Customer Retention

Author: Ibrahim A. Shukoor

Problem Statement

Word of mouth is regarded as one of the most organic techniques of sales generation, requiring no additional cash from companies (E-tailers) for promotion. On the contrary, greater expenditures in the company's services are required to improve the website's efficiency and performance in order to gain more trust and speed up transactions. But which is more significant, hedonistic or utilitarian values? A case study of Indian e-commerce customers brought to light a list of E retail factors that may possibly play a role in customer activation and retention. The dataset contained a total of 71 columns, the first 10-11 columns contain questions aimed to extract individual user details such as age, gender, and city *etcetera*. The next 36 questions include questions which are designed to understand user's shopping habits and to assess the importance of E-retail factors for individual users. For example:

i. How frequently do you abandon (selecting an item and leaving without making payment) your shopping cart? {user shopping habit}
ii. Displaying quality Information on the website improves satisfaction of customers. {importance of e-retail factor for the user}

The remaining 24 questions were open-ended and carefully selected to extract the name of the online retailer that the user believed would be the greatest match for each question. It included questions like:

i. Easy to use website or application
ii. Reliability of the website or application
iii. Limited mode of payment on most products (promotion, sales period)

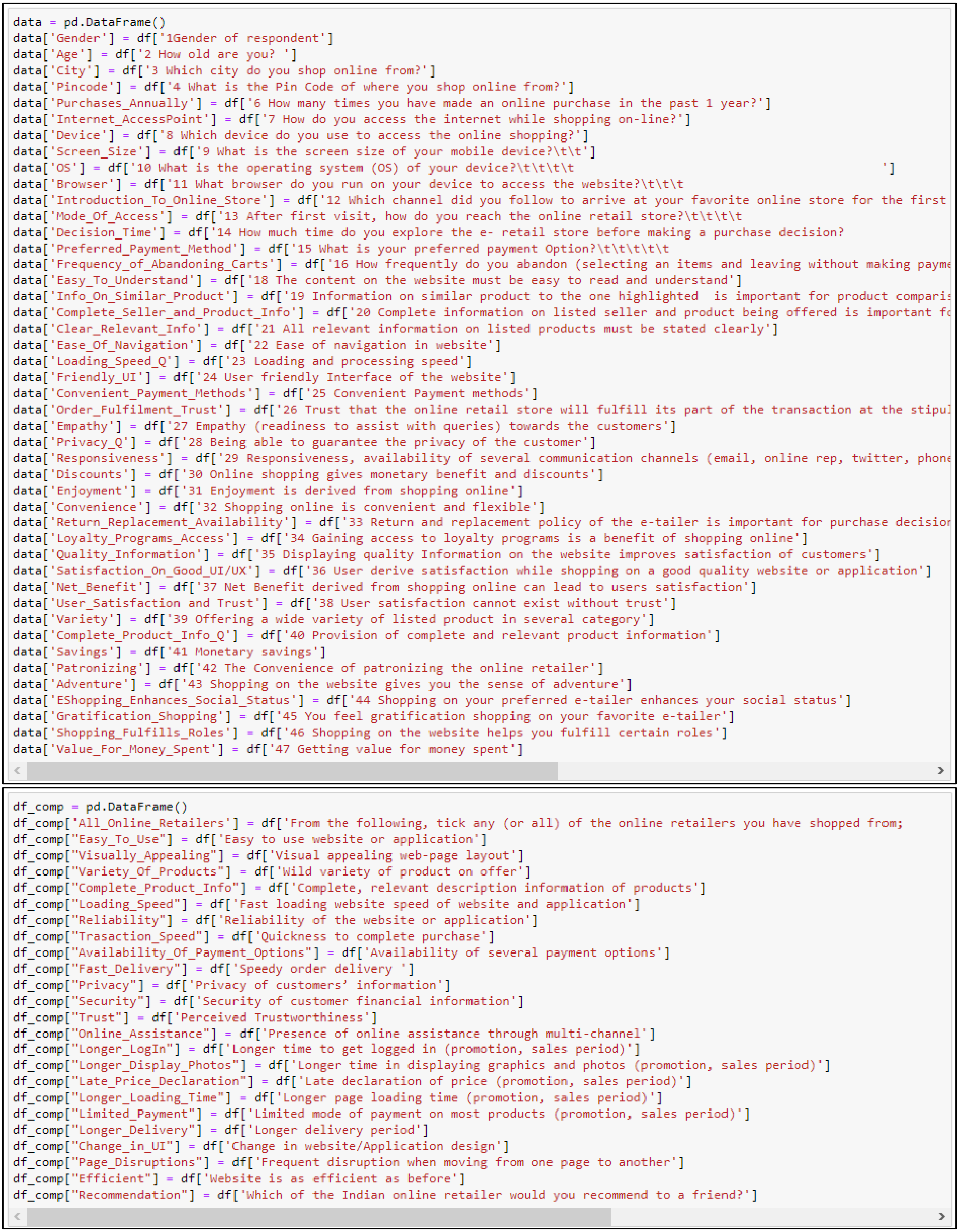
The provided dataset was subjected to data analysis, and numerous questions were addressed.

Scope of the study

The main objective was to evaluate how important hedonistic and utilitarian values were to the questioned user population. It also looked to see whether there were any gender-related differences in user opinions. Additionally, it revealed which E-tailer was believed to be more well-liked and which were considered unsatisfactory according to the users.

Exploratory Data Analysis and Data Visualization

The dataset was examined for null values before being divided into two and having the column names modified to something more succinct and clearer. The initial 47 questions were included in the first dataset, while the remaining subjective questions were included in the second dataset.



```
data = pd.DataFrame()
data['Gender'] = df['1Gender of respondent']
data['Age'] = df['2 How old are you? ']
data['City'] = df['3 Which city do you shop online from?']
data['Pincode'] = df['4 What is the Pin Code of where you shop online from?']
data['Purchases_Annually'] = df['6 How many times you have made an online purchase in the past 1 year?']
data['Internet_AccessPoint'] = df['7 How do you access the internet while shopping on-line?']
data['Device'] = df['8 Which device do you use to access the online shopping?']
data['Screen_Size'] = df['9 What is the screen size of your mobile device?\t\t']
data['OS'] = df['10 What is the operating system (OS) of your device?\t\t\t\t                                          ']
data['Browser'] = df['11 What browser do you run on your device to access the website?\t\t\t
data['Introduction_To_Online_Store'] = df['12 Which channel did you follow to arrive at your favorite online store for the first
data['Mode_Of_Access'] = df['13 After first visit, how do you reach the online retail store?\t\t\t\t
data['Decision_Time'] = df['14 How much time do you explore the e- retail store before making a purchase decision?
data['Preferred_Payment_Method'] = df['15 What is your preferred payment Option?\t\t\t\t\t
data['Frequency_of_Abandoning_Carts'] = df['16 How frequently do you abandon (selecting an items and leaving without making payme
data['Easy_To_Understand'] = df['18 The content on the website must be easy to read and understand']
data['Info_On_Similar_Product'] = df['19 Information on similar product to the one highlighted  is important for product comparis
data['Complete_Seller_and_Product_Info'] = df['20 Complete information on listed seller and product being offered is important f
data['Clear_Relevant_Info'] = df['21 All relevant information on listed products must be stated clearly']
data['Ease_Of_Navigation'] = df['22 Ease of navigation in website']
data['Loading_Speed_Q'] = df['23 Loading and processing speed']
data['Friendly_UI'] = df['24 User friendly Interface of the website']
data['Convenient_Payment_Methods'] = df['25 Convenient Payment methods']
data['Order_Fulfilment_Trust'] = df['26 Trust that the online retail store will fulfill its part of the transaction at the stipul
data['Empathy'] = df['27 Empathy (readiness to assist with queries) towards the customers']
data['Privacy_Q'] = df['28 Being able to guarantee the privacy of the customer']
data['Responsiveness'] = df['29 Responsiveness, availability of several communication channels (email, online rep, twitter, phone
data['Discounts'] = df['30 Online shopping gives monetary benefit and discounts']
data['Enjoyment'] = df['31 Enjoyment is derived from shopping online']
data['Convenience'] = df['32 Shopping online is convenient and flexible']
data['Return_Replacement_Availability'] = df['33 Return and replacement policy of the e-tailer is important for purchase decisio
data['Loyalty_Programs_Access'] = df['34 Gaining access to loyalty programs is a benefit of shopping online']
data['Quality_Information'] = df['35 Displaying quality Information on the website improves satisfaction of customers']
data['Satisfaction_On_Good_UI/UX'] = df['36 User derive satisfaction while shopping on a good quality website or application']
data['Net_Benefit'] = df['37 Net Benefit derived from shopping online can lead to users satisfaction']
data['User_Satisfaction and Trust'] = df['38 User satisfaction cannot exist without trust']
data['Variety'] = df['39 Offering a wide variety of listed product in several category']
data['Complete_Product_Info_Q'] = df['40 Provision of complete and relevant product information']
data['Savings'] = df['41 Monetary savings']
data['Patronizing'] = df['42 The Convenience of patronizing the online retailer']
data['Adventure'] = df['43 Shopping on the website gives you the sense of adventure']
data['EShopping_Enhances_Social_Status'] = df['44 Shopping on your preferred e-tailer enhances your social status']
data['Gratification_Shopping'] = df['45 You feel gratification shopping on your favorite e-tailer']
data['Shopping_Fulfills_Roles'] = df['46 Shopping on the website helps you fulfill certain roles']
data['Value_For_Money_Spent'] = df['47 Getting value for money spent']
```

```
df_comp = pd.DataFrame()
df_comp['All_Online_Retailers'] = df['From the following, tick any (or all) of the online retailers you have shopped from;
df_comp["Easy_To_Use"] = df['Easy to use website or application']
df_comp["Visually_Appealing"] = df['Visual appealing web-page layout']
df_comp["Variety_Of_Products"] = df['Wild variety of product on offer']
df_comp["Complete_Product_Info"] = df['Complete, relevant description information of products']
df_comp["Loading_Speed"] = df['Fast loading website speed of website and application']
df_comp["Reliability"] = df['Reliability of the website or application']
df_comp["Trasaction_Speed"] = df['Quickness to complete purchase']
df_comp["Availability_Of_Payment_Options"] = df['Availability of several payment options']
df_comp["Fast_Delivery"] = df['Speedy order delivery ']
df_comp["Privacy"] = df['Privacy of customers’ information']
df_comp["Security"] = df['Security of customer financial information']
df_comp["Trust"] = df['Perceived Trustworthiness']
df_comp["Online_Assistance"] = df['Presence of online assistance through multi-channel']
df_comp["Longer_LogIn"] = df['Longer time to get logged in (promotion, sales period)']
df_comp["Longer_Display_Photos"] = df['Longer time in displaying graphics and photos (promotion, sales period)']
df_comp["Late_Price_Declaration"] = df['Late declaration of price (promotion, sales period)']
df_comp["Longer_Loading_Time"] = df['Longer page loading time (promotion, sales period)']
df_comp["Limited_Payment"] = df['Limited mode of payment on most products (promotion, sales period)']
df_comp["Longer_Delivery"] = df['Longer delivery period']
df_comp["Change_in_UI"] = df['Change in website/Application design']
df_comp["Page_Disruptions"] = df['Frequent disruption when moving from one page to another']
df_comp["Efficient"] = df['Website is as efficient as before']
df_comp["Recommendation"] = df['Which of the Indian online retailer would you recommend to a friend?']
```

Figure 1: Renaming columns

Analysis

1. Hedonistic Values

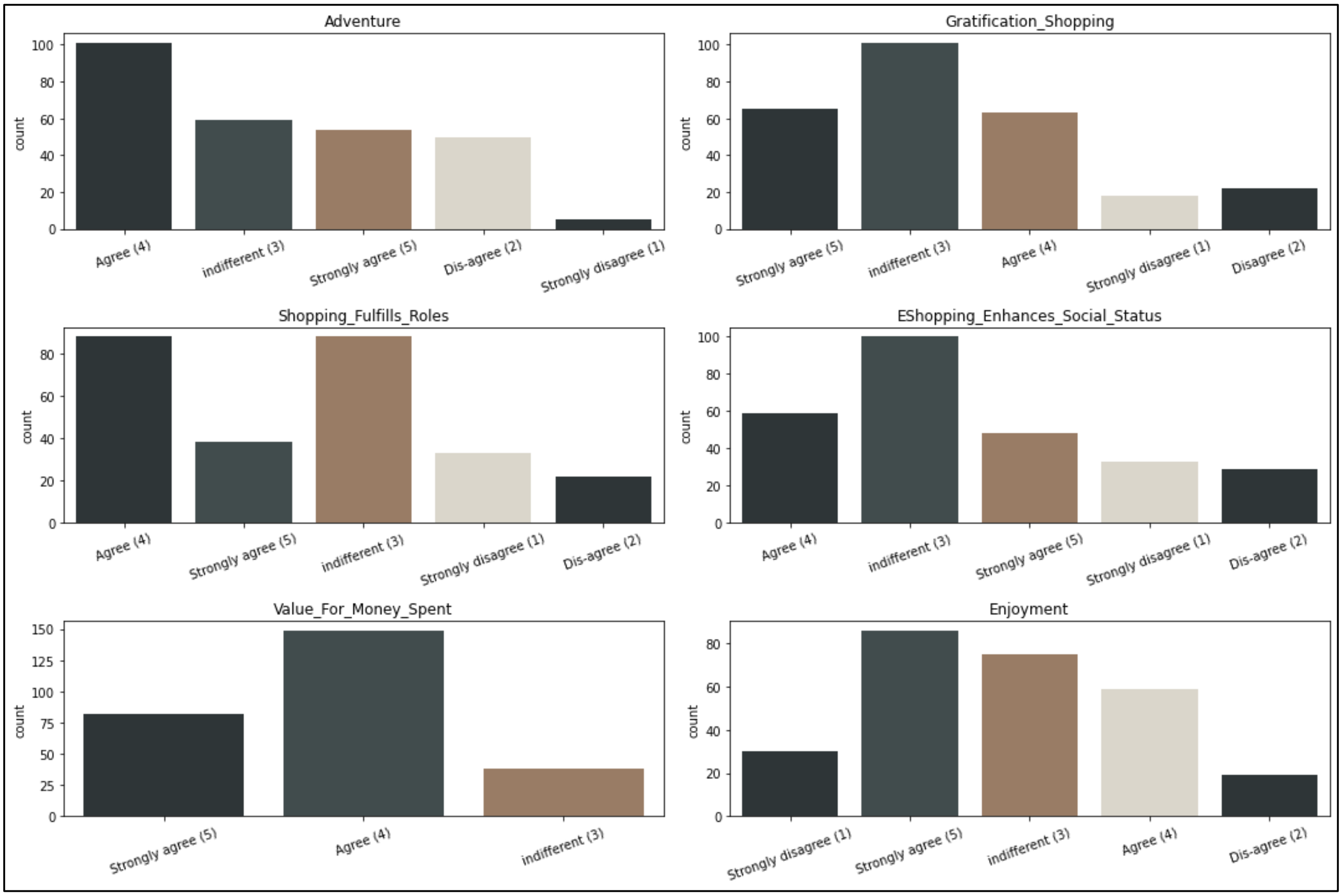


Figure 2: Importance of Hedonistic values among users

Customers appear to agree wholeheartedly that elements like a sense of adventure, receiving value for the money invested, and enjoyment play a role in their purchasing decision. However, they find that factors including sense of gratification, fulfilling roles, and enhancing social status are irrelevant. We can examine this case in more detail to identify differences in opinion in terms of gender.

2. Hedonistic Values and Gender

The hedonistic values vary depending on the gender. Most often, this conflict/difference can be seen in:

- In the case of gratification, most females opt for 'indifferent' and most guys choose to 'agree'.
- Shopping Fulfils Roles: The majority of women choose 'indifferent', and the majority of men 'agree' that shopping helps them fulfil certain roles.
- E Shopping Enhances Social Status: Both the genders opt for 'indifferent'
- When it comes to enjoyment, most Females chose 'strongly agree', whereas the majority of males selected 'indifferent'.

We can safely say that Hedonistic values are not constant in terms of gender, and they vary quite a lot.

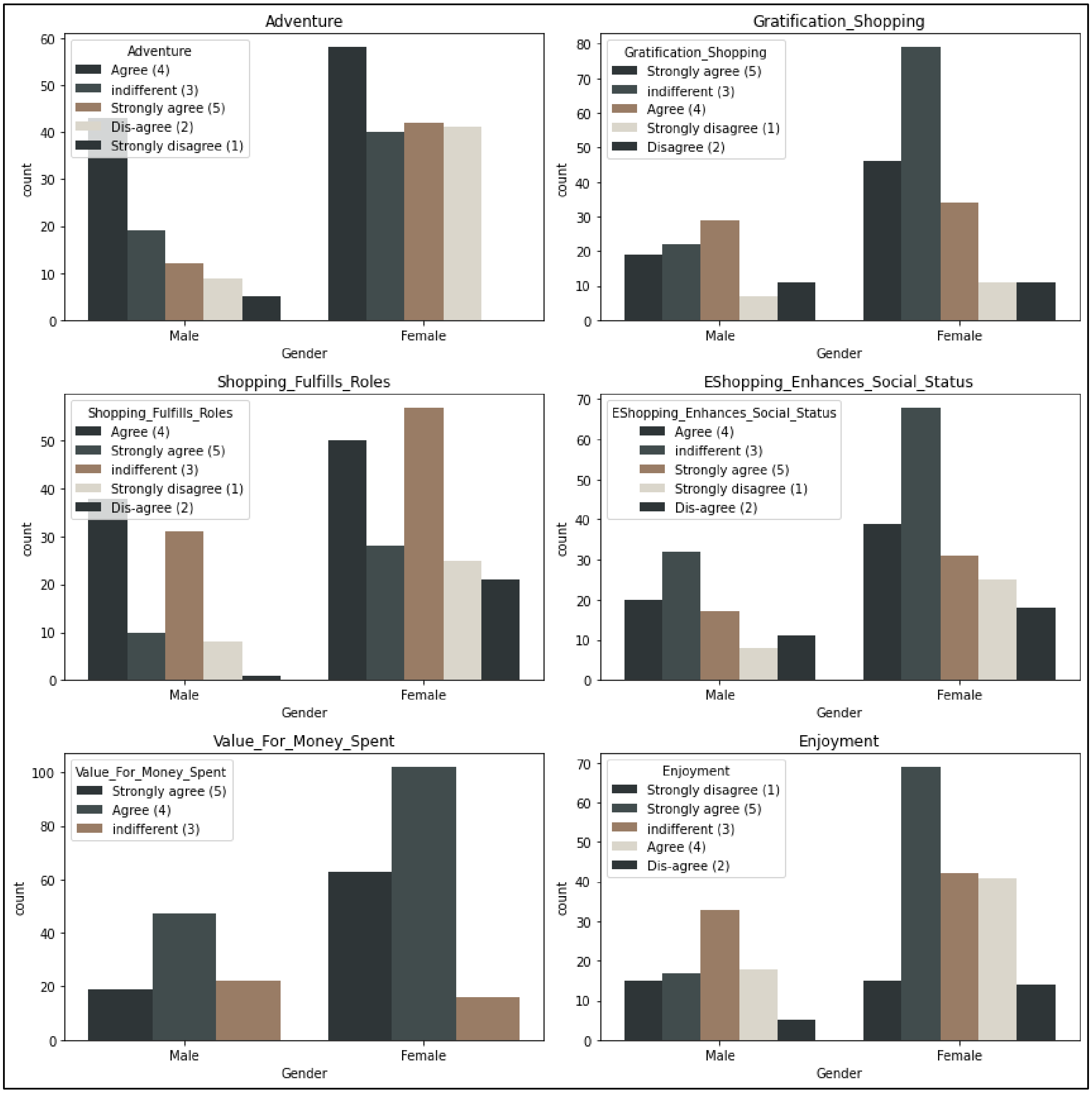


Figure 3: Hedonistic values with respect to Gender

3. Utilitarian Values

From the illustration below (Figure 4), it is clear that variables like complete product information, monetary savings, discounts, value for money spent, information on similar products, and convenience are do play a role in users' purchase decisions. Thus, we can conclude that utilitarian principles are essential.

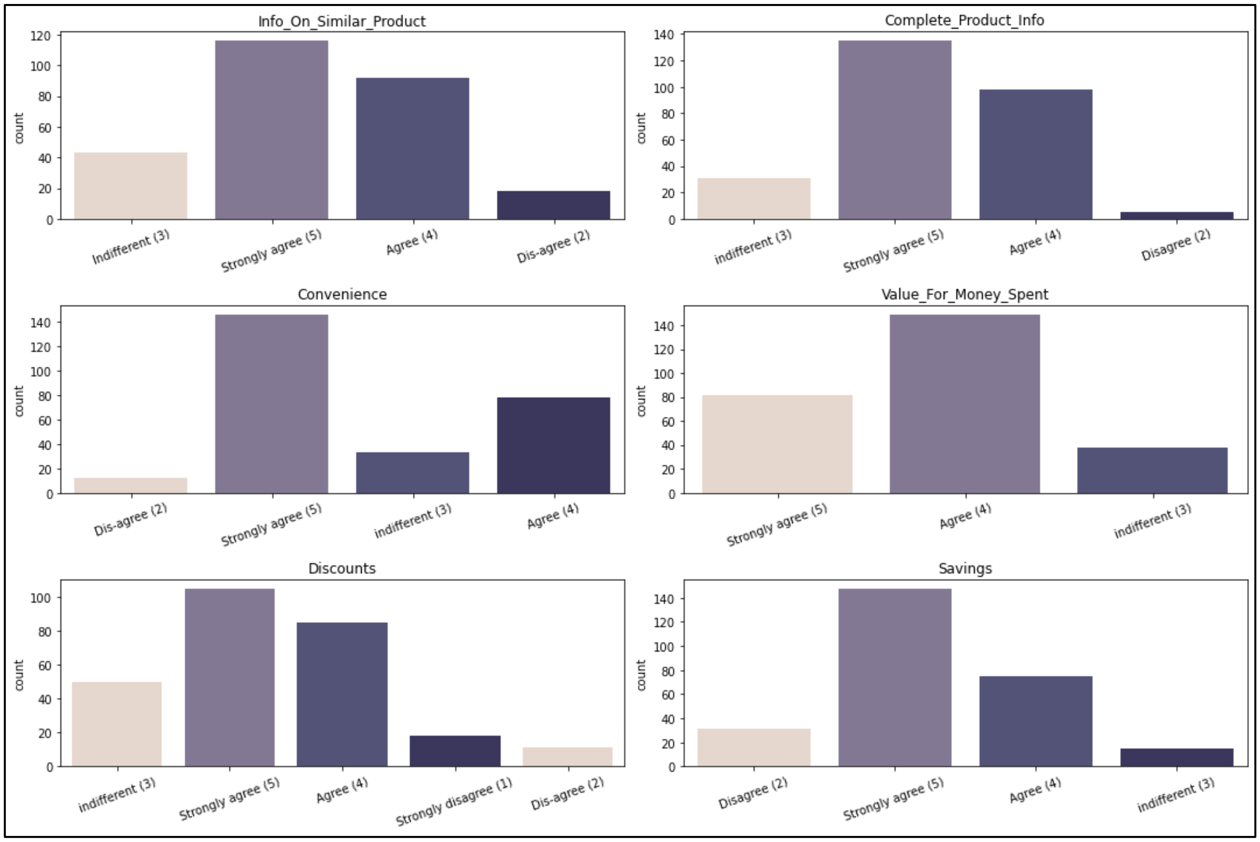


Figure 4: Importance of Utilitarian values among users

4. Utilitarian Values and Gender

There appears to be no distinction between the genders in terms of utilitarian values. The following elements—complete product information, financial savings, discounts, value for money spent,

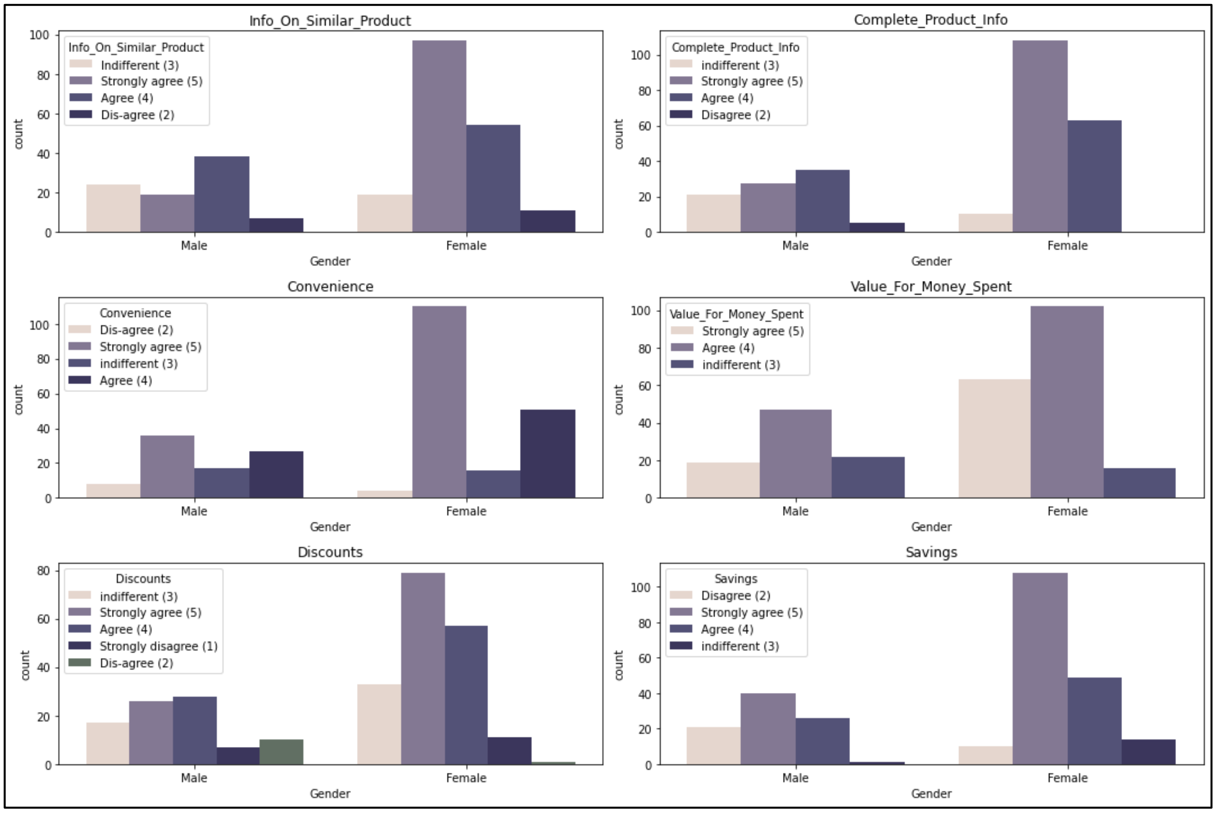


Figure 5: Utilitarian values and with respect to Gender

information on similar products, and convenience—appears to be deemed important by both the genders.

5. Service Quality

Empathy, privacy, responsiveness, availability of returns and replacements, quality information, ease of navigation, responsiveness, friendliness of the user interface, order fulfilment trust, and satisfaction with good UI/UX are all service qualities that are critical to boosting sales and boosting customer loyalty, as can be seen in the figure below. (Figure 6A, 6B)

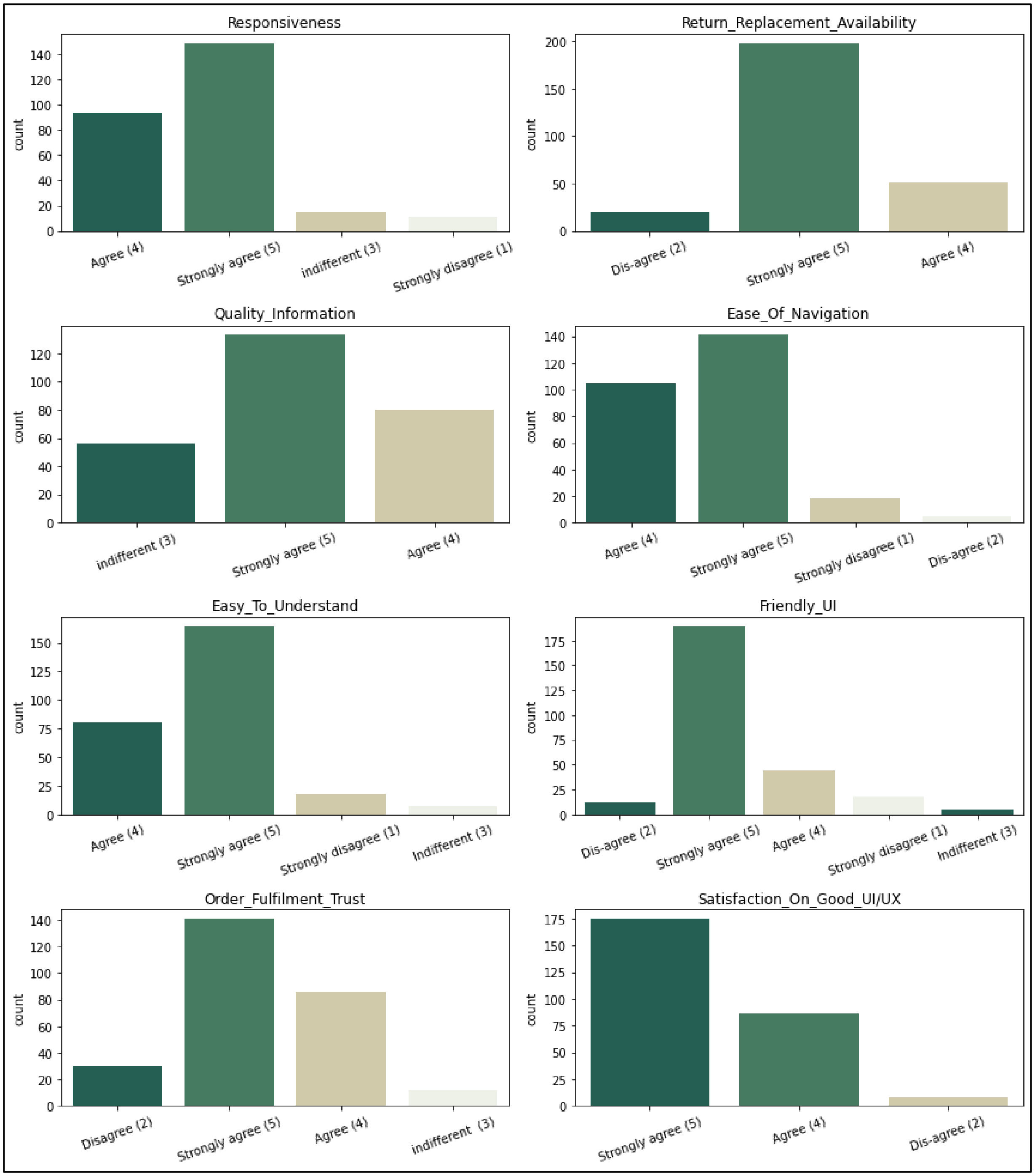


Figure 6A: Importance of Service Quality among the users

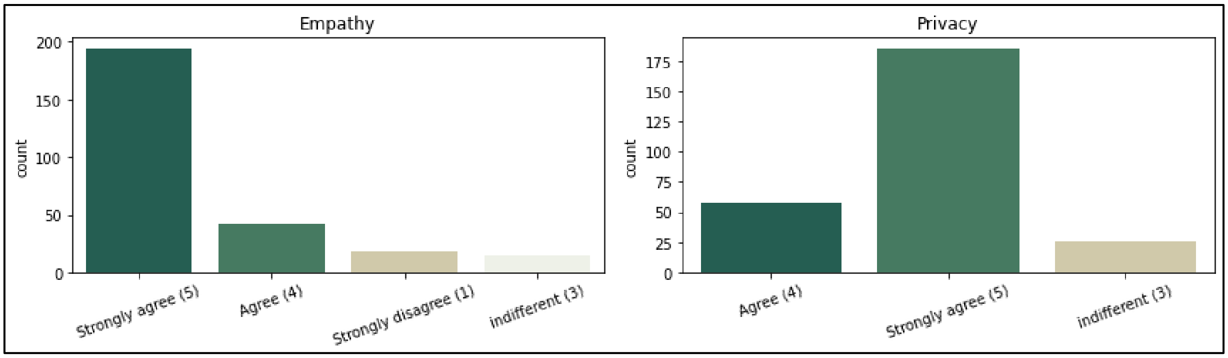
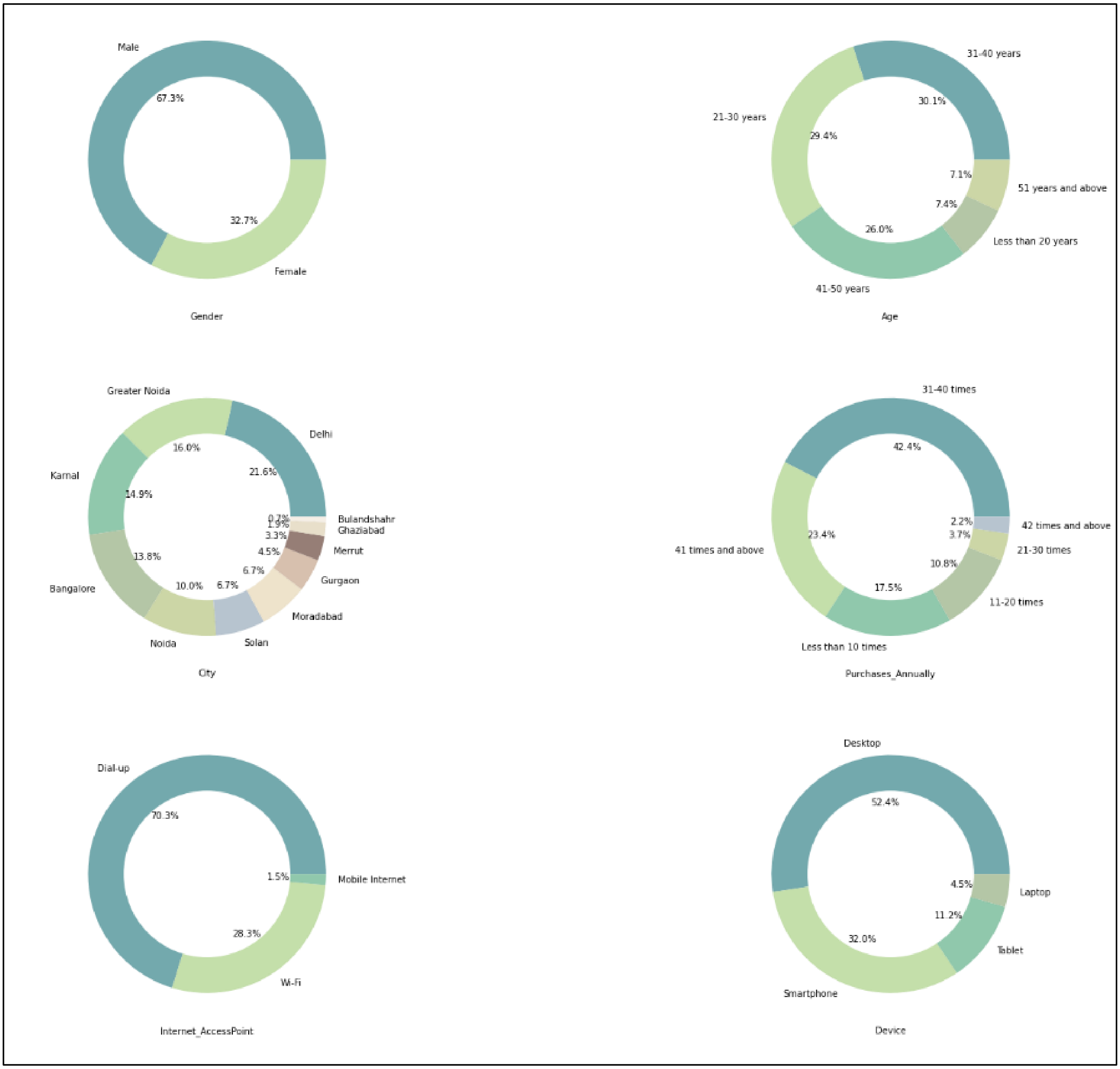


Figure 6B: Importance of Service Quality among the users

6. Population Characteristics

Analysis was done to understand the population characteristics (Figure 7)



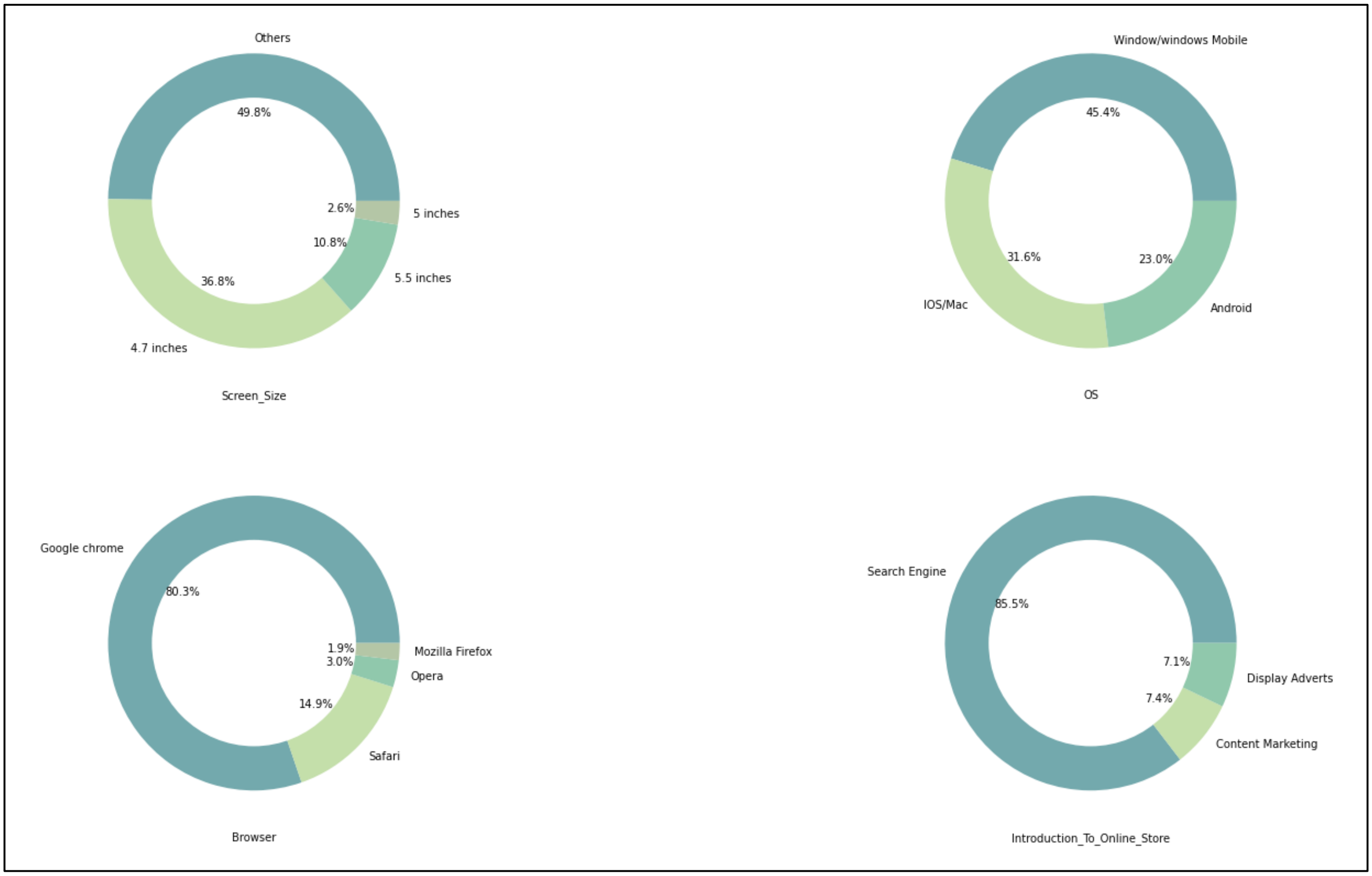
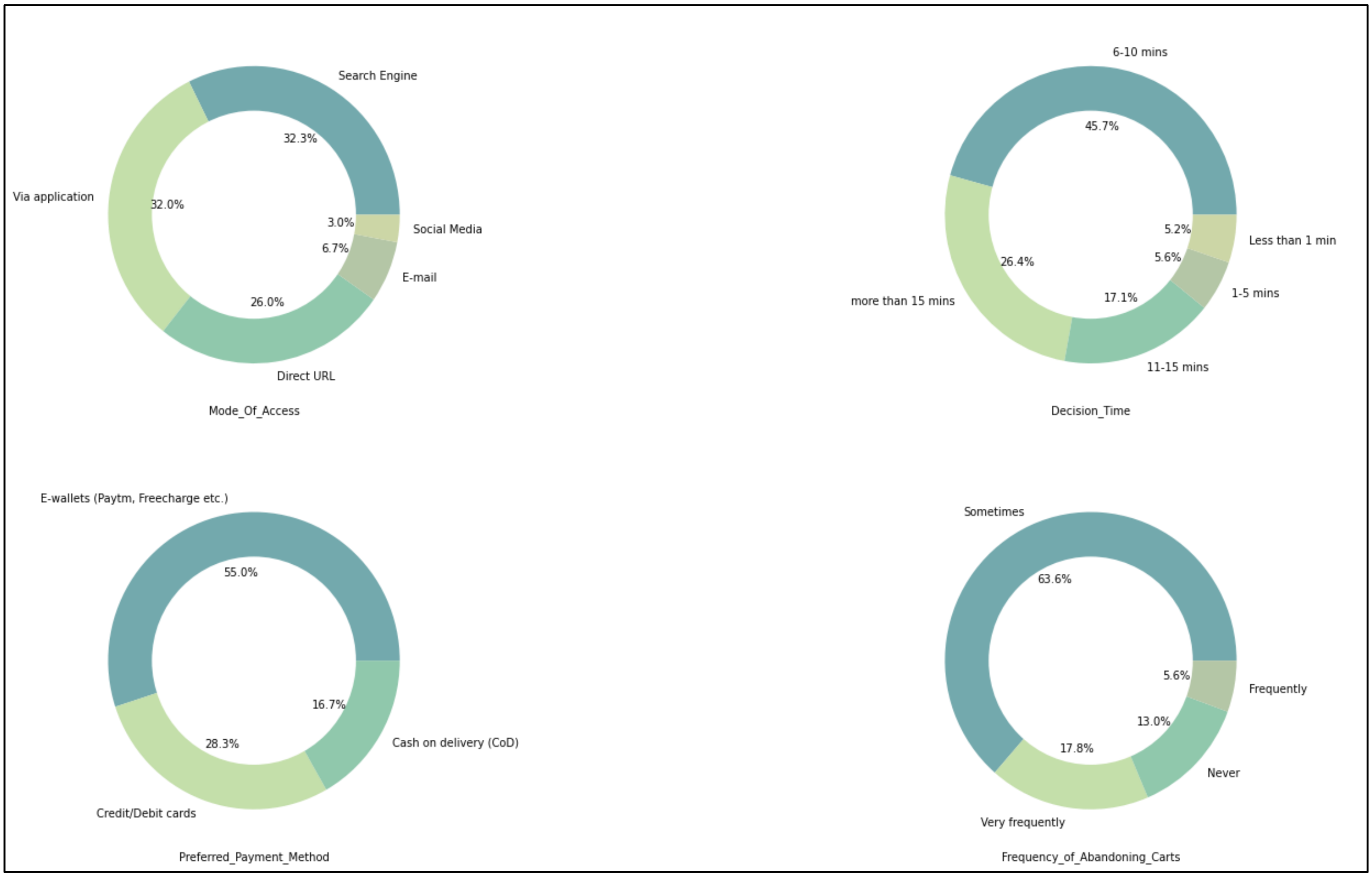


Figure 7: Population characteristics



7. Additional Analysis

In order to obtain more information, further analysis was done. It was discovered throughout this analysis that

a. The purchase decision-making process varied by gender, with the majority of males requiring less than one minute and the majority of females taking more than 15 minutes to make a purchase decision. (Figure 8)

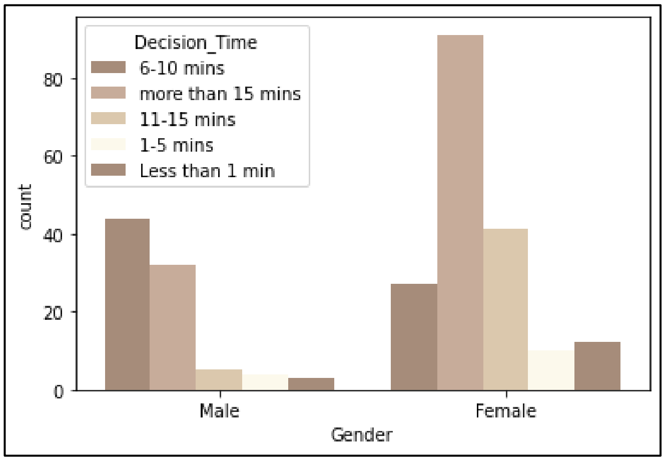


Figure 8: Decision time with regards to Gender

b. The best decision-making window was between 1–5 and 6–10 minutes. Users who made purchases during this window had greater annual purchases. Users who took longer than 15 minutes to complete a transaction often made less than 10 purchases per year.

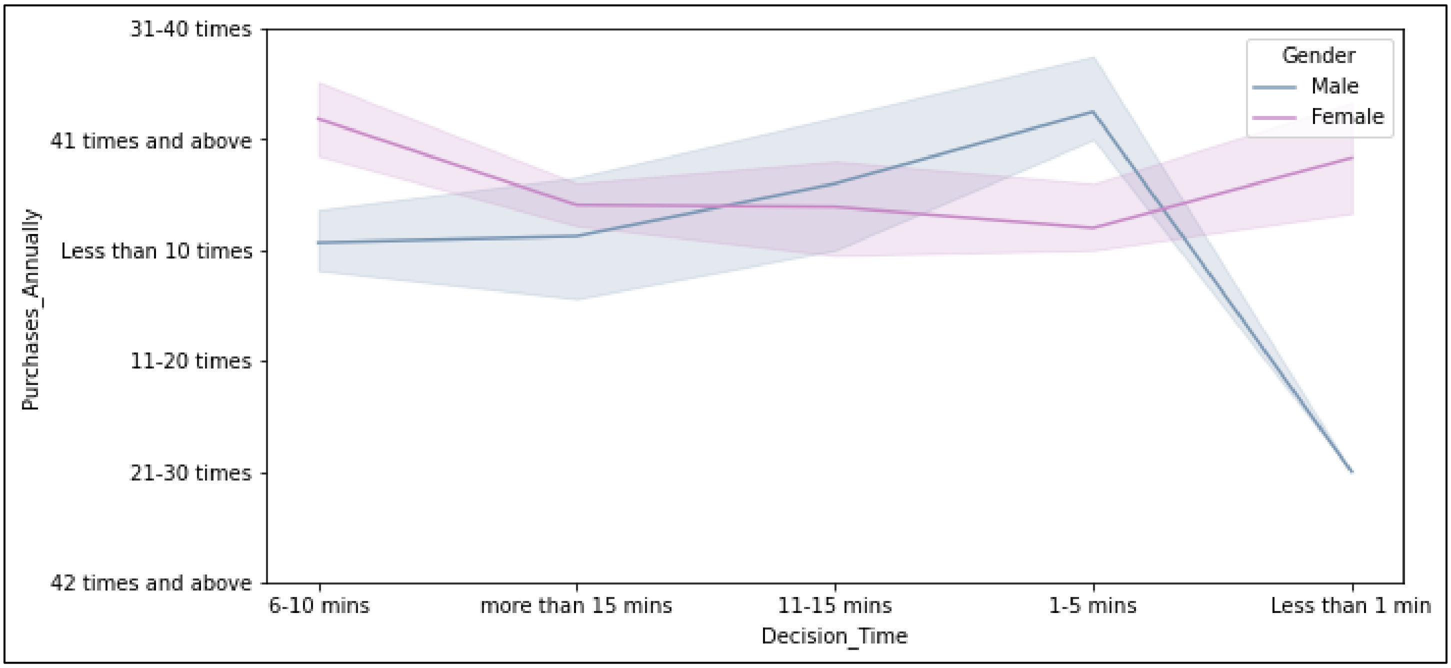


Figure 9: Decision time vs Purchases Annually

c. Additionally, we also saw that Credit/Debit cards and Window/windows Mobile are popular amongst Females and E-wallets (Paytm, Freecharge etc.) and IOS/Mac are more popular amongst Males.

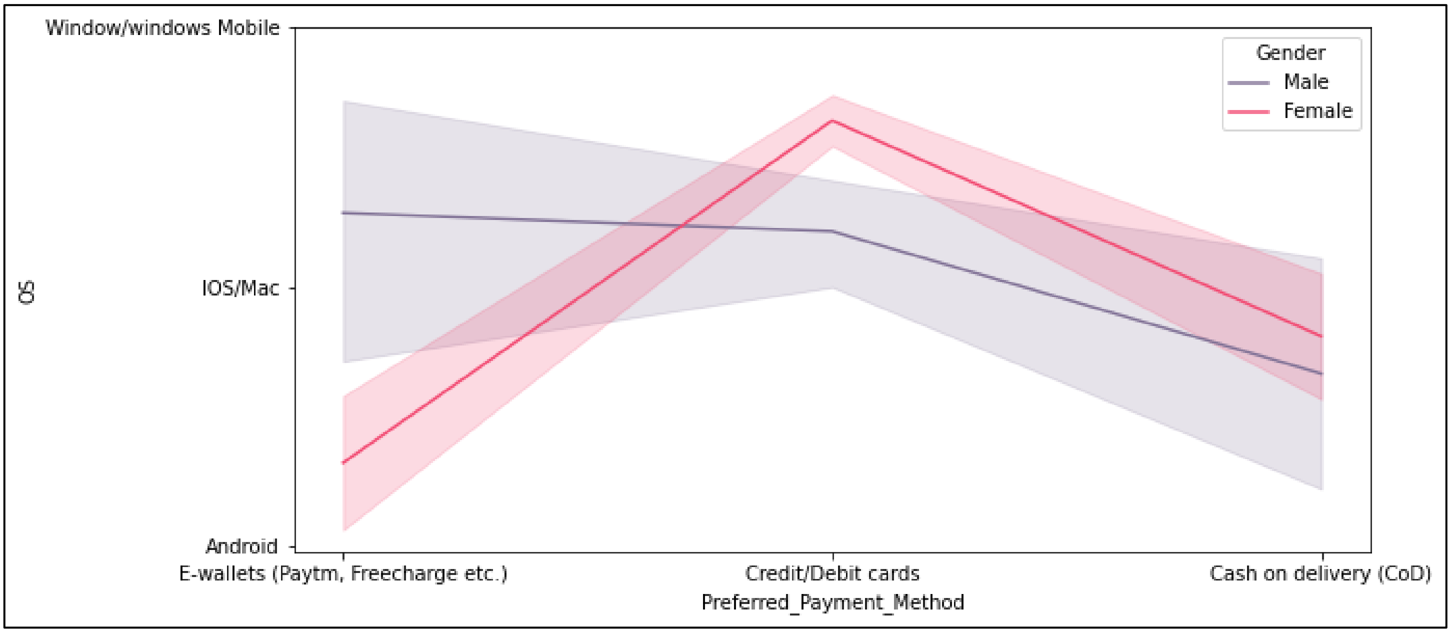


Figure 10: Preferred Payment Method vs

d. The most popular devices for making purchases are smartphones, followed by laptops and then desktops.

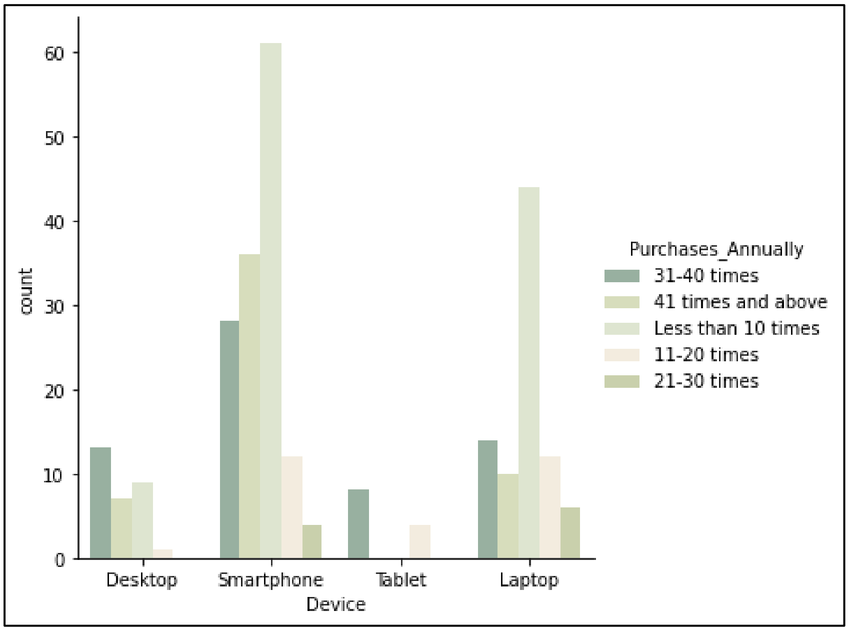


Figure 11: Device with regards to Purchases Annually

8. Best and Least Satisfactory E-tailers (according to the users)

The following were the factors considered while assessing the best online store according to the users:

Online Retailers Used, Visually Appealing, Easy to Use, Variety of Products, Complete Product Info, Loading Speed, Reliability, Transaction Speed, Availability of Payment Options, Fast Delivery, Privacy, Security, Trust, Online Assistance, Efficient and Recommendation

Rank	E-tailer Name	# of times mentioned
1	Amazon	3860
2	Flipkart	2774
3	Myntra	1436
4	Paytm	1208
5	Snapdeal	1128

Table 1: Best online store assessment

The following were the factors considered while assessing the least satisfactory according to the users:

Longer Log In, Longer time to display photos, Late price declaration, longer loading time, limited payment options, longer delivery, changes in user interface and frequent page disruptions.

Rank	E-tailer Name	# of times mentioned
1	Amazon	745
2	Snapdeal	551
3	Flipkart	539
4	Paytm	524
5	Myntra	388

Table 2: Least satisfactory online store assessment

As seen above (Table 1 and Table 2) customers appreciated Amazon in terms of its system, trust, reliability, and utilitarian values. Because Amazon excels in these areas, users appear to downplay the unpleasant aspects, such as slower loading times, fewer payment options, etc.

EDA Conclusion and Remarks:

We investigated the hedonistic values and discovered that the user base appears to be divided between competing ideals. It is more difficult for an E tailer to satisfy the hedonistic desires of the user community because males and females tend to have different hedonistic values.

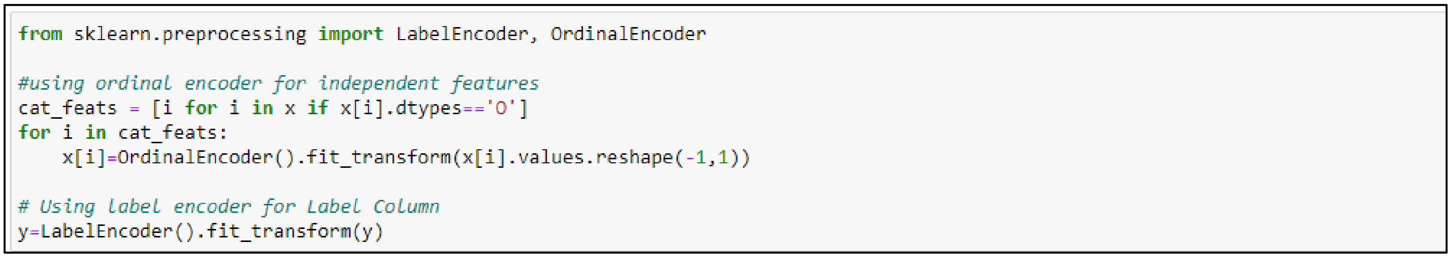
The utilitarian values analysis revealed that both genders reach a general understanding (i.e., similar values), making it simpler for an E tailer to concentrate on enhancing its utilities, system qualities, and service qualities. In this way, consumer loyalty can be maximized.

Given that 80% of people use smartphones and laptops, online retailers can adapt their advertisements to be focused on these devices to attract more customers. Additionally, since 55% of the user base uses e-wallets to finish their purchases, they can offer more deals on e-wallets.

Moreover, the analysis revealed that shorter decision times—typically under 10 minutes—significantly boost the number of transactions a user makes, online retailers can therefore work to reduce the time customers take to make decisions. This can be accomplished through speeding up loading times, log-in times, price disclosures, and UI improvements.

Model Building:

The number of unique values was reduced through feature engineering, and the column headers were made brief and concise. This helped in the understanding and visualizing the columns. Once the data was divided into x (independent variables) and y (target variable). The data was encoded using Ordinal Encoder and Label Encoder from the sklearn library.



```
from sklearn.preprocessing import LabelEncoder, OrdinalEncoder

#using ordinal encoder for independent features
cat_feats = [i for i in x if x[i].dtypes=='O']
for i in cat_feats:
    x[i]=OrdinalEncoder().fit_transform(x[i].values.reshape(-1,1))

# Using label encoder for Label Column
y=LabelEncoder().fit_transform(y)
```

*Figure 12: Encoding **x** and **y***

The values were plotted using "boxplots" and "kdeplots" to visually identify outliers and column skewness.

i. Outliers: The outliers were treated using Z-score because IQR resulted in huge amounts of data loss.
ii. Skewness: Outliers were removed, and the data was normalized using Min Max Scaler.

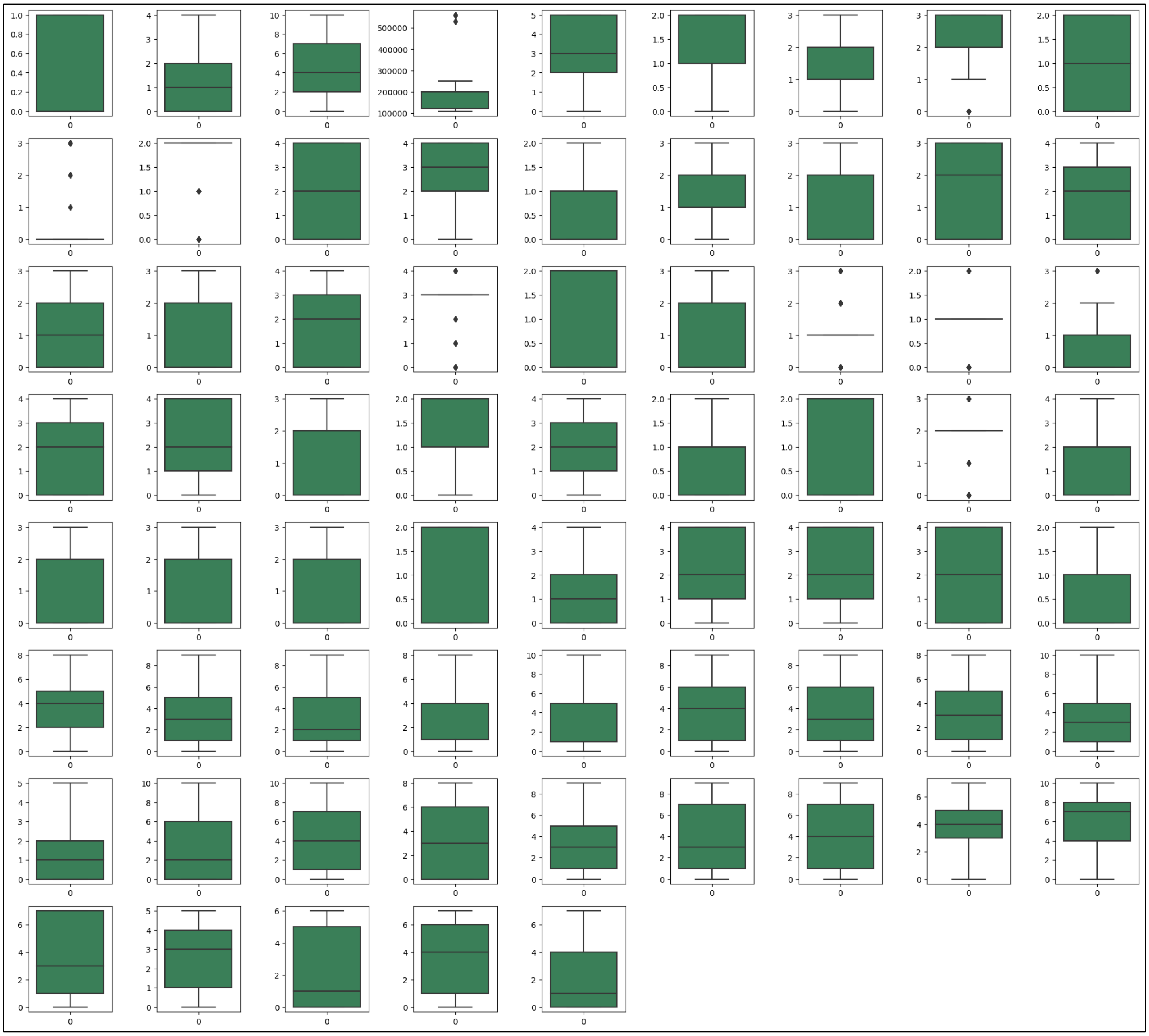
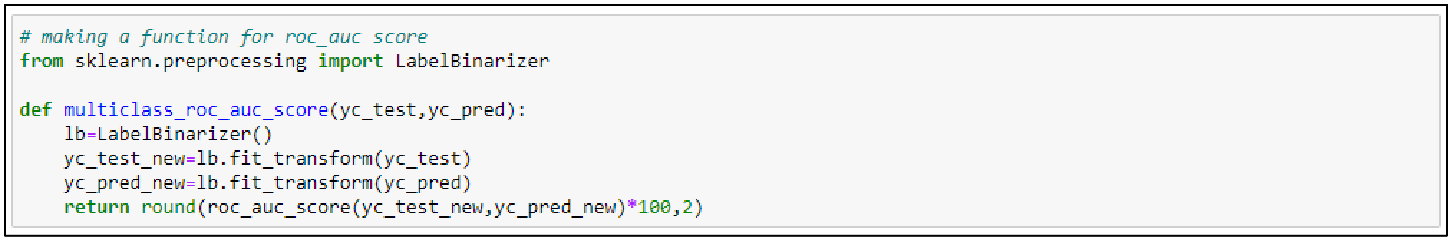


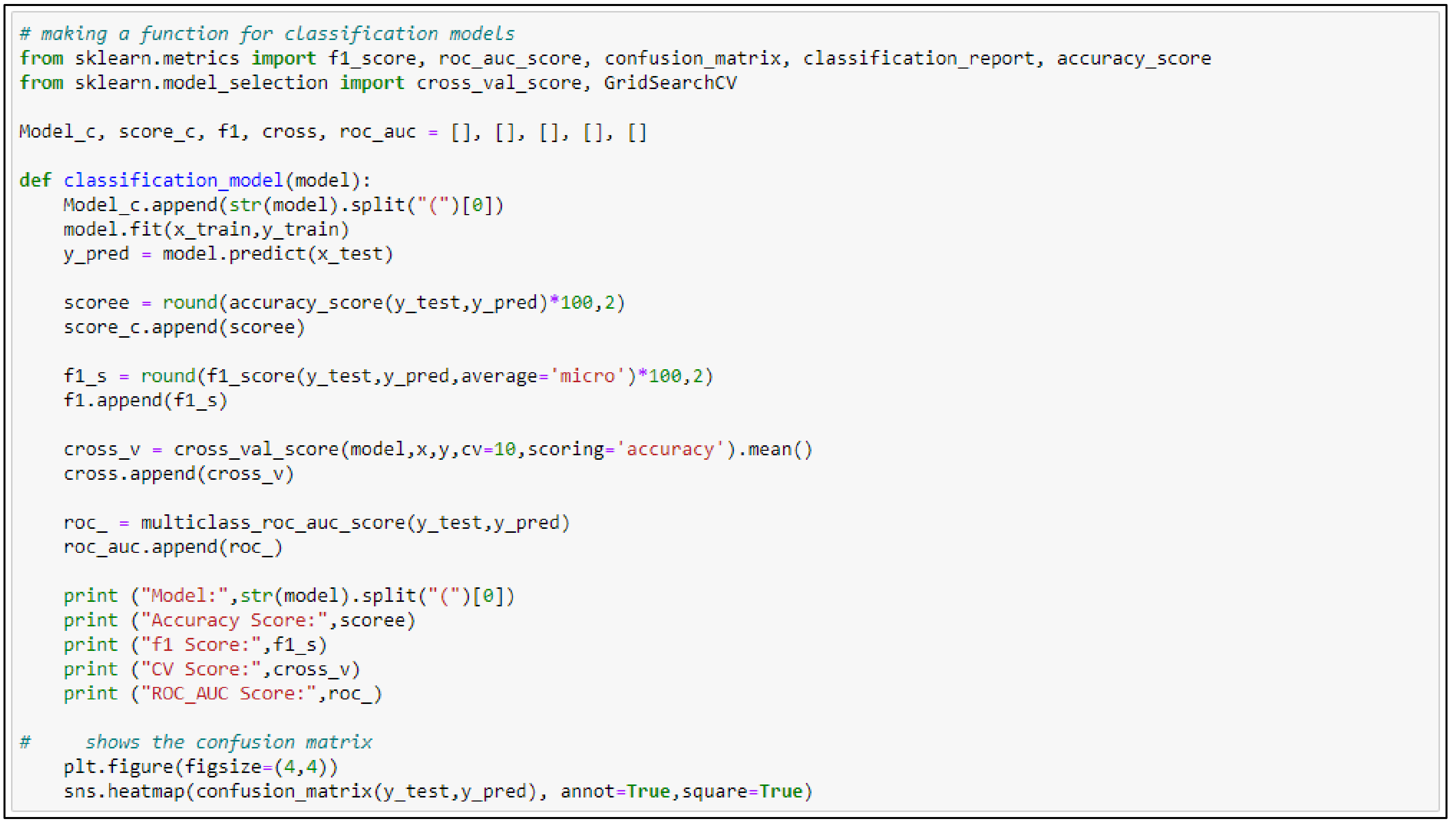
Figure 13(A): outlier visualization

Skewness and multicollinearity were handled after the outliers were removed. To balance the target variable, SMOTE over sampling was used. Following that, different classification models were applied to the dataset in order to compare and select the best model.



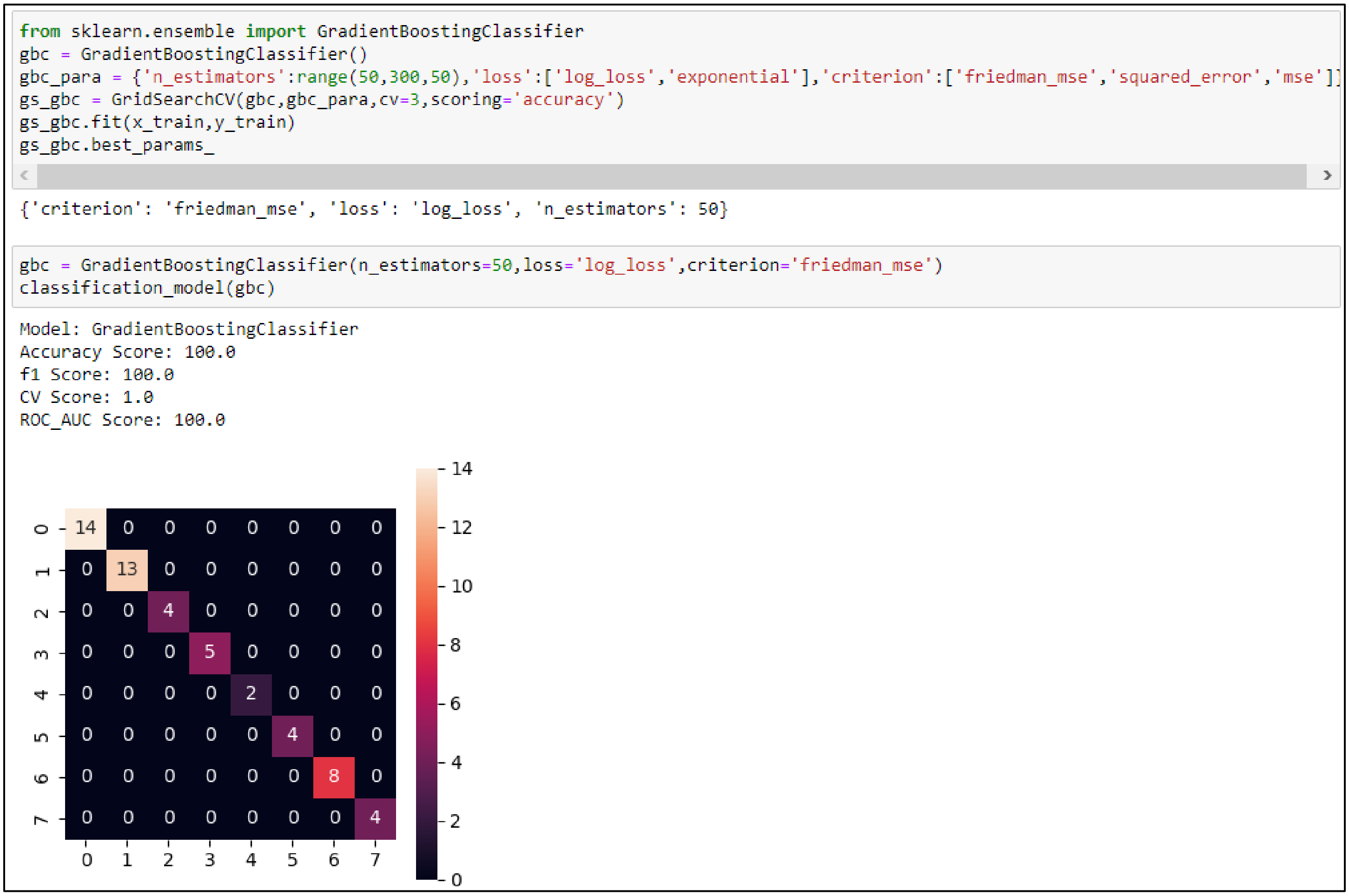
```python
# making a function for roc_auc score
from sklearn.preprocessing import LabelBinarizer

def multiclass_roc_auc_score(yc_test,yc_pred):
    lb=LabelBinarizer()
    yc_test_new=lb.fit_transform(yc_test)
    yc_pred_new=lb.fit_transform(yc_pred)
    return round(roc_auc_score(yc_test_new,yc_pred_new)*100,2)
```

Figure 14: user-defined function for obtaining the ROC-AUC score



```python
# making a function for classification models
from sklearn.metrics import f1_score, roc_auc_score, confusion_matrix, classification_report, accuracy_score
from sklearn.model_selection import cross_val_score, GridSearchCV

Model_c, score_c, f1, cross, roc_auc = [], [], [], [], []

def classification_model(model):
    Model_c.append(str(model).split("(")[0])
    model.fit(x_train,y_train)
    y_pred = model.predict(x_test)

    scoree = round(accuracy_score(y_test,y_pred)*100,2)
    score_c.append(scoree)

    f1_s = round(f1_score(y_test,y_pred,average='micro')*100,2)
    f1.append(f1_s)

    cross_v = cross_val_score(model,x,y,cv=10,scoring='accuracy').mean()
    cross.append(cross_v)

    roc_ = multiclass_roc_auc_score(y_test,y_pred)
    roc_auc.append(roc_)

    print ("Model:",str(model).split("(")[0])
    print ("Accuracy Score:",scoree)
    print ("f1 Score:",f1_s)
    print ("CV Score:",cross_v)
    print ("ROC_AUC Score:",roc_)

#     shows the confusion matrix
    plt.figure(figsize=(4,4))
    sns.heatmap(confusion_matrix(y_test,y_pred), annot=True,square=True)
```

Figure 15: user-defined function for obtaining the performance metrics of the models



```python
from sklearn.ensemble import GradientBoostingClassifier
gbc = GradientBoostingClassifier()
gbc_para = {'n_estimators':range(50,300,50),'loss':['log_loss','exponential'],'criterion':['friedman_mse','squared_error','mse']]
gs_gbc = GridSearchCV(gbc,gbc_para,cv=3,scoring='accuracy')
gs_gbc.fit(x_train,y_train)
gs_gbc.best_params_
```

```
{'criterion': 'friedman_mse', 'loss': 'log_loss', 'n_estimators': 50}
```

```python
gbc = GradientBoostingClassifier(n_estimators=50,loss='log_loss',criterion='friedman_mse')
classification_model(gbc)
```

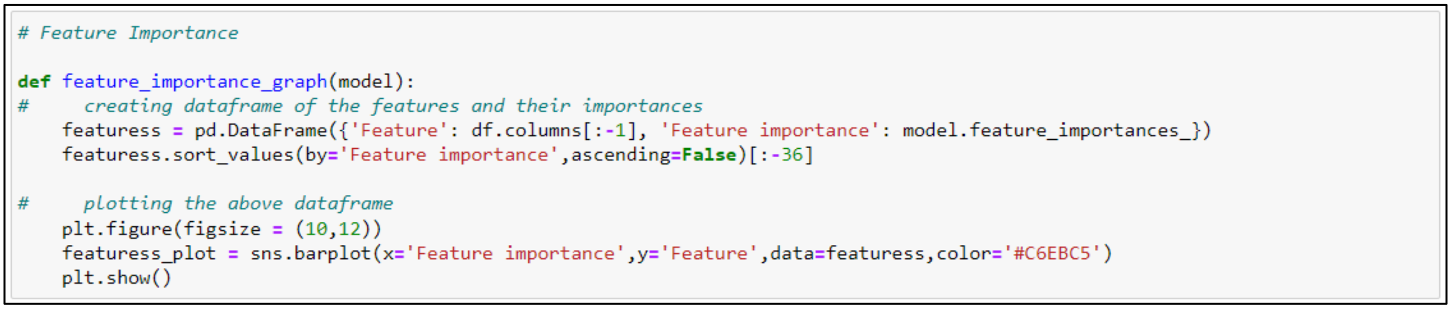
```
Model: GradientBoostingClassifier
Accuracy Score: 100.0
f1 Score: 100.0
CV Score: 1.0
ROC_AUC Score: 100.0
```

	0	1	2	3	4	5	6	7
0	14	0	0	0	0	0	0	0
1	0	13	0	0	0	0	0	0
2	0	0	4	0	0	0	0	0
3	0	0	0	5	0	0	0	0
4	0	0	0	0	2	0	0	0
5	0	0	0	0	0	4	0	0
6	0	0	0	0	0	0	8	0
7	0	0	0	0	0	0	0	4

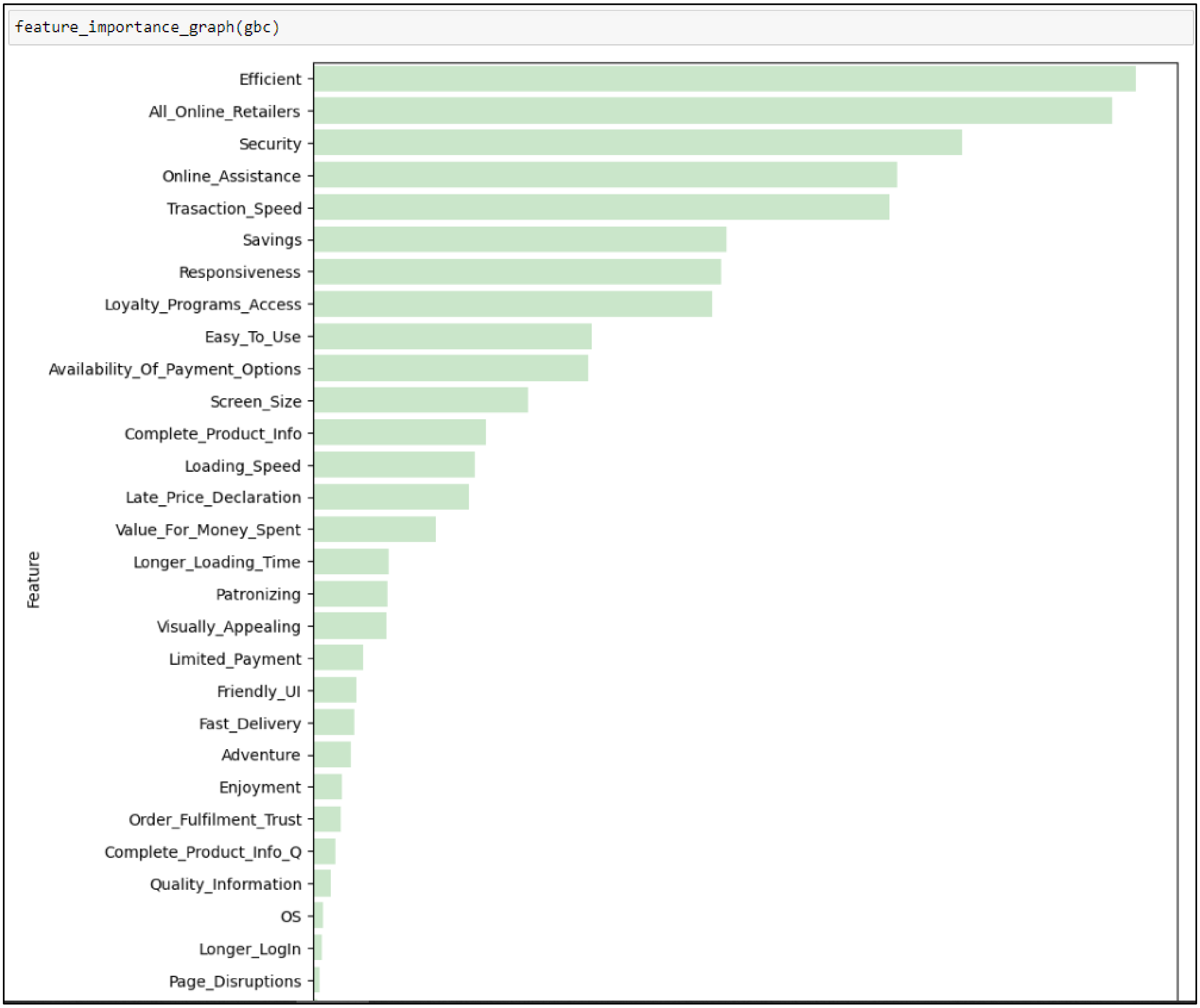
As a result Gradient Boosting Classifier had the best score of 100% accuracy, CV score and ROC_AUC score.

Feature Importance

A feature importance plot is one way to interpret model results. It tells us what features are important for the machine learning model in deciding which variables are important for more recommendations.



```
# Feature Importance

def feature_importance_graph(model):
#    creating dataframe of the features and their importances
    featuress = pd.DataFrame({'Feature': df.columns[:-1], 'Feature importance': model.feature_importances_})
    featuress.sort_values(by='Feature importance',ascending=False)[:-36]

#    plotting the above dataframe
    plt.figure(figsize = (10,12))
    featuress_plot = sns.barplot(x='Feature importance',y='Feature',data=featuress,color='#C6EBC5')
    plt.show()
```



```
feature_importance_graph(gbc)
```

Figure 16: Feature importance

Model Conclusion:

The top 10 essential features that can effectively predict the "recommended e-tailer," according to the model with the highest accuracy, are:

i. Efficiency of the website
ii. Retailers used by the customer
iii. Security of the E-tailer
iv. Online assistance provided by the E-tailer
v. Transaction speed
vi. Offers and Discounts (Savings)
vii. Responsiveness of the website
viii. Loyalty programs
ix. Easy-to-use
x. Availability of payment options

Jupyter Notebook

The jupyter notebook with the visualizations performed and models built can be found here: https://github.com/ibvhim/Blog-Notebooks/blob/main/Customer_Retention_ML%20version.ipynb