CUSTOMER RETENTION-PROJECT

Submitted by:

ANJANA.P

ACKNOWLEDGMENT

I would like to express my appreciation to team Flyprobo for giving such a realistic data for analysis, with a full-length description of the project. My mentor Ms.Khushboo Garg has helped me in many stages of this project where I was stuck with problems. I use this opportunity to thank him for helping me at the right time without any delay.

I also thank DataTrained academy team for their wonderful classes and also their live support team who have been there at any time to help.

Also, this project made me search for a lot of data's in several webpages and sites, that helped me to rectify my doubts and, I was able to study more about data analysis.

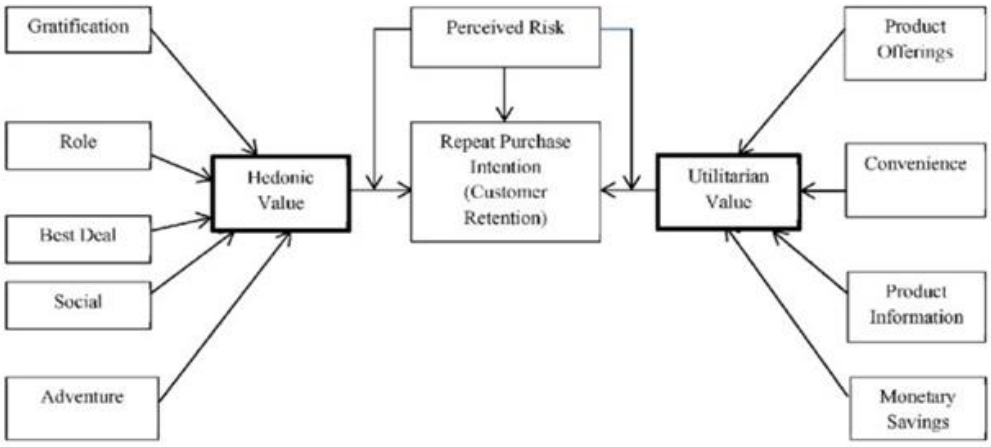
INTRODUCTION

• Business Problem Framing

This is a project based on E-retail factors for customer activation and retention from Indian e-commerce customers. Customer satisfaction has emerged as one of the most important factors that guarantee the success of online store; it has been posited as a key stimulant of purchase, repurchase intentions and customer loyalty.

A comprehensive review of the literature, theories and models have been carried out to propose the models for customer activation and customer retention. Five major factors that contributed to the success of an e-commerce store have been identified as: service quality, system quality, information quality, trust and net benefit. The research furthermore investigated the factors that influence the online customers repeat purchase intention. The combination of both utilitarian value and hedonistic values are needed to affect the repeat purchase intention (loyalty) positively. The data is collected from the Indian online shoppers. Results indicate the e-retail success factors, which are very much critical for customer satisfaction.

The factors which influence the customers to repeat purchase are utilitarian values and hedonic values. It can be understood from the below figure.



• Conceptual Background of the Domain Problem

This project is based on unsupervised learning. As the name suggests it is a machine learning technique in which models are not supervised using training dataset. They are commonly used for finding meaningful patterns and groupings inherent in data, extracting generative features and exploratory purpose. We have to perform EDA analysis and also form clusters in our analysis. Clustering involves segregating data based on similarity between data instances. Various types of clustering methods are Heirarchial clustering, centroid based clustering, distribution based clustering, density based clustering, fuzzy clustering etc. Here we used K-means algorithm which is centroid based algorithm.

Points to Remember:

There are no null values in the dataset.

There are 269 rows and 71 columns

Data is based on real survey among e-customers about online shopping

There are both integer and object datatypes in the dataset

• Review of Literature

First of all the data is saved in a csv file. There were two datasets , one is original object datatype datas and the other one is integer type data.Here I chose integer data which is easier for our analysis.Then its shape, datatypes, column value counts are all checked to get an outline of the data collected. Null values and correlation between the columns are checked using heatmap. Univariate and bivariate analysis are done for more clarification. Boxplot method is used to check the presence of outliers.Standard scaler is used to scale the data,since K-Means is a distance based algorithm the difference of magnitude can create problems.So all the magnitudes are brought to the same magnitude.Then K-Means clustering method is used to find the inertia of the model.Elbow ethod is used to find the number of clusters in the model.

• Motivation for the Problem Undertaken

The main objective behind doing this project is to make an understanding of online services that are widely accepted nowadays. E-commerce overcome geographical limitations,provide goods at lower cost, provide abundant information, eliminate travel time and cost, provide comparison shopping etc.There are so many e-retailers today which provide better shopping experiences. From this analysis of data we get a clear picture of customers needs and their satisfaction in online shopping.

ANALYTICAL PROBLEM FRAMING

- ## Mathematical/ Analytical Modelling of the Problem

In the describe function we have checked mean, std.deviation, minimum, maximum, 25 percentile, 50 percentile, 75 percentile of each attribute columns.

Mean is the average, median is the central value and mode is the frequency.

Percentile is the value below which the percentage of data falls.

We also checked value counts of each attributes for better understanding.

EUCLIDEAN

The K-means algorithm aims to choose centroids that minimise the **inertia**, or **within-cluster sum-of-squares criterion**:

$$\sum_{i=0}^{n} \min_{\mu_j \in C} (\|x_i - \mu_j\|^2)$$

EUCLIDEAN DISTANCE=

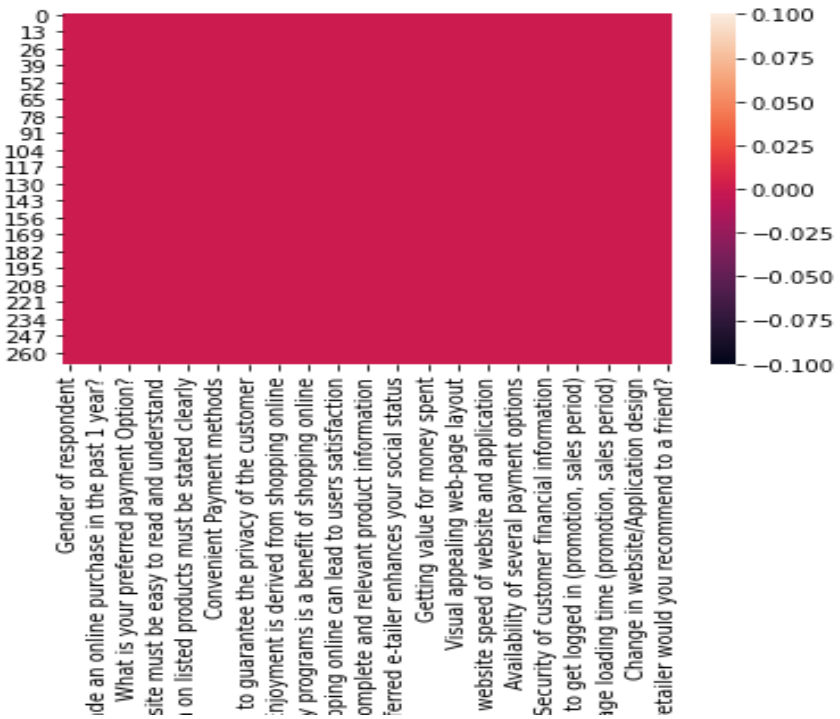
$$d(p,q) = d(q,p)$$
$$= \sqrt{(q_1 - p_1)^2 + (q_3 - p_3)^2 + \cdots + (q_n - p_n)^2}$$
$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

- ## Data Sources and their formats

FlipRobo technologies have provided this dataset for detailed analysis which was collected from realtime customers who purchase online. The data collected was in an excel sheet with a very detailed description of each columns with it. The data is converted into csv file and loaded in Jupyter notebook first. There are 71 columns and 269 rows in this dataset. The data was in integer and object datatypes. The data collected was from a survey of realtime online customers, who tried various online sites.There is no train and test datas to learn the model, thus it is unsupervised machine learning.

- ## Data Pre-processing Done

First the important libraries for pre-processing and also csv file for analysis is imported. Shape and datatypes of columns are checked. There are 269 rows and 71 columns in our dataset. There are object and integer datatypes. The object datatypes should be converted to integer datatypes using LabelEncoder for the analysis. Unnecessary columns like city of customer, pin code, their browser ,device, screen size ,OS, etc. are dropped as they provide no necessary information for our analysis. Null values are checked and should be cleared if there is any.



The uninterrupted lines show there are no null values in our dataset.

The value count of each attribute is checked which makes better understanding of the data. Label encoder is used to convert object datatype columns to integer. Describe functions provide information with minimum, maximum,mean,std.deviation,25th percentile,50th percentile 75th percentile of each column. We can see that there is only less difference between min and maximum,25% and 50% and 75%, 75% and maximum. This shows there are only less outliers or no outliers in our dataset. Standard scaler is used to scale the data.K-means algorithm is used to cluster the data and elbow method is used to find the number of clusters.

- ## Hardware and Software Requirements and Tools Used

I have used intel core i3 processor, 4GB RAM and 64 bit operating system as hardware and windows 10, MS excel, MS word and python 3 Jupyter notebook as software for the completion of this project. In jupyter notebook various libraries are also used. They include pandas, numpy, matplotlib , seaborn , imblearn and sklearn.

MODEL/S DEVELOPMENT AND EVALUATION

- ## Identification of possible problem-solving approaches

The major problems we dealt with this dataset is

The dataset is unsupervised. There is no train and test data to learn the machine.We have to create clusters and group the datas based on their similarity.I used elbow method to find the number of clusters in the model.

- ## Testing of Identified Approaches (Algorithms)

After the EDA analysis standard scaler is used to scale the data,as kmeans is a distance based algorithm.The difference of magnitude can create a problem, so it is necessary to bring all the variables to same scale.Then Kmeans algorithm is used to find the inertia of the model. Elbow method is used to find the number of clusters needed for our model.

- ## Evaluation of selected models

i. Inertia-

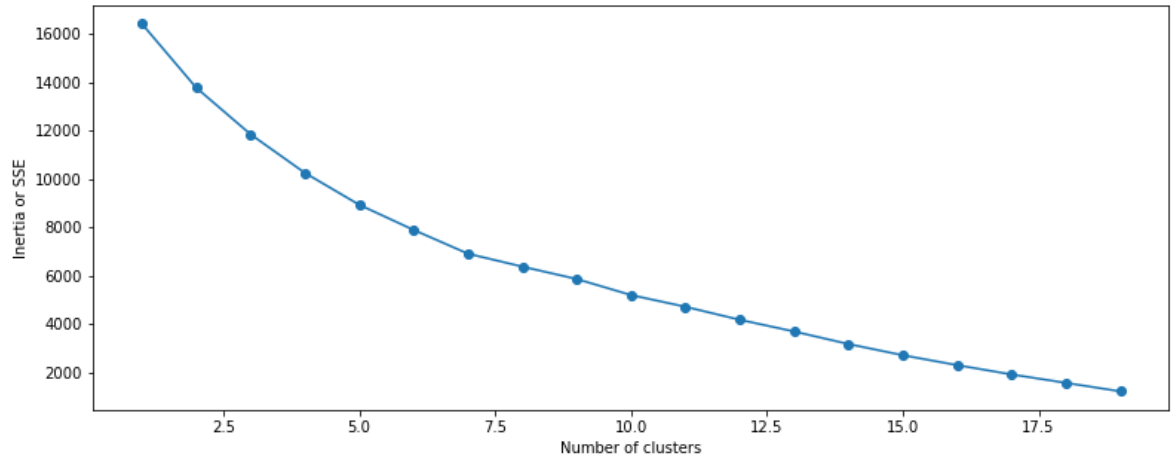
```
#defining k-means function with initialization as k-means++
from sklearn.cluster import KMeans
model = KMeans(n_clusters=3,
               init='k-means++',
               n_init=10,
               max_iter=300,
               tol=0.0001,
               precompute_distances='auto',
               verbose=0,
               random_state=42,
               copy_x=True,
               n_jobs=None,
               algorithm='auto')

model.fit(df)
model.inertia_
```

```
11837.169964495277
```

K-means algorithm clusters data by trying to separate samples in n groups of equivalent variance, minimizing a criterion known as inertia .Inertia is the sum of squared distances of samples to their closest cluster centre.Inertia makes the assumptions that clusters are convex and isotropic.Inertia tells how far away the points within a cluster are.The range of inertia starts from zero and goesup.A small of inertia should be aimed.

ii. elbow method-

```
from sklearn.cluster import KMeans
clusters = range(1, 20)
sse=[]
for cluster in clusters:
    model = KMeans(n_clusters=cluster,
                init='k-means++',
                n_init=10,
                max_iter=300,
                tol=0.0001,
                precompute_distances='auto',
                verbose=0,
                random_state=42,
                copy_x=True,
                n_jobs=None,
                algorithm='auto')

    model.fit(df)
    sse.append(model.inertia_)

sse_df = pd.DataFrame(np.column_stack((clusters, sse)), columns=['cluster', 'SSE'])
fig, ax = plt.subplots(figsize=(13, 5))
ax.plot(sse_df['cluster'], sse_df['SSE'], marker='o')
ax.set_xlabel('Number of clusters')
ax.set_ylabel('Inertia or SSE')
```



Fundemental step for any unsupervised algorithm is to determine the optimal number of clusters into which data may be clustered.Elbow method is one of the popular methods to determine this optimal value of k.When using K-means clustering,users need some way to determine whether they are using the right number of clusters.Then plot a line chart of SSE for each value of K.If the line chart looks likes an arm then elbow of the arm is the value of k that is the best.

Kmeans++ is a smart centroid initialization technique and algorithm for choosing the initial values.

n_init - is the number of times k-means algorithm will be running with different centroid seeds

max_iter-maximum number of iterations of k-means algorithm for a single run.

tol-the relative tolerance of the difference in cluster centres of two consecutive iteratons to declare convergence.

Verbose- verbosity mode

Random state-determines the random number generation for centroid initialization

Copy_x-when precomputing distances it is more numerically accurate to center the data first.If copy_x is true then original data is not modified

N_iter-no.of iterations run

iii. K-means clustering

```
model = KMeans(n_clusters=7,
               init='k-means++',
               n_init=10,
               max_iter=300,
               tol=0.0001,
               precompute_distances='auto',
               verbose=0,
               random_state=42,
               copy_x=True,
               n_jobs=-1,
               algorithm='auto')

model.fit(df)
```

```
KMeans(n_clusters=7, n_jobs=-1, precompute_distances='auto', random_state=42)
```

```
print('SSE: ', model.inertia_)
print('\nCentroids: \n', model.cluster_centers_)

pred = model.predict(df)
d1['cluster'] = pred
print('\nCount in each cluster: \n', d1['cluster'].value_counts())
```

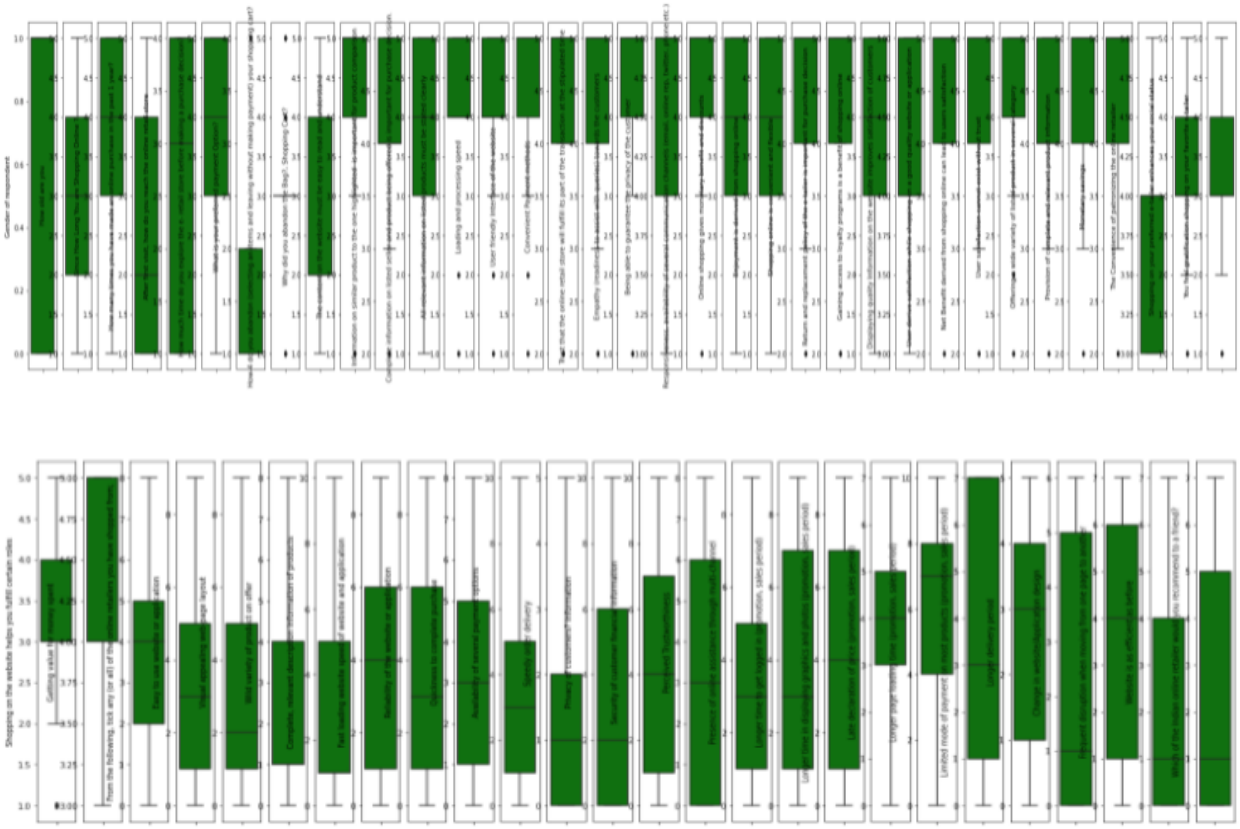
```
SSE:  6900.39636897185


Count in each cluster:
 3    72
0    67
5    41
1    30
6    29
2    18
4    12
Name: cluster, dtype: int64
```

4th cluster has maximum number of samples and 5th cluster has minimum

From elbow method,we get the value of k as 7,that is we chose number of clusters as 7and we get the count of datapoints in each clusters as above.4th cluster has maximum number of datapoints,ie 72 and 5th cluster has minimum number of datapoints,ie 12

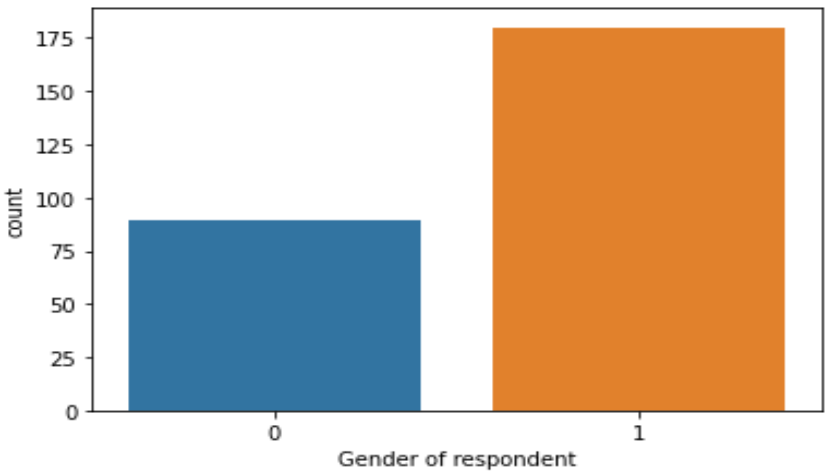
- ## Visualizations

1) Boxplot- Boxplots are the best methods to check for the presence of outliers



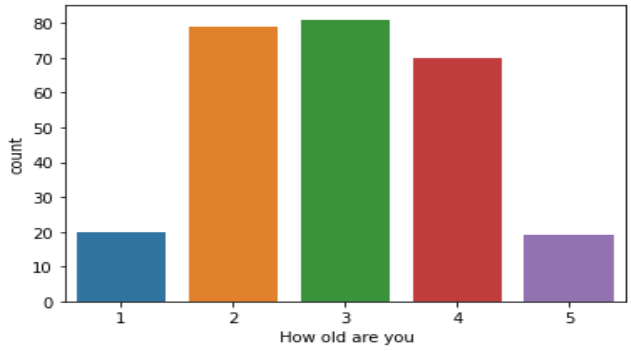
From this we can see that there are only few outliers

2) UNIVARIATE ANALYSIS



0-Male

1-Female



#2,3-maximum customers are between age group ->31 to 40 and 21 to 30.

#4-around 70 customers are from 41 to 5o age group

#1,5-less customers are from below 20 and above 50 age group

#5-maximum customers shop for more than 4 years

#3-around 60 customers shop for 2-3 years

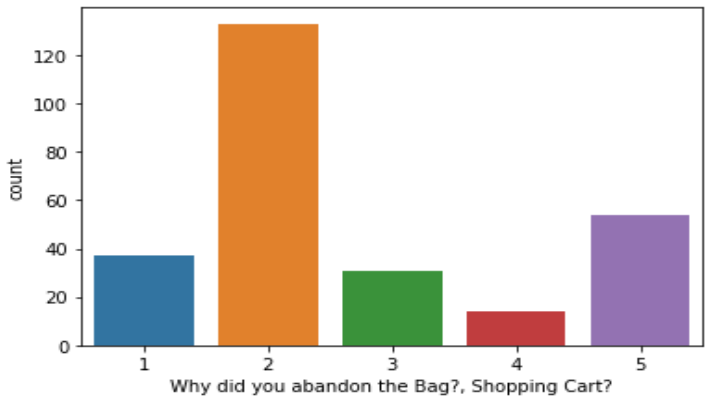
#1-around 40 customers are new to online shopping,ie,less tahn 1 year

#2-below 20 customers shop for 1 to 2 years

#1-maximum customers shop below 10 timesin past year

#4,5- around 50 to 60 customersshop above 30 times

#2,3- only few shop for 11 to 30 times



#2-above 120 abandon the bag due to better alternative option

#4-very few abandon due to promocode not applicable

#3-around 30 abandon due to lakh of trust

#0-amazon

#1-amazon,flipkart

#2-amazon,flipkart,myntra

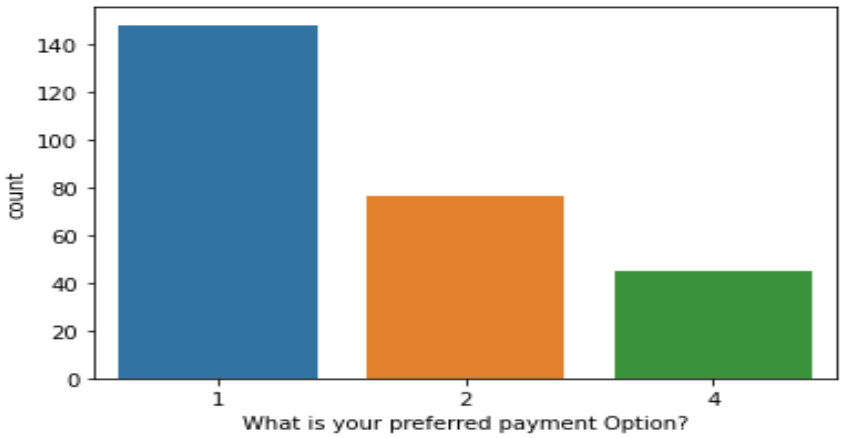
#3-amazon,myntra

#4-amazon,paytm

#5-amazon,paytm,myntra

#6-flipkart

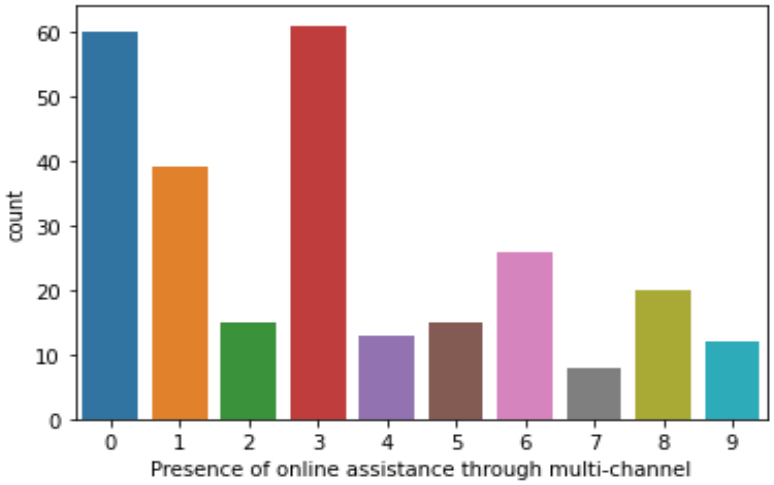
#7-flipkart,paytm,myntra,snapdeal



#1- majority shop using credit/debit card

#2-around 70 shop using cash on delivery

#4-around 40 shop using e-wallet



#0-amazon

#1-amazon,flipkart

#2-amazon,flipkart,myntra

#3-amazon,flipkart,myntra,snapdeal

#4-amazon,flipkart,paytm

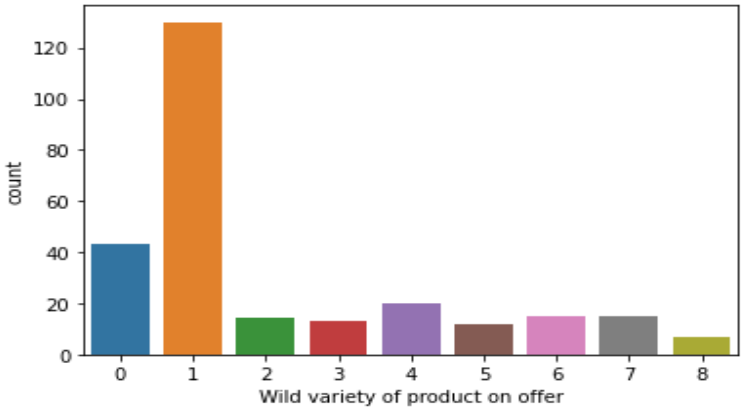
#5-amazon,myntra

#6-amazon,snapdeal

#7-flipkart

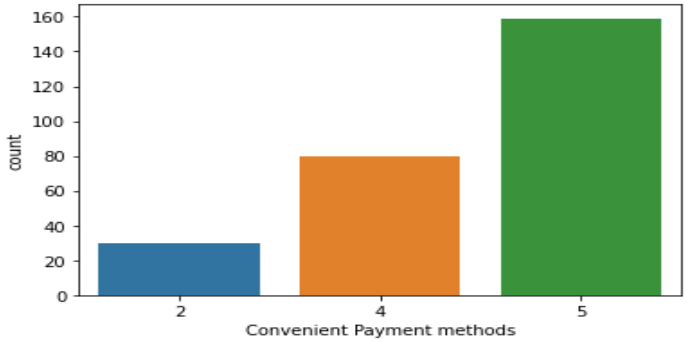
#8-myntra

#9-paytm



#1-amazon,flipkart

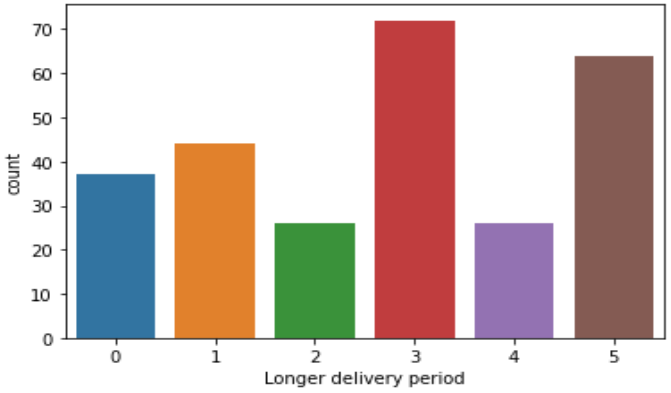
#0-amazon



#5-strongly agree

#4-agree

#2-disagree

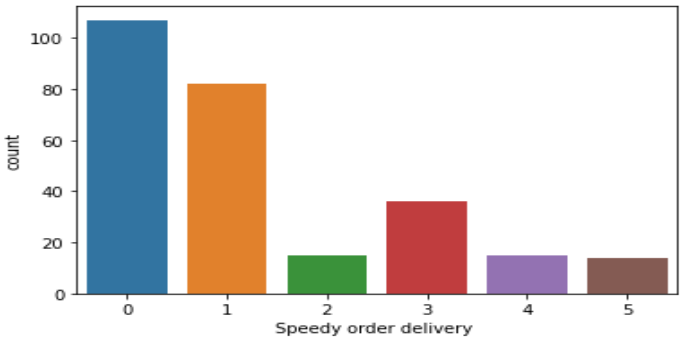


#3-paytm

#5-snapdeal

#1-flipkart

#2,4-paytm,snapdeal,myntra



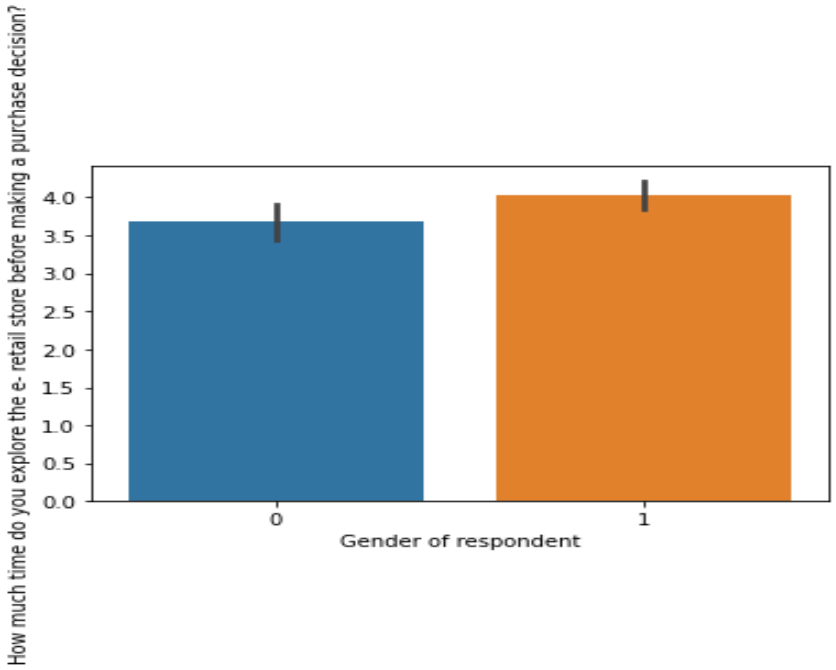
#0-amazon

#1-amazon,flipkart

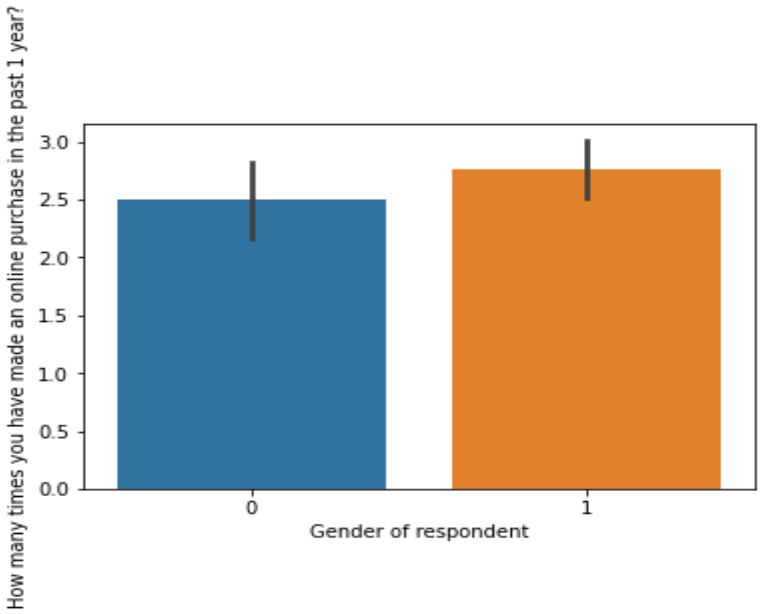
#3-amazon,flipkart,snapdeal

#2-flipkart,myntra,snapdeal

3) BIVARIATE ANALYSIS



#comparatively female take more time than men



#female purchased more no.of times than men

CONCLUSION

From this analysis we understood that customers depend on various factors for their purchase through online. There are customers with more than 4 years of history. This proves that online purchase gives them satisfaction and also they get better results than from ongoing purchase. We can see that females ageing from 20 to 40 are comparatively more attracted to online purchase. They purchase for more than 10 times in a year. Also customers prefer credit card options than e wallet and cash on delivery. There are many online sites nowadays in which amazon and flipkart are in the leading position. There are many factors which made them best among customers, they are presence of online assistance, many offers, convenient payment options, less delivery period, easy return policy etc. Also when compared e-retailers which is less popular and less frequently used among customers are paytm and snapdeal. From this we can understand that customers look more for their convenience, brands, value for their cash, offers, trust etc. All these must be improved to get a better or increased sale. Customers can be retained with some techniques like implementing a customer feedback loop, maintain customer communication calendar, customer education programe, build trust with customers, offer unique services, start customer retention programe, provide better payment options, decrease shipping cost.

REFERENCE

1)https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html

2)https://scikit-learn.org/stable/modules/clustering.html

3)https://towardsdatascience.com/machine-learning-vi-unsupervised-learning-k-means-kaggle-dataset-with-k-means-1adf5c30281b

4)https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

5)https://www.analyticsvidhya.com/blog/2021/02/simple-explanation-to-understand-k-means-clustering/

6)https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/

7)https://www.google.com/search?q=inertia+of+kmeans+clustering&oq=inertia+of+kmeans+clustering&aqs=chrome..69i57j0i22i30.10043j0j7&sourceid=chrome&ie=UTF-8

8)https://towardsdatascience.com/k-means-clustering-from-a-to-z-f6242a314e9a