Customer Segmentation based on Behavioral Data in Marketplace

A Project Report

Submitted by

AASHI PATNI, AMISHA DUBEY, AVISHA VIJAYVARGIYA, DHRUV JAIN

Under the Guidance of

PROF. GAURAV PALIWAL

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

COMPUTER SCIENCE AND BUSINESS SYSTEMS

At

SCHOOL OF TECHNOLOGY MANAGEMENT AND ENGINEERING

APRIL 2023

Declaration

We, Aashi Patni, Amisha Dubey, Avisha Vijayvargiya, Dhruv Jain Roll No. O501, O507, O513, O515, B.Tech. (Computer Science and Business Systems), VII-VIII semester understand that plagiarism is defined as anyone or combination of the following:

1. Un-credited verbatim copying of individual sentences, paragraphs, or illustration (such as graphs, diagrams, etc.) from any source, published or unpublished, including the internet.

2. Un-credited improper paraphrasing of pages paragraphs (changing a few words phrases, or rearranging the original sentence order)

3. Credited verbatim copying of a major portion of a paper (or thesis chapter) without clear delineation of who wrote what. (Source:IEEE, The institute, Dec. 2004)

4. We have made sure that all the ideas, expressions, graphs, diagrams, etc., that are not a result of my work, are properly credited. Long phrases or sentences that had to be used verbatim from published literature have been clearly identified using quotation marks.

5. We affirm that no portion of my work can be considered plagiarism, and we take full responsibility if such a complaint occurs. We understand fully well that the guide of the seminar/ project report may not be able to check for the possibility of such incidents of plagiarism in this body of work.

Signature of the Students:

Name: Aashi Patni, Amisha Dubey, Avisha Vijayvargiya, Dhruv Jain

Roll No. O501, O507, O513, O515

Place: NMIMS Indore

Date: 15/04/2023

Certificate

This is to certify that the project entitled "Customer Segmentation" is the bonafide work carried out by Aashi Patni, Amisha Dubey, Avisha Vijayvargiya, Dhruv Jain of B.Tech (Computer Science and Business Systems), STME (NMIMS), Indore, during the VIIth and VIIIth semester of the academic year 2022-23, in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology as per the norms prescribed by NMIMS. The project work has been assessed and found to be satisfactory.

Prof. Gaurav Paliwal

Internal Examiner

External Examiner

Dr. Aaquil Bunglowala
Associate Dean

Acknowledgement

We are highly indebted to our respected Director **Dr. Niranjan Shastri** and Associate Dean **Dr. Aaquil Bunglowala** for providing us with this opportunity to work on the project. It would not have been possible without the kind support and help of the faculty members **Prof. Gaurav Paliwal** and **Dr. Dharmendra Sharma**.

We have heartfelt gratitude for their leadership, ongoing oversight, and provision of the project's vital information as well as for their assistance in seeing the project through to completion.

We would like to convey our sincere appreciation and thanks to the professionals in the business for their time and attention, as well as for giving us the information we needed and assisting us in finishing the project.

We also express our gratitude and appreciation to our project coworker and the individuals who volunteered their skills in order to assist us. We also want to express our gratitude to our family members for their support and encouragement throughout the process.

Aashi Patni
Amisha Dubey
Avisha Vijayvargiya
Dhruv Jain

Table of Contents

List of Figures

List of Tables

Abstract

Customer segmentation has gained a lot of traction recently as a means of retaining customers and generating revenue from them. Customers are divided into groups based on behavioural traits like spending and income. Since every customer is unique, we cannot be certain of their purchasing habits or interests. However, by using a variety of algorithms on the dataset, one can sort the data and identify the target group. Without this, it will be extremely challenging to identify a group of individuals with similar characters and interests in a huge dataset. By using the clusters, the company may target particular customers and provide them with the content they are actually interested in through advertising campaigns and social media platforms. After our analysis we were able to find that the sales for the company were highest in 2013. Different campaigns launched by the company had different results for different groups. The campaign 3 was successful for customers whose age is less than 53 and campaign 4 was successful for customers with age more than 53. We were also able to find out the different groups and their spending patterns like the most amount spent on wine is by Widow.

Chapter 1

Introduction

1.1 Problem Definition

Effective decision making is mandatory for any company to generate good revenue. Data plays a pivotal role in gathering valuable and actionable insights to get a fruitful outcome. Every individual has diverse thoughts and opinions depending upon their likings to choose any product or service.

The problem is to identify different groups of customers based on their shared characteristics and behaviors in order to develop more targeted marketing strategies and improve customer satisfaction. By segmenting customers, businesses can tailor their communication, product offerings, and services to specific groups, resulting in increased customer loyalty and higher revenue. However, determining the most effective segmentation strategic approach and identifying the key factors that differentiate each customer segment can be a challenging task. Therefore, the goal is to analyze the historical data so as to get insights and increase the sales of the company by focusing on different segments of the customers by observing their buying behavior as well as the effectiveness of the campaigns.

1.2 Domain Introduction

Domain: Data Science - Machine Learning

Modern technologies and techniques are used in data science to analyze enormous amounts of data in order to uncover hidden patterns, produce valuable information, and assist in commercial decision-making. It creates predictive models using sophisticated machine learning algorithms.

Client segmentation is the practice of dividing up a customer base into distinct groups based on shared traits or characteristics utilizing data science methodologies. In the digital era of intense competition, improving customer experience is the main issue and a huge necessity. Machine learning techniques can be used to analyze consumer data in order to find insights and patterns. Companies must implement ML systems that can process vast amounts of client data if they want to improve customer connections. The ML aids in obtaining razor-sharp customer analytics of the entire buying process.[7]

They can precisely identify customer types, which is far more challenging to do manually or with conventional analytical methods. To understand what makes the customers unique, data mining for customer segmentation is helpful. Additionally, it is beneficial to have a thorough understanding of your viewers' makeup.

Personalization is crucial, but it can be costly and challenging to scale. Customer management is becoming a necessity rather than a luxury that businesses can effectively implement by using several analysis approaches. The first stage towards personalization is segmentation. Marketing efforts can be prioritized, and a company's conversion rate can be increased by employing machine learning techniques to create subgroups depending on the likelihood of engagement.

1.3 Technical specifications and requirements

- **Programming Languages:** Python, R, SQL, HTML, CSS and JavaScript
- **Tools and Softwares:** Streamlit, Microsoft PowerBI, Apache Superset, Figma
- **Recommended Operating Systems:**

  Windows: 10

  MAC: OS X v10.7 or higher

  Linux: Ubuntu
- **Hardware Requirements:**

  Processor: Minimum 1 GHz; Recommended 2 GHz or more

  Hard Drive: Minimum 32 GB; Recommended 64 GB or more

  Memory (RAM): Minimum 4 GB; Recommended 8 GB or above

1.4 About Dataset

This dataset is a survey that a general store conducted to compile a database for a response model that can significantly improve the effectiveness of a marketing campaign by increasing responses or cutting costs. The objective is to predict who will respond to an offer for a product or service.[1]

The data set, which includes information of 2240 customers on:

- Campaign successes and failures.
- Product inclinations
- Televised performances
- Using customer profiles based on purchasing patterns

People

- *ID*: Unique column to identify and differentiate the customer
- *Year_Birth*: Represents the customer's year of birth
- *Education*: Represents the highest level of customer's education
- *Marital_Status*: Represents the current marital status of the customer (YOLO - You only live once, category/attribute)
- *Income*: yearly household income of the customer
- *Kidhome*: Count of children in customer's household
- *Teenhome*: Count of teenagers in customer's family
- *Dt_Customer*: Represents the joining date of customer in the company
- *Recency*: Number of days since the customer made the last purchase
- *Complain*: Represents if the customer made any complaint in the last 2 years. The value is 1 if the customer complained and 0 otherwise

Products

- *MntWines*: Amount spent on wine in last 2 years
- *MntFruits*: Spending on fruits over the past two years
- *MntMeatProducts*: Spending on meat products over the past two years

- *MntFishProducts*: Spending on fish products over the past two years
- *MntSweetProducts*: Spending on sweet products over the past two years
- *MntGoldProds*: Spending on gold products over the past two years

Promotion

- *NumDealsPurchases*: Number of purchases made with a discount
- *AcceptedCmp1*: 0 if the customer did not accept the offer during the first campaign or else 1
- *AcceptedCmp2*: 0 if the customer did not accept the offer during the second campaign or else 1
- *AcceptedCmp3*: 0 if the customer did not accept the offer during the third campaign or else 1
- *AcceptedCmp4*: 0 if the customer did not accept the offer during the fourth campaign or else 1
- *AcceptedCmp5*: 0 if the customer did not accept the offer during the fifth campaign or else 1
- *Response*: 0 if the customer did not accept the offer during the last campaign or else 1

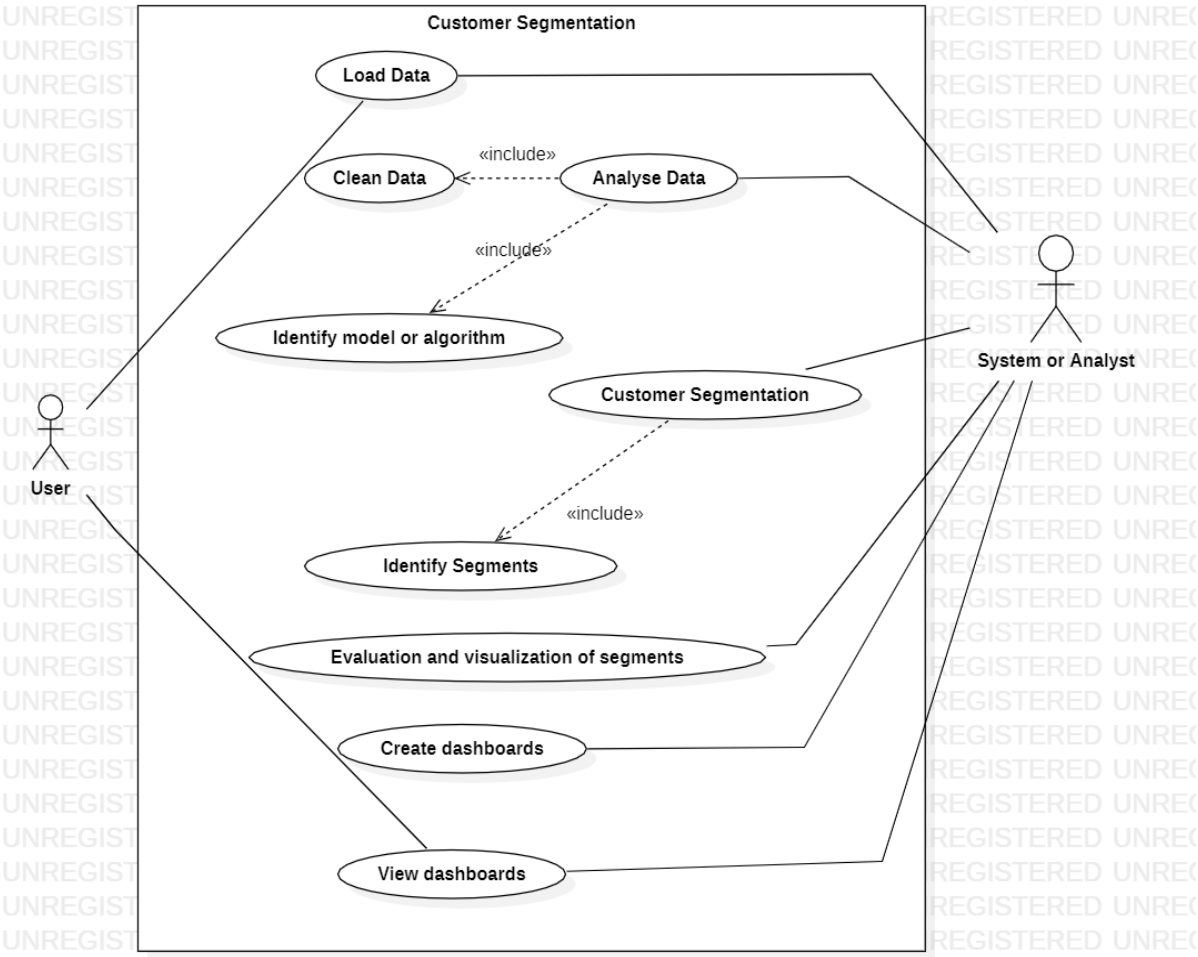
Place

- *NumWebPurchases*: Represents the total number of website purchases
- *NumCatalogPurchases*: Represents the total number of purchases through catalogs
- *NumStorePurchases*: Represents the total number of purchases made by visiting stores
- *NumWebVisitsMonth*: Represents the total number of customer visits in a month

1.5 Use Case Diagram

This illustrates how a system behaves from the perspective of the user when it answers a request. This describes how users will use your program to carry out tasks. User and System or Analyst are the two actors in our use case. The use case graphic demonstrates the steps involved in customer segmentation, from loading the data to creating and viewing dashboards.

Fig 1.1: Use Case Diagram



1.6 Use Case Description

Table 1.1: Use Case Description - Load data

Use case ID	1
Use case name	Load Data
Actor	User, Analyst/ System
Trigger	This use case will start when the customer visits the website and select their option to load data.
Pre-condition	Open the application
Post-condition	Analyze the data i.e., clean data and Identify Model or Algorithm
Main Flow	User opens the website/application Selects the option to load data User Loads the data Data loaded successfully
Alternate Flow	The data will not be loaded because an exception was thrown while loading the data.

Table 1.2: Use Case Description – Analyze Data

Use case ID	2
Use case name	Analyze Data
Actor	Analyst/ System
Trigger	This use case will start as soon as the user loads the data on the home page.
Pre-condition	Data is loaded
Post-condition	Cleaning of data with customer segments
Main Flow	Data loaded from the database is available to analysts. The analyst initially cleans the data as part of the data analysis. The best model or algorithm will be chosen after data cleaning.
Alternate Flow	Data is not loaded correctly. The analyst lacks the necessary access to view the data. Data cannot be accessed by analysts Inconsistent data, useless values, a dearth of data points, etc. cause data cleansing to fail.
List related use case names	Clean data Identify model or algorithm

Table 1.3: Use Case Description - Customer Segmentation

Use case Id	3
Use case Name	Customer Segmentation
Actor	Analyst/System
Trigger	This use case will start when the data is analyzed and a relevant model is selected.
Pre-condition	Data is analyzed i.e data is cleaned and Model or Algorithm is identified.
Post-condition	Evaluation and Visualization of Segments
Main flow	Process the analyzed data Using the information, identify segments of similar traits and characteristics Split the data to gain meaningful insights
Alternate Flow	Customer segmentation cannot be performed because there is not enough data to construct segments, there are not enough data points, or one segment dominates the dataset.
List related use case names	Identify segments

Table 1.4: Use Case Description - Evaluation and Visualization of Segments

Use case Id	4
Use case Name	Evaluation and Visualization of Segments
Actor	Analyst/System
Trigger	This case will initiate when the data is split into relevant segments.
Pre-condition	Customer Segmentation
Post-condition	Create dashboards
Main Flow	The processed data is evaluated. Based on this, data is visualized. Graphs and charts are formed.
Alternate Flow	Evaluation and visualization is not possible as the segments are not segmented properly.

Table 1.5: Use Case Description - Create Dashboards

Use case Id	5
Use case Name	Create Dashboards
Actor	Analyst/System
Trigger	This particular use case will start once the segments are visualized and charts and graphs are formed.
Pre-condition	Evaluation and Visualization of data.
Post-condition	View Dashboards
Main Flow	Analysts create a dashboard based on the charts and graphs
Alternate Flow	Please try again after some time.

Table 1.6: Use Case Description - View dashboards

Use case Id	6
Use case Name	View Dashboards
Actor	User, Analyst/System
Trigger	This case will be initiated when the users chooses to view dashboard
Pre-condition	Create dashboards
Post-condition	Exit the application/website.
Main Flow	Open the application, click on the view dashboards option.
Alternate Flow	Dashboard cannot be loaded currently either due to the wrong URL being called.

Chapter 2

Literature Review

It is quite challenging to determine and satisfy each customer's demands and requirements in a corporate setting. This is due to the fact that different clients may have different demands, wants, demographics, body types, tastes, features, and so forth. Currently, it is not a good business strategy to serve every consumer equally. Due to this problem, the idea of market or customer segmentation has been adopted, in which consumers are split up into smaller groups or segments, each of which comprises consumers who share certain market characteristics or behaviors. Customer segmentation is hence the process of separating the market into local populations.

Since there is fierce rivalry in the business world nowadays, businesses must increase their earnings and business by meeting customer requests and luring in new clients in accordance with their requirements. The process of identifying clients and meeting their individual needs is extremely difficult and complex. This is due to the fact that various clients may have varied needs, tastes, and preferences. Customer segmentation separates customers into groups that have common traits or behaviors as opposed to taking a "one size fits all" approach. Customer segmentation, according to, is a tactic for splitting the market into uniform groups.

Customer segmentation has gained a lot of traction recently as a means of retaining customers and generating revenue from them. In the study that follows, clients of various businesses are divided into groups based on behavioural traits like spending and income. These techniques outperform others because they take behavioural considerations into account. Based on their behavioural traits, customers are identified using a machine learning technique known as the k means clustering method. By using these clusters, the company may target particular customers and provide them with the content they are actually interested in through advertising campaigns and social media platforms.[8]

This tutorial offers a step-by-step procedure for locating, ranking, and focusing on the ideal existing client segments, but simply adhering to it won't ensure accomplishment. To be effective, one must anticipate and plan for the different difficulties and obstacles that may arise at each step, and one must always make sure to modify the procedure considering any fresh information

criticism that might alter the results. Furthermore, business cannot be forced to use this approach. The outputs generated by the best existing customers segmentation method will be largely meaningless if the key stakeholders who will be touched by it do not fully buy in.

However, if the best customer segmentation process is maintained, it can have a huge impact on all aspects of the business, including sales, marketing, product development, customer service, etc. The company will have a sharper focus on its customers and a clearer understanding of the market, which will enable it to scale in a far more controlled and effective way.[9]

That ultimately entails concentrating on a specific segment of consumers who present the most lucrative chances and effective use of resources rather than needing to accept every buyer who is willing to pay for the good or service. Every business needs that, but when a company is expanding, it may frequently be the difference between incredible success and certain failure.

Segmentation is one of the most important marketing concepts. Businesses can serve different consumer demographics. The market should therefore be segmented rather than enterprises trying to compete in it as a whole. Market and customer segmentation can help businesses identify the market segments that they can best serve.

Chapter 3

System Analysis & Design

3.1 ML Pipeline Architecture

This illustrates the input and output of data into machine learning models. All the following are included: the raw data input, features, outputs, machine learning model and model parameters, and prediction outputs in the form of segments and clusters.

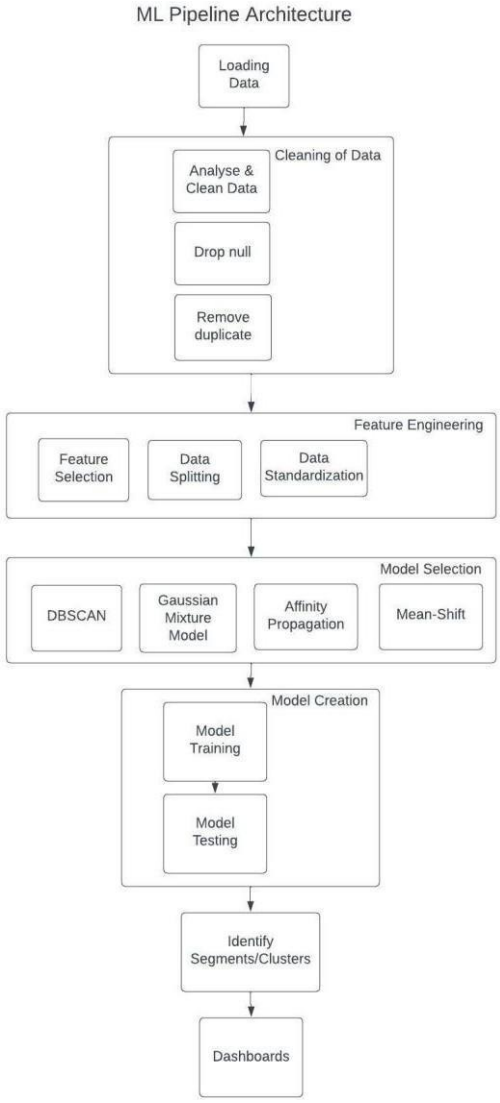


Fig 3.1: ML Pipeline Architecture

3.2 Activity Diagram

Following is a visual representation of step-by-step processes with assistance for the user's and analyst's decision-making and iteration.

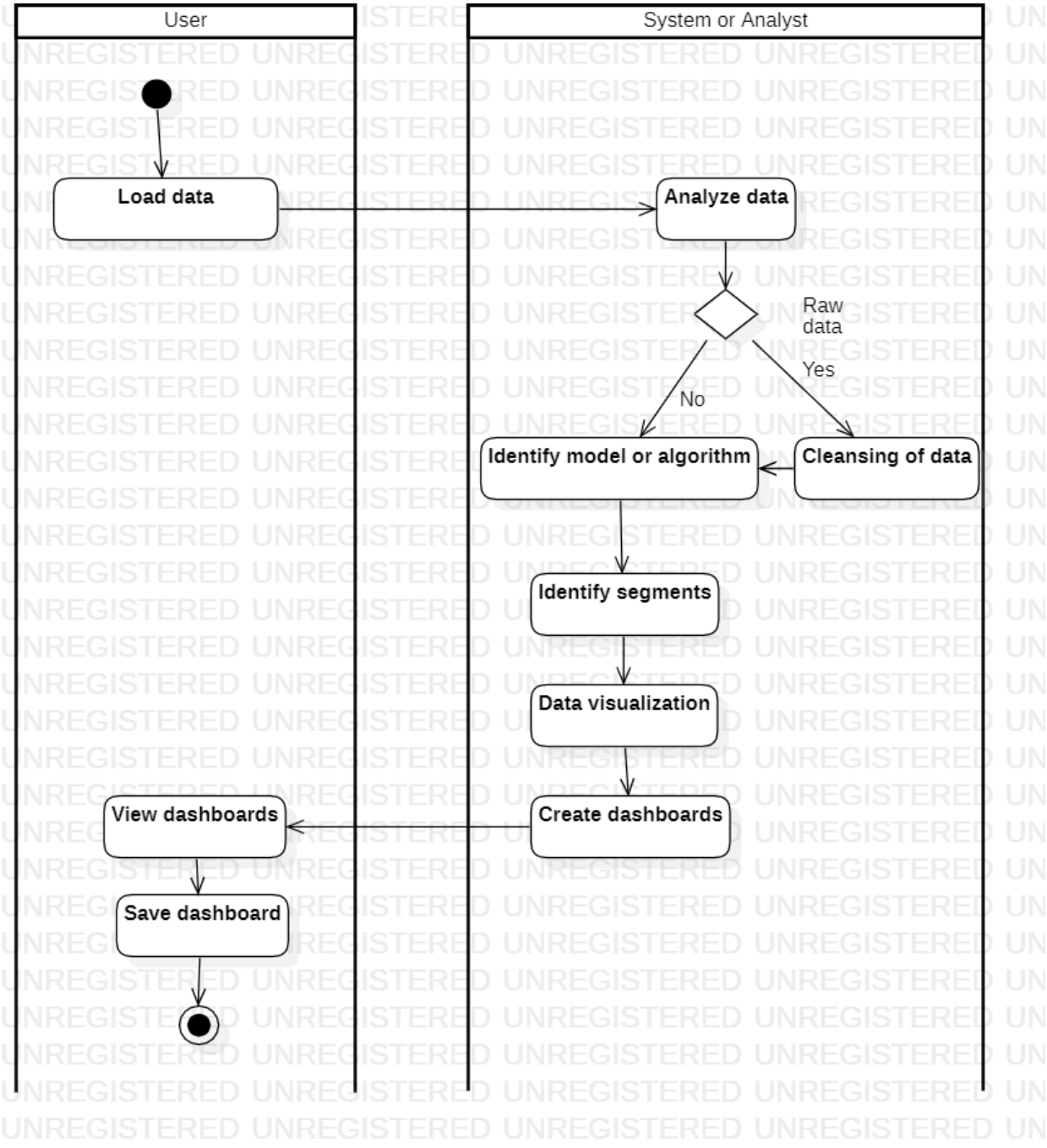


Fig 3.2: Activity Diagram

Chapter 4

System Implementation, Result and Discussion

4.1 Methodology

Customer segmentation is a crucial aspect of marketing strategy as it enables businesses to identify and target specific customer groups with tailored marketing campaigns[3]. Here is a methodology for a customer segmentation project:

- Define the objectives: Before starting the project, it is essential to define the objectives of customer segmentation. For instance, a business might want to identify customer groups with the highest potential for repeat purchases, or those that are most likely to buy premium products.
- Collect data: Next, gather data on customer demographics, behavior, and preferences. This data can be collected from various sources such as CRM systems, customer surveys, social media, and website analytics.
- Identify relevant variables: Based on the objectives of the segmentation project, identify the variables that are most relevant to segment customers. For instance, if the objective is to target premium product buyers, variables such as income, spending habits, and product preferences might be relevant.
- Analyze data: Use statistical analysis techniques such as cluster analysis or factor analysis to identify segments of customers with similar characteristics. Cluster analysis groups customers based on similarities between them, while factor analysis identifies underlying factors that drive customer behavior.
- Validate segments: Once customer segments are identified, validate them by examining their characteristics and behaviors. This can be done by conducting surveys or focus groups with customers in each segment to understand their needs and preferences.
- Develop marketing strategies: Finally, develop targeted marketing strategies for each customer segment. These strategies might include personalized messaging, product recommendations, or special offers.

4.2 Code Snippets

Figure 4.1 highlights a list of extensive libraries like datetime, numpy, pandas, matplotlib, seaborn etc used for this project ranging for the process of data manipulation, visualization, training, and testing models, clustering , regression analysis, plotting, data simplification.

```
import numpy as np
import pandas as pd
pd.set_option('display.max_column', None)

from datetime import datetime,timedelta
import matplotlib
import matplotlib.pyplot as plt
from matplotlib import colors
from matplotlib.colors import ListedColormap
import seaborn as sns
from mpl_toolkits.mplot3d import Axes3D

import scipy.stats
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from yellowbrick.cluster import KElbowVisualizer
from sklearn.cluster import KMeans
from sklearn.cluster import AgglomerativeClustering
from sklearn.model_selection import train_test_split
from sklearn.model_selection import KFold, cross_val_score
from sklearn.metrics import confusion_matrix, accuracy_score

from sklearn.ensemble import RandomForestClassifier
from sklearn import metrics
import warnings
import sys
if not sys.warnoptions:
    warnings.simplefilter("ignore")
```

Fig 4.1: All libraries implemented

Figure 4.2 depicts the use of command for reading and displaying the initial rows of the database

```
[2] data = pd.read_csv("marketing_campaign.csv")
```

```
[3] data.head()
```

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	MntFruits	MntMeatProducts
0	5524	1957	Graduation	Single	58138.0	0	0	04-09-2012	58	635	88	546
1	2174	1954	Graduation	Single	46344.0	1	1	08-03-2014	38	11	1	6
2	4141	1965	Graduation	Together	71613.0	0	0	21-08-2013	26	426	49	127
3	6182	1984	Graduation	Together	26646.0	1	0	10-02-2014	26	11	4	20
4	5324	1981	PhD	Married	58293.0	1	0	19-01-2014	94	173	43	118

Fig 4.2: Displaying top 5 rows

Figure 4.3 shows the usage of data.columns command which displays the names of all the columns in the dataset.

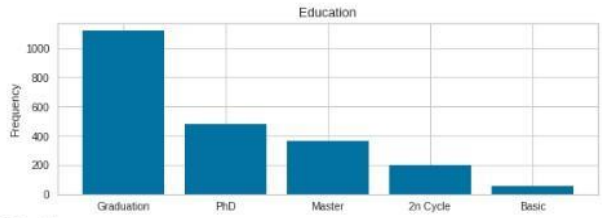
```
[4]  data.columns

Index(['ID', 'Year_Birth', 'Education', 'Marital_Status', 'Income', 'Kidhome',
       'Teenhome', 'Dt_Customer', 'Recency', 'MntWines', 'MntFruits',
       'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts',
       'MntGoldProds', 'NumDealsPurchases', 'NumWebPurchases',
       'NumCatalogPurchases', 'NumStorePurchases', 'NumWebVisitsMonth',
       'AcceptedCmp3', 'AcceptedCmp4', 'AcceptedCmp5', 'AcceptedCmp1',
       'AcceptedCmp2', 'Complain', 'Z_CostContact', 'Z_Revenue', 'Response'],
      dtype='object')
```

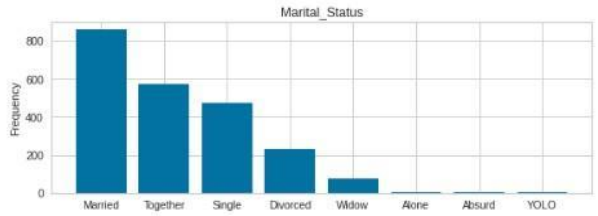
Fig 4.3: Column Names

In Figure 4.4, using functions and libraries, the different categories of Marital status and Education column and the number in each type are displayed. There are 1116 customers who belong to the Graduation category, 481 in PhD, 365 in Master, 200 in 2n Cycle and 54 in Basic category. Similarly for Marital Status, 857 belong to Married, 573 Together, 471 Single, 232 Divorced, 75 Widow, 3 Alone, 2 Absurd and 2 YOLO.

```
[ ]  categorical1 = ["Education", "Marital_Status"]
     for c in categorical1:
         bar_plot(c)
```




```
Education:
 Graduation    1116
PhD             481
Master          365
2n Cycle        200
Basic            54
Name: Education, dtype: int64
```




```
Marital_Status:
 Married     857
Together     573
Single       471
Divorced     232
Widow         76
Alone          3
Absurd         2
YOLO           2
Name: Marital_Status, dtype: int64
```

Fig 4.4: Plot for Education and Marital Status

Figure 4.5 shows the usage of functions and libraries using which the histogram for Year_Birth, Kidhome, Teenhome, Recency, and Complain is displayed. The histogram depicts the frequency of customers born in different years and the number of customers having 0, 1, or 2 kids in each home.

```
[ ]  def plot_hist(variable):
         plt.figure(figsize = (9, 3))
         plt.hist(data1[variable], bins = 50)
         plt.xlabel(variable)
         plt.ylabel("Frequency")
         plt.title("{} distribution with hist".format(variable))
         plt.show()
```

```
[ ]  people1 = ["Year_Birth", "Kidhome", "Teenhome", "Recency", "Complain"]
     for n in people1:
         plot_hist(n)
```

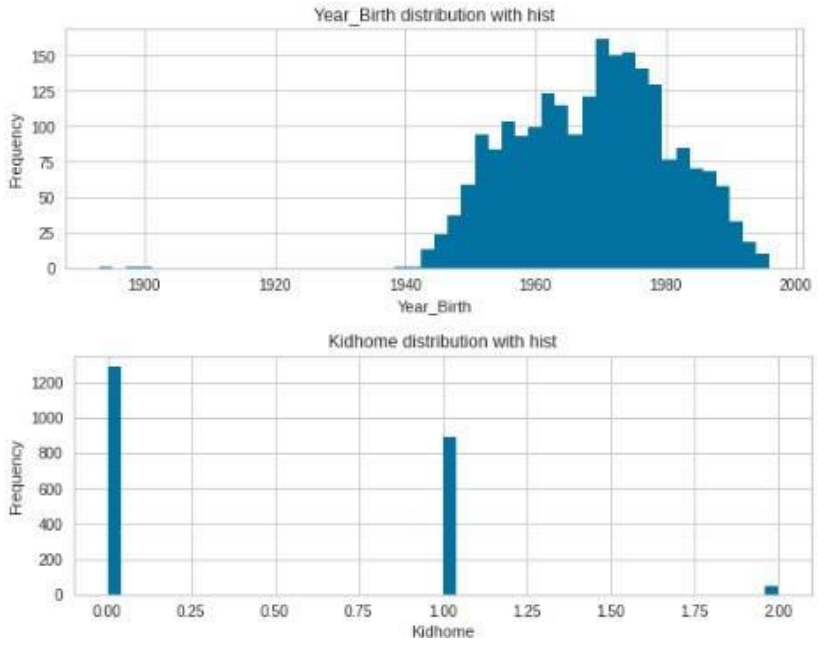


Fig 4.5: Year of birth and Kidhome

Figure 4.6 shows a heatmap which displays the correlation between different variables MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds, Response.

```
promotion = data1[["NumDealsPurchases", "AcceptedCmp1", "AcceptedCmp2", "AcceptedCmp3", "AcceptedCmp4", "AcceptedCmp5", "Response"]]
product = data1[["MntWines", "MntFruits", "MntMeatProducts", "MntFishProducts", "MntSweetProducts", "MntGoldProds", "Response"]]
place = data1[["NumWebPurchases", "NumCatalogPurchases", "NumStorePurchases", "NumWebVisitsMonth", "Response"]]
```

```
sns.heatmap(product.corr(), fmt = ".2f",annot = True)
plt.show
```

```
<function matplotlib.pyplot.show(*args, **kw)>
```

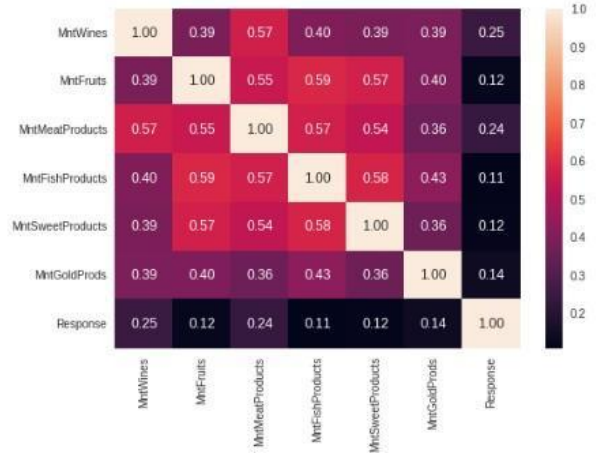


Fig 4.6: Heatmap

In figure 4.7, the age of the customer is being extracted from Year_Birth and the number of kids is being extracted from KidHome and TeenHome.

Feature Engineering

Extracting Age a Customer From Year_Birth Feature

```
data1["Age"] = 2022-data1["Year_Birth"]
```

```
data1["Total_kids"] = data1["Kidhome"] + data1["Teenhome"]
```

Feature For Customer's Family Size

```
data1["Family_members"] = data1["Relationship_status"].replace({"Single":1,"Partner":2}) + data1["Total_kids"]
```

Creating Spent Feature which indicates Total spendings of a customer on items

```
#Total spendings on all items
data1["Spent"] = data1["MntWines"]+ data1["MntFruits"]+ data1["MntMeatProducts"]+ data1["MntFishProducts"]+ data1["MntSweetProducts"]+ data1["MntGoldProds"]
```

```
data1.head()
```

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth	Accepte
0	5524	1957	Graduation	Single	58138.0	0	0	2012-04-09	58	635	88	546	172	88	88	3	8	10	4	7	
1	2174	1954	Graduation	Single	46344.0	1	1	2014-08-03	38	11	1	6	2	1	6	2	1	1	2	5	
2	4141	1965	Graduation	Together	71613.0	0	0	2013-08-21	26	426	49	127	111	21	42	1	8	2	10	4	
3	6182	1984	Graduation	Together	26646.0	1	0	2014-10-02	26	11	4	20	10	3	5	2	2	0	4	6	
4	5324	1981	PhD	Married	58293.0	1	0	2014-01-19	94	173	43	118	46	27	15	5	5	3	6	5	

Fig 4.7: Feature Engineering

Figure 4.8 depicts a box plot having the outlier for Year_Birth and Income.

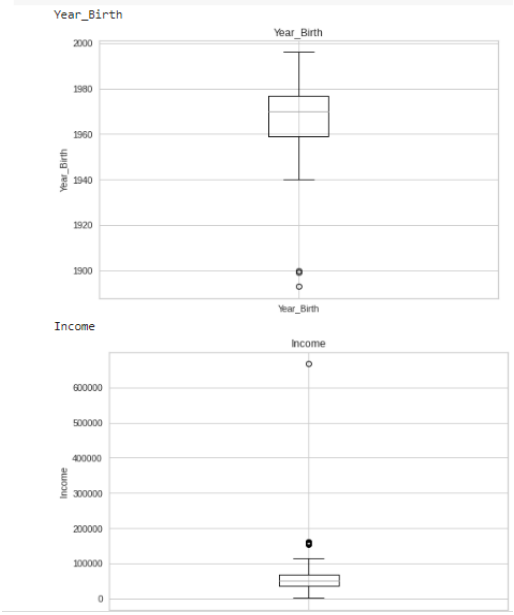


Fig 4.8: Outliers using box plot

Figure 4.9 shows the relative plot of some selected features and shows the various plots between parameters like Income, Recency, Age and Spent.

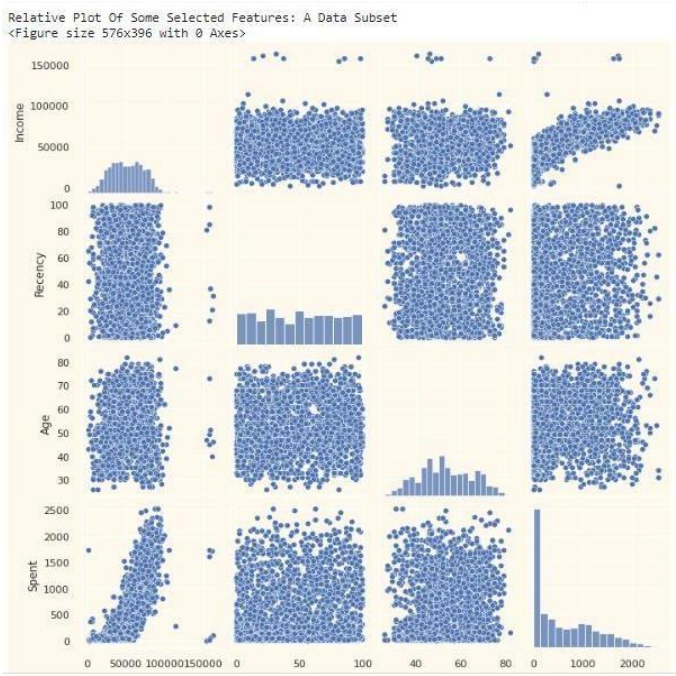


Fig 4.9: Relative Plot Of Some Selected Features

Figure 4.10 depicts a heatmap representing the correlation between different variables. The darker the color in the chart, the more is the correlation.

Fig 4.10: Heatmap 2

The Elbow Method is used to find the number of clusters to be formed. Figure 4.11 shows the output having a distortion score of 6709.843 and elbow at k=4 clusters.

```
[ ]  # Quick examination of elbow method to find numbers of clusters to make.
     print('Elbow Method to determine the number of clusters to be formed:')
     Elbow_M = KElbowVisualizer(KMeans(), k=10)
     Elbow_M.fit(PCA_ds)
     Elbow_M.show()
```

```
Elbow Method to determine the number of clusters to be formed:
```


```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7935976550>
```

Fig 4.11: Elbow Method.

Figure 4.12 shows the distribution of clusters using Agglomerative Clustering.

```
[ ]  #Plotting countplot of clusters
     pal = ["#682F2F","#B9C0C9", "#9F8A78","#F3AB60"]
     pl = sns.countplot(x=df["AggClusters"], palette= pal)
     pl.set_title("Distribution Of The Clusters")
     plt.show()
```

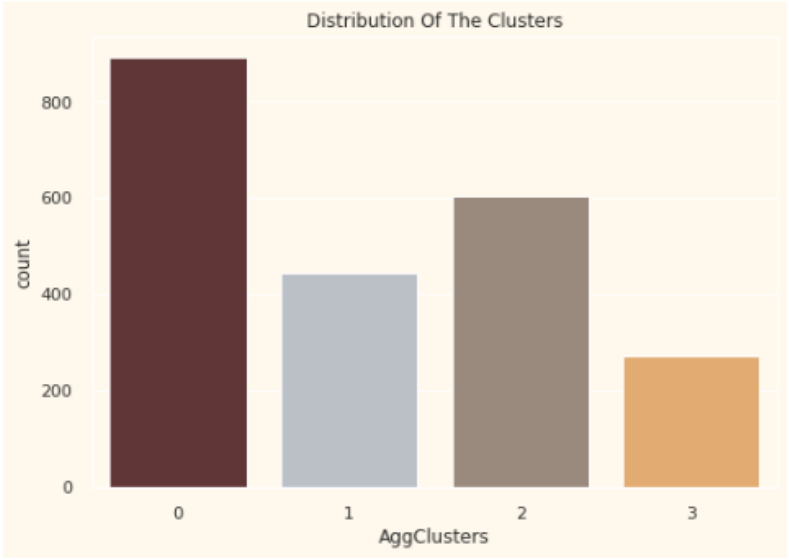


Fig 4.12: Agglomerative Clustering

Figure 4.13 shows a 3D Plot of clusters done through DBScan Clustering. Principal component analysis or PCA is also used.

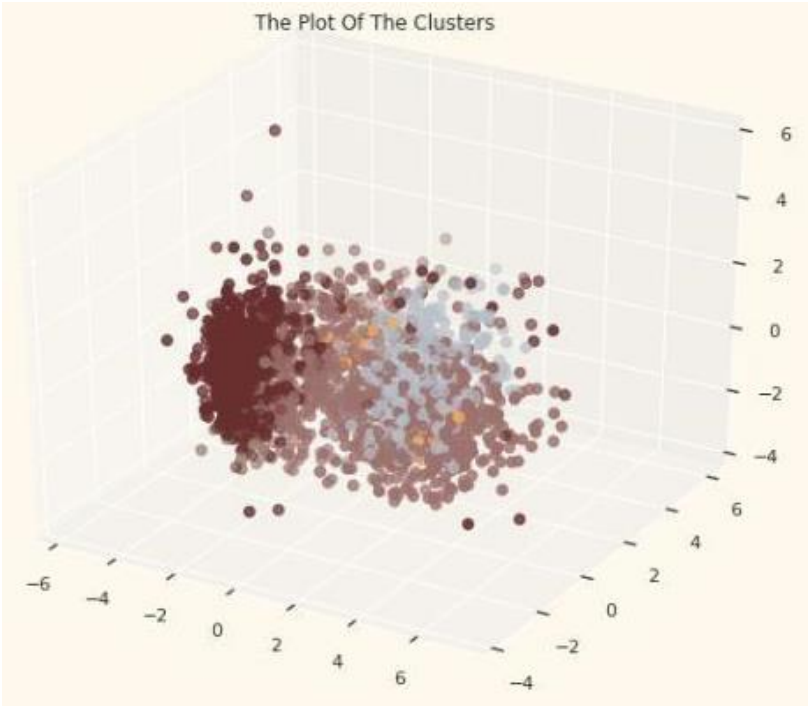


Fig 4.13: 3D Plot of DBScan Clustering

Figure 4.14 is the screenshot of the dashboard created using Power BI. Here, we have plotted the output which we received while formulating the code in such a way that the user will get an overview of the areas in which they need to work on to make more profit by just looking at this dashboard and getting the information of the customers as well.

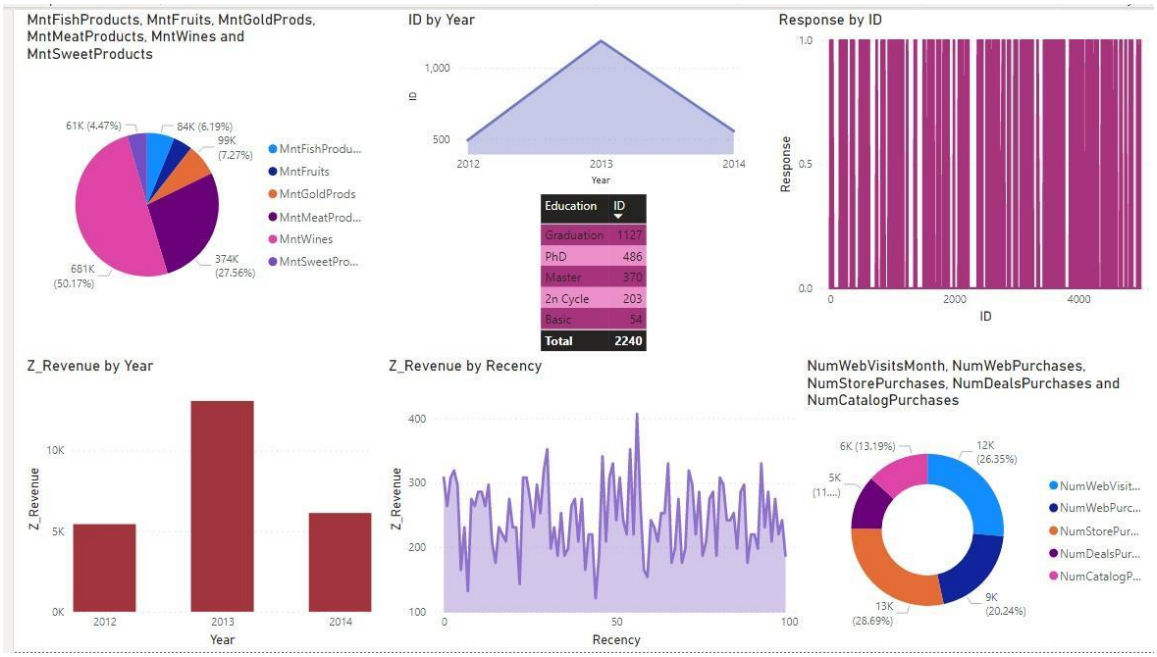


Fig 4.14: Dashboard 1 using Power BI

Figure 4.15 shows the dashboard created in Tableau depicting the marital expenses, expenses, education, Mnt Gold purchases and Mnt Wine Sales. It also contains the chart for the types of purchases vs Average Age.

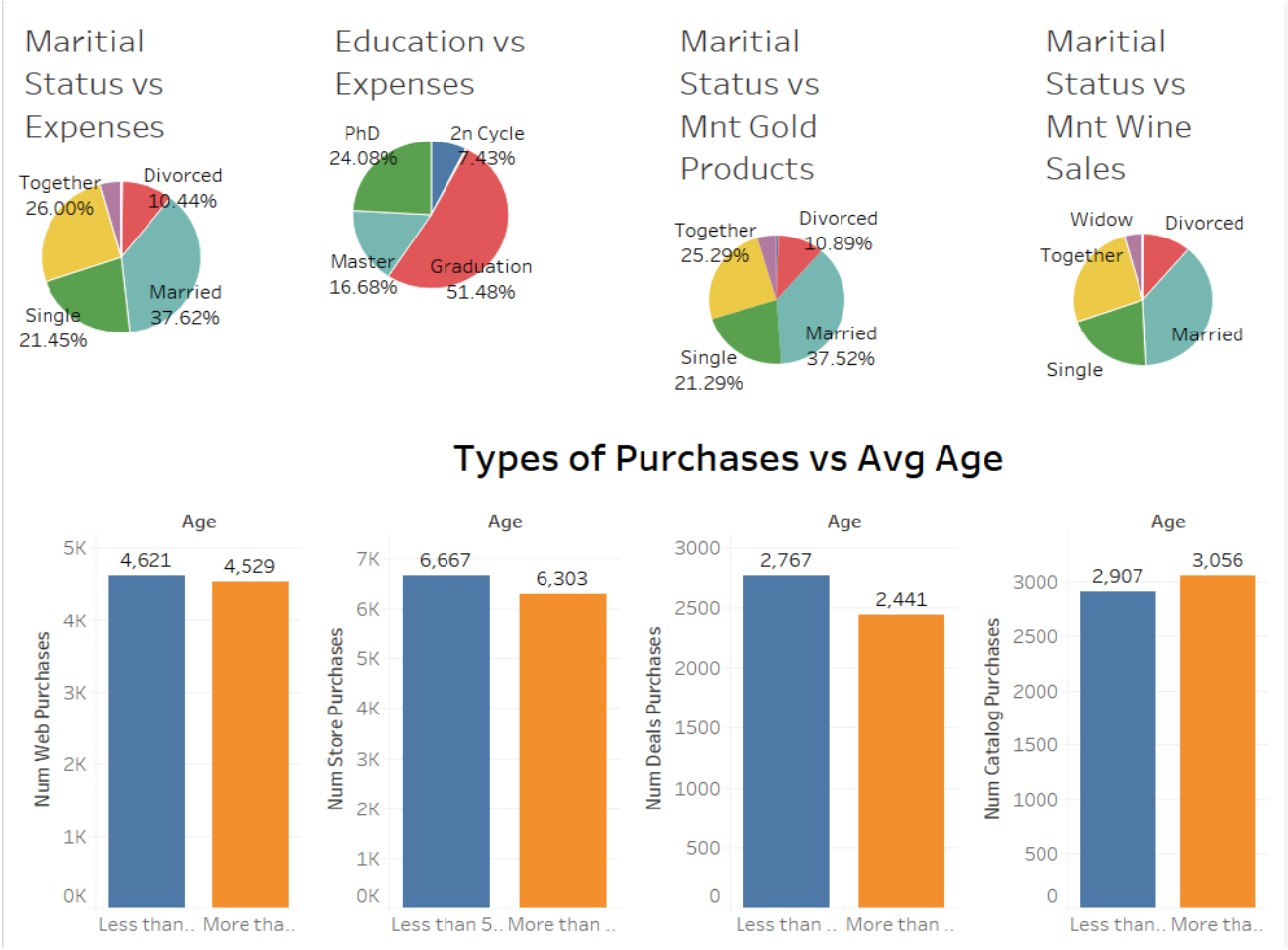


Fig 4.15: Dashboard 2 using Tableau

Figure 4.16 is the dashboard depicting the number of different purchases against Education and Number of web purchases vs Age. The highest purchases are of Education and the number of web purchases are 300 made by 57 year old people.

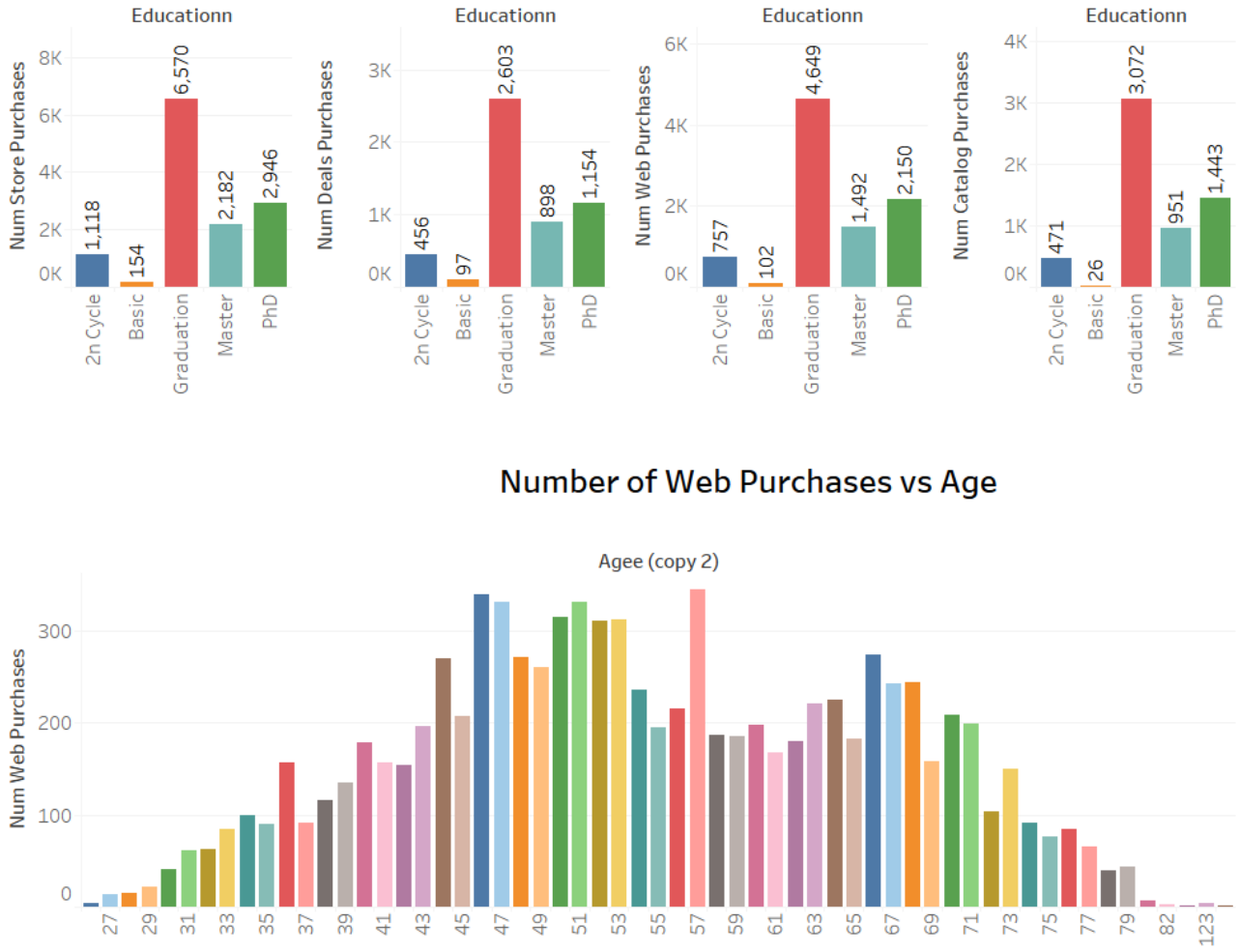


Fig 4.16: Dashboard 3 using Tableau

Figure 4.17 shows the dashboard depicting the Amount of Sweet and Age and also the accepted campaigns according to the different age groups. The amount of sweet purchases are 2766 made by 47 year old people.

Fig 4.17: Dashboard 4 using Tableau

Figure 4.18 is a dashboard depicting the Accepted campaigns against Education and the count of year wise expenses. The highest campaigns are accepted by the Graduation category and the highest expenses are 715425 that were made in 2013.

Fig 4.18: Dashboard 5 using Tableau

Figure 4.19 shows the code for the distribution of marital status expenses for the various categories and the pie chart for it respectively. The highest percentage is for the YOLO category and the lowest is for the Alone category.

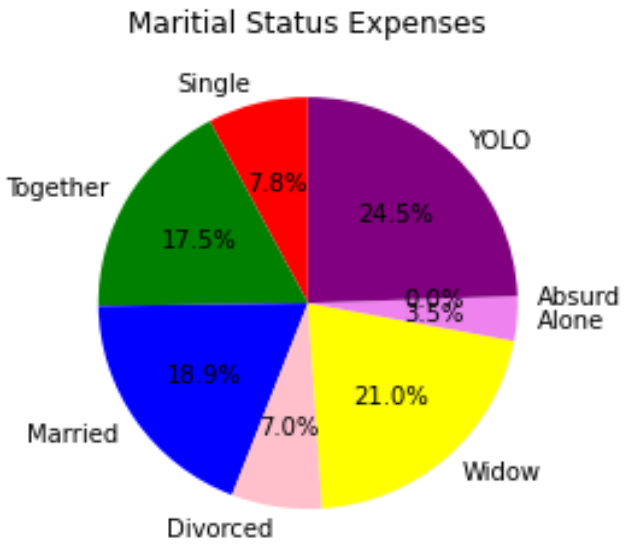


Fig 4.19: Pie chart for Marital Status Expenses

Figure 4.20 shows the code and pie chart for Education Expenses which is the highest for 2nd Cycle.

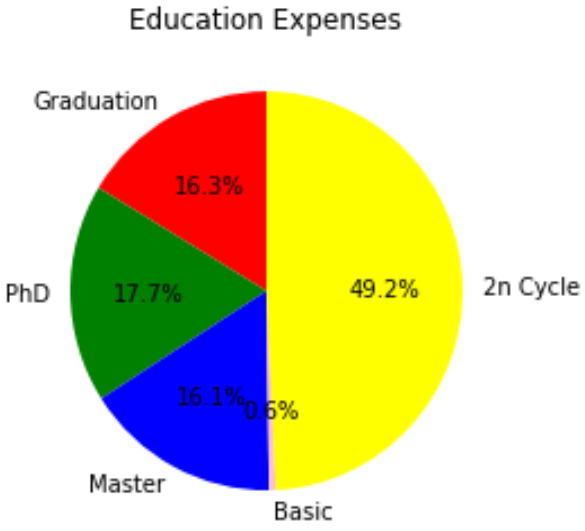


Fig 4.20: Pie chart showing the Education Expenses

Figure 4.21 depicts the code and Bar Graph distribution for total number of web purchases, total number of catalog purchases and total number of store purchases. The highest count is of the store purchases which is between 6000 and 7000.

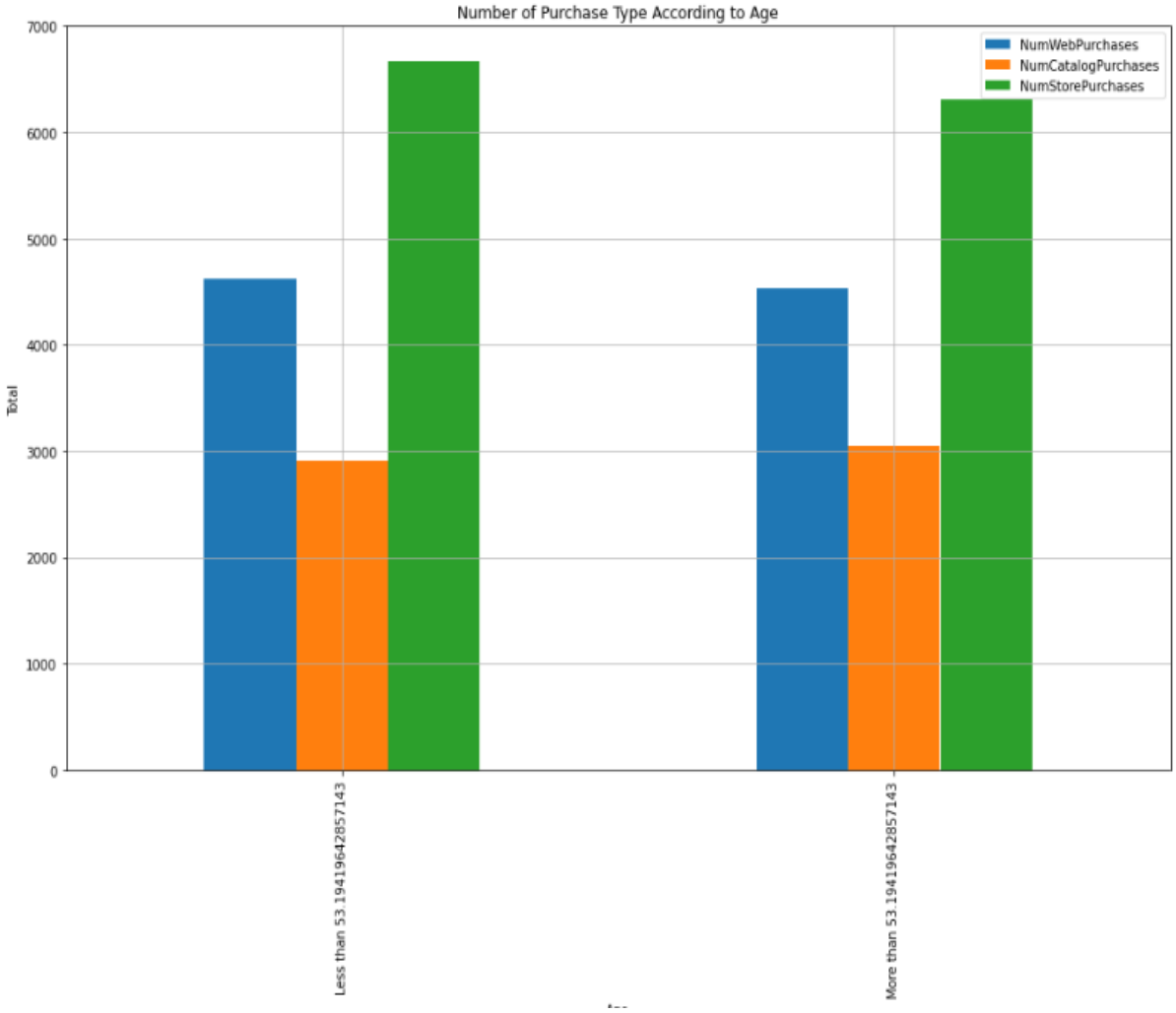


Fig 4.21: Bar chart of Number of Purchases according to Age

Figure 4.22 depicts the bar chart for total number of web purchases, total number of catalog purchases and total number of store purchases according to Education.

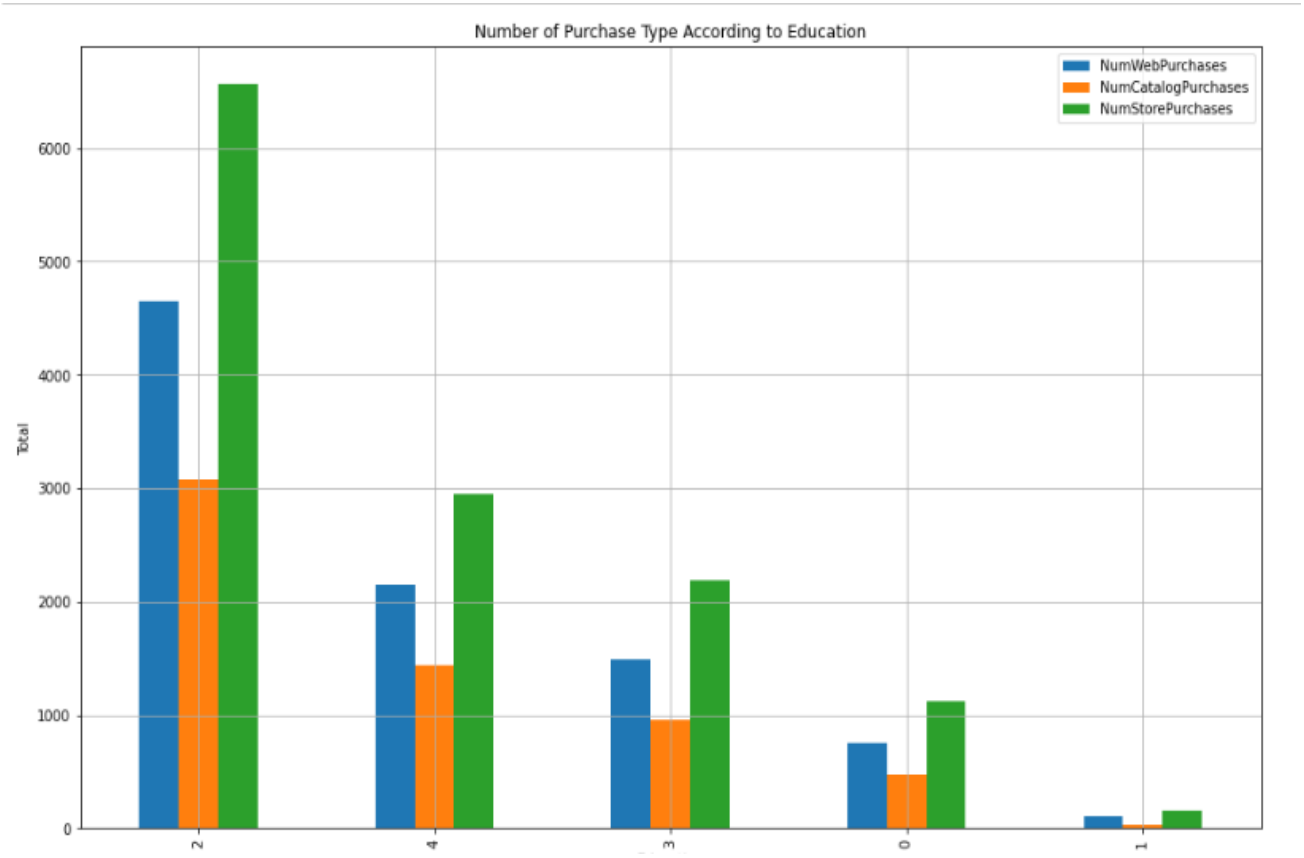


Fig 4.22: Bar chart of Number of Purchase Type according to Education

Figure 4.23 shows the count of all the accepted campaigns according to age group. The highest number is of campaign 3 for the people with age less than 53 years and the lowest number is of campaign 2. The highest number is of campaign 4 for the people with age more than 53 years old and the lowest number is of campaign 2.

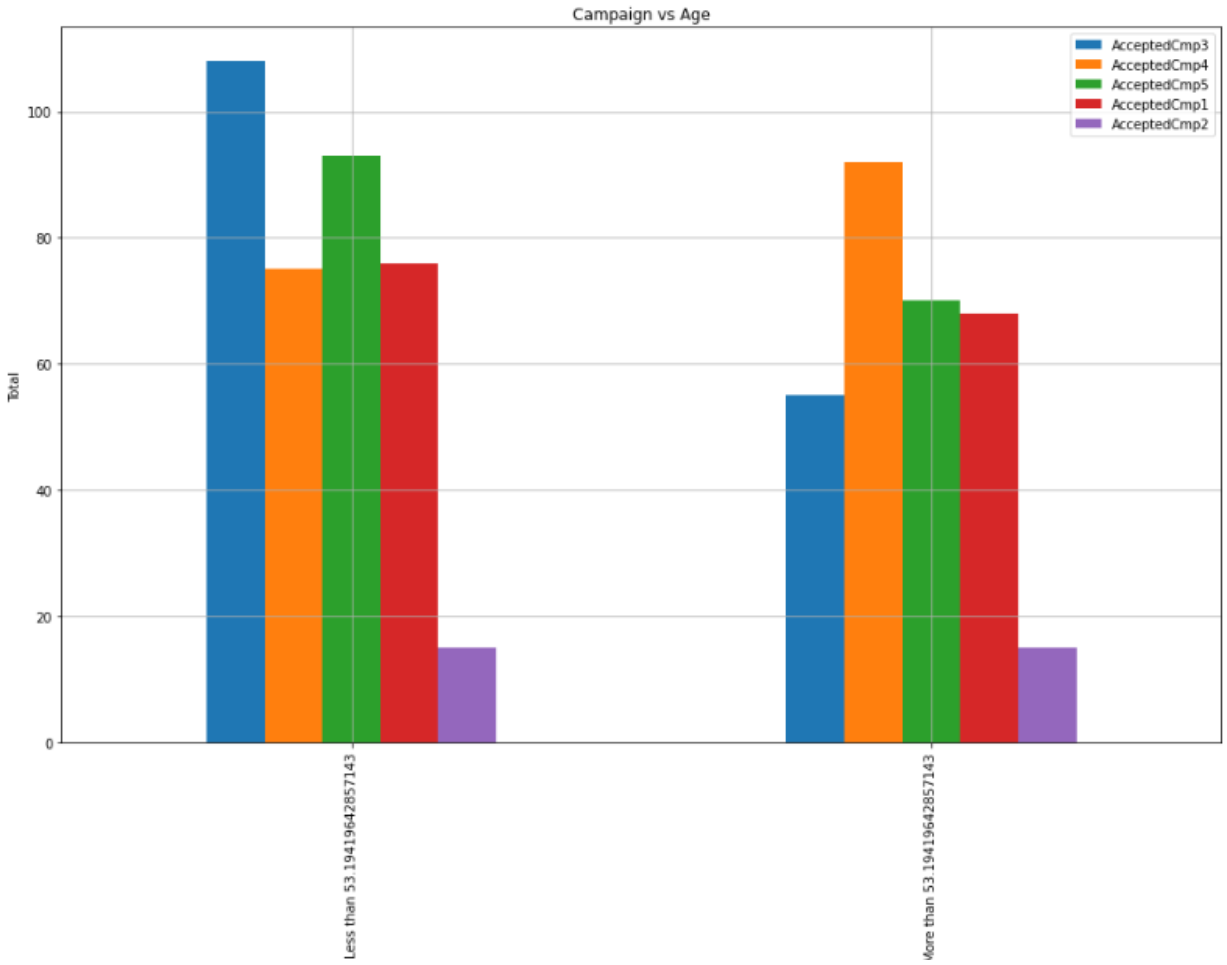


Fig 4.23: Bar chart for different campaigns vs Age

Figure 4.24 and Figure 4.25 shows the percentage of campaign responses based on Income and Education and Children and Marital Status respectively. Figure 4.24 helps us to interpret the response of people with different ranges of income and education level. Figure 4.25 helps us to interpret the response to campaign based on Marital Status and Children

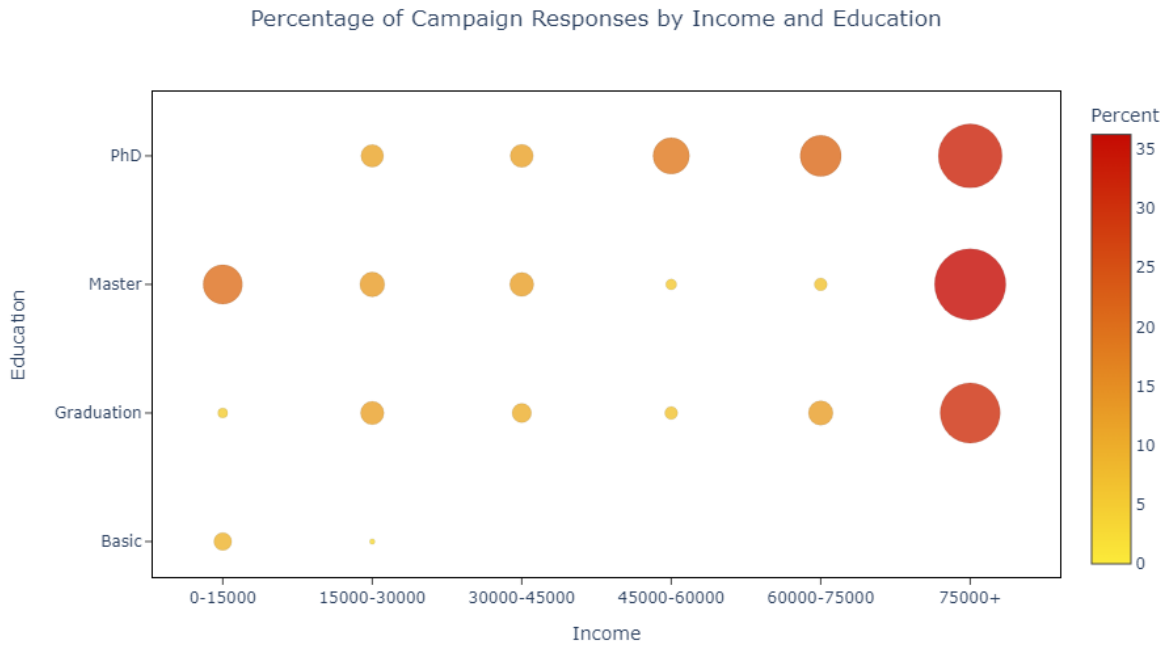


Fig 4.24: Percentage of campaign responses by Income and Education

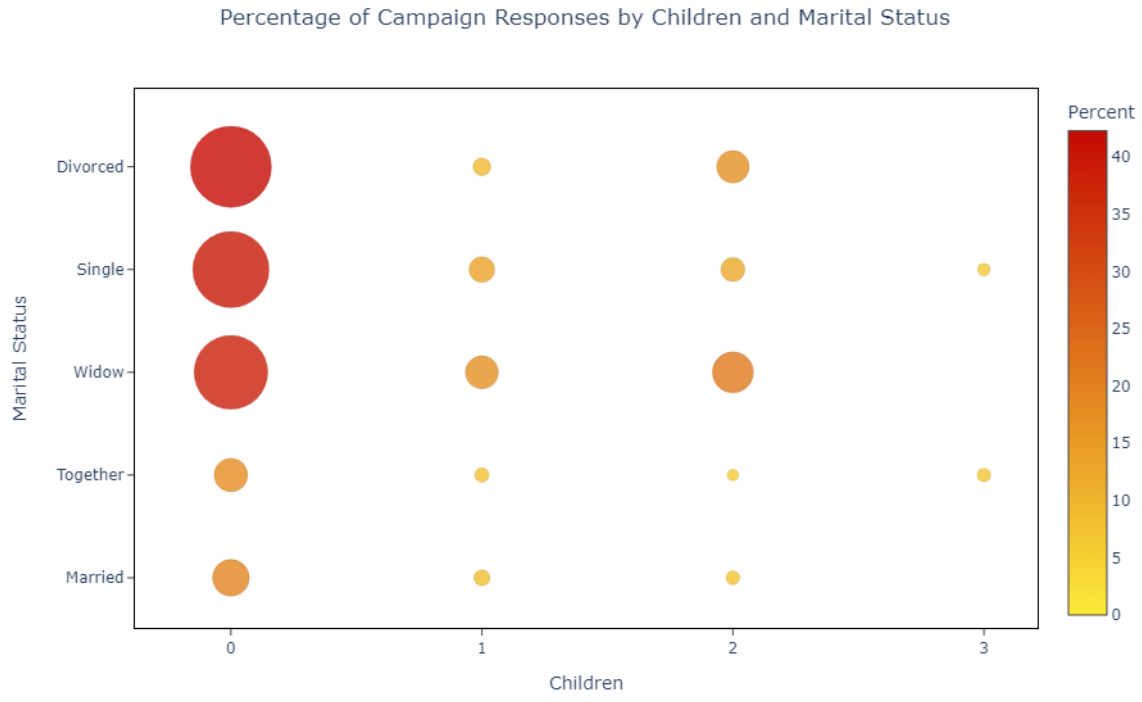


Fig 4.25: Percentage of campaign responses by Children and Marital Status

Figure 4.26 represents the percentage of campaign responses by Recency and Frequency Analysis. We cannot lose 28.57% of customers at any cost.

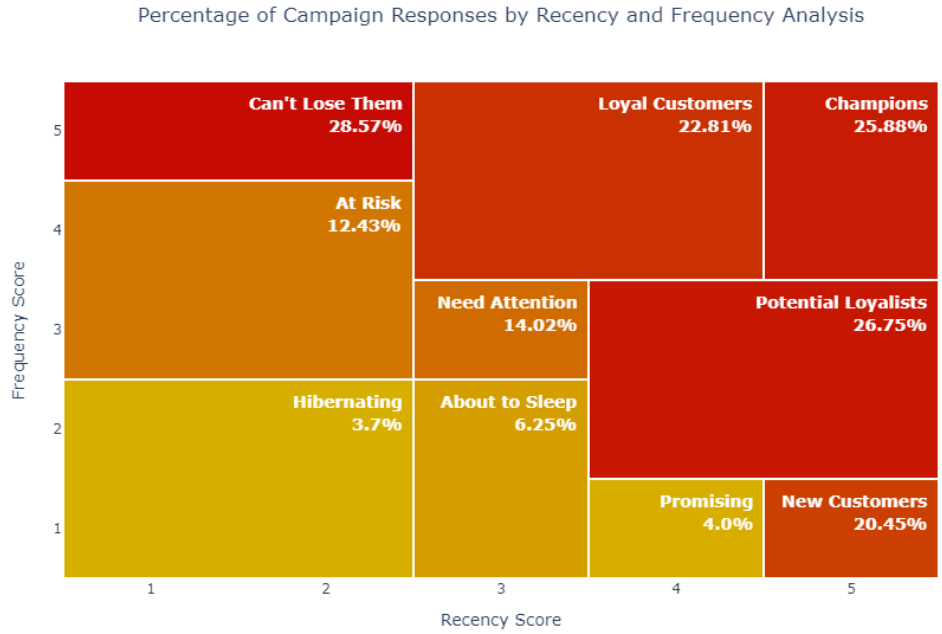
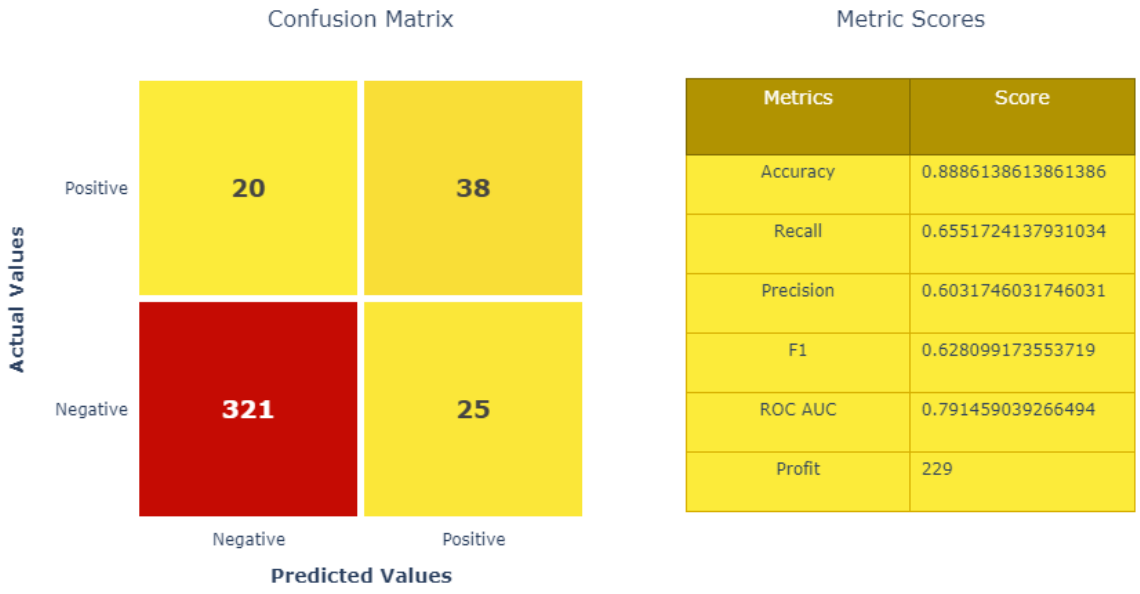


Fig 4.26: Percentage of campaign responses by Recency and Frequency Analysis

Figure 4.27 represents the Confusion Matrix and Metric Scores. Here confusion matrix is allowing and helping us to measure recall, precision and accuracy. It is also representing a table of different outcomes of prediction and helps visualize its outcomes.



Metric Scores

Metrics	Score
Accuracy	0.8886138613861386
Recall	0.6551724137931034
Precision	0.6031746031746031
F1	0.628099173553719
ROC AUC	0.791459039266494
Profit	229

Fig 4.27: Confusion Matrix and Metric Scores

Figure 4.28 represents importance of each feature. We can understand that Store purchases is an important feature.

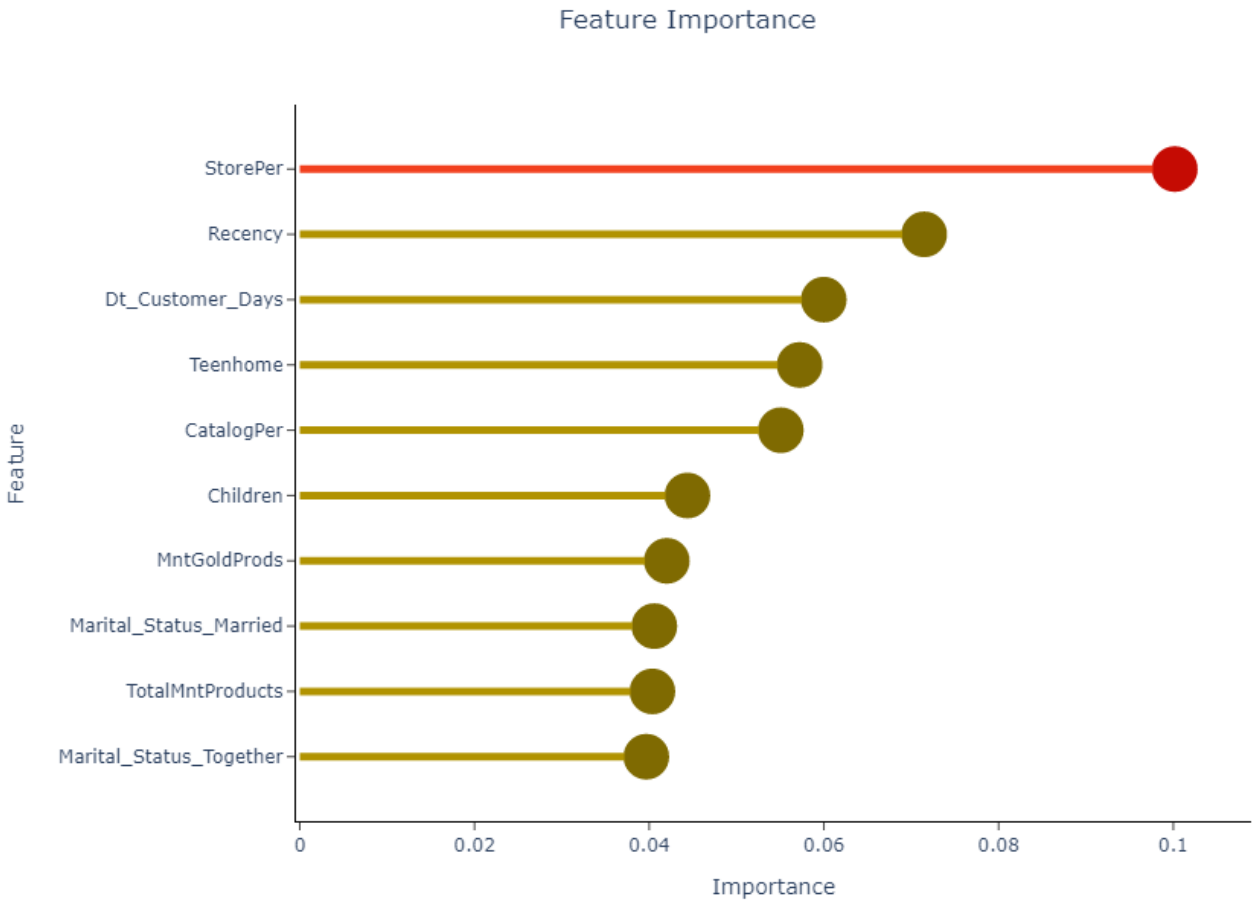


Fig 4.28: Feature Importance

Figure 4.29 represents Profit Score based on Random Forest Classifier. It helps us to get the optimum result by choosing the majority among them as the best value.

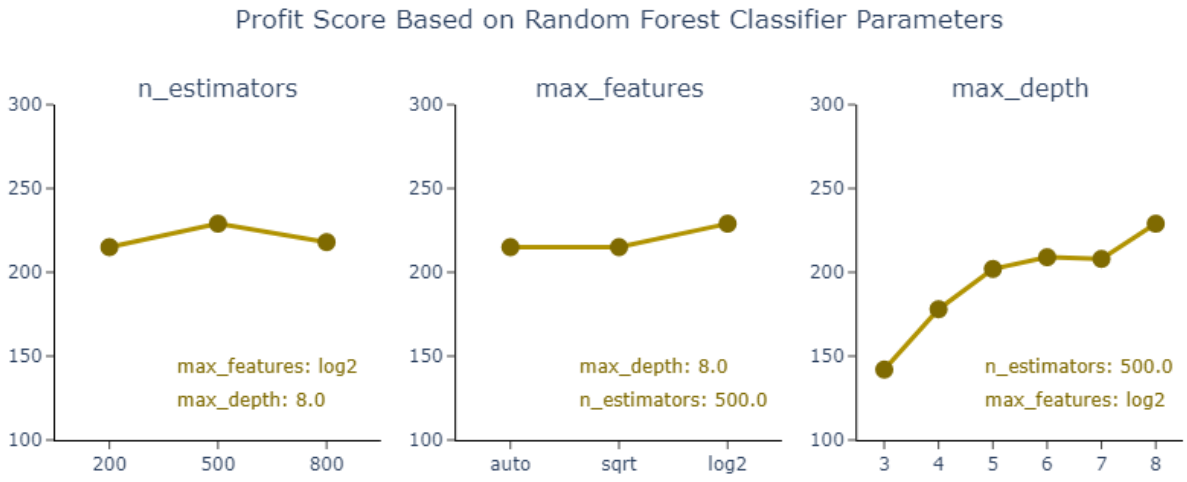


Fig 4.29: Profit Score based on Random Forest Classifier

Figure 4.30 represents Cluster's Profile Based on Income and Spending. Based on the clusters we can see the income and the spending pattern. We can see Cluster 0 has high spending & average income.
Cluster 1: high spending & high income
Cluster 2: low spending & low income
Cluster 3: high spending & low income

Fig 4.30: Cluster's Profile Based on Income and Spending

Figure 4.31 represents Purchasing style of customers depending upon clusters. Different clusters prefer different type of purchasing methods and spend amount according to their preference.

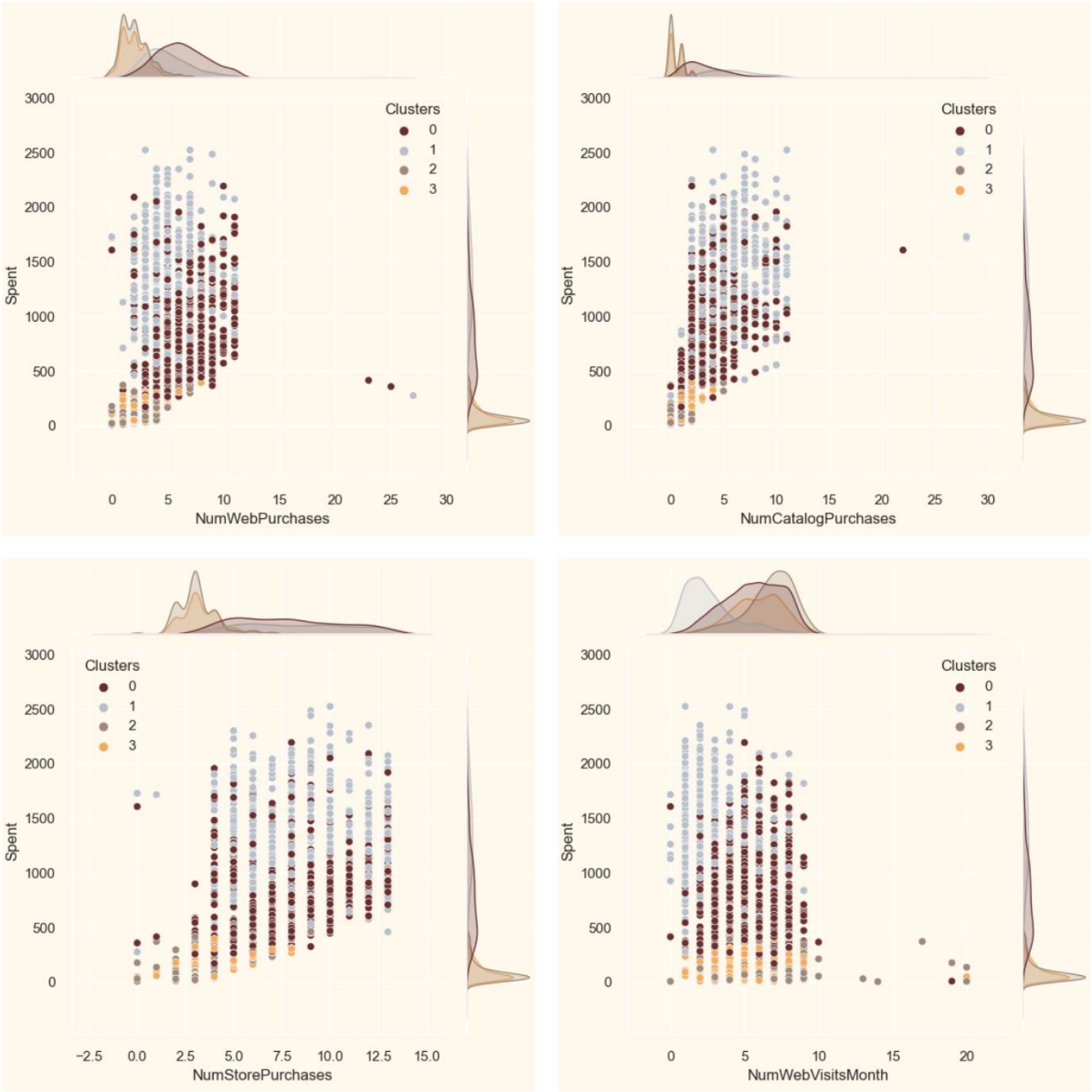


Fig 4.31: Purchasing style depending upon clusters

4.3 Test Cases Report

Table 4.1: Test Cases

S.no.	Action	Input	Expected Output	Actual Output	Test Browser	Test Result
1.	Checking and printing the null values	Dataset	Should display Count and percentage of null values in the dataset	Count and percentage of null values in the dataset is displayed	Google Chrome	Pass
2.	Calculating total number of kids	Dataset (Kidhome & Teenhome)	Number of children in the family should be printed in numeric forms	Number of children in the family in numeric form is printed	Google Chrome	Pass
3.	Calculating amount spent on all items	Dataset	The total amount spent on all items should be printed	The total amount spent on all items is printed	Google Chrome	Pass
4.	Calculate the number of web purchases according to age	Dataset	Should display a graph showing the number of web purchases according to age	Number of web purchases depending on age is displayed	Google Chrome	Pass
5.	Find the effectiveness of campaign depending on the Education	Dataset	Should display a graph representing the number of people that accepted different campaigns depending on education	Number of people that accepted different campaigns depending on education is displayed	Google Chrome	Pass

```
[8]  data1 = data.copy()

[9]  col_info=(pd.DataFrame(data1.isnull().sum()).T.rename(index={0:'null values'})) # to calculate null values
     col_info=col_info.append(pd.DataFrame(data1.isnull().sum()/data1.shape[0]*100).T.rename(index={0:'null values (%)'}))
     #to calculate percentage of null values

     print
     display(col_info)
```

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts
null values	0.0	0.0	0.0	0.0	24.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
null values (%)	0.0	0.0	0.0	0.0	1.071429	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Fig 4.32: Test Case: Checking and printing the null values

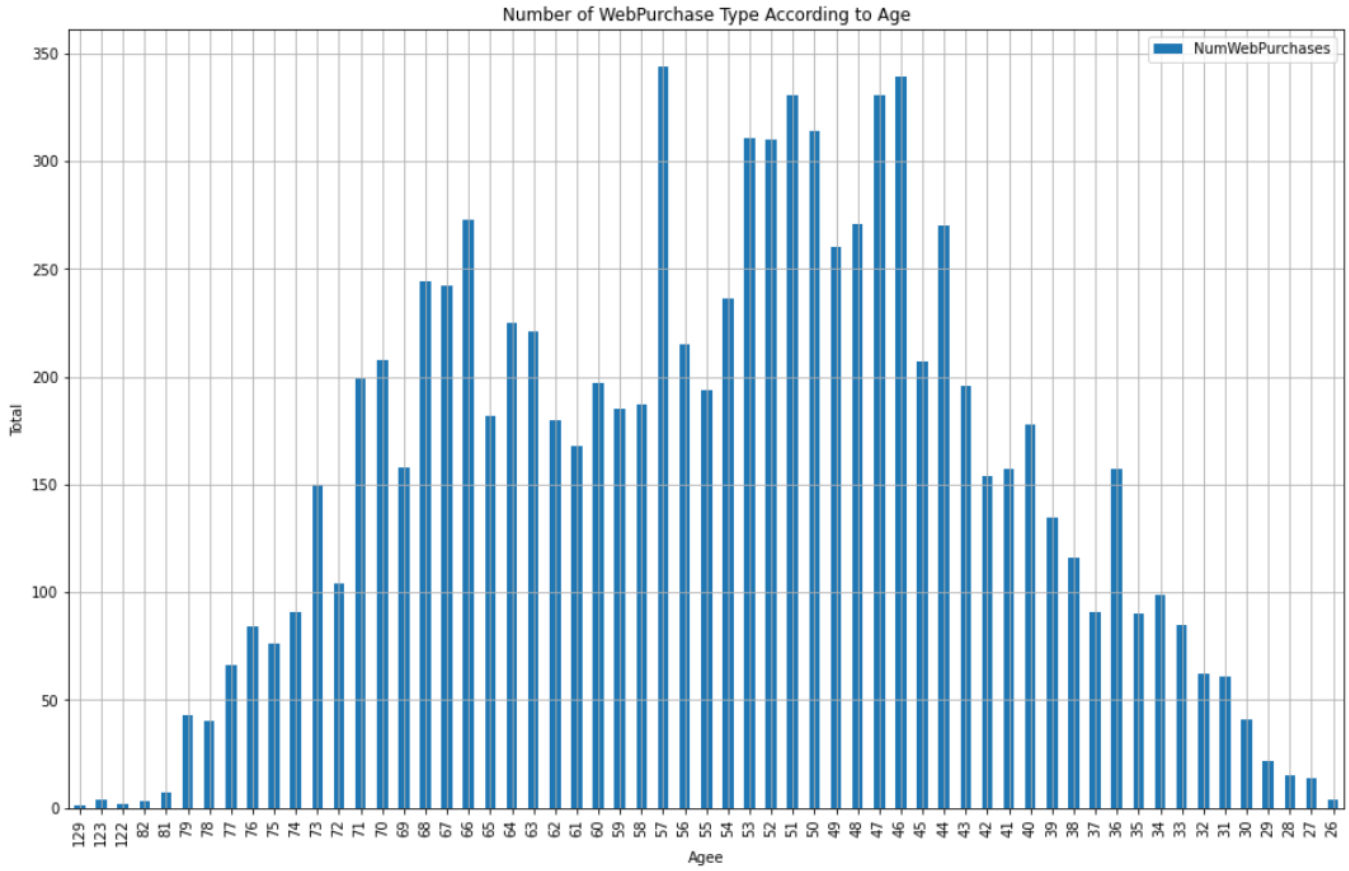


Fig 4.33: Test Case: Calculate the number of web purchases according to age

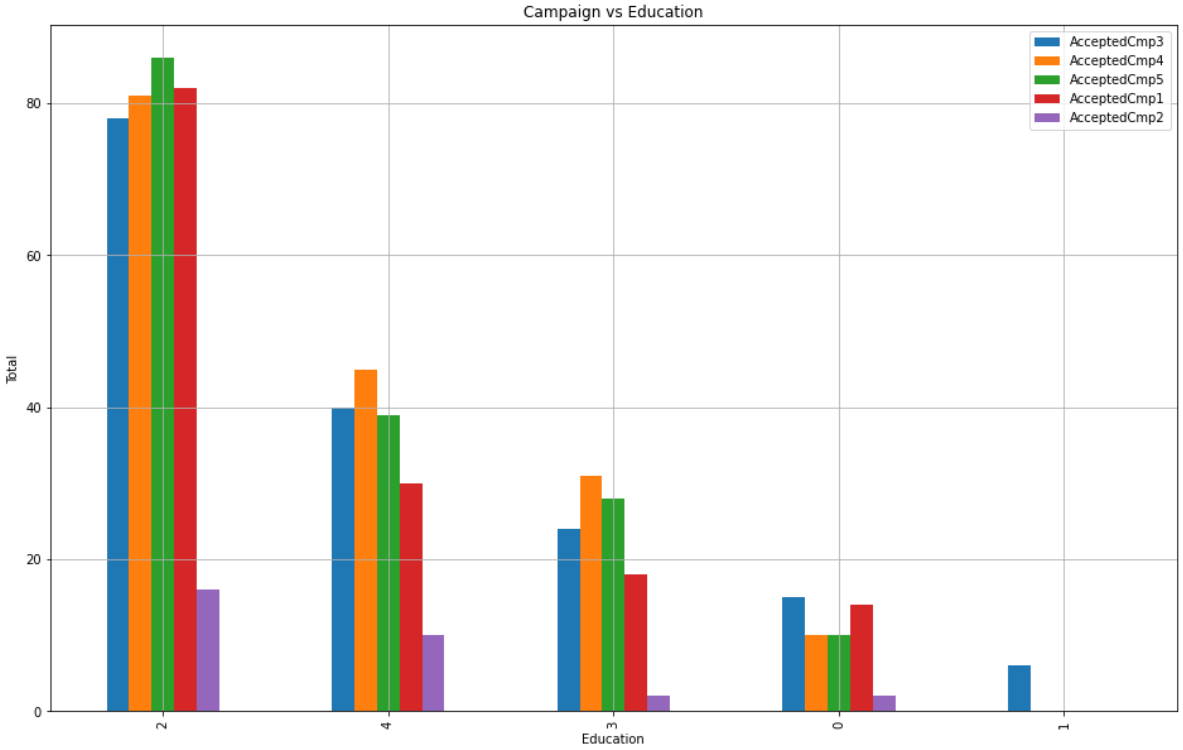


Fig 4.34: Test Case: Find the effectiveness of campaign depending on the Education

4.4 Result

After data analysis and customer classification based on features and annual income, the marketing team can use these clusters to develop strategies for specific consumer ideas to offer value to them. To solve the existing customer segmentation problem, segmenting customers based on behavioral attributes is a superior strategy. The K-means clustering algorithm is recommended for this strategy. Using parameters like total spending and annual income, clients are segmented based on the company's customer data using agglomerative clustering and the elbow method machine learning algorithm. By segmenting the market, we could find out more about the goods that consumers favored, chose, and purchased. Using clustering algorithms, patterns in the data were found and then developed into groupings.[11]

During our analysis we found out some key points:

- It is found that 49.2% of expenditure for education was spent on 2n cycle.
- The most spent amount on wine was by the widow group.
- The most spent amount on gold was by YOLO.
- Graduates prefer spending the most amount on shopping and they prefer in store purchases.
- The average age of the customers is 53 years. When dividing the customer base on the average age it is found that whether the customers are above the age of 53 or below them they prefer web purchases.
- The most spent amount on sweet products is by people with age greater than 53.
- Campaign 3 was successful for customers whose age is less than 53.
- Campaign 4 was successful for customers whose age is more than 53.
- For the basic level of Education, campaign 3 was successful.
- No campaign was accepted by the YOLO group.
- Most sales were made in the year 2013.

Based on the analysis and the clusters formed we can identify the features of different clusters

- **About Cluster Number: 0**
  1. Are definitely a parent
  2. At the max have 4 members in the family and at least 2
  3. Single parents are a subset of this group
  4. Most have a teenager at home
  5. Relatively older

- **About Cluster Number: 1**
  1. Are definitely not a parent
  2. At the max are only 2 members in the family
  3. A slight majority of couples over single people
  4. Span all ages
  5. A high income group

- **About Cluster Number: 2**
  1. The majority of these people are parents
  2. At the max are 3 members in the family
  3. They majorly have one kid
  4. Relatively younger

- **About Cluster Number: 3**

  1. They are definitely a parent
  2. At the max are 5 members in the family and at least 2
  3. Majority of them have a teenager at home
  4. Relatively older
  5. A lower income group

Chapter 5

Conclusion and Future Work

5.1 Conclusion

Serving every consumer with the identical product model, email, text message, or advertisement is not a good idea and hard to implement. The best existing client categories are found, prioritized, and targeted using a step-by-step procedure. Customer segmentation technique helped to find an optimal number of unique customer groups which results in better understanding of the specific needs and requirements of customers. The best existing client categories are found, prioritized, and targeted using a step-by-step procedure. Through the analysis, we were able to identify that customer segmentation depends on various parameters.

A company needs to keep in mind various attributes such as age group, salary, gender and marital status, etc. of their customers to identify the specific needs and requirements of their customers. Since widows spend the most on Wine, they should be targeted for increasing the sales of Wine. Similarly for Gold, the target group should be YOLO. Since no campaign was accepted by the YOLO group, a campaign should be launched focusing on the YOLO Group.

5.2 Future Scope

In the coming future, we will be creating a dashboard using streamlit in which when the user inputs relevant information, the name of the segment to which the customer belongs will be displayed. We will also be applying different models in the future to segment the data and then compare which model will produce the best results. The future scope of customer segmentation projects is quite promising, as businesses continue to collect large amounts of customer data through various channels, including social media, e-commerce platforms, and mobile apps. With the help of advanced analytics and machine learning algorithms, businesses can analyze this data to identify patterns and trends that can inform their customer segmentation strategies. As businesses continue to collect more data and adopt more advanced analytics tools, the future scope of customer segmentation projects will continue to evolve and expand.

5.3 System Limitations

- The primary limitation is data as the customer segmentation heavily relies on data in order to gain fruitful results. So the data must be accurate, sufficient and recent to derive a useful result. As after a period, the analysis performed will be of less significance to us as the choices of customers will change in the future timeline. [10]
- While implementing and considering the outcomes of customer segmentation, the sample considered may lead to biases which will make incomplete segments and the accuracy won't reflect customer behavior properly.
- It is observed that the preferences of customers change in the course of time. Customer segments are based on the past preferences of customers which will not align with expected predictions according to their new preferences.
- Customer segmentation may involve collecting and analyzing sensitive customer data, which raises ethical concerns around privacy and data protection. Companies must ensure they are collecting and using customer data in a responsible and transparent manner.
- Once customer segments have been created, it can be challenging to implement targeted marketing campaigns and strategies that effectively reach each segment. This requires resources and expertise, which may not be readily available.

References

[1] https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis (accessed 4th Nov. 2022)

[2] https://blog.hubspot.com/service/customer-segmentation (accessed 20th Oct. 2022)

[3] https://towardsdatascience.com/clustering-algorithm-for-customer-segmentation-e2d79e28cbc3 (accessed 20th Oct. 2022)

[4] https://www.business-science.io/business/2016/09/04/CustomerSegmentationPt2.html (accessed 20th Oct. 2022)

[5] https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html (accessed 20th Oct. 2022)

[6] https://powerbi.microsoft.com/en-au/ (accessed 5th Nov. 2022)

[7] https://www.simplilearn.com/tutorials/data-science-tutorial/what-is-data-science#:~:text=with%20an%20example%3F-,Data%20science%20is%20the%20domain%20of%20study%20that%20deals%20with,assess%20creditworthiness%20and%20loan%20risk. (accessed 3rd Oct. 2022)

[8] http://e-journal.uajy.ac.id/6688/3/EMI218464.pdf

[9] https://odr.chalmers.se/server/api/core/bitstreams/983356e9-5c72-4df7-bcef-3dc96f179131/content

[10] Aman Banduni, Prof Ilavendhan A. School Of Computing Science & Engineering, Galgotias University, Greater Noida, U.P Customer Segmentation Using Machine Learning

[11] Prof. Nikhil Patankar a ,1, Soham Dixit a, Akshay Bhamare a, Ashutosh Darpel and Ritik Raina Dept. Of Information Technology Sanjivani College of Engineering, Kopargaon-423601 (MH), India. Customer Segmentation Using Machine Learning