Scientific Programming with Python - Final Project

Dana Daniella Aloni – 207907742

Customer grouping by shopping behaviour and personal information to improve targeted marketing.

1) Introduction

The subject matter of the data set is concerned with respect to the shopping behaviour of customers to improve the target marketing. Different customers are grouped into different classes based on their shopping behaviour. The data set consists of 8120 instances with 15 attributes including the ID.

The different attributes are ID, Gender, is the person ever married, Age, if he is graduated, his profession, work experience, spending score, Family size, the day during which he shops, Customer deviation from average store customer spending on non-specified products, Customer deviation from average store customer spending on dairy products, Customer deviation from Average store customer on household products. The target variable is the group and can be classified into either A, B, C or D. Most of the variables are categorical in nature. However, there are a few features which are continuous as well.

2) Initial Data Analysis

The description of the data is as shown in as shown in Figure (a)

```
[158] 1 df.describe()
```

	Age	Work_Experience	Family_Size	Shop_Day	Shop_Other	Shop_Dairy	Shop_Household	Shop_Meat
count	8000.000000	7082.000000	7784.000000	8078.000000	7869.000000	7827.000000	8120.000000	8120.000000
mean	43.423500	2.627648	2.851233	5.086779	7.911617	4.766589	5.589339	5.103642
std	16.724207	3.385906	1.530861	1.503912	5.042374	4.267057	4.869441	3.808877
min	8.000000	0.000000	1.000000	0.000000	-6.055000	-9.070000	-9.581000	-8.670000
25%	30.000000	0.000000	2.000000	4.000000	4.404000	1.992500	2.309750	2.453750
50%	40.000000	1.000000	3.000000	6.000000	7.175000	4.341000	4.999000	4.642500
75%	53.000000	4.000000	4.000000	6.000000	11.179000	7.121500	8.646500	7.461750
max	89.000000	14.000000	9.000000	7.000000	28.181000	22.588000	24.215000	20.562000

Figure (a) Description of the data set

We can see from the description of the data set that there are few negative values as well for features like Shop_Dairy, Shop_Household, etc. This is since the features were normalised.

Data Fixes: There were several null values present in the data set (as shown in Fig 2) that was handled accordingly. The rows of the categorical features that contained null values was removed from the data set.

```
Gender              220
Ever_Married        179
Age                 120
Graduated           325
Profession          124
Work_Experience    1038
Spending_Score        0
Family_Size         336
Shop_Day             42
Shop_Other          251
Shop_Dairy          293
Shop_Household        0
Shop_Meat             0
Group                 0
dtype: int64
```

Figure (b) Number of null values for each feature

Shop_day had to be between 1 and 7. But there were rows where the Shop day was 0. All these rows were removed from the data set. For the other features, the null values were replaced with its median.

3) Exploratory Data Analysis

Exploratory Data Analysis is an approach to summarise the main characteristics of the data set.

Plotting of Histograms

Histograms are somewhat like bar charts but are used to check the characteristics of continuous variables. Figure (c) and (d) shows the histogram for Age and Shop_Dairy

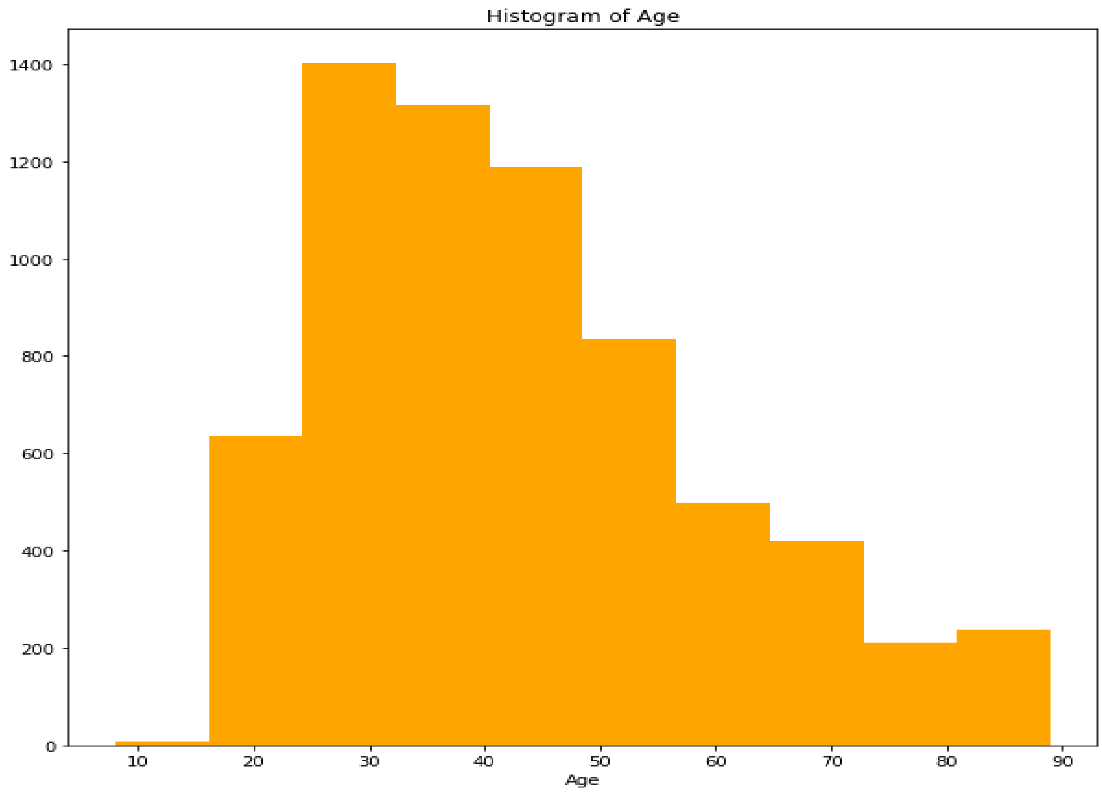
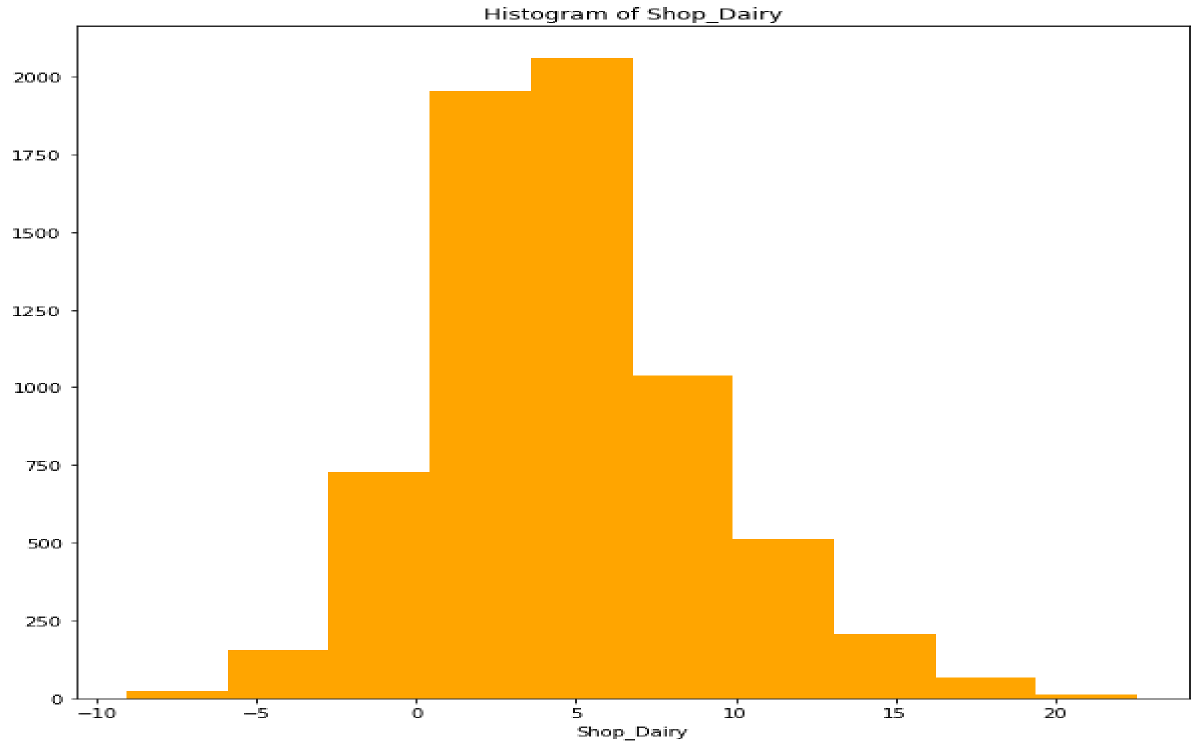


Figure (c): Histogram for Age

We can check that the median age is around 40. Most of the people's age is concentrated around 25-45.



Figure(d) Histogram for Shop Dairy

We can see that the Shop Dairy is feature is Normally Distributed and both its mean and median is equal and is around 5

Plotting of Boxplots

Boxplots are used to check for the outliers in the data set. Black dots represent outliers. Figure e and Figure (f) shows the boxplots for Age and Work Experience

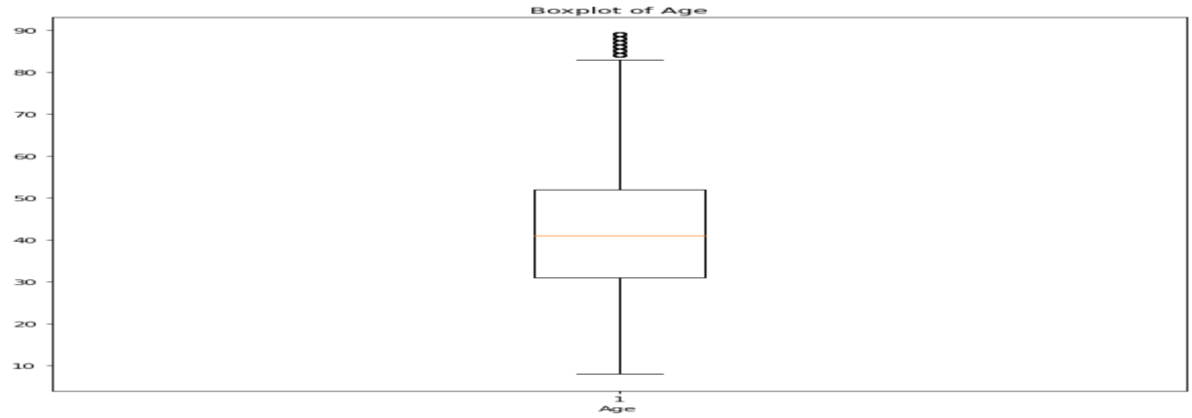


Figure e: Boxplot for Age

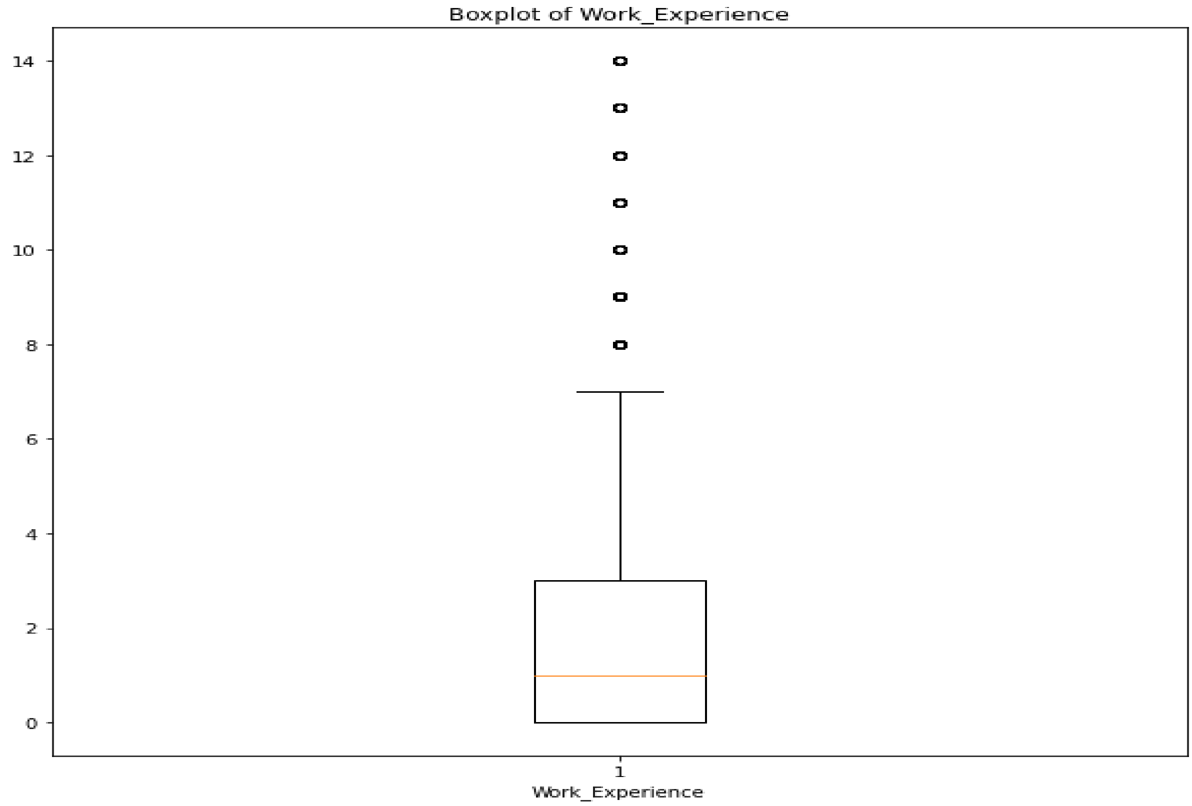


Figure f: Boxplot for Work Experience

There aren’t many outliers present in the data set.

Data Visualization

There are many categorical features present in the data set. Let us do some data visualisation on them using Bar plots. Let’s do a value count for the target feature (number of instances of each group). Similarly lets do a value count for Gender, value counts by profession (as shown in Figure h and Figure i)

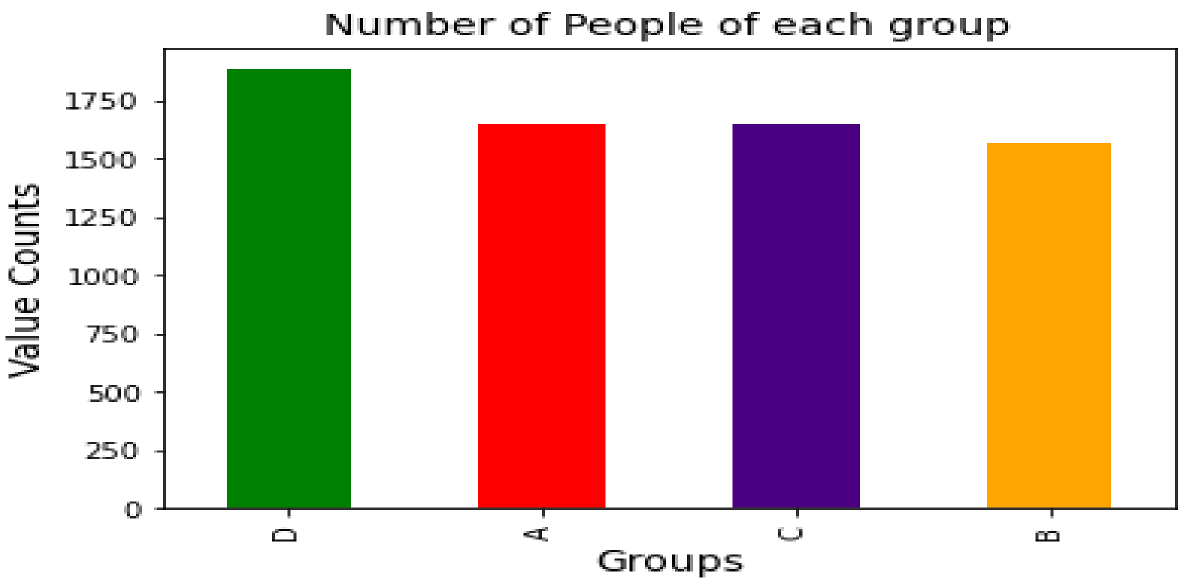


Figure g: Number of People of Each Group in the data set

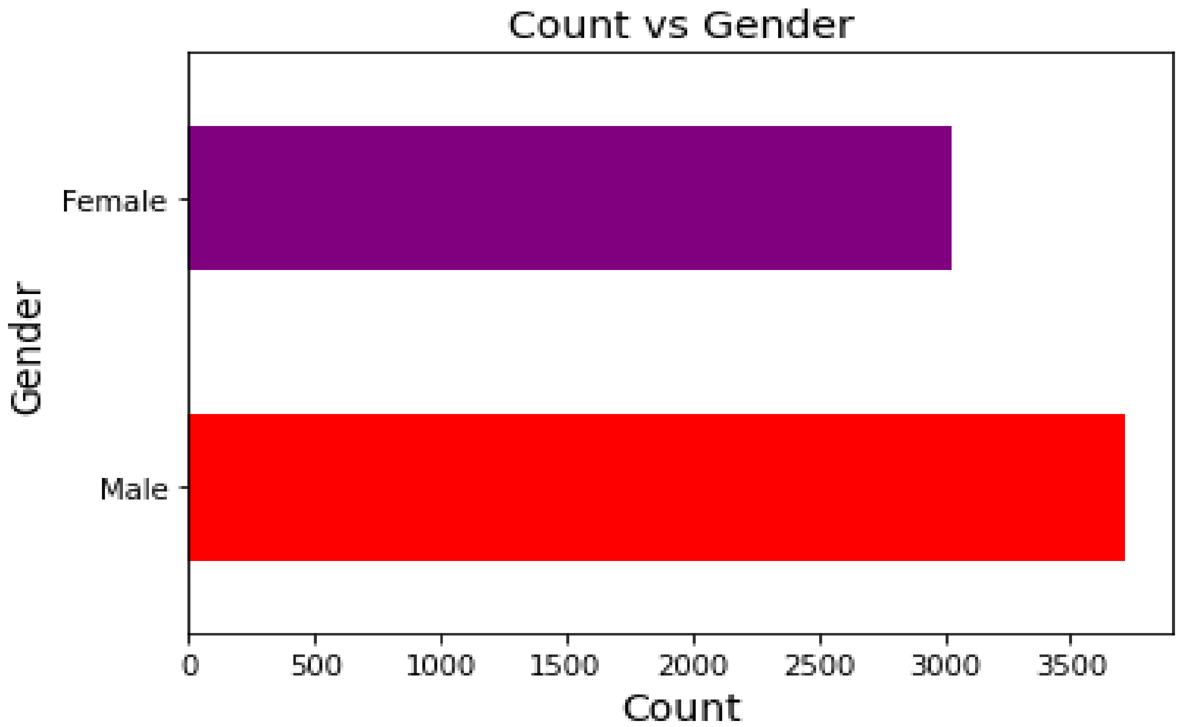


Figure h: Number of People for each gender

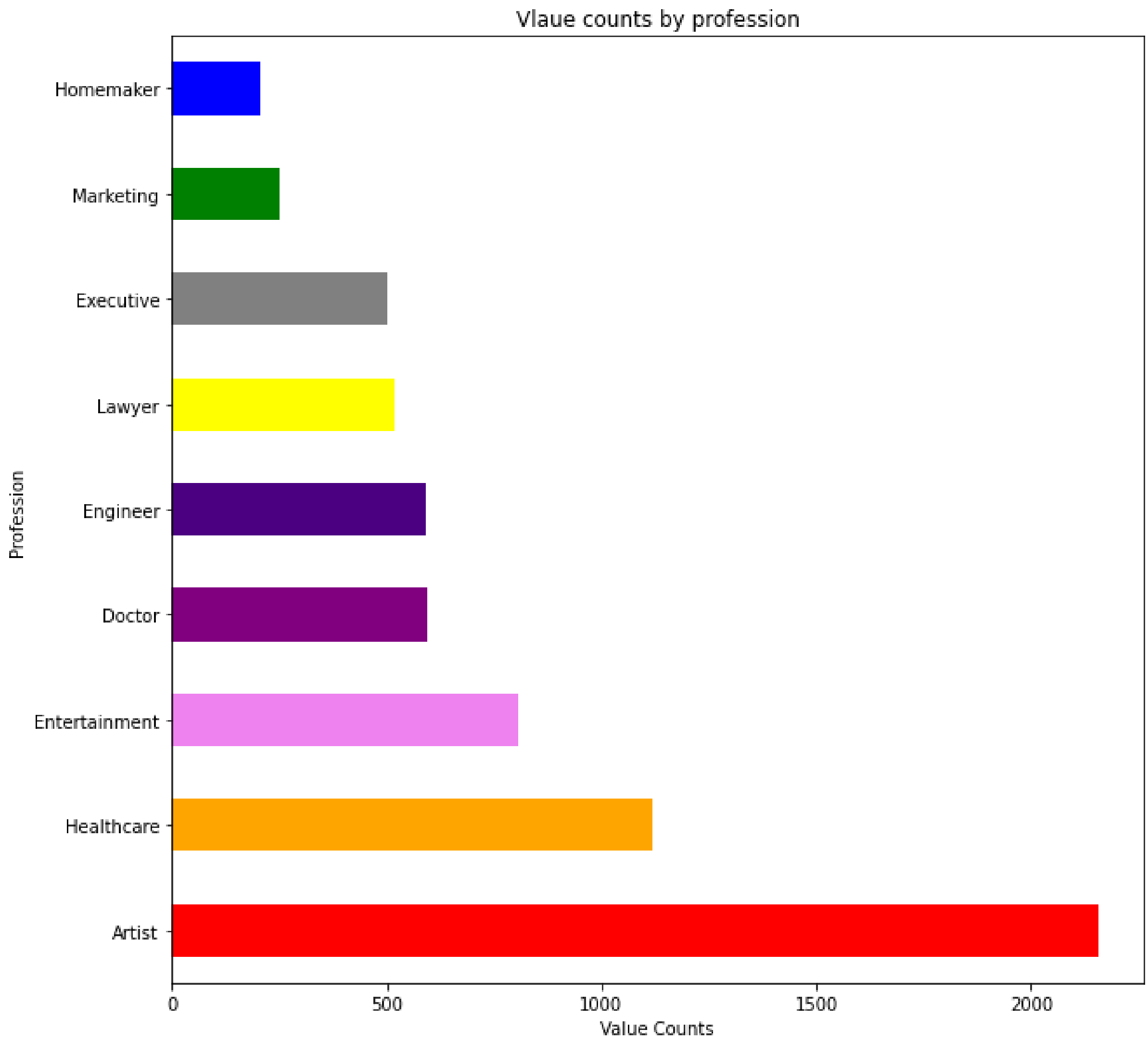


Figure I: Value Counts by profession

Additional Data Processing

The next step in the process is to convert the categorical variables using Label Encoding. For this purpose, we will Label Encoder from sk-learn pre-processing library.

4) Classification Model

a) Using Gaussian Naïve Bayes

We extract the 2 important features using permutation importance. The two important features turn out to be shop other and shop household. 2D plot and classification report is shown for the same in Figure J and Figure K

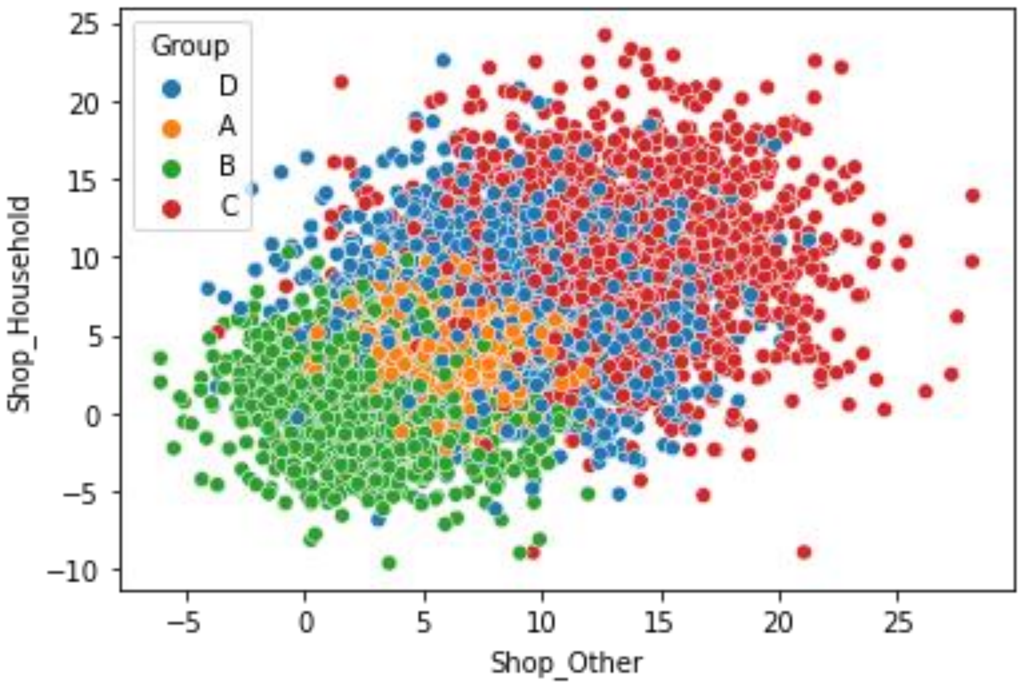


Figure J : 2-D plot

```
The predicted values are ['D' 'A' 'A' ... 'A' 'D' 'A']
The classification report is               precision    recall  f1-score   support

           A       0.76      0.60      0.67       523
           B       0.72      0.75      0.73       365
           C       0.65      0.74      0.69       359
           D       0.52      0.58      0.55       441

    accuracy                           0.66      1688
   macro avg       0.66      0.67      0.66      1688
weighted avg       0.67      0.66      0.66      1688
```

Figure K: Classification Report for Gaussian Naïve Bayes using 2 features

It is worth to mention here that we get better results when we use all features for Gaussian Naïve Bayes. When we use only 2 features based on Feature Importance, we get an accuracy of around 66 percent.

b) Decision Tree Classifier Baseline Model

When we use the Decision Tree Classifier on the Baseline Model, we obtain the following classification report as shown in Figure L.

```
the predicted values are ['D' 'B' 'A' ... 'A' 'B' 'C']
The classification report is               precision    recall  f1-score   support

           A       0.67      0.68      0.68       418
           B       0.77      0.71      0.74       399
           C       0.83      0.86      0.85       422
           D       0.73      0.74      0.73       459

    accuracy                           0.75      1698
   macro avg       0.75      0.75      0.75      1698
weighted avg       0.75      0.75      0.75      1698
```

Figure L: Decision Tree Baseline Model Classification Report

We can see that we obtain an accuracy of around 75 percent.

c) Decision Tree Model on Modified Data Set

When we use decision tree on our modified data set, we get the classification report as shown in Figure M.

```
the predicted values are ['D' 'A' 'C' ... 'C' 'B' 'B']
The classification report is               precision    recall  f1-score   support

           A       0.69      0.68      0.68       503
           B       0.74      0.72      0.73       497
           C       0.80      0.85      0.82       458
           D       0.74      0.73      0.73       567

    accuracy                           0.74      2025
   macro avg       0.74      0.74      0.74      2025
weighted avg       0.74      0.74      0.74      2025
```

Figure M: Classification Report on Modified Data Set

The feature importance graph is as shown in Figure N

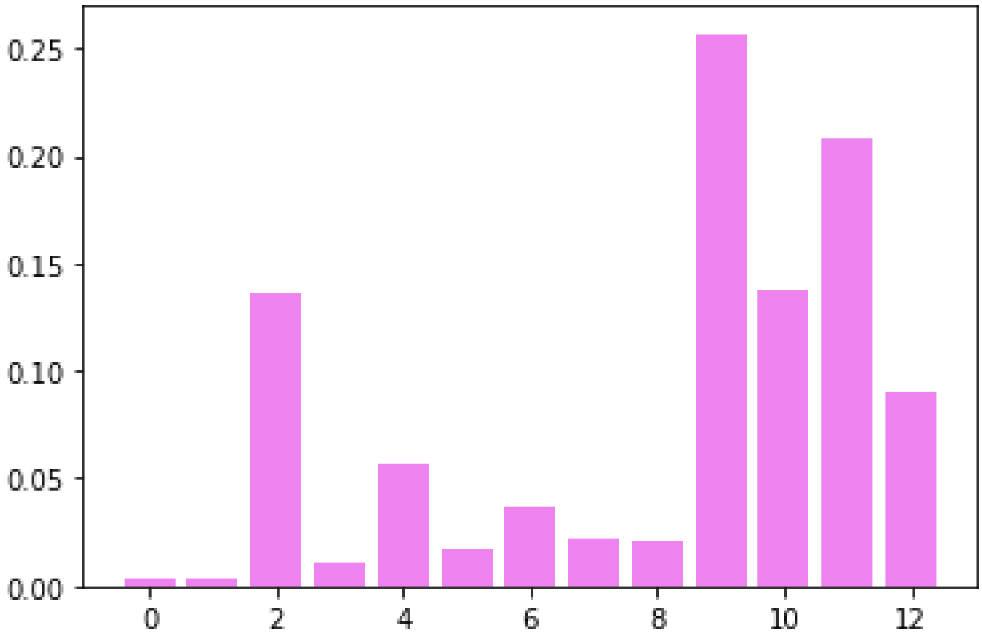


Figure N: Feature Importance graph

Summary

We were able to effectively classify the customers into groups using ML techniques. GNB classier gave satisfactory results with an accuracy of around 80 percent when all features were used. On the other hand, when only 2 features were used, its accuracy dropped to 66 percent. Moreover, the baseline model and the model on the modified data set on decision tree classifier gave similar results with around, 74 percent accuracy. Data set would have been much more effective if more instances were available for model training. Model performance could have been improved by trying out different ML algorithm with appropriate hyper parameter tuning techniques. Moreover, usage of K -cross fold validation during training might have given better results as well.