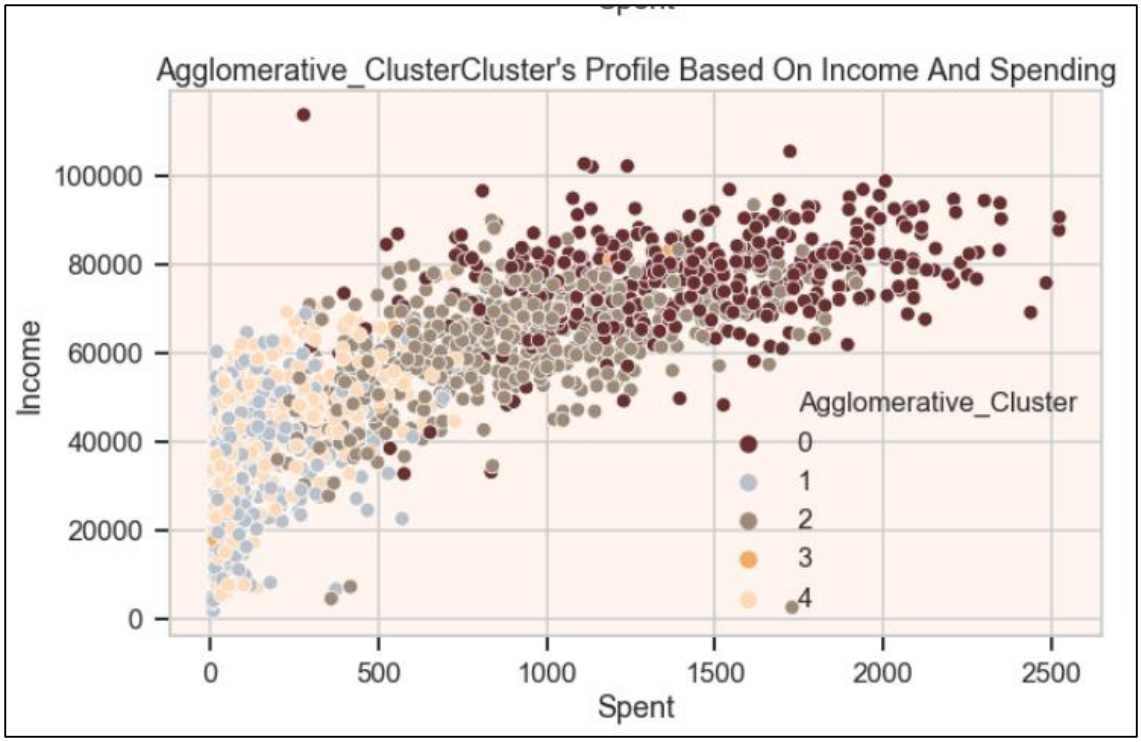


DBSCAN – I looked into DB scan as well using different values of eps and min_samples but was not able to find optimum clusters it was always putting most of the values into one cluster or creating multiple small clusters. Finally, I decided to move ahead with the K mean and Agglomerative clusters at this point



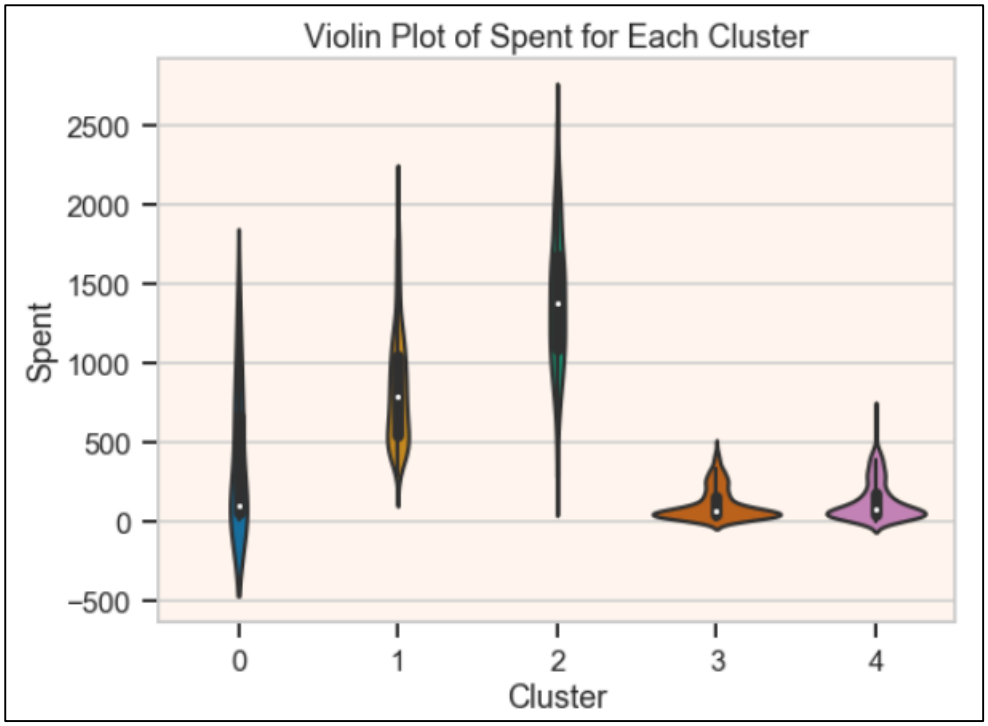
Next to examine the clusters formed I tried to look at them with different dimensions, firstly looked at them using Income Vs Spend

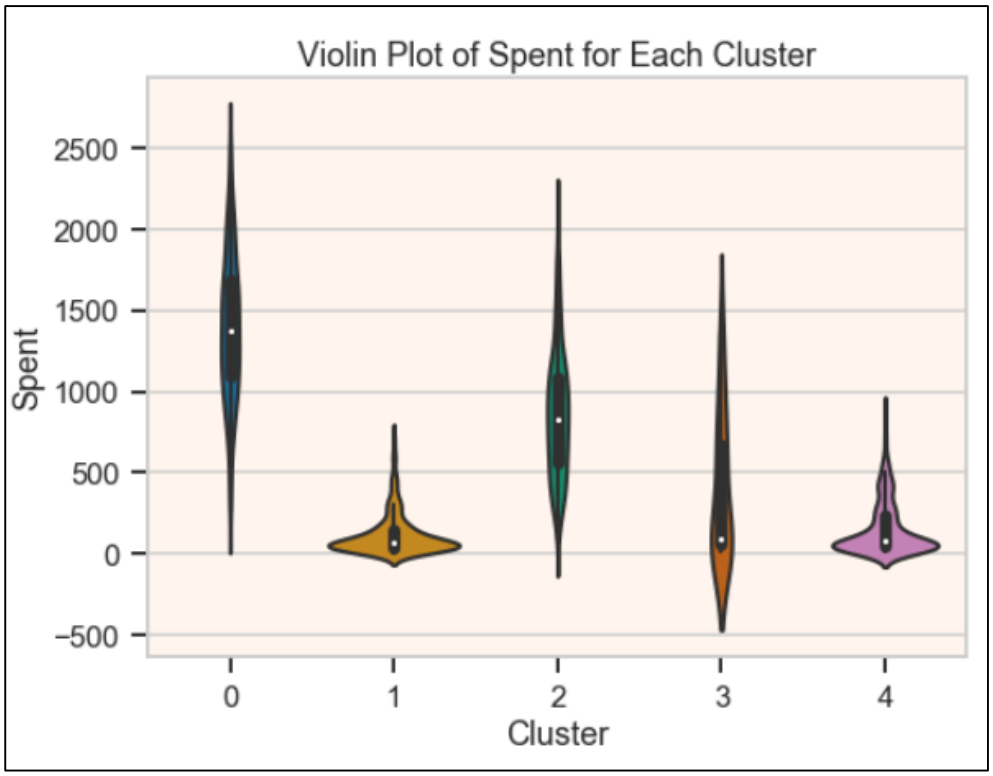




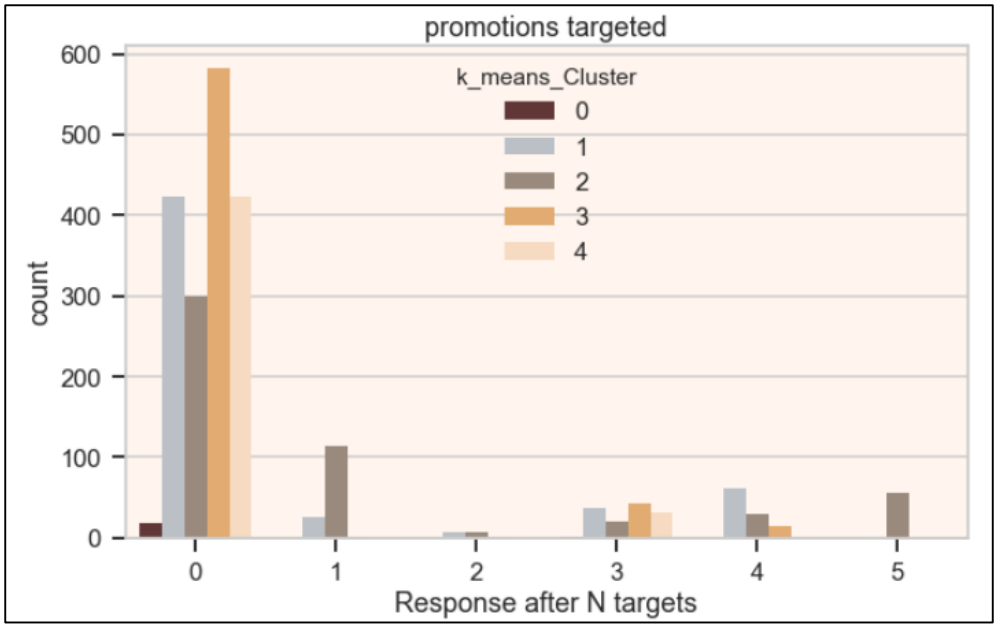
In the above plot we can see there are nearly 4 type of spending and income groups, and the clusters have identified them both in K means and agglomerative clusters

- Cluster 4 in Kmeans and 4 in Agglomerative seem be showing low income and medium spend category
- Cluster 3 in Kmeans and 1 in Agglomerative seem be showing low income and low spend category
- Cluster 1 in Kmeans and 2 in Agglomerative seem be showing medium income and medium spend category
- Small Cluster 0 in Kmeans and 3 in Agglomerative seem be showing medium income and high spend category
- Cluster 2 in Kmeans and 0 in Agglomerative seem be showing high income and high spend category





It can be observed clearly from both the plots that we have 2 clusters predominantly that are spending more compared to the other 3, Next I tried to look into some more variables before starting with the profiling



Overall it can be observed that there is not much response to the different campaigns, but of the response received, looks like cluster 1 and 2 tend to respond more, these are the high and medium income groups who tend to spend more.

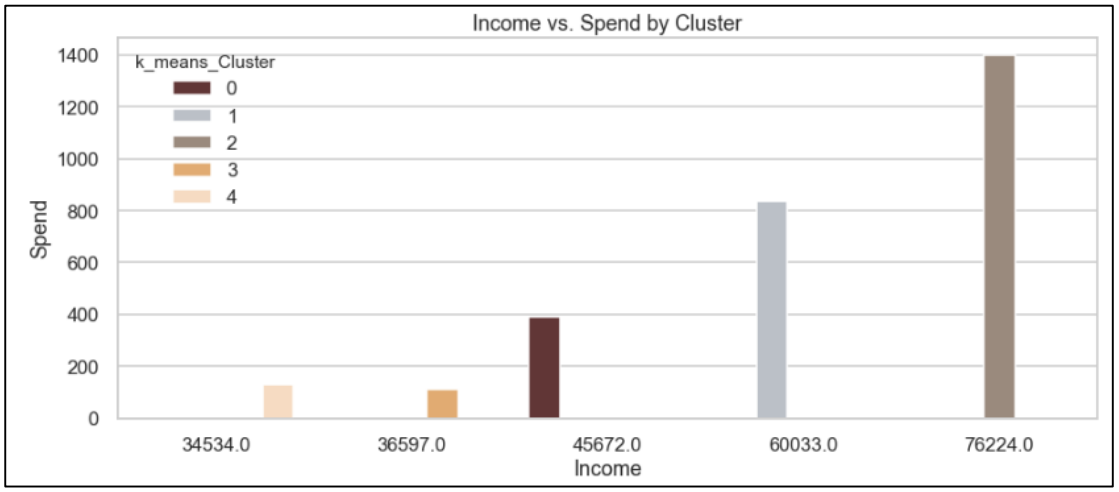
Since Kmeans clustering and agglomerative clustering were both giving similar results then it could be decided that either of the two can be used for the cluster profiling. In My cas I went ahead with Kmeans clusters

Cluster Profiling –

With my Clusters now ready, next thing comes is to learn about these cluster behaviors, to understand the underlying customers. Which of these customers are the best and which customers need more attention. Maybe certain different strategies that needs to be taken to understand their behavior.

Spend Vs Income:

First thing that I wanted to do is to look at the average spend of these clusters and also the average income

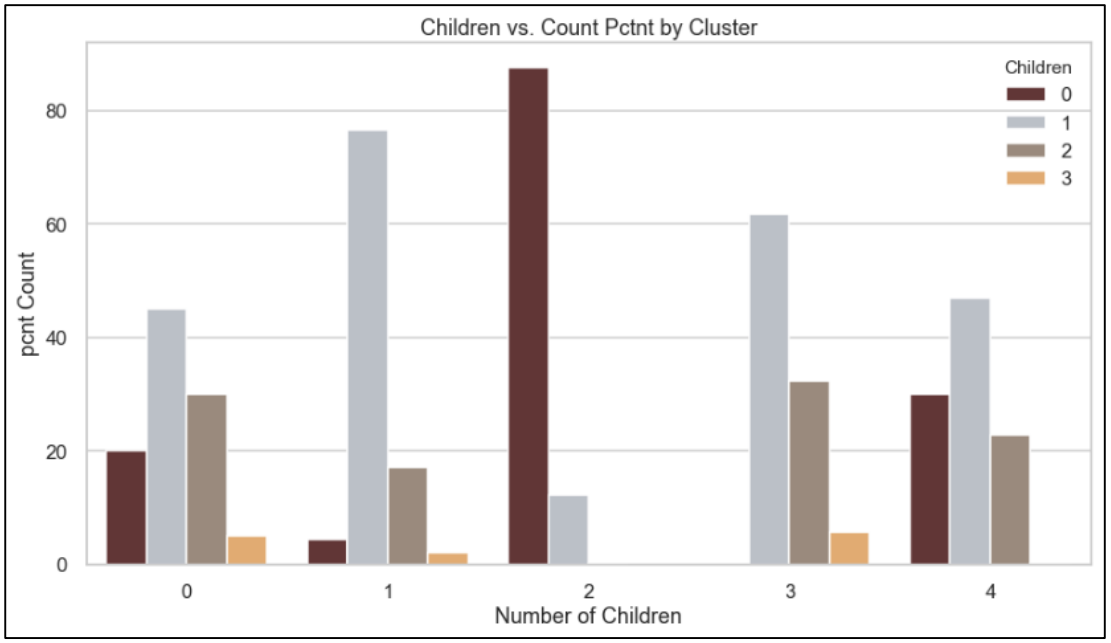


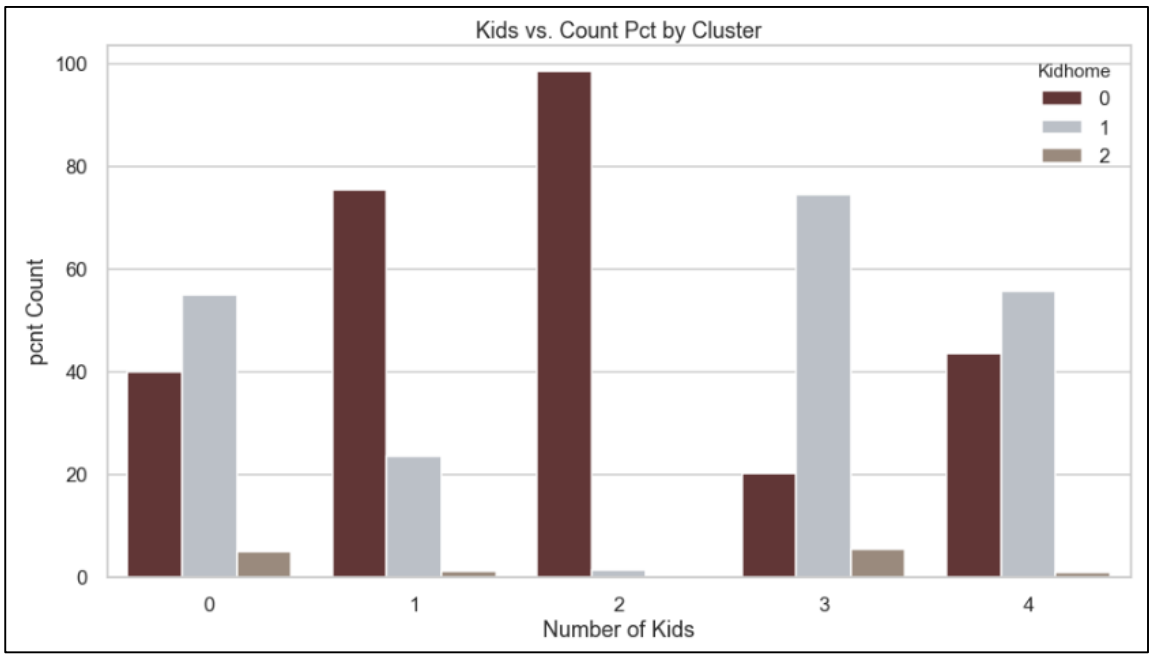
Based on the above plot it can be said pretty easily that cluster 2 is our star cluster that tends to spend more and also the average income is higher above the rest, followed closely by cluster 1

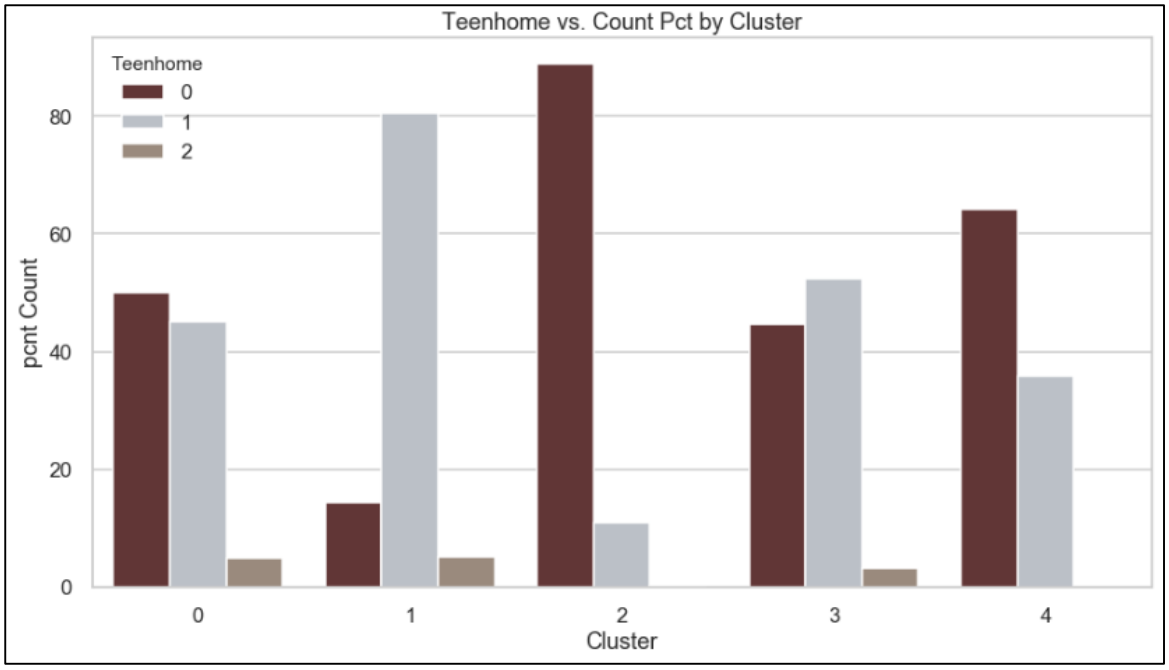
Rest of the clusters 0 3 and 4 have low and medium incomes and also spend less

Children/ Kids/ Teens –

During the EDA portion of the project we had observed there is some stark differences in spending based on a customer being a parent or not a parent, keeping that in mind I wanted to look at the cluster behaviors based on Children.





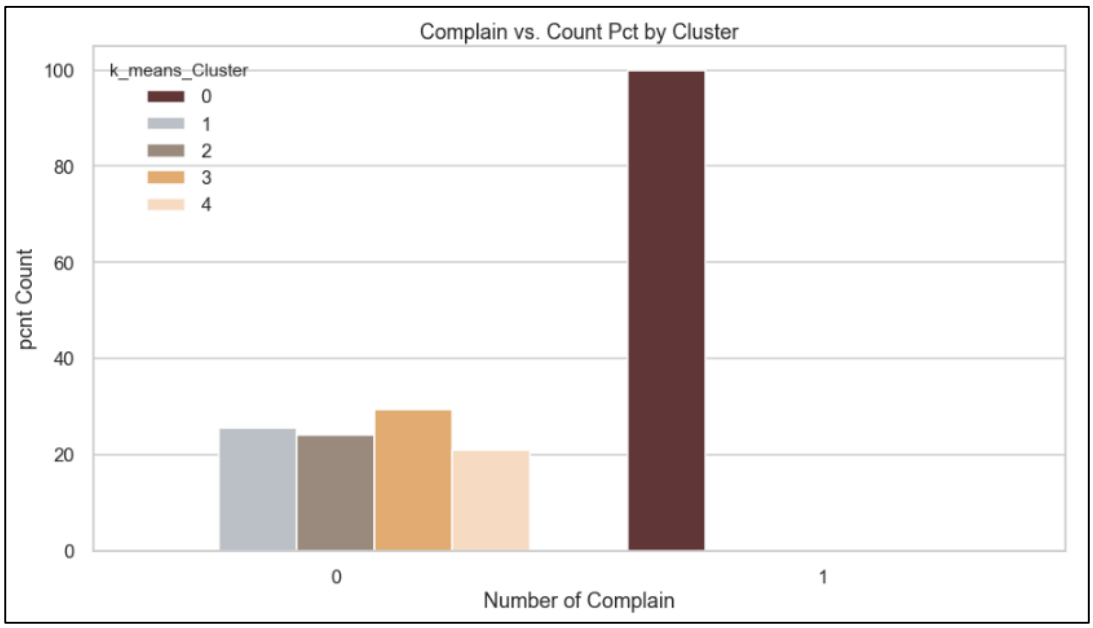


Based on the above plots it can be confirmed that

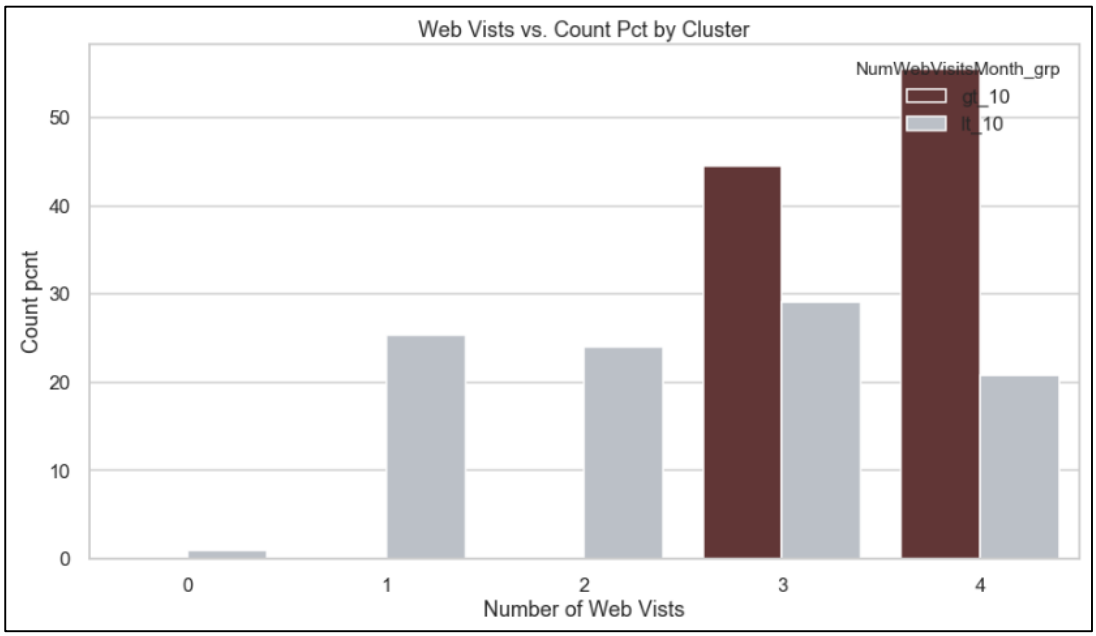
- Cluster 0 – tends to have 1 kid and 1 teen at home, sometimes 3 kids
- Cluster 1 – tends to have 1 teen at home and less chance of having kids
- Cluster 2 – tends to be customers with 0 kids or teens
- Cluster 3 – tends to have atleast 1 kid, sometimes 3
- Cluster 4 – tends to have 1 or 2 kids at home

Complain –

Using the variable complain I was looking into the clusters that are not happy with the service and I found that all of the complains came from cluster 0 which means they are not happy with the service and needs to be looked at by customer service



Number of Web Visits:

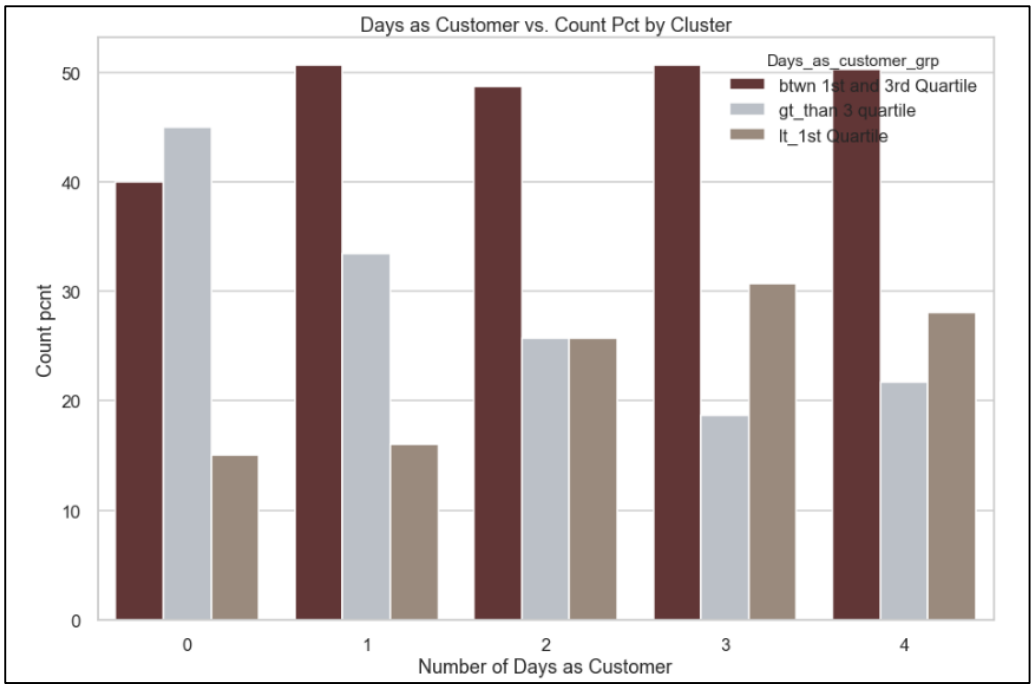


- Cluster 3 and 4 tend to make more than 10 web visits in the per months
- Cluster 0, 1 and 2 don't tend to make much of web visits

Customers from cluster 3 and 4 can be targeted with ads to direct them to the web pages and also try and make the product quality look good in the web pages

Days as Customers:

I looked into the quartile distribution of the numbers of days as customer and divided the data into 3 categories, i.e. less than 1st quartile, 1 to 3 quartile and greater than 3rd quartile

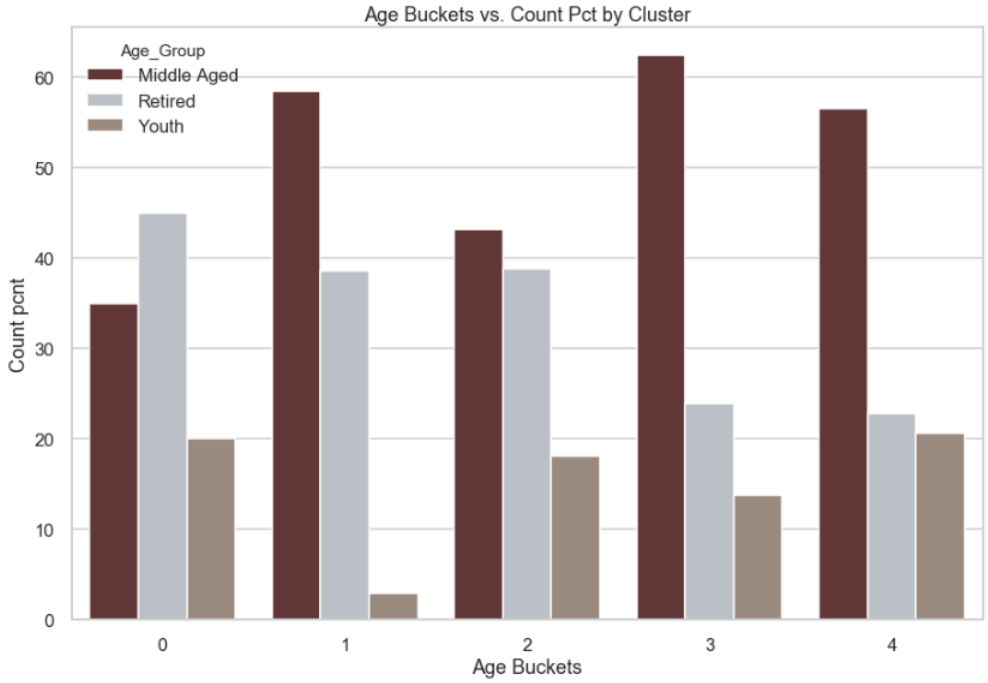


With this plot it can be said that

- Cluster 0 and 1 are more old time customers and needs to be made sure that their lifetime value increases further by providing coupons or gift cards
- Cluster 2 and 3 are more in between, these customers can be targeted with price discounts and promos to increase their LTV
- Cluster 4 are more new customers and we need to make sure that they stick around longer

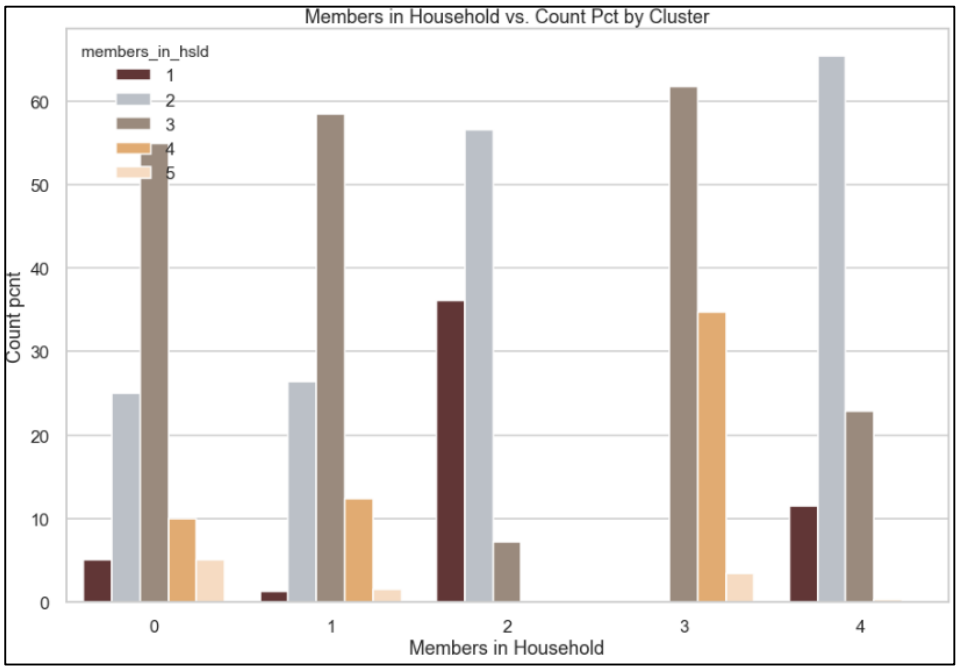
Age:

I divided age as well into 3 buckets, anything less than 40 as youngsters, 40-60 as middle aged and above 60 as retired and looked into the clusters



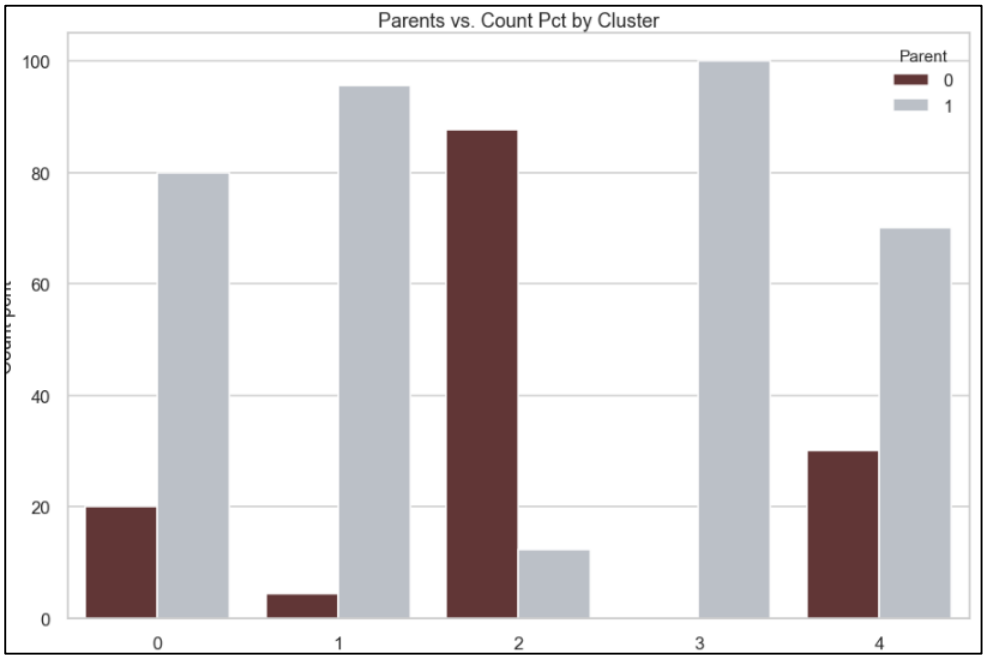
- Cluster 0 dominated by retired customers but has a mix of middle-aged and youngsters
- Cluster 1 is mostly mix of middle-aged and retired people
- Cluster 2 is mix of retired and middle-aged
- Cluster 3 and 4 are mostly middle-aged

Members in Household:



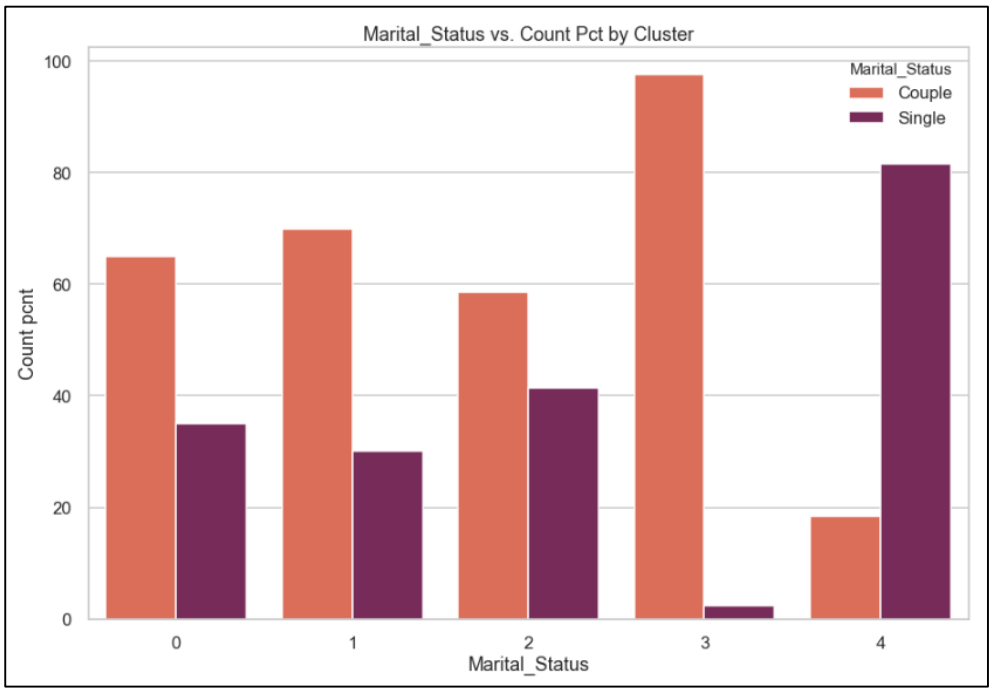
- Cluster 0 has bigger families – 4 to 5 members
- Cluster 1 has mostly 3 member families
- Cluster 2 is smaller either single or 2 members' household
- Cluster 3 has no single or dual family household, mostly 3-4
- Cluster 4 is a 2 to 3-member family

Parent:



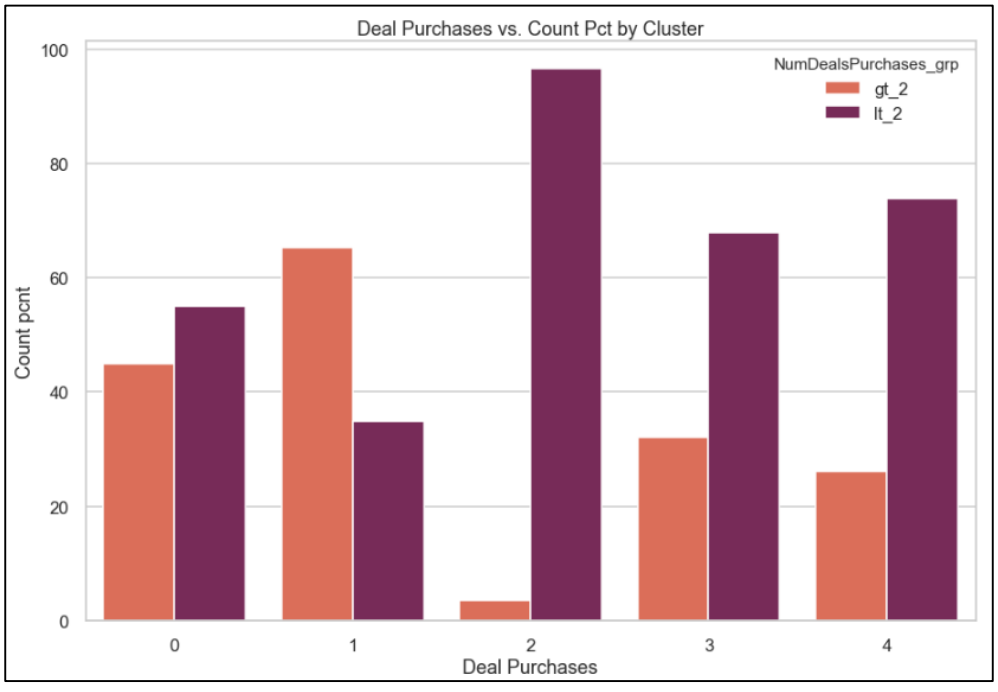
- First thing that we see is that –
- Cluster 0 and 4 are mostly mix of parent and non-parent
- Cluster 2 is mostly non-parent
- Cluster 3 is all parent

Marital Status:

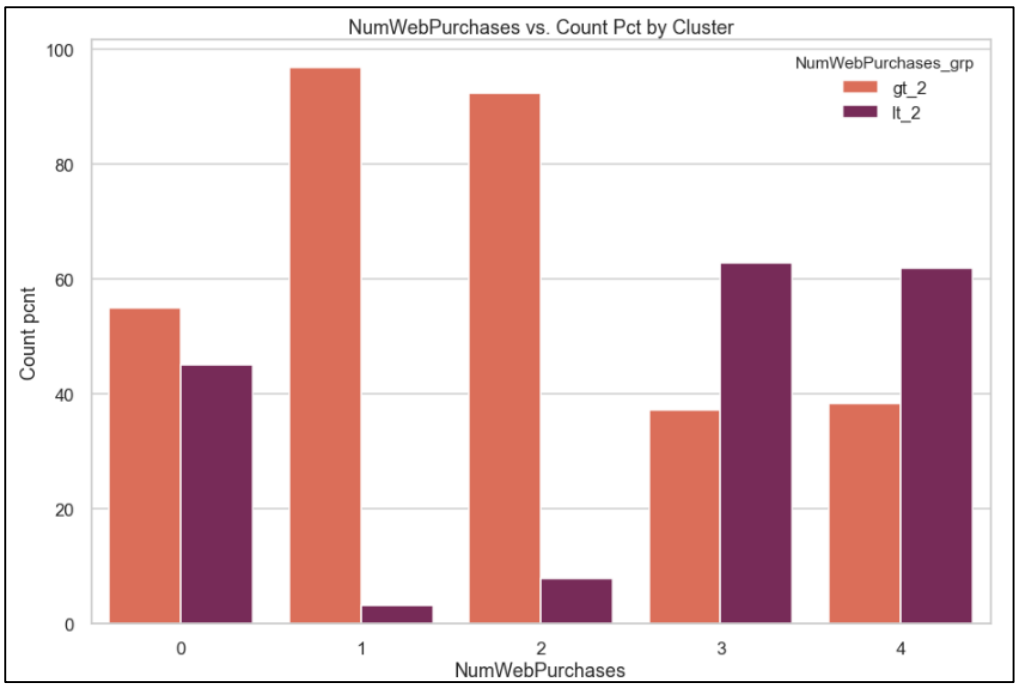


- Cluster 0 1 and 2 are a mix of single and couples
- Cluster 3 is mostly couples
- Cluster 4 is mostly singles

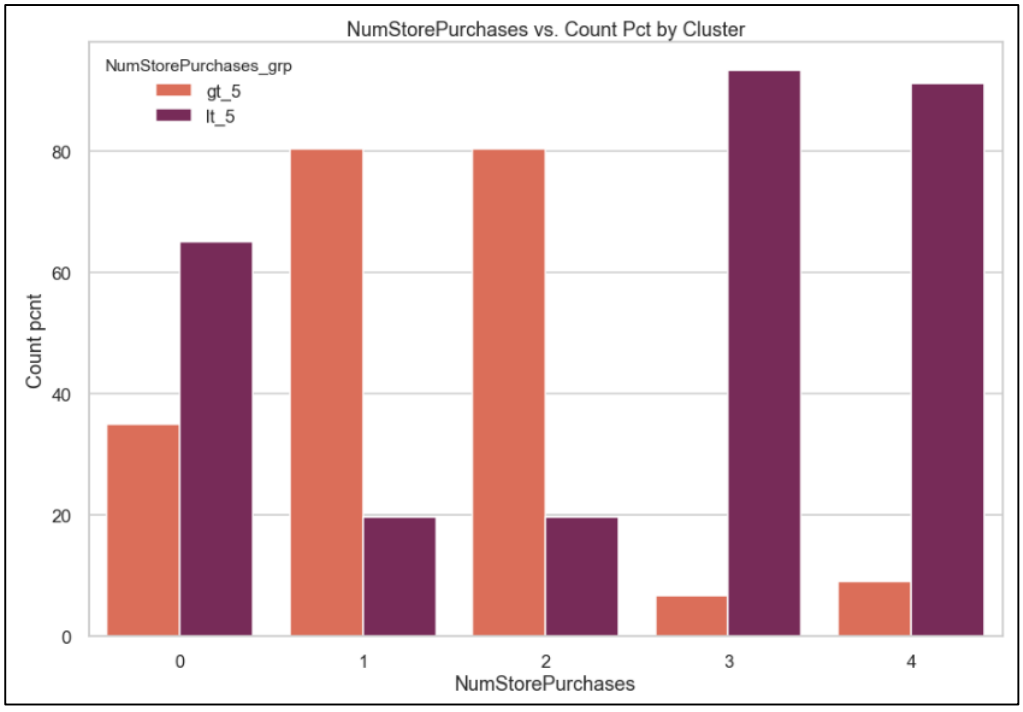
Purchases:



- Cluster 0 and 1 tend to purchase more on deals
- Cluster 2,3,4 purchase less deals



- Cluster 0,1,2 make more web purchases
- Cluster 3 and 4 does not make much web purchases



Cluster 0,1 and 2 make more store purchases whereas cluster 3 and 4 make less store purchases

Summary of clusters

Cluster 0: Older couples with big families' middle income middle spend -

- Income – Medium
- Spend – Medium
- Number of people in the cluster – 20 least
- Mostly Bigger families with one or more teens and kids
- This cluster also complains the most
- Mostly older couples and retired
- These are long term customers
- Do not respond to promotions, deals but make lot of web visits

Resolutions – These are usually older couples with big families –

- Provide them with big family size products
- Makes a lot of web visits maybe we can advertise such products and deals in the web
- Long term customers provide them with gift coupons
- Retired people provide senior citizen discounts
- Customer support need to engage with them as they tend to complain the most

Cluster 1: Middle aged couple with smaller families' middle income high spend:

- Income – Medium
- Spend – High
- 3 member families and mostly kids are teen
- Middle-aged couples
- Old time customers
- Most educated among all the clusters
- Makes a lot of web Visits
- Makes a lot of in store visits
- Responds to Promotions and Deals but after 4rth attempt

Resolutions Middle aged couples with smaller families–

- These customers are already spending high
- Has lot of teenagers in the house can be targeted with new age products
- Respond to deals and promotions can be targeted with deals for smaller families
- Old time customers can be provided with gift coupons

Cluster 2: Middle aged to retired couples Income High spends high

- Income – High
- Spend – High
- Smallest families mostly single
- Middle aged to retired people
- Makes a lot of web visits
- Makes a lot of in store purchases
- Responds to promotions

Resolutions Middle Aged Singles

- Should not be targeted with Kids products
- Target with small products for single people
- Responds to promotions so can be targeted with deals though web pages
- More in-between customers in terms of loyalty
- Rich category can be targeted with luxurious products and apparels

Cluster 3 – Middle aged less income less spend

- Income – Less
- Spend - Less
- Smaller families
- Mostly Married people with 2 kids
- Middle aged
- Mostly newer customers
- Education background has a mix of Graduate and High school people
- Does not make in store purchases
- Does not make web Visits
- Does not respond to promotions

Resolution:

- These are one of the two clusters that are not buying much
- Do not respond well to promotions
- May spend less on targeting this cluster and focus more on people from other clusters

Cluster 4 – Young couples less income less spend

- Income – Least
- Spend – Less
- Compact families with kids mostly
- Younger people
- Less educated mostly High school
- Newer customers
- Does not purchase from Stores, Web
- Does not respond to offers or deals

Resolution:

- These are younger couples with small means of income, can be attracted with large discounts
- Harder to convert but are similar to Cluster 0 and 3 so some of them can become long term customers
- Deals need to be specifically made keeping in mind lower income customers

Next Steps –

Overall I was able to determine good information about the clusters our smallest cluster of 20 people actually was quite a bit different from other customers and provided some good information. Maybe right now due to lack of data this cluster is not having a lot of people but that can change

Cluster 2 and 1 are our star customers and keeping that in mind some good promotions and deals can be made to attract these customers but further more information is required to be able to do that. Right

now all the information about the products sold in the shop is too low to be able to confirm which products might look good to the customers with teens vs customers who are single

Cluster 3 are customers that I felt could be left from targeting as they have very less chances of converting and we can focus ourselves to the other customers

Cluster 4 being the most younger set of couples can be targeted specifically with products popular to younger people but currently the data does not provide us with a lot of information about products