Project Title – CampaignOptiSales

Measuring marketing campaign peformance to optimize future marketing strategies and enhance product placement for increase sales

Abdul Haseeb Mohammed
dept. Business Intelligence and Systems Infrastructure
Algonquin College of Arts & Technology
Ottawa, Ontario

Md Irteza Chowdhury
dept. Business Intelligence and Systems Infrastructure
Algonquin College of Arts & Technology
Ottawa, Ontario

Mohamed AbdelSalam
dept. Business Intelligence and Systems Infrastructure
Algonquin College of Arts & Technology
Ottawa, Ontario

John Tacco-Melendez
dept. Business Intelligence and Systems Infrastructure
Algonquin College of Arts & Technology
Ottawa, Ontario

***Abstract*—Business organizations invest huge capital to conduct marketing. It is important for an organization to assess the performance of its marketing strategies along with evaluating the performance of a historical marketing campaign with different customer segments. Furthermore, product placement is crucial to sales of products and for this reason market basket analysis was explored. However, due to the nature of dataset market basket analysis proved unfeasible.**

Keywords—Marketing Campaign Performance Assessment, Customer Segmentation via RFM (Recency, Frequency, Monetary) Analysis, customer demographics and k-means, Market Basket Analysis.

I. INTRODUCTION

Motivation and Importance:

Business organizations spend capital on marketing strategies to increase sales. However, an organization cannot simply do marketing without a defined target customer base. Marketing strategies without a defined target customer base are blindly trying to attract customers and have less probability of attracting customers. Moreover, businesses engage in marketing efforts during specific occasions or events like Christmas, Men's Day, Women's Day, Children's Day, or seasonal campaigns during summer, winter, etc. And even during such occasions a business probably targets a selected customer, such as during children's day marketing could be family-oriented towards children, parents etc. For these reasons it is vital that a business assesses the performance of marketing campaigns in relation to different customer bases. This can help decide future marketing strategies for specific customer bases. Lastly, product placement is a huge factor affecting customer spending habits. A business needs to analyze customer spending habits to place products together in such a way that customers likelihood of purchasing products increases.

Data Source:

The dataset has been retrieved from Kaggle website. The name of the dataset is Marketing Campaign - Boost the profit of a marketing campaign. The dataset has sales data from the year 2012 to 2014. The dataset was posted on Kaggle by Rodolfo Saldanha. However, the provenance of the dataset is from SAS Institute and the collection methodology employed was Business Analytics Using SAS Enterprise Guide and SAS Enterprise Miner.

Data Characteristics:

The dataset has 29 columns and 2240 rows when it was fetched from the source. A special characteristic of the dataset is that every record in the dataset represents a unique customer and all the information about the customer in the record is aggregated information about the customer over the period of three years. Of the 29 columns in the dataset 26 are quantitative columns and 3 are qualitative columns. The quantitative columns and their descriptions are as follows:

- ID - Unique ID of the customer.
- Year_Birth – The year of birth of the customer.
- Income - customer's yearly household income.
- Kidhome - number of children in customer's household.
- Teenhome - number of teenagers in customer's household.
- Recency - number of days since the last purchase by customer.
- MntWines - amount spent on wine products in the last 3 years by customer.
- MntFruits - amount spent on fruits products in the last 3 years by customer.
- MntMeatProducts - amount spent on meat products in the last 3 years by customer.
- MntFishProducts - amount spent on fish products in the last 3 years by customer.
- MntSweetProducts - amount spent on sweet products in the last 3 years by customer.
- MntGoldProds - amount spent on gold products in the last 3 years by customer.
- NumDealsPurchases - number of purchases made with discount by customer in the last 3 years.
- NumWebPurchases - number of purchases made through the company's web site by customers in the last 3 years.
- NumCatalogPurchases - number of purchases made using catalogue by the customer in the last 3 years.

- NumStorePurchases - number of purchases made directly in stores by customer in the last 3 years.
- NumWebVisitsMonth - number of visits to company's web site by customer in the last month.
- AcceptedCmp3 - 1 if customer accepted the offer in the 3rd campaign, otherwise 0.
- AcceptedCmp4 - 1 if customer accepted the offer in the 4th campaign, otherwise 0.
- AcceptedCmp5 - 1 if customer accepted the offer in the 5th campaign, otherwise 0.
- AcceptedCmp1 - 1 if customer accepted the offer in the 1st campaign, otherwise 0.
- AcceptedCmp2 - 1 if customer accepted the offer in the 2nd campaign, otherwise 0.
- Complain - 1 if a customer has complained in the last 3 years, otherwise 0.
- Z_CostContact - The metadata about this column has not been provided.
- Z_Revenue – The metadata about this column has not been provided.
- Response - 1 if customer accepted the offer in the last campaign, otherwise 0.

The qualitative columns of the dataset and their descriptions:

- Education – The educational level of the customer.
- Marital_Status – The marital status of the customer.
- Dt_Customer – the date when customer first enrolled with the company.

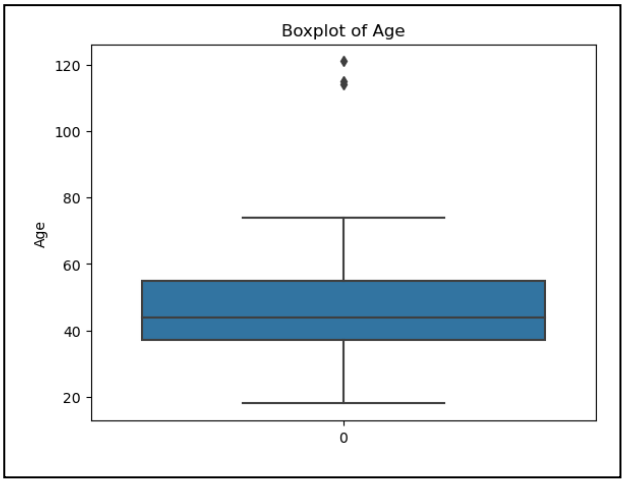
II. Data Cleaning & Preparation steps taken:

- Many columns such as AcceptedCmp1, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5, Response and Complain are flag columns. By using descriptive statistics, we found the min value for these columns was 0 and max value was 1 so additional exploration of distribution and cleaning of these columns was not required.

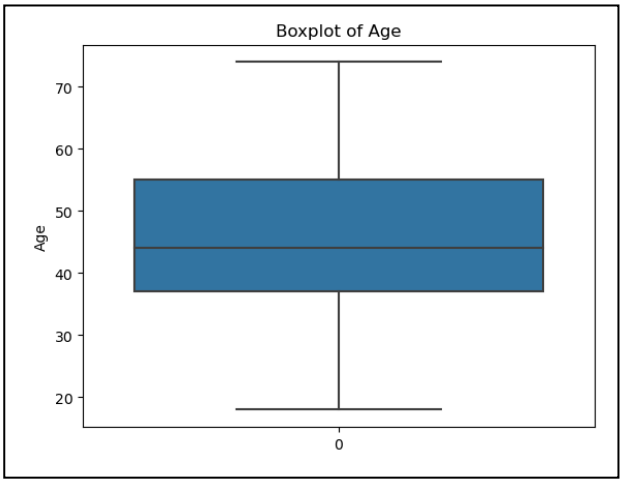
Dropping unncessary columns from dataframe

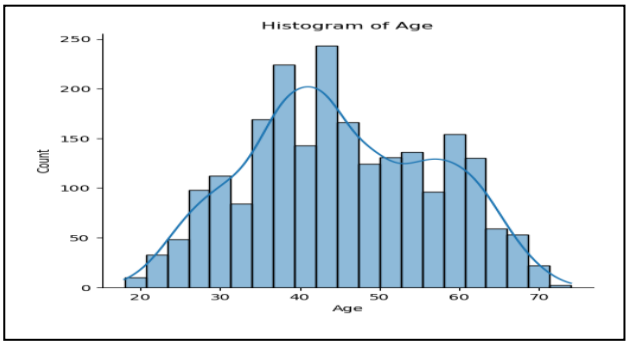
```
In [9]: df['Z_Revenue'].value_counts()
Out[9]: Z_Revenue
        11    2240
        Name: count, dtype: int64

In [10]: df['Z_CostContact'].value_counts()
Out[10]: Z_CostContact
         3    2240
         Name: count, dtype: int64
```

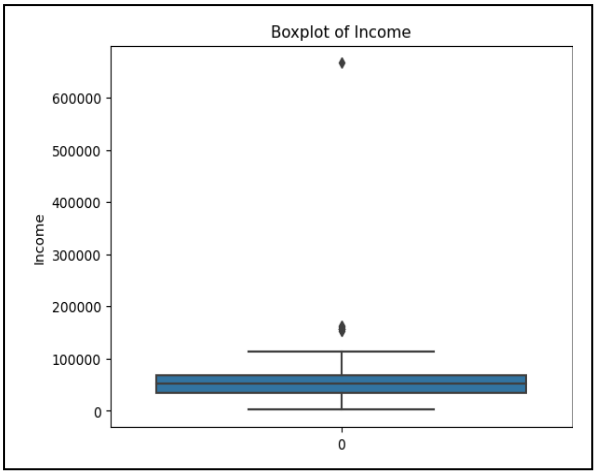
- The columns Z_CostContact and Z_Revenue were unnecessary columns adding no value to the analysis as their metadata was not provided by the source additionally these columns had only 1 unique value, so these columns were dropped.
- The Dt_Customer column was of object type and was converted to appropriate date format using dateutil.parser.
- The year birth column was used to create an Age column, instead of using the difference from current year to year of birth of the customer. The difference was calculated from the year 2014. As the dataset had customer data from 2012 to 2014.

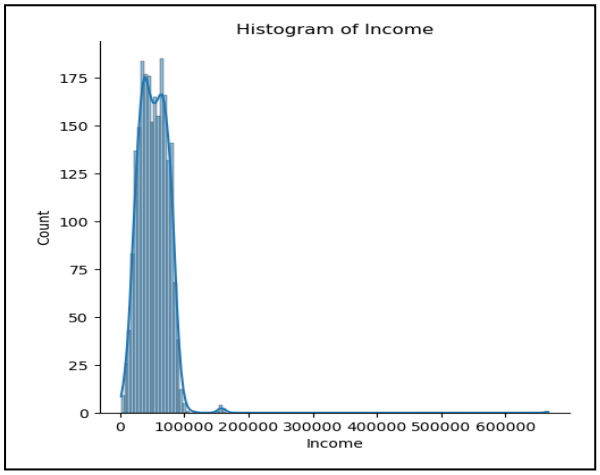


- The Age column had 3 extreme data points which were removed as the ages of these customers were greater than 110 years. Even though the year 2014 was used to calculate the age of the customers which was slightly affecting the distribution of the column. Also, the distribution of age column was normally distributed after removing these data points.

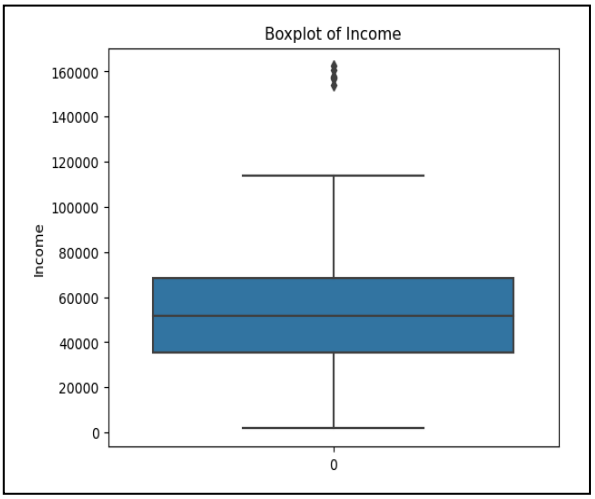




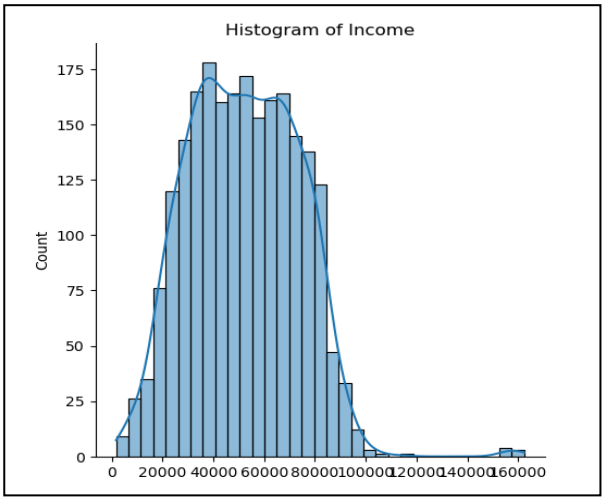




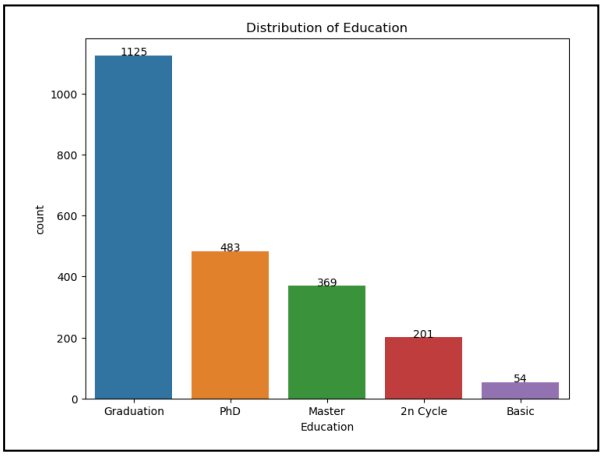
- The income column was explored using boxplot and we found only one customer whose income was greater than 600k and the rest of the customer's income was under 200k, also, there was almost no spending by these customers on any of our products. Also, the income column was extremely skewed due to these extreme data points. so, this extreme data point was removed for being an outlier.



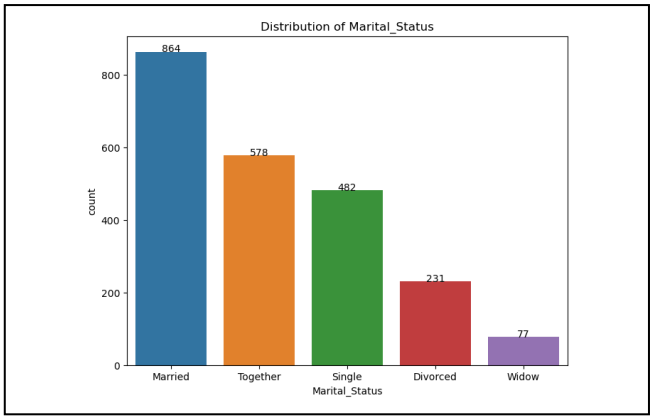
- On exploratory analysis of dataset, we found that only income column had missing values. To handle the missing values, we choose to impute them. The method of imputing chosen was to find the average income of a customer's educational level and replacing the missing income of said customer with the average income of their respective educational level. The Income column was again checked for its distribution, and we found 7 customers with higher income but their incomes were very close and the distribution of income columns seemed close to symmetric so these points were not removed.



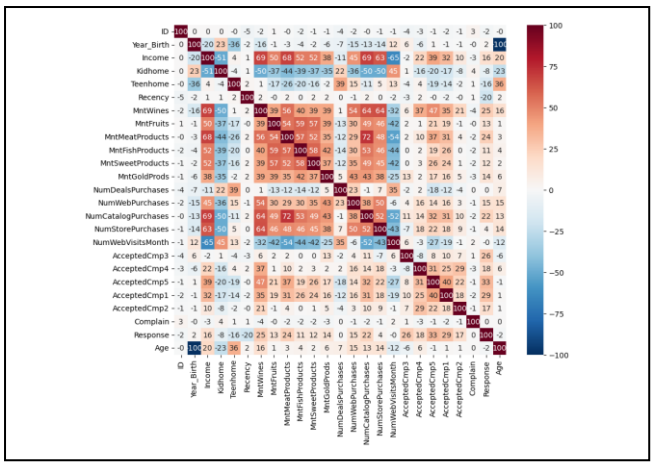
- The education level was checked for inconsistent values using count plots and we found classifications which made sense.



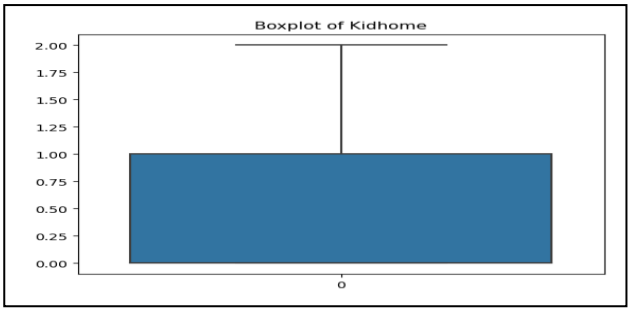
- The above image shows distribution of education column.
- The Marital Status was checked for inconsistent values, and we found a few classifications which made no sense such as Absurd, YOLO. The dataset had only 4 customers classified as absurd and YOLO, so these records were removed. Also, a redundant classification for single was used by the name alone. The 3 records with alone marital status were re-classified as single.

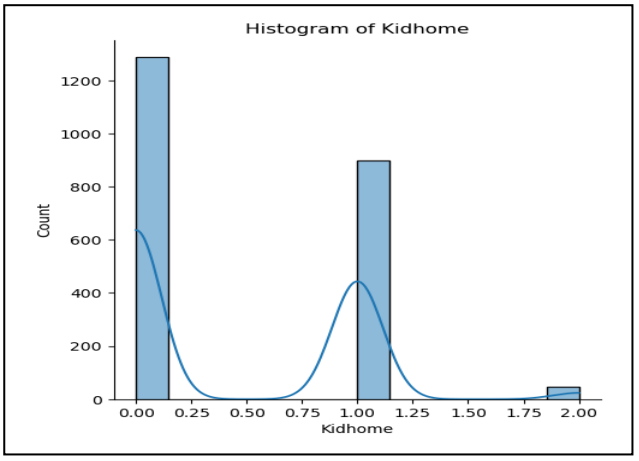


- The above image shows the distribution of Marital_Status column.

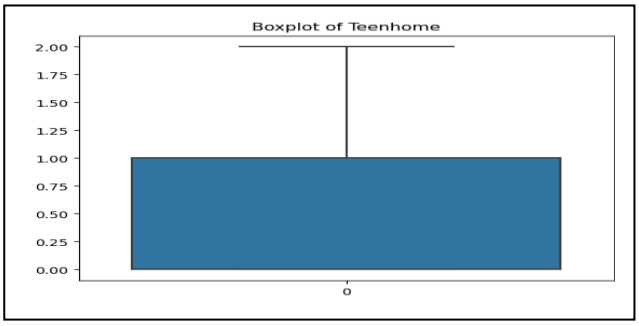


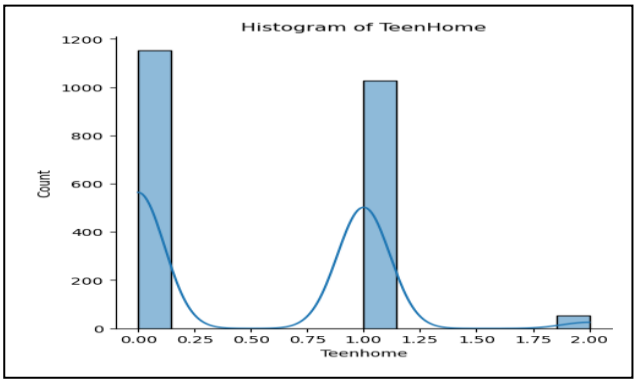
- We can observe the Income column has moderate correlation with the six-product categorical column in the dataset. Here the columns are Mnt meaning amount spent on followed by product category name such as amount spent on Wines, Fruits, Meat Products, Fish Products, Sweet Products, and weak correlation with amount spent on gold products column indicating as income increases spending of customers on products increases which makes sense.
- Also, the Kidhome column seems to have negative correlation with amount spent on products which doesn't make sense as large families would need to spend more on groceries. Indicating the customers with children do not choose our store for groceries. This is a huge blow to the business. The marketing team needs to attract customers with children. Also, customers with Kidhome moderate negative correlation with NumCatalog purchases, NumStore purchases indicate customers with young children have some issue with shopping in our stores. TeenHome also has weak and almost no negative correlation with the amount spent on products. Meaning customers with children are not spending money on our business.
- The ID column is only used to uniquely identify customers and doesn't show any inconsistent values.



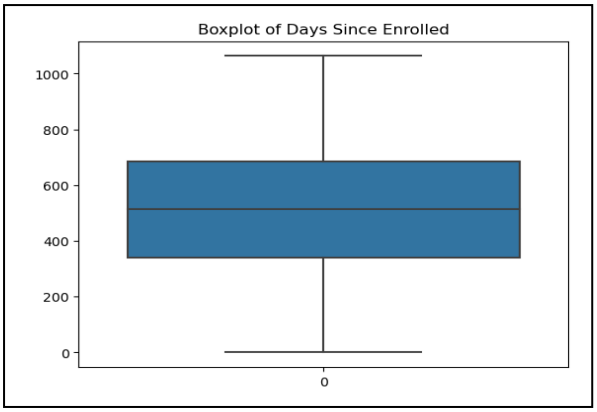


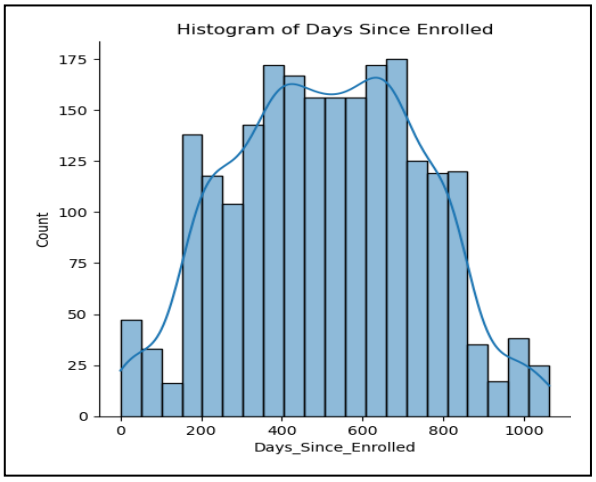
- The Kidhome column doesn't have any outliers but was slightly positively skewed. Its skewness was 0.63 and for this reason it was transformed using square root and its skew became close to symmetrical.



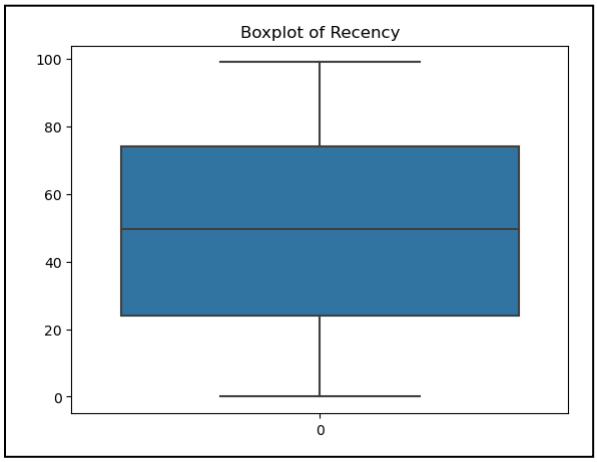


- The TeenHome column had no outliers and had symmetrical distribution, so it wasn't transformed.

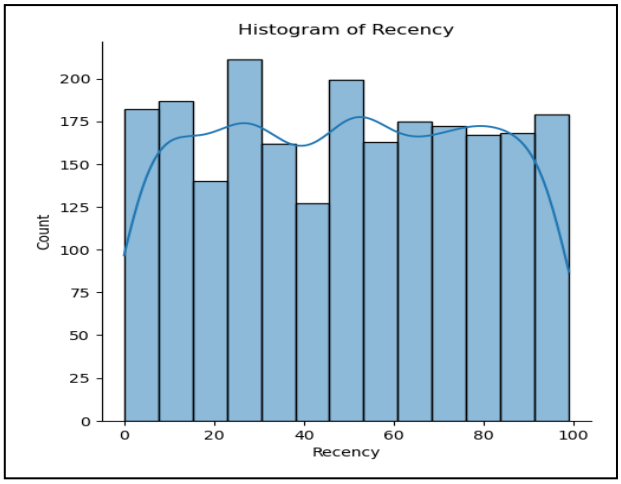




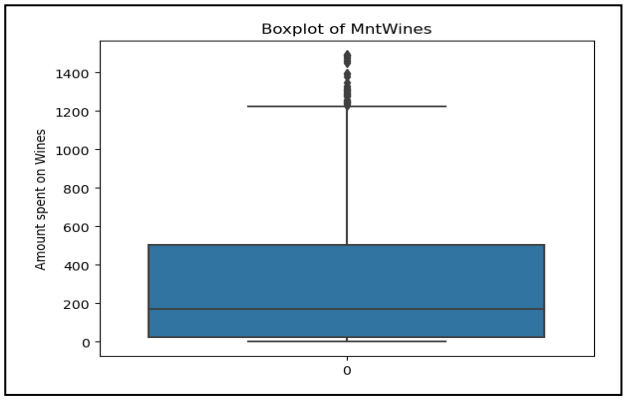
- The Dt_Customer column cannot be used directly with models, so it was used to create a new column called Days_Since_Enrolled which showed number of days since customer enrolled with company. The calculation was done by finding the difference between the customer's most recent enrollment date to said customer's enrollment date.

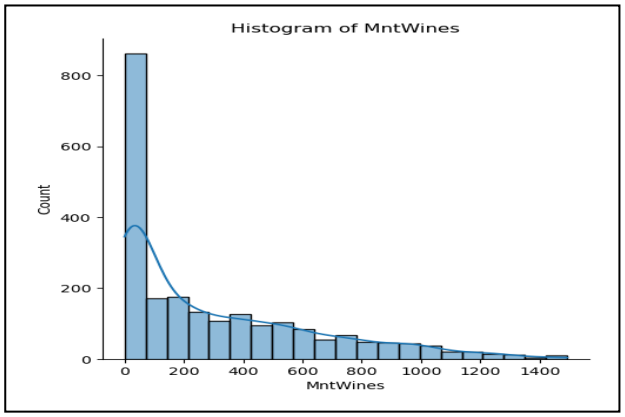


- The recency column had no outliers.

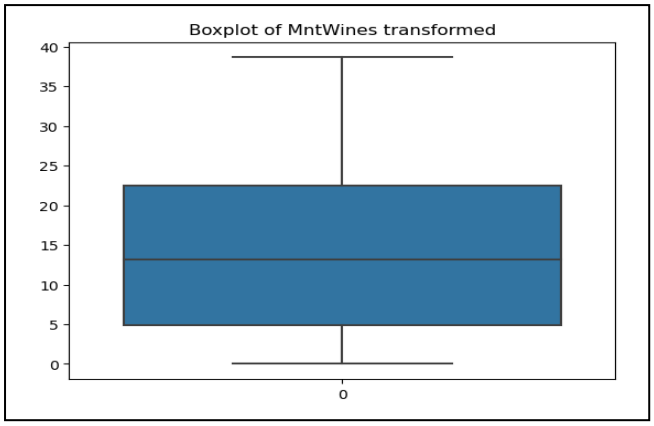


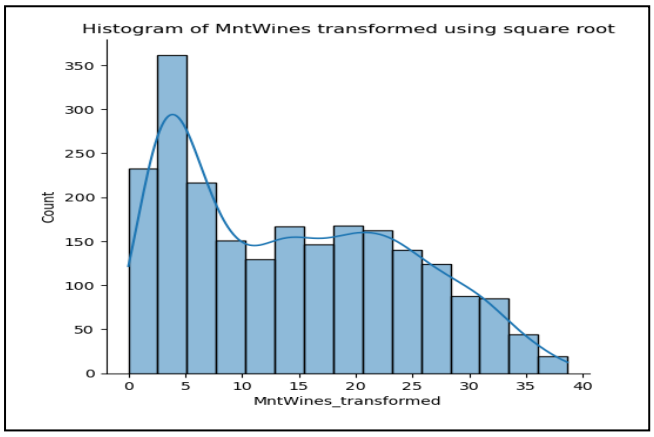
- The distribution of the recency column was symmetrical.



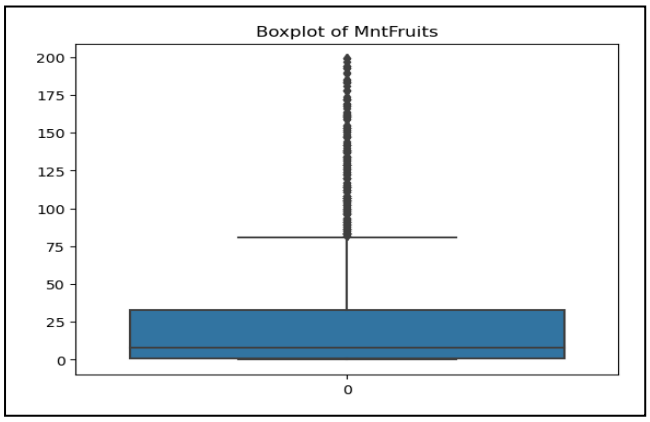


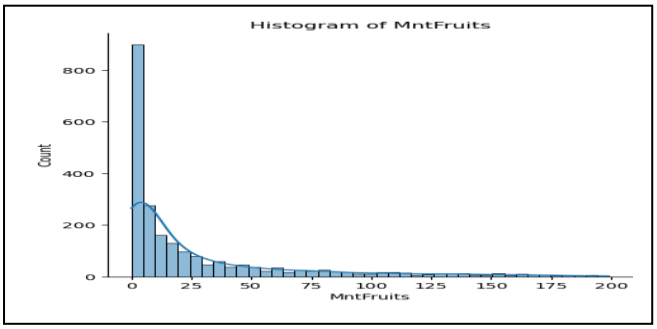
- The MntWines column had some extreme points as these data points were close to each other they were not removed. However, the MntWines column was extremely positively skewed and was transformed.



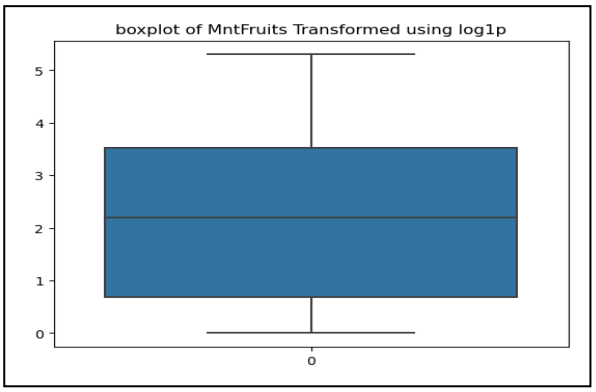


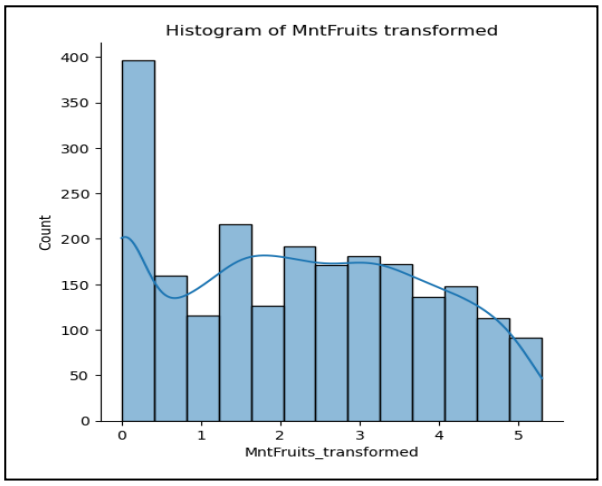
- The boxplot and histogram of MntWines column transformed using square root is shown above. The skewness of MntWines column has reduced to 0.394 which is close to symmetric distribution.



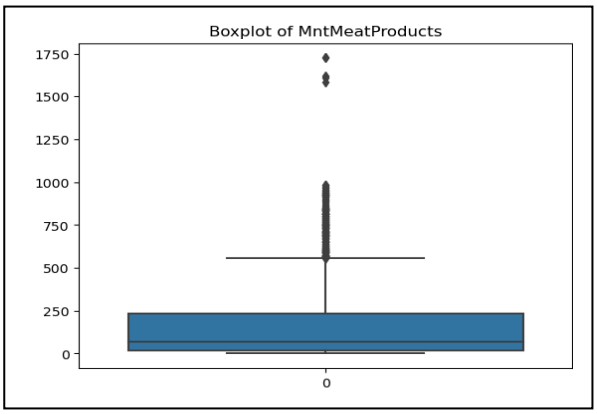


- MntFruits has many data points which are potential outliers but as they are close to each other they were not removed. Also, The MntFruits column was extremely positively skewed and was transformed.





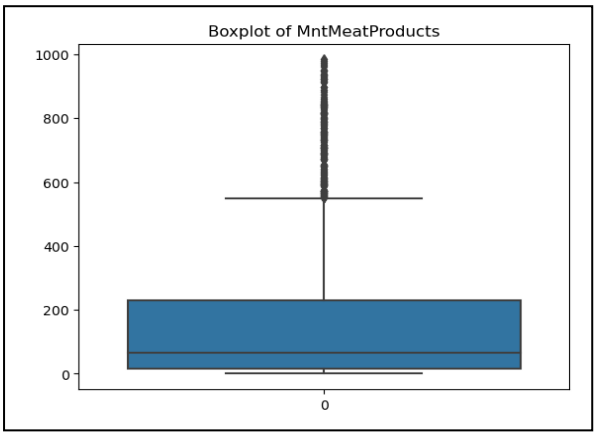
- The box plot and histogram of MntFruits transformed using log1p is shown above.

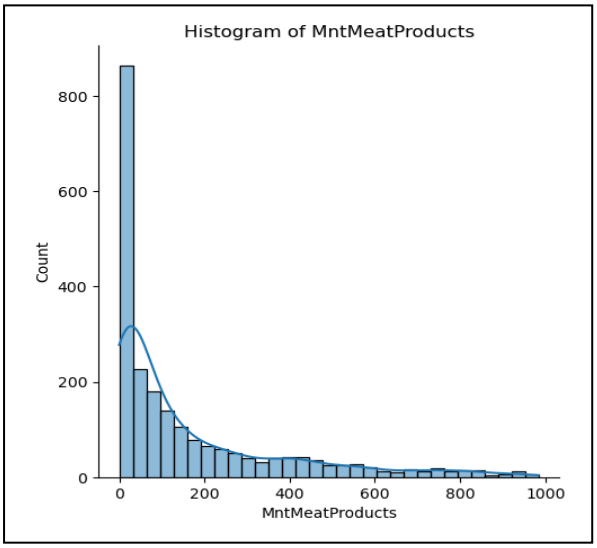


- The MntMeatProducts column had many potential outliers out of which 4 were extreme data points compared to the rest of the data points, these 4 extreme data points were removed to avoid an extreme skew in distribution of the column.

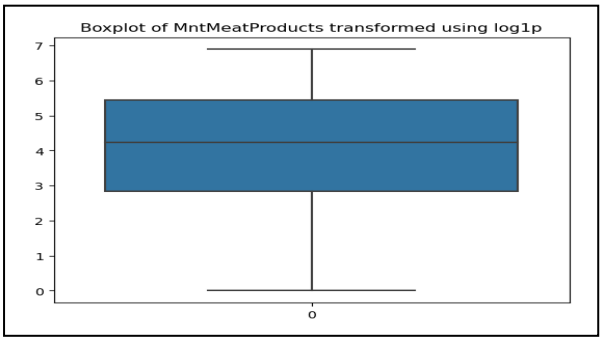
```
MntMeatProducts
1725    2
1582    1
1622    1
1607    1
Name: count, dtype: int64
```

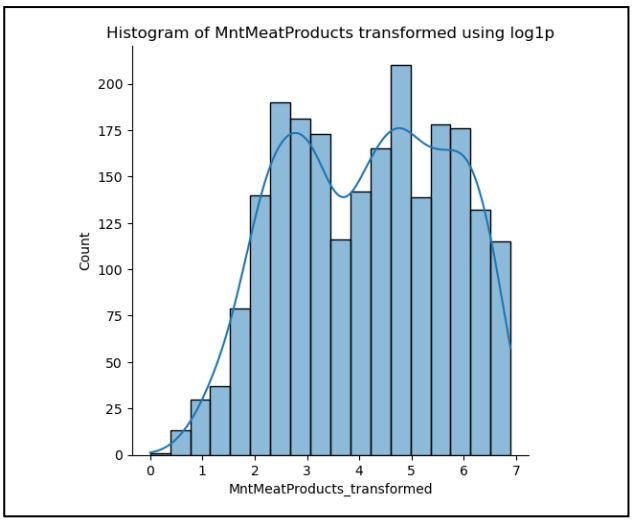
- These were the extreme points as it can be observed they are not close to each other neither are these many data points so we can afford to remove these extreme data points.



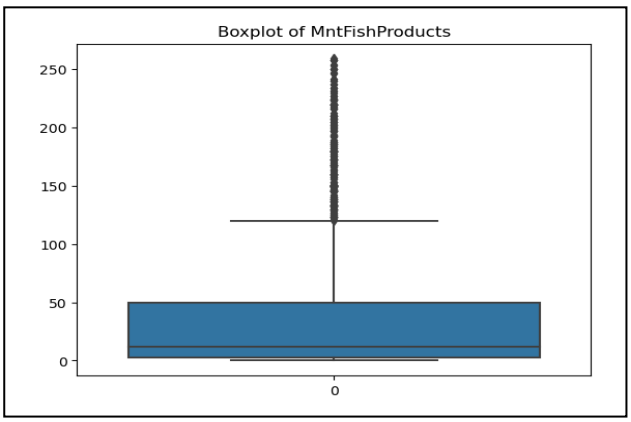


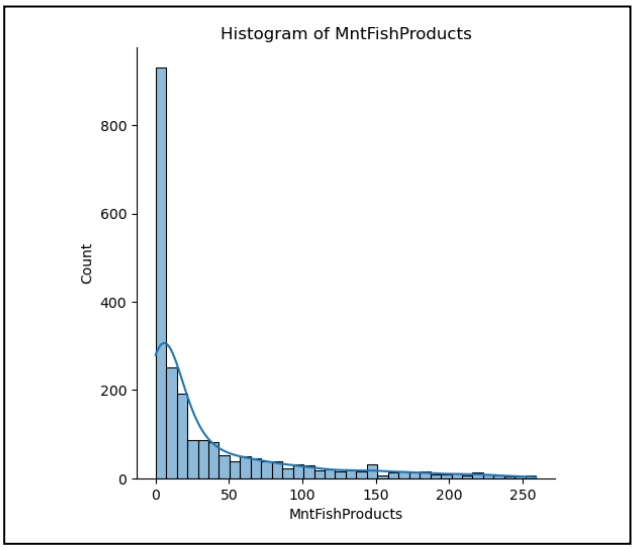
- After removing these data points, the box plot and histogram of MntMeatProducts is shown above. However, the distribution is still extremely positively skewed. So, it was transformed.



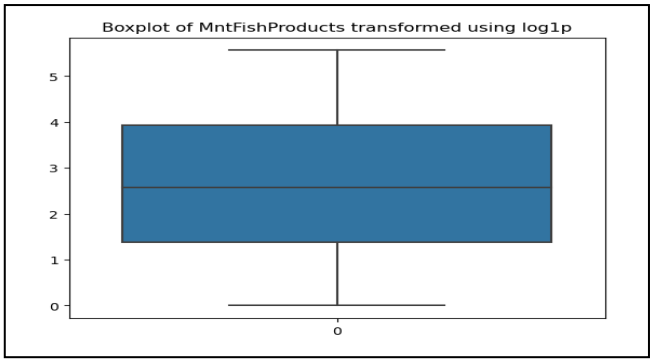


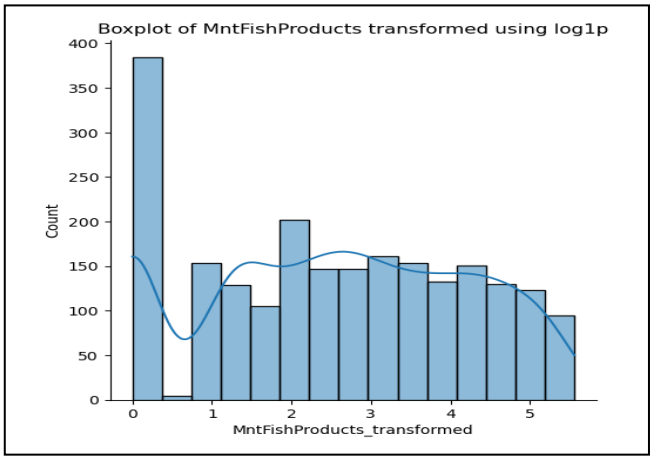
- The MntMeatProducts column was transformed using log1p and it can be observed the skewness has reduced and due to transformation, the outliers are gone.



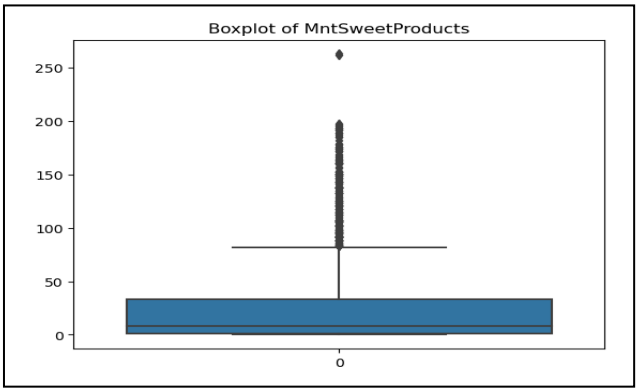


- The boxplot and histogram of MntFishProducts shows many extreme data points, extremely positively skewed distribution of MntFishProducts column.



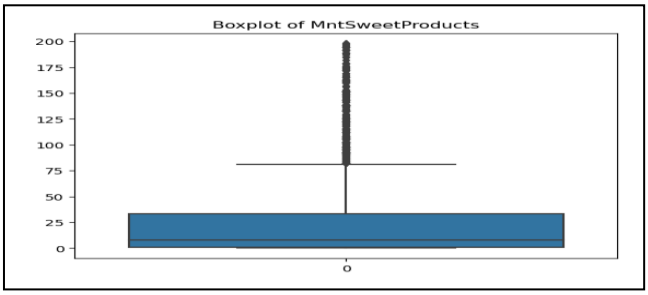


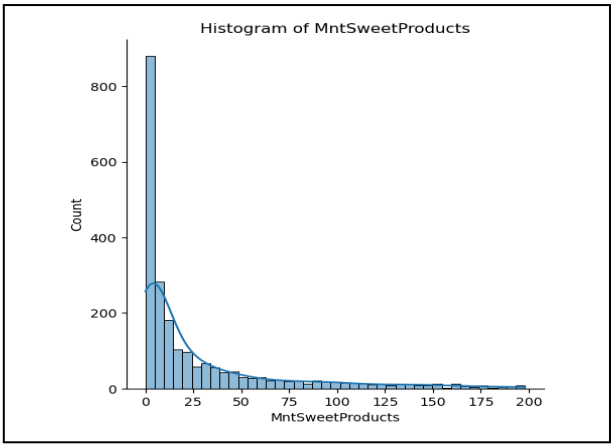
- The boxplot and histogram of transformed MntFishProducts is shown above where outliers are gone, and distribution is close to symmetrical.



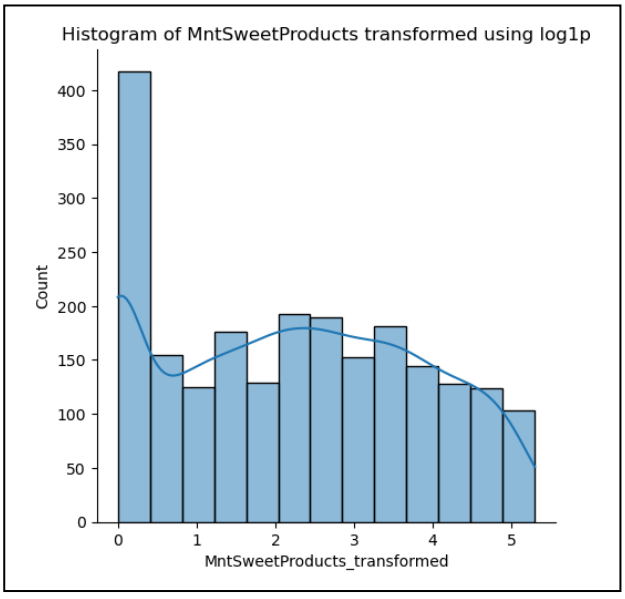

```
MntSweetProducts
263    1
262    1
Name: count, dtype: int64
```

- From the boxplot of MntSweetProducts two extreme data points can be seen. These two extreme data points were removed as they would create a skew in the distribution of column.

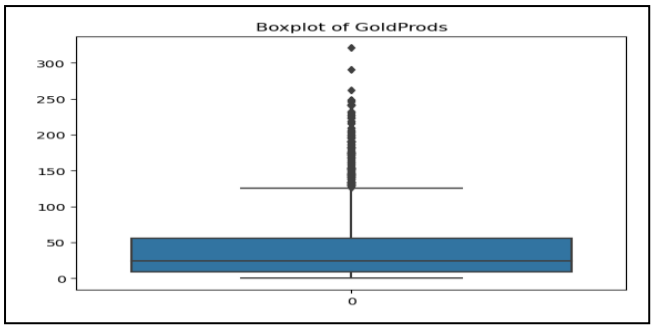




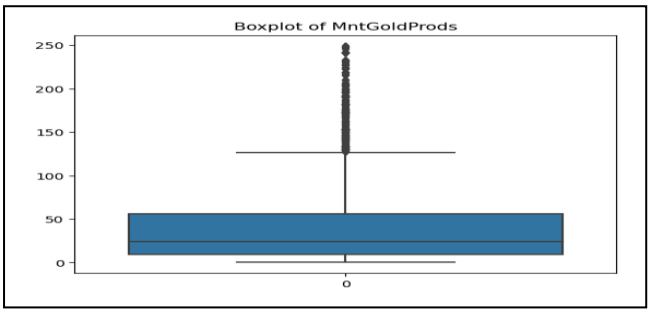
- Even after removing the extreme points the distribution of MntSweetProducts is still skewed and the many potential outliers will be handled by transformation.

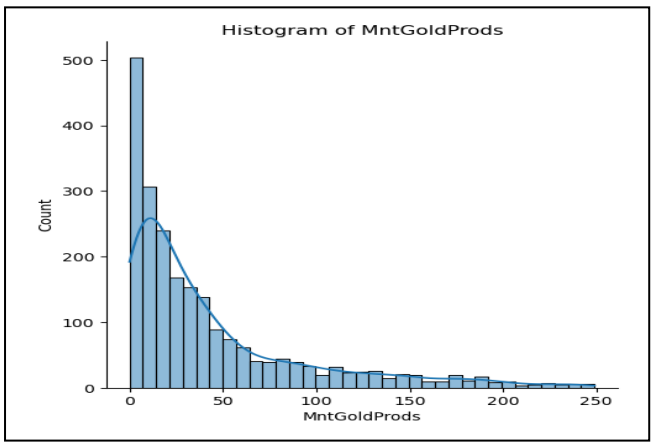


- After transforming MntSweetProducts, the extreme points are reduced, and distribution is symmetrical.

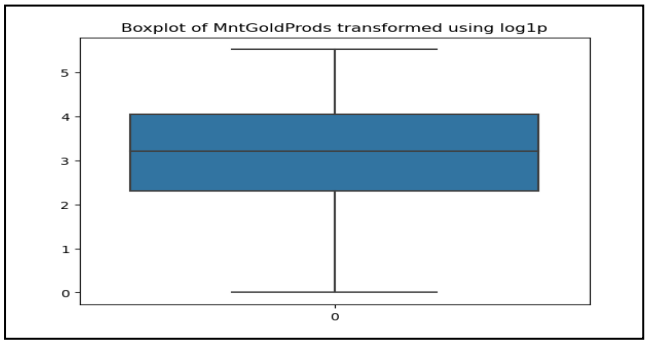


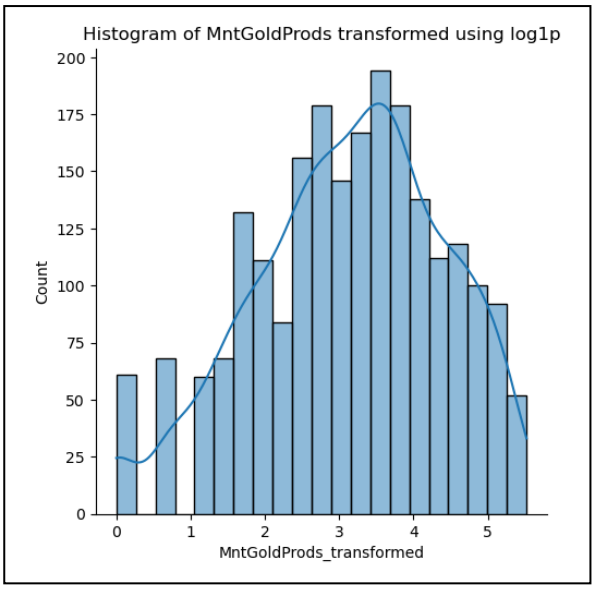
- The MntGoldProds column has many potential outliers, out of which the 3 extreme points were removed as other points are still close to each other.



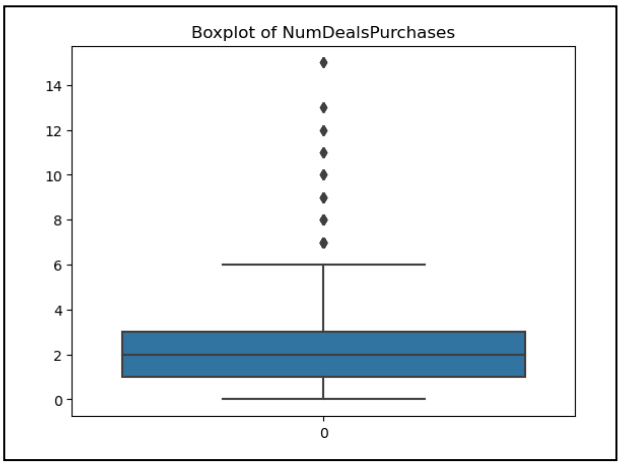


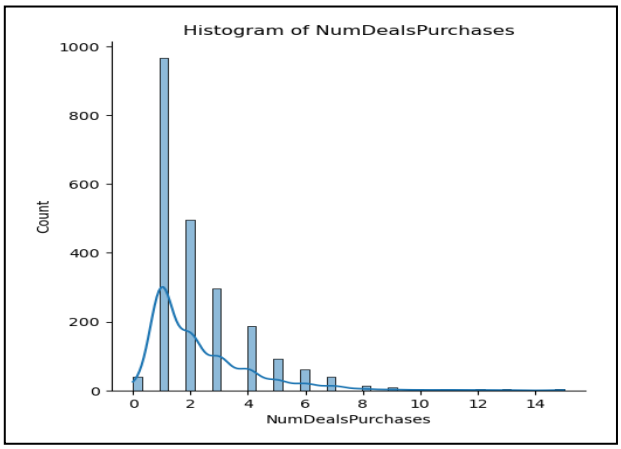
- The MntGoldProds extreme points will need to be reduced, and distribution is extremely positively skewed and will need to be transformed.



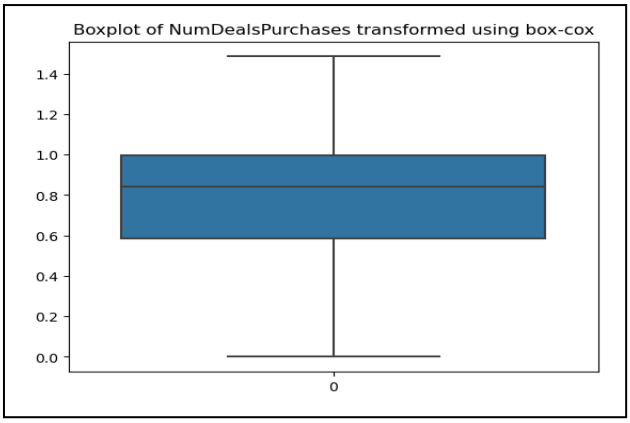


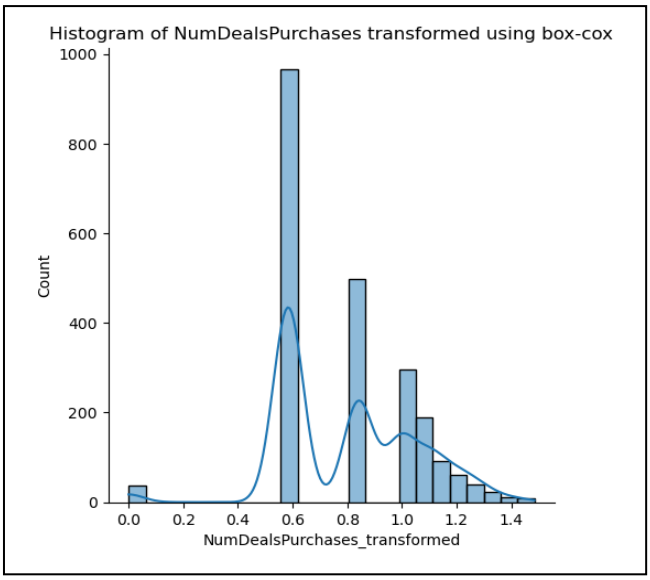
- The extreme points of MntGoldProds have been reduced and distribution is close to symmetrical after transformation.



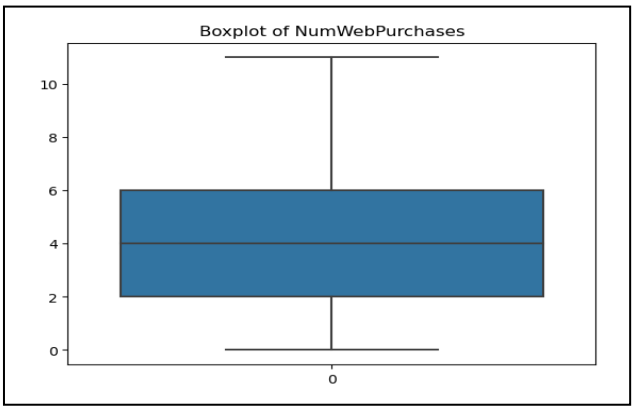


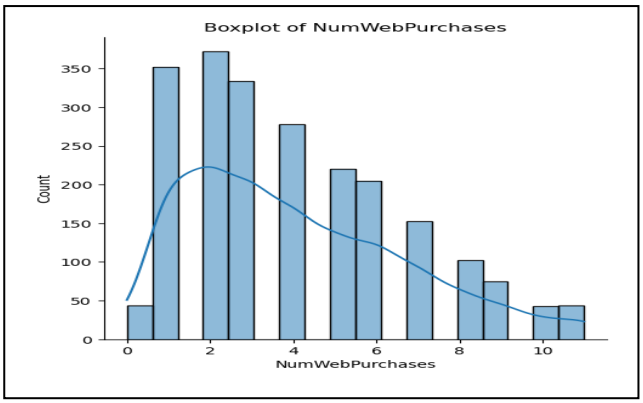
- The Boxplot of NumDealsPurchases which is purchases made with discount had some potential outliers, but the points are clustered together and were not removed, but the distribution of column is still skewed and will need to be transformed.



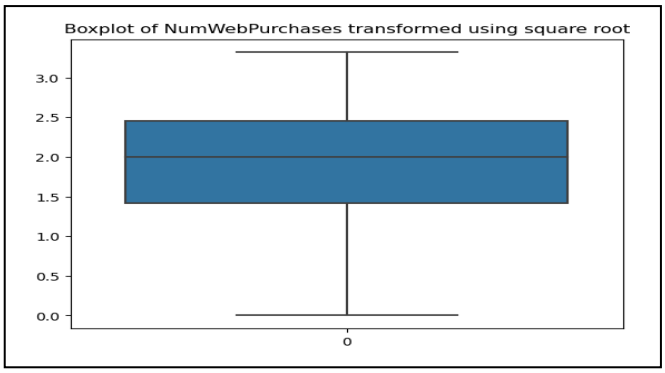


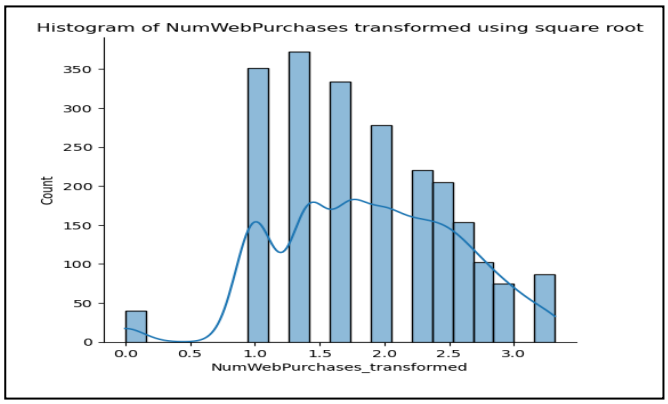
- The NumDealsPurchases column was transformed using box-cox transformation, usual transformation was not sufficient for the distribution of this column, Also, the lambda value for the transformation was -0.5096551481877226.
- The Boxplot and histogram of NumDealsPurchases after transformation are shown above. After transformation, it can be observed that the extreme points were reduced, and distribution is symmetric.



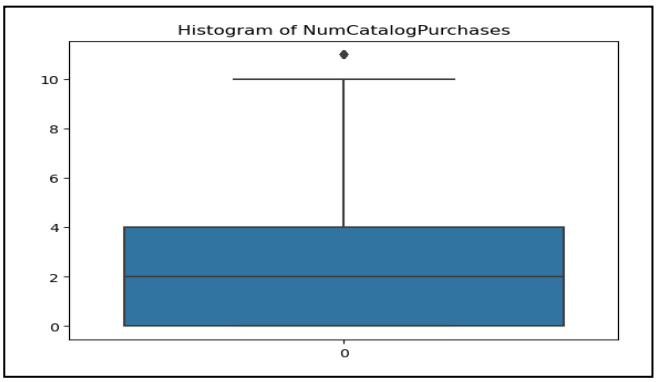


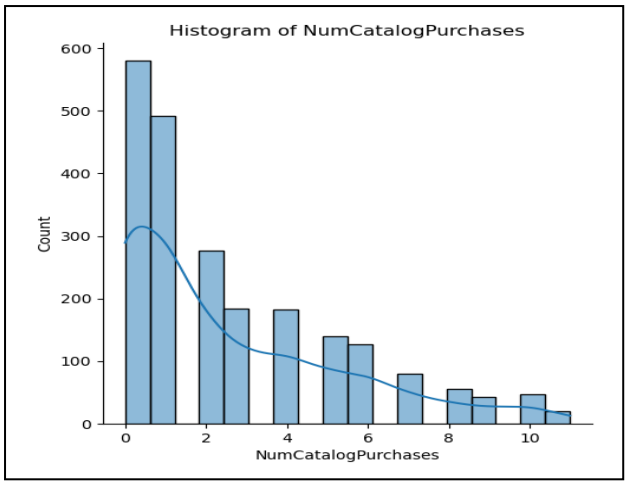
- The box plot of NumWebPurchases shows no outliers but the distribution is slightly skewed and will need to be transformed.



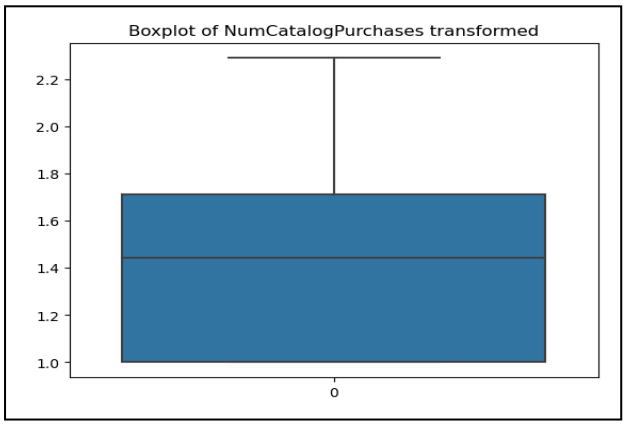


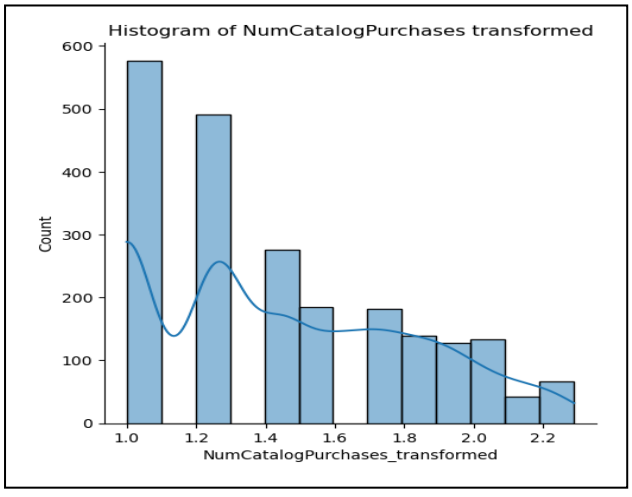
- The NumWebPurchases distribution is close to symmetric after transformation using square root.



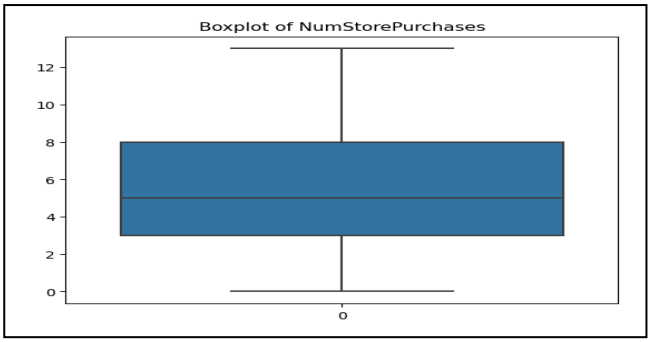


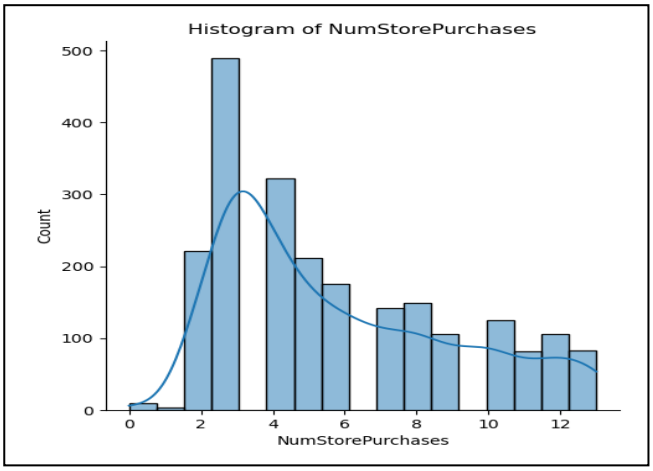
- NumCatalogPurchases has 19 data points clustered together so they are not outliers, but the distribution of columns is extremely skewed and will need to be transformed.



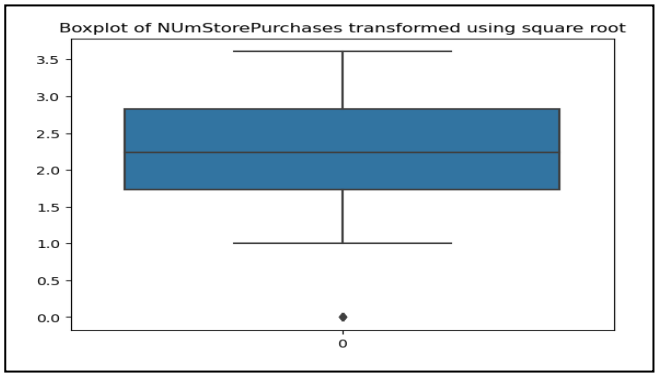


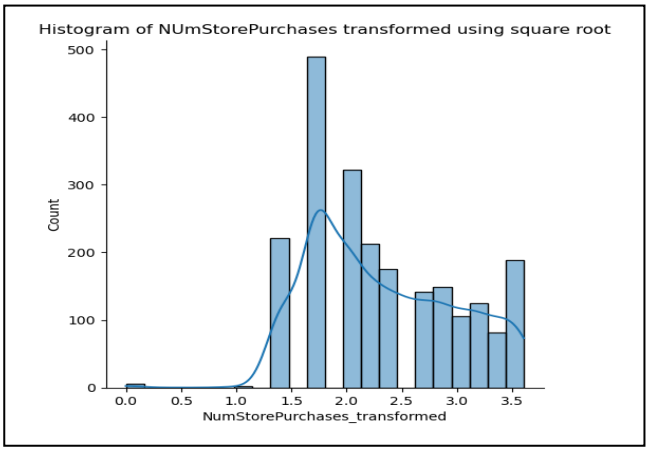
- NumCatalogPurchases was transformed using cube root plus one and the distribution of column is symmetric and the clustered extreme data points have been reduced.



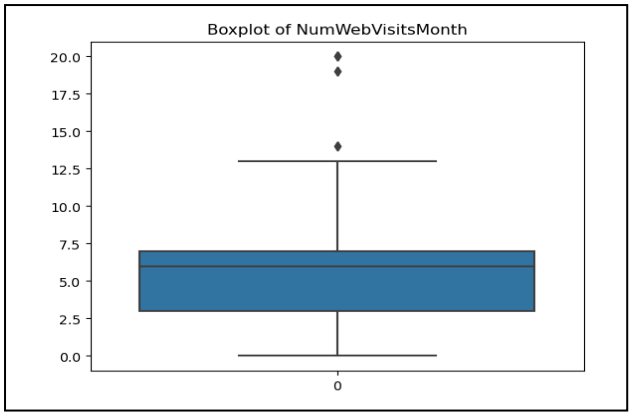


- The NumStorePurchases column doesn't have extreme data points but is slightly skewed and needs to be transformed.



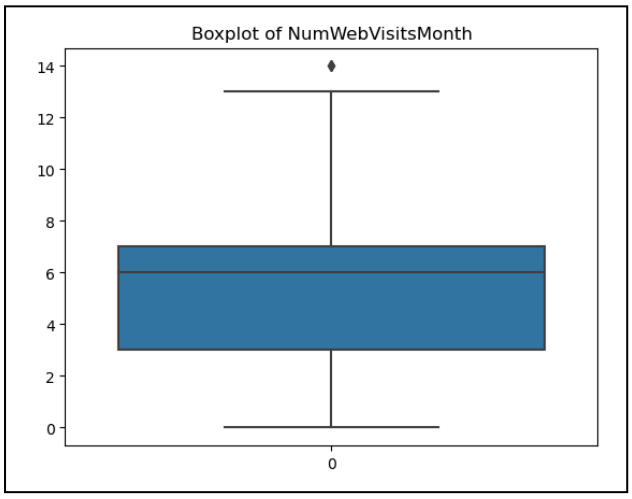


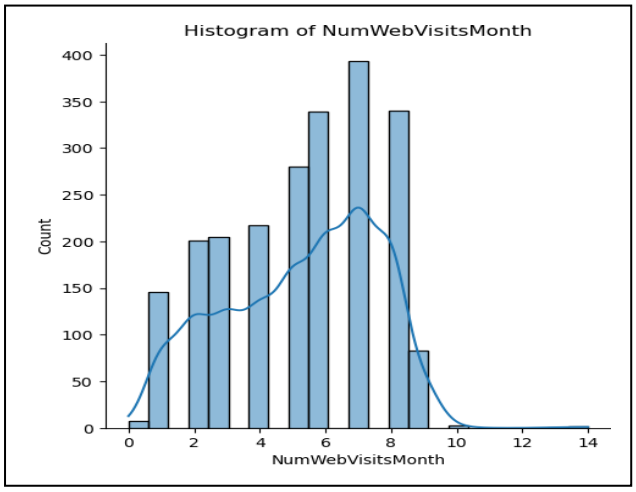
- The NumStorePurchases was transformed using square root and the distribution of column is symmetric and extreme data point was created after transformation.




```
NumWebVisitsMonth
20    3
14    2
19    2
13    1
Name: count, dtype: int64
```

- The NumWebVisitsMonth Column has few extreme data points which would affect distribution of column, so these 19,20 were removed as they are much bigger than rest of distribution while 13,15 are still close.





- The NumWebVisitsMonth column after transformation had a symmetric distribution even though few extreme points are present.

III. Methodology/Discussion

RFM Analysis:

RFM stands for Recency, Frequency and Monetary. It is used to divide customers into different segments. It can be customized to divide customers into chosen number of segments based on RFM score which is calculated based on customers performance in the forementioned three metrics, they are:

- Recency: How many days it has been since the customer last made a purchase with us?
- Frequency: What is the total number of transactions the customer has made with us?
- Monetary: What is the total spending of the customer on our products?

The scale of three metrics will not be same, so three columns have been created recency score, frequency score and monetary score.

- These columns have been used to scale the columns into the range 1 to 5.
- After creating these columns RFM Score column has been created in which 15% weight has been given to recency, 28% to frequency and 57% to monetary which will result in giving an RFM score to the customer in the range 1 to 5.
- Finally, a Customer Segment column has been created, which is a qualitative column classifying customers into different segments by using RFM score.

Implementation of RFM Analysis:

- The dataset had the column recency, frequency, but the monetary column was created by taking the total amount spent on the six product categories in the dataset such as MntWines, MntFruits, MntFishProducts, MntMeatProducts, MntSweetProducts, MntGoldProds.
- As mentioned previously, the scale of columns was different so the qcut function from pandas library was used to scale the three metrics of RFM analysis into the range 1 to 5.
- The recency column shows the number of days since last purchase so when scaling it into the range 1 to 5, the customer whose recency value is high is less valuable to us since it has been many days since customer made purchase with us. For this reason, the bins have been created in reverse order that is 5 to 1. This will result in putting customers who have made purchases with us recently a higher value and customers who haven't made purchases with us a lower value. As the algorithm doesn't understand such details this is necessary to not lead to biased analysis.
- The frequency and monetary column have also been put into the range 1 to 5. No reverse ordering has been used here since higher frequency and monetary are valuable for the business.

- The three columns created by scaling recency, frequency, and monetary were then ranked by using rank function from pandas.
- Finally, as mentioned previously, the three metrics of RFM were used by assigning them a weightage and calculating RFM score with 15% weight on recency, 28% on frequency and 57% on monetary.
- Six customer segments have been created. They are:
    - Top Customers
    - High Value Customers
    - Medium Value Customers
    - Low Value Customers
    - At Risk
    - Lost Customers
- These customer segments can be used to assess the performance of the marketing campaign.
- Also, these segments can be used to identify customers based on their value to the business and direct campaign efforts towards these customers.
- As mentioned previously, an organization has limited funds and wants to focus its marketing efforts on customers who bring value to the business. No business can afford to use marketing strategies for all customers without assessing the success of marketing campaigns.
- The results of this segmentation are provided in the Insights/Findings section.

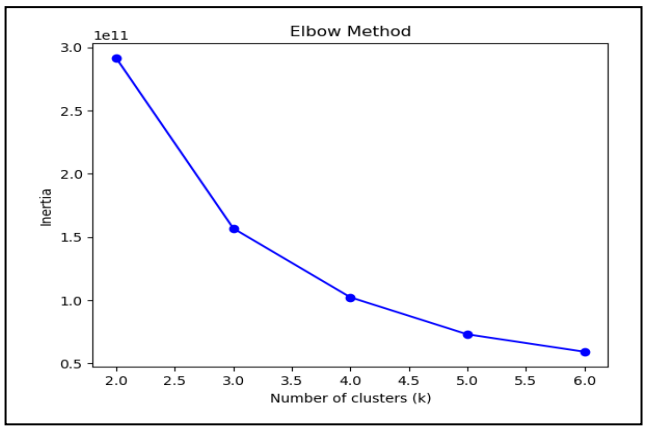
K-Means Algorithm:

K-means algorithm is an unsupervised clustering algorithm, which calculates the distance between data points to groups or separate similar data points into clusters. The number of clusters is a hyperparameter for this algorithm which is chosen to get minimum intra-class variance and maximum inter-class variance. Here intra-class variance refers to how close data points are to each other within a cluster and inter-class variance refers to how far or separated the data points are in different clusters. These metrics help determine the performance of k-means algorithm on the data. Additionally, there are methods such as Elbow method and Silhouette Score which help determine the optimal number of clusters.

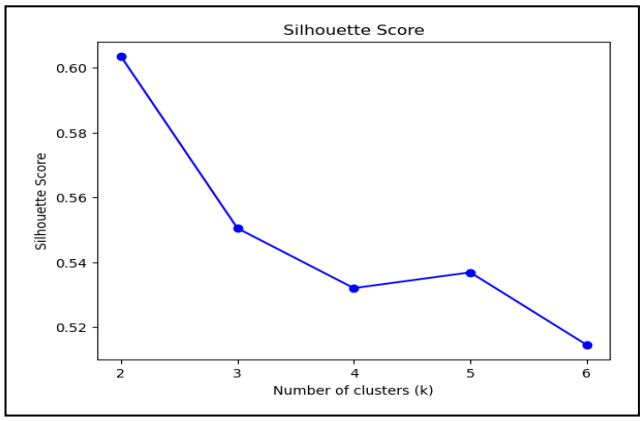
For customer segmentation, features whose scatter plots have obvious patterns and relationships can be utilized to segment customers into different groups. However, the interpretability and merit of using clusters created from such logic is not always useful. On top of this, the clusters created from k-means may not make sense from a logical point of view as k-means looks for distance and the features selected for k-means mainly affects the segments for instance, if features which have a relationship in data but no meaningful relationship logically from a business point of view can be used to create the clusters but the stakeholders may be hesitant to use strategies they do not have confidence upon so it is incumbent upon the analyst to look for features which not only have a relationship in data but also from a business perspective so stakeholders can have confidence in the results of the algorithm. Finally, the clusters made by k-means are optimal in terms of class separation evaluated through inter class variance and still the usefulness, accuracy, and effect of using the clusters made through k-means cannot measured or evaluated without using the clusters for some marketing strategy in the business.

K-Means Implementation:

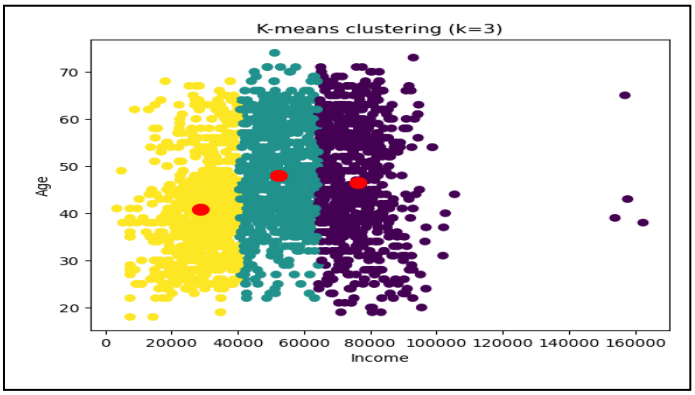
- Age and Income column have been used to plot k-means segmentation.
- By using the Elbow method, 3 clusters seem to be optimal.



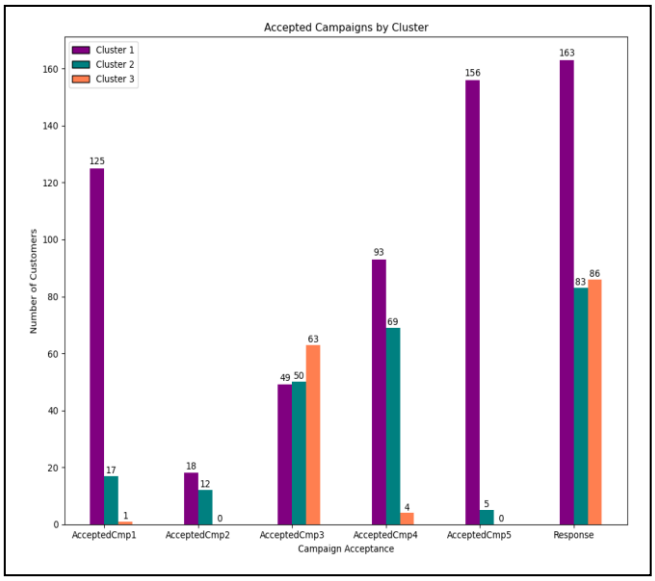
- However, by using silhouette score 2 clusters seems to be optimal.



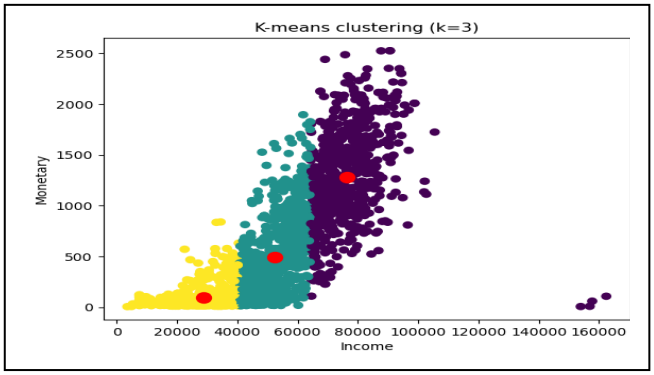
- The clustering results for 3 clusters with Age and Income:



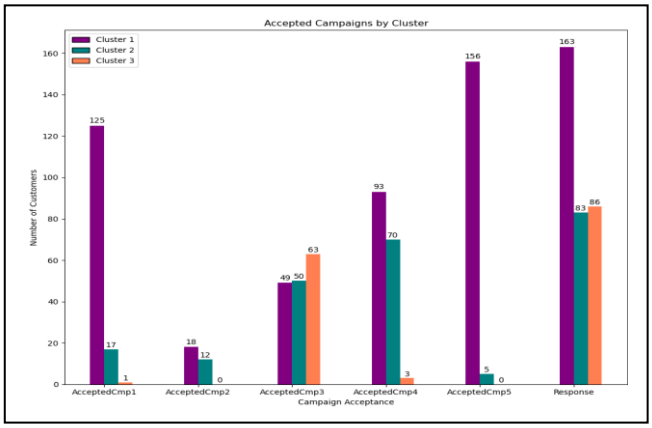
- The clustering results with Age and Income using 3 clusters seems like it can be used from business perspective by dividing customers from different income levels.
- The Marketing Campaign performance with customers of different income levels is:



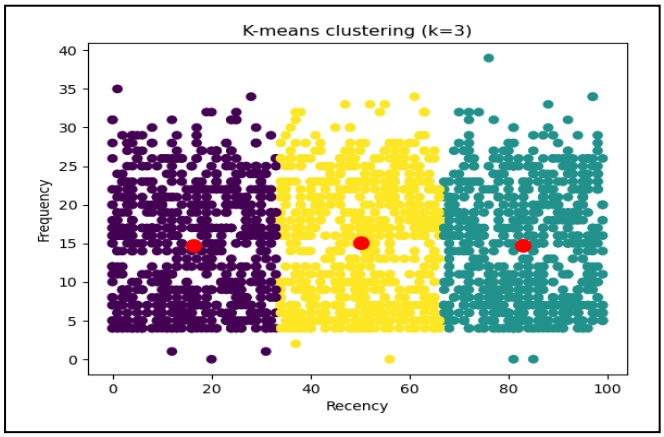
- Additional segmentations were tried, and 3 clusters were optimal like previous analysis, the results are as follows:
- The clustering results with Income and Monetary:



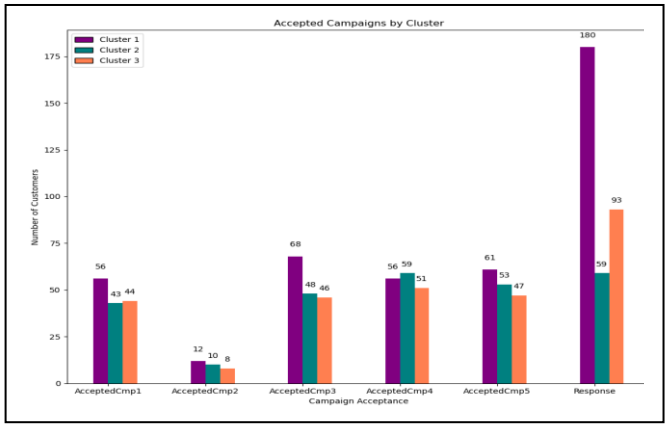
- Marketing Campaign performance with cluster from Income and Monetary:



- The clustering results with Frequency and Recency with 3 clusters:



- The marketing campaign performance with clusters from frequency and recency:



- The segmentation of Income with Age and Income with Monetary has similar results in marketing campaign performance assessment.
- However, recency with frequency has different results.
- These customer segmentations can be used by the marketing teams to target the customer base based on seasonality.

Market Basket Analysis:

MBA, also known as association-rule mining, gathers and researches sales data, which has developed rapidly in the past years. When a customer purchases products in a store, all these transactions are recorded in the store's system. This gives the store owners a better understanding of the customer's purchasing habits and tries to identify patterns in the consumer's behavior. Market basket analysis main use areas are: Cross-selling, recommendation engines, product placement, affinity promotion, costumer behavior, inventory management, store and website traffic. The main concept of MBA is that if a customer purchases some items, then the customer is more or less likely to purchase other items. For instance, if a customer purchases bread and milk, it is very likely that the customer also purchases butter. [1]

The group of items a customer purchases is known as an itemset. MBA attempts to identify relationships and patters of

the customer's purchases from the itemset. Market basket analysis obtains interesting association rules among products.

Apriori Algorithm:

This algorithm produces the most significant group of rules from a given transaction data. Support, confidence and lift are the three main measures that are applied to choose the strength of each rule. [2]

Lets consider the rule A => B, so we can calculate these metrics:

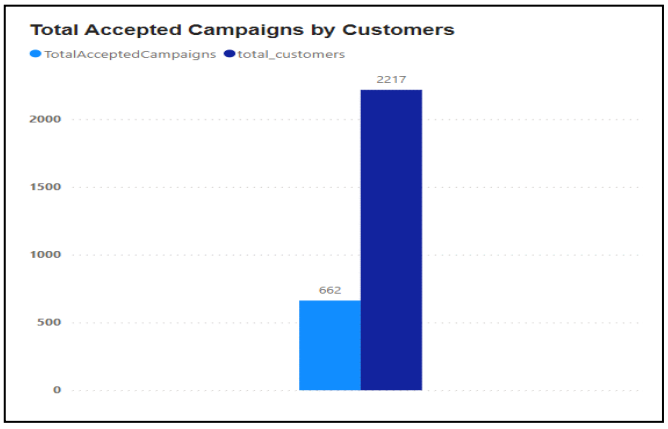
- Support: The ratio of number of transactions of A and B divided by the total number of transactions.
- Formula: Support = P(A ∩ B) / Total number of transactions
- Confidence: The ratio of number of transactions of A and B divided by the total number of transactions with A.
- Formula: Confidence = P(A ∩ B) / P(A)
- Lift: The ratio of number of transactions of A and B divided by A and B. It is the influence by which the co-occurrence of A and B goes beyond the expected probability of A and B co-occurring. If the value of the lift is higher, then the chance of A and B occurring together is also higher.
- Formula: Lift = P(A ∩ B) / P(A) . P(B)

Challenges with MBA using Apriori Algorithm:

As mentioned in data characteristics section, every record in the dataset belongs to one customer which represents aggregate information of the customer. The Apriori algorithm finds associations between products being bought together. For this Apriori algorithm needs individual transactions of the customer while the dataset has aggregate transactions. For these reasons the results of Apriori algorithm showed most products were being bought together. In summary, due to aggregate nature of dataset the Apriori algorithm was not suitable for analysis on this dataset. It could also be said MBA cannot be conducted on this dataset. The data source did not specify the aggregated nature of dataset. As the team had finished the cleaning, preparation and modeling only then it was discovered and led to understanding the true nature of the problem. The team views this as a opportunity to learn and is motivated to focus on parts which are interpretable.

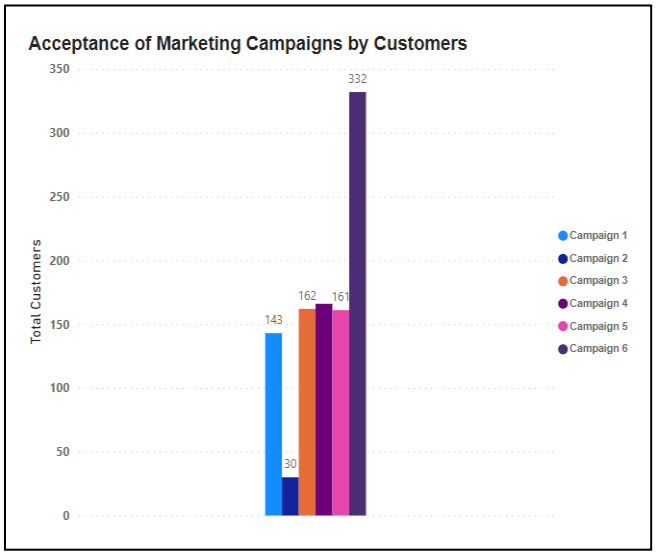
Insights/Findings:

Power BI has been used to explore and analyze the dataset with the analysis so far serving as the base to look for insights which can help assess the performance of marketing campaigns with different segments of customers.



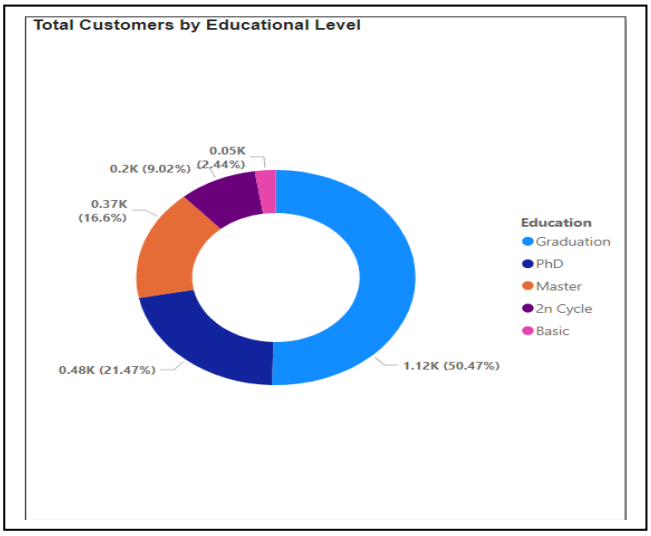
Total Accepted Campaign by Customers: The bar chart illustrates the relationship between the number of accepted marketing campaigns and the total number of customers.

- The total number of customers in the dataset is 2217.
- The total number of customers who accepted the campaigns in the dataset is 662.



Acceptance of Marketing Campaigns by Customers: The bar chart depicts customer acceptance of various marketing campaigns. The x-axis displays six different marketing campaigns, while the y-axis shows the number of customers who accepted each campaign.

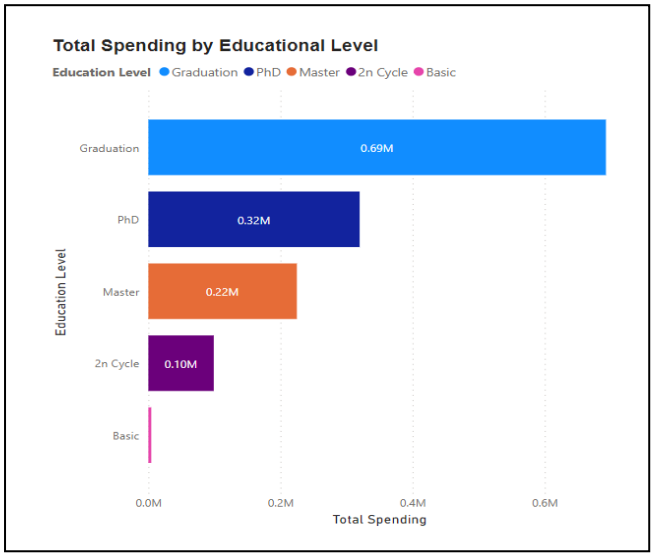
The bar chart illustrates the acceptance of various marketing campaigns by customers. On the x-axis, there are six distinct marketing campaigns, while the y-axis represents the number of customers who accepted each campaign. Campaign 6 exhibits the highest acceptance rate, while campaign 2 demonstrates the lowest acceptance rate. The remaining campaigns has close performance to each other. Specifically, campaign 3 garnered 162 acceptances, campaign 4 received 166 acceptances, and campaign 5 had 161 customer acceptances.



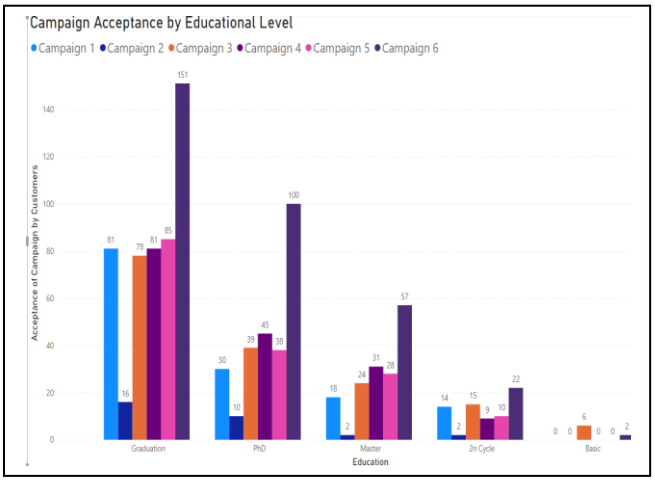
Total customers by Education Level: The chart shows the distribution of customers across different education levels.

It is important to note that the number of customers in each category is represented in the thousands. For instance, there are roughly 1,120 customers with a graduation level education.

Overall, the data shows that most of the customers (over 50%) have a graduation level education and with PhD and Master and 2nCycle it shows over 90% of our customer have strong educational background.

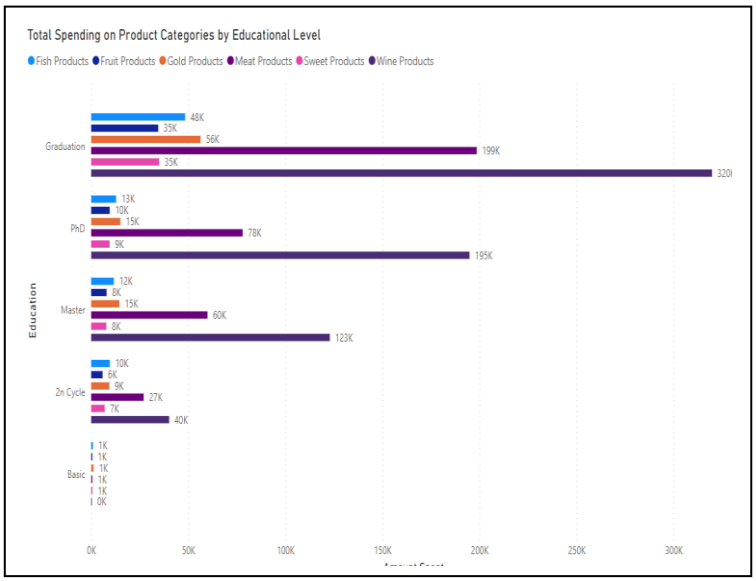


Total Spending by Customers of different educational levels: The graph is a bar chart of total spending by customers of different educational levels. The chart indicates that the highest educational level with most spending on our business is graduation level customers, which makes sense as over 50% of our customers are graduates. It is important to note an average graduate is not spending more on the business rather the aggregate spending of graduates is high due to half of our customer base being graduates.



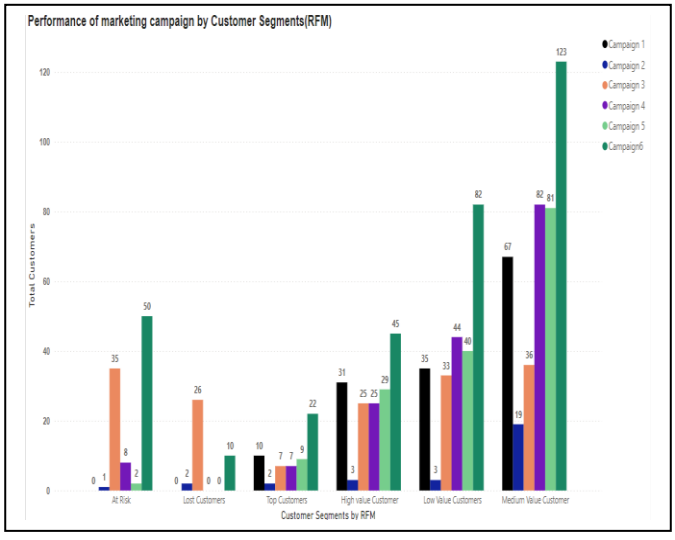
Education level with Campaigns: The Clustered bar chart shows Campaign Acceptance by Educational Level: The x-axis of the graph shows educational level, which ranges from "Basic" to "PhD". The y-axis shows the number of customers who participated in the marketing campaign. There are six campaigns represented in different colors.

The breakdown of campaign acceptance by educational level reveals that customers with Graduation-level education exhibit the highest acceptance rate, followed by those with a PhD, Masters, 2nd Cycle, and Basic education, in descending order of acceptance rates.



Total Spending on Product Categories by Customers of different Educational Levels: The chart illustrates diverse spending patterns among customers based on their educational levels.

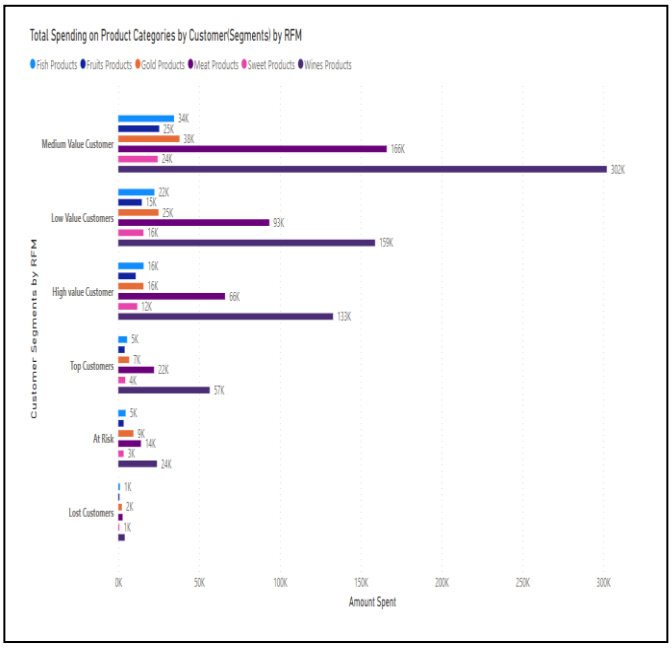
It can be observed there isn't a clear varying pattern among spending habits based on educational level, for instance wine is the most spent on product category followed by meat products and so on. Since the dataset has over 50% graduates it may seem that the spending of graduate level customers is high but the expenditure of graduate level customers in not higher than the other education levels.



Customer Segments (RFM):

The graph displays the performance of marketing campaigns across various customer segments identified through RFM analysis. Notably, Campaign 6 is the most successful campaign among these segments. Given the priority on valuable customers, the marketing team may opt to utilize campaigns that demonstrate strong performance with this group for future marketing endeavors. Similarly, for products

targeting high-income customers, campaigns showing effective performance within this segment would be preferable choices for marketing initiatives.



Total Spending on product categories by Customer segments made with RFM:

The Cluster Bars in the graph shows the total spending on product categories by customer segments. The data is segmented by RFM analysis, which is a method used in marketing to identify customer value based on their recency, frequency, and monetary value of purchases.

The graph shows six customer segments:

Top Customers – The top customer plays a significant role in driving business success. According to the data, the highest expenditure among top customers is observed in wine, totaling $57k. Conversely, the number of top customers is lower in the meat product category, with an expenditure of $66k. Additionally, sweet products account for $4k in expenditures, while gold products total $7k, and fish products amount to $16k.

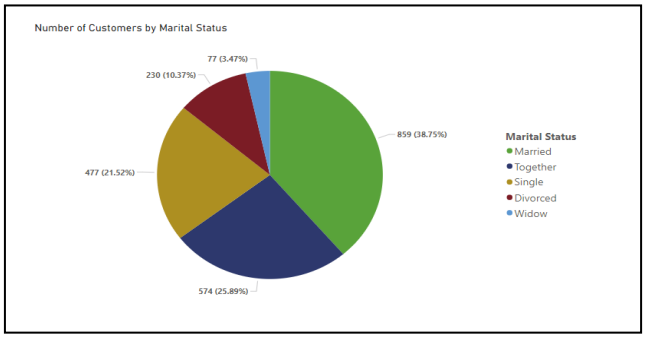
High Value Customers - These customers spend a significant amount but not quite as much as Top Customers. They spend the most on Wines at $133,000, followed by Fish Products at $66,000 and Fruits Products at $66,000.

Medium Value Customers - These customers spend a moderate amount across all categories. They spend the most on Wines at $302,000, followed by Fish Products at $166,000 and Sweet Products at $24,000. But these segments are high in numbers and generate the highest profits of all campaigns.

Low Value Customers - These customers spend the least overall. They spend the most on Wines at $159,000, followed by Meat Products at $93,000, Gold Product $25000 and Fish Product at $22000.

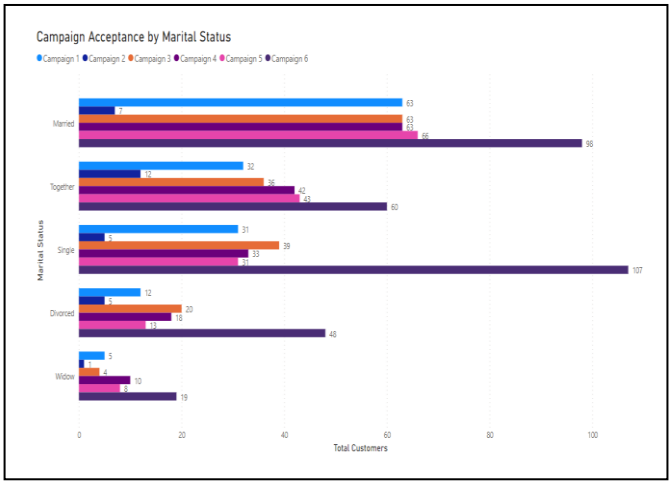
At Risk - These customers have not purchased very frequently. They spend the most on Wines at $24,000, followed by Meat Products at $14,000, Gold Products at 7k and Fish Products at $5,000.

Lost Customers - These customers have not purchased in a while. They spend the most on Wines at $4000, followed by Gold Products at $2000, Meat products at $1000 and Fish Products at $1024.



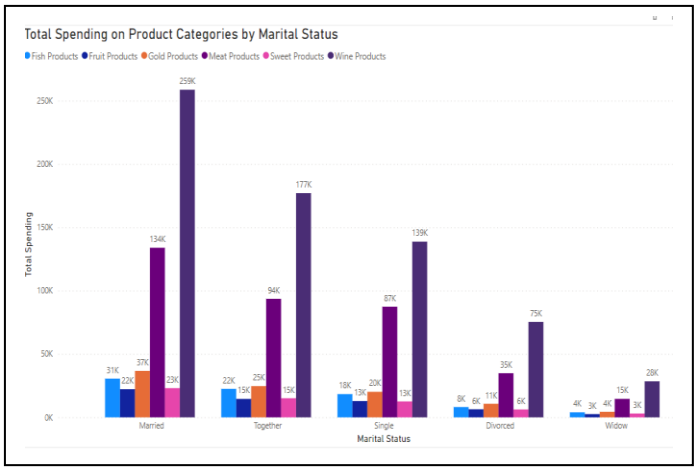
Number of customers by Marital Status:

The pie chart shows over 50% of the customers are married or in a relationship. While the remaining 50% of the customers are currently single.



Campaign Acceptance by Marital Status:

The clustered bar chart shows Campaign Acceptance by customers of different marital status. acceptance rates for married clients are evenly distributed across campaigns, with Campaign 2 standing out with only 7 acceptances.

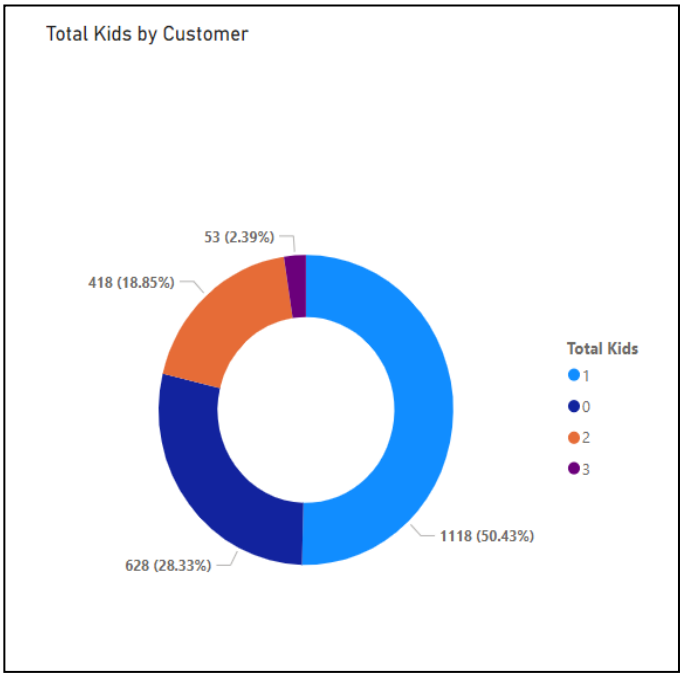


Total Spending on product categories by marital status:

Married customers exhibit the highest spending, notably investing significantly in wine products, totaling 259K.

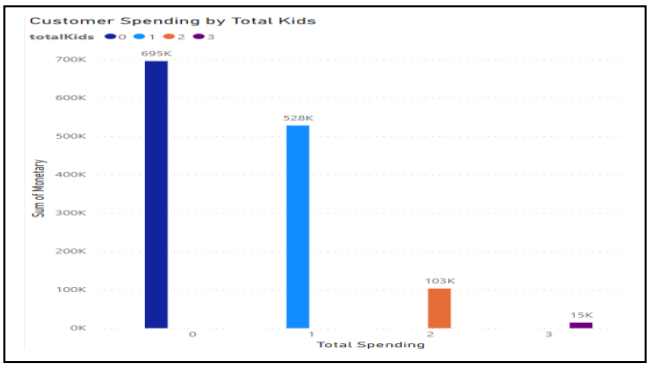
Conversely, Widows show the lowest expenditure, with the highest spending observed on wine products at 28K. The data makes sense as the expenditure by marital status is in proportion to number of customers of said marital status.

Additionally, divorced customers spend approximately 25% less compared to the total spending of married customers on sweet products. Single and Together customers closely match the total spending across various product categories.



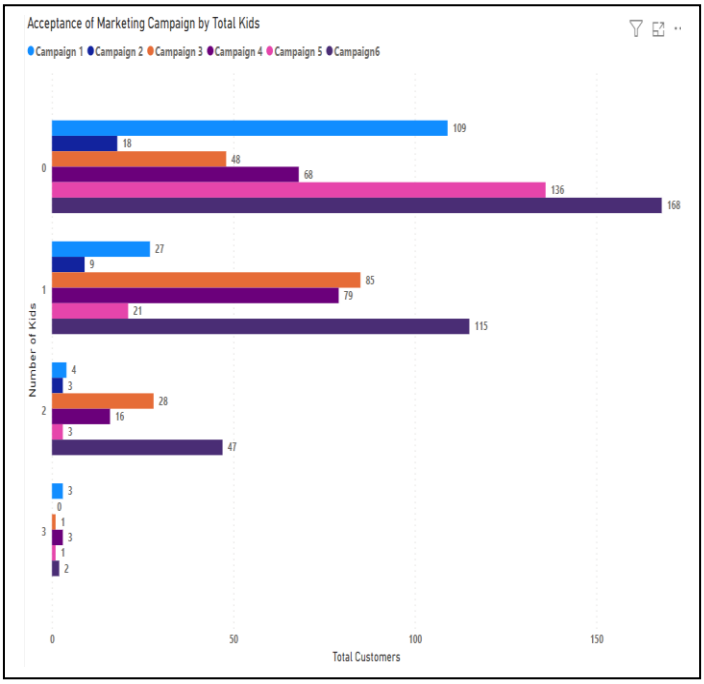
The pie chart illustrates the distribution of total kids among customers, segmented by the number of children they have.

- The largest slice, constituting 1118 children, represents 50.43% of the total kids among customers with one child.
- The second-largest slice, comprising 628 children, represents 28.33% of the total kids among customers with no children.
- The third-largest slice, accounting for 418 children, represents 18.85% of the total kids among customers with two children.
- The smallest slice, indicating 53 children, represents 2.39% of the total kids among customers with three children.



Total kids by Expenditure of Customers: The graph shows the total spending of customers who have a certain number of kids. The x-axis represents the total number of kids (total Kids), which goes from 0 to 3. The y-axis represents the total spending in dollars (Sum of Monetary).

-
- There are four data points plotted on the graph. The first data point shows that customers with 0 kids spend around $695000 in total. The second data point shows that customers with 1 kid spend around $528,000 in total. The third data point shows that customers with 2 kid spending around $103000k and the fourth data point customers with 3 kid spending around $15000k.
- IA decreasing trend in total spending as the number of kids increases can be seen. This suggests that customers with fewer children tend to participate in the campaign less.

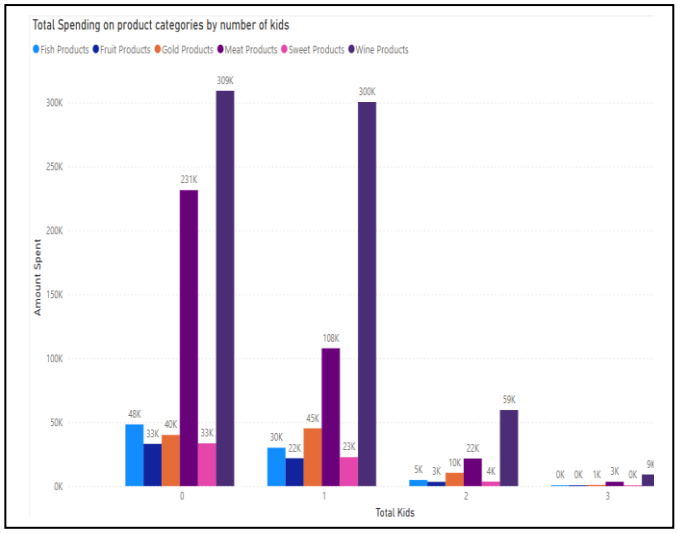


Acceptance of marketing Campaign by customer with Kids: The carts depict the acceptance rates of various marketing campaigns among customers with no children, one child, two children, and three children.

Customers with 0 kid: Among customers with zero children, there are data points available for Campaign 1, which has 109 customer, Campaign 2, which has 18 customers, Campaign 3, which has 48 customer, Campaign 4, which has 68, Campaign 5 has 136 and campaign 6 has 168. It shows Campaign 1,5 and 6 has the good response among no kids' customers.
Customers with 1 kid: Customers with one child showed positive responses in Campaigns 1, 3, and 6.

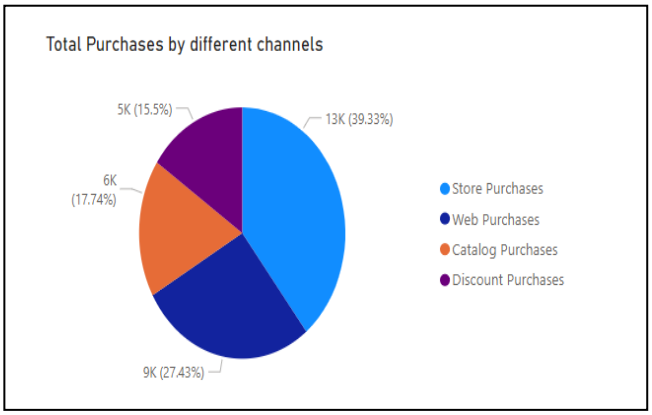
Customers with 2 kids: Customers with two children exhibited favorable responses in Campaign 3 and the highest response rate was observed in Campaign 6.

Customers with 3 kids: Customer have 3 kids has lowest response in the campaigns.



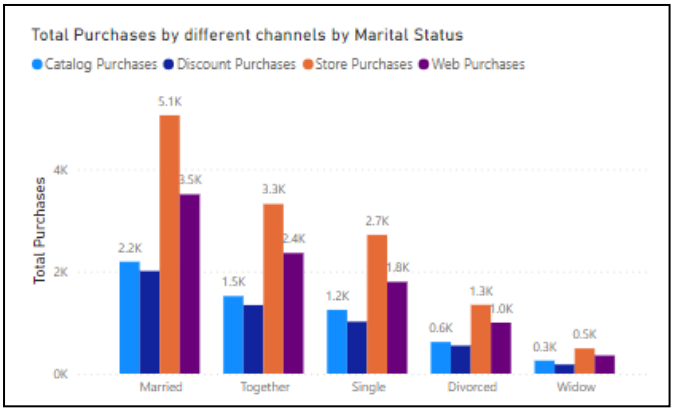
Total Spending on product categories by Number of kids:

A simple look at the data may show spending decreases as the number of kids in the customers family increases however the proportion of customers having 0 kids and 1 kids is high and it may appear look as though the spending decreases with number of kids but as the data has majority of customers with no kids or 1 kid if the reason the spending is high for customers with 0 or 1 kids.



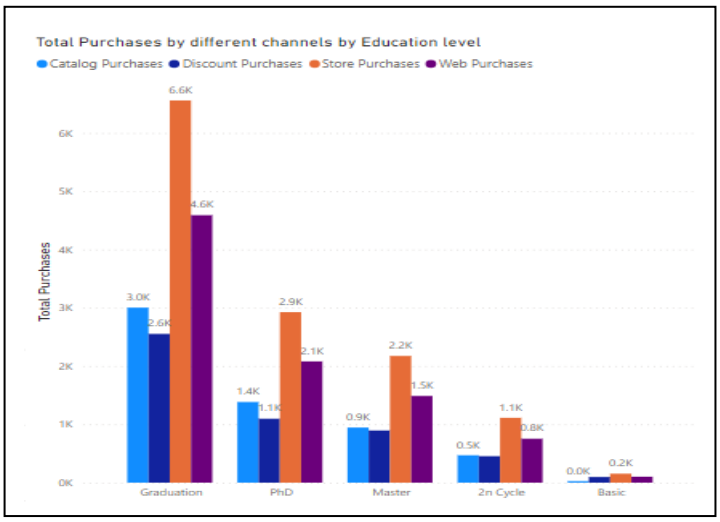
Total Purchases by different channels:

The pie chart shows the most preferred channel of purchase by customers is in store purchases followed by web purchases and catalog purchases. And as discounts are offered only in some products maybe that's the reason for discount purchases being the lowest here.



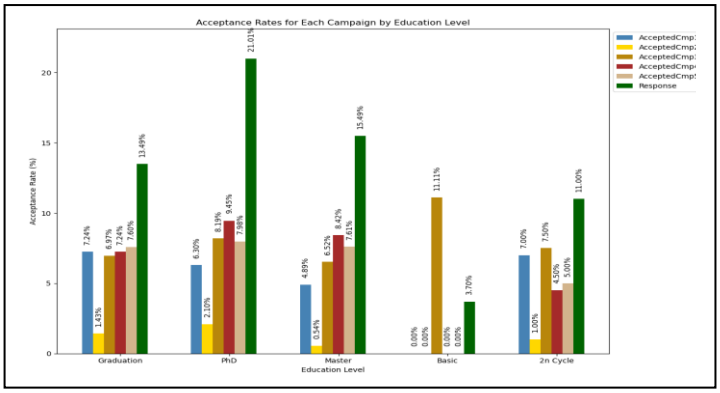
Total Purchases by different channels:

The most preferred form of purchase seems to be Store purchases, followed by web purchases and the least preferred is discount purchases.



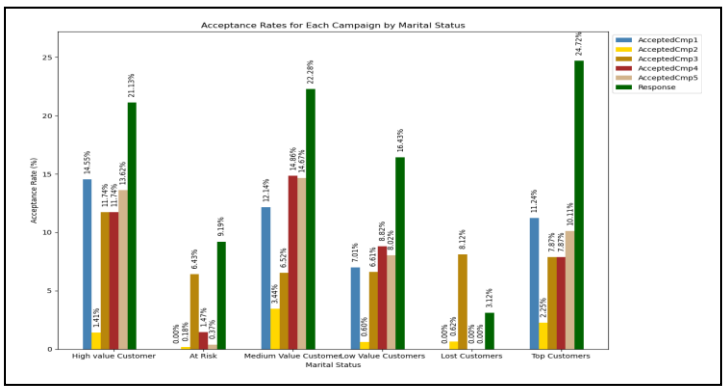
Total Purchases by different channels by Education level:

The most preferred form of purchase seems to be Store purchases, followed by web purchases and the least preferred is discount purchases. This shows similar results to marital status preferences.



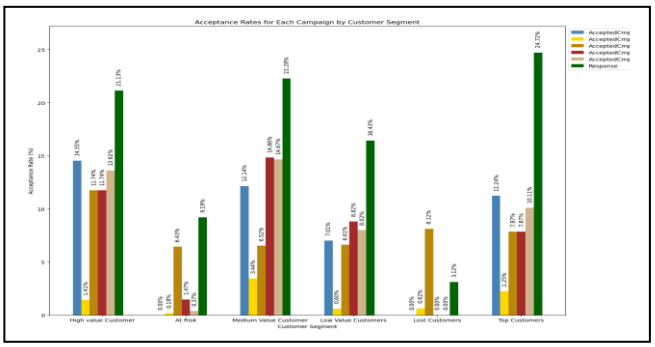
Acceptance Rates for each campaign by Education Level:

The plot shows the percentage of customers who accepted a campaign from each education level. For instance, from graduation the percentage of Customers with graduation educational level who accepted Campaign 1 by total number of customers with graduation as their education level is used to find the acceptance rate for every campaign with every educational level.



Acceptance Rates for each campaign by Marital Status:

The plot shows the percentage of customers who accepted a campaign from each Marital Status level. For instance, from married marital status the percentage of Customers with married as their marital status who accepted Campaign 1 by total number of customers with married as their marital status is used to find the acceptance rate every campaign with every other marital status.



Acceptance Rates for each campaign by Customer Segments created with RFM analysis:

The plot shows the percentage of customers who accepted a campaign from each Customer segment. For instance, from "High value customer" customer segment the percentage of Customers with high value customers as customer segment who accepted Campaign 1 by total number of customers with high value customer as their customer segment is used to find the acceptance rate for every campaign with every customer segment.

Conclusion:

The main objective of the analysis was to assess the performance of the 6 marketing campaigns launched by the company with various customer segments in order to find effectiveness of these marketing campaigns with the customer segments which were already available in the dataset along with the customer segments created through RFM analysis. This has been accomplished successfully. Although due to the nature of dataset Market Basket Analysis cannot be done on this dataset.

References:

[1] T. Raeder, N. Chawla. "Market basket analysis with networks". pp 1:97–113. 2011

[2] F. Kurniawan, B. Umayah, J. Hammad, S. Nugroho & M. Hariadi. "Market Basket Analysis to Identify Customer Behaviors by Way of Transaction Data". pp. 20–25. 2018

Sources which helped understand & do the analysis.

- *Kadlaskar, "Market basket analysis: A comprehensive guide for businesses," Analytics Vidhya, https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-on-market-basket-analysis/ (accessed Apr. 8, 2024).*
- R. Saldanha, "Marketing campaign," Kaggle, https://www.kaggle.com/datasets/rodsaldanha/arketing-campaign/
- *Murphy, "What is recency, Frequency, monetary value (RFM) in marketing?," Investopedia, https://www.investopedia.com/terms/r/rfm-recency-frequency-monetary-value.asp (accessed Apr. 8, 2024). code (accessed Apr. 8, 2024).*
- GfG, "RFM analysis analysis using Python," GeeksforGeeks, https://www.geeksforgeeks.org/rfm-analysis-analysis-using-python/ (accessed Apr. 8, 2024).