FLIP ROBO

CUSTOMER RETENTION PROJECT

Submitted by:

RAUSHAN KUMAR

ACKNOWLEDGMENT

I would like to express my special thanks to my mentor Khusboo Gargwho gave me his support and assistance throughout the project. I amthankful to Fliprobo Technologies and Data Trained Institute to give me guidelines and knowledge to complete this project.

Links and websites that I preferred:
https://towardsdatascience.com/oversampling-and-undersampling-5e2bbaf56dcf

https://career-resource-center.udacity.com/portfolio/data-science-reports#:~:text=A%20data%20science%20report%20is,the%20legitimacy%20of%20your%20process.

https://towardsdatascience.com/top-3-methods-for-handling-skewed-data-1334e0debf45

INTRODUCTION

➢ Business Problem Framing

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.

➢ Conceptual Background of the Domain Problem

It is helpful for those who have experience of e-shopping and on which basis they use to order the products. It will help in analyzing the data as there are 71 columns in the dataset the domain knowledge and experience can help to understand the correlated things and changing trends over the period of time.

➢ Review of Literature

This study combines factors that other studies have done that will influence the consumer’s purchasing decision in online Shopping.

Online shopping indicates electronic commerce to buy products or services directly from the seller through the Internet. Internet-based or Click and Order business model has replaced the traditional Brick and Mortar business model. More people than before are using the web to shop for a wide variety of items, from house to shoes to airplane tickets. Now people have multiple options to choose their products and services while they are shopping through an online platform.

➢ Motivation for the Problem Undertaken

Online shopping is the easy solution for busy life in today’s world. In the past decade, there had been a massive change in the way of customer’s shopping. Despite consumers’ continuation to buy from a physical store, the users or buyers feel very convenient to online shopping. Online shopping saves crucial time for modern people because they get so busy that they cannot or unwilling to spend much time shopping.

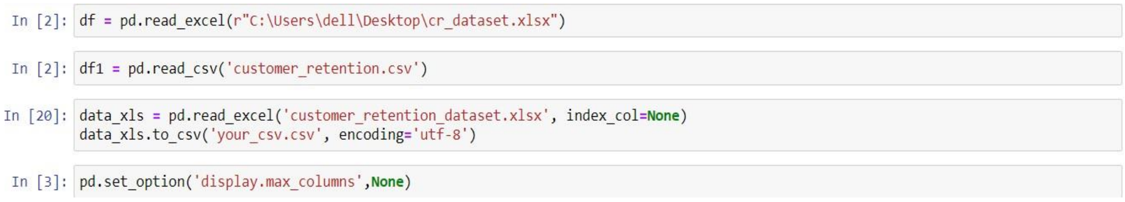
Analytical Problem Framing

Mathematical/Analytical Modelling of the Problem

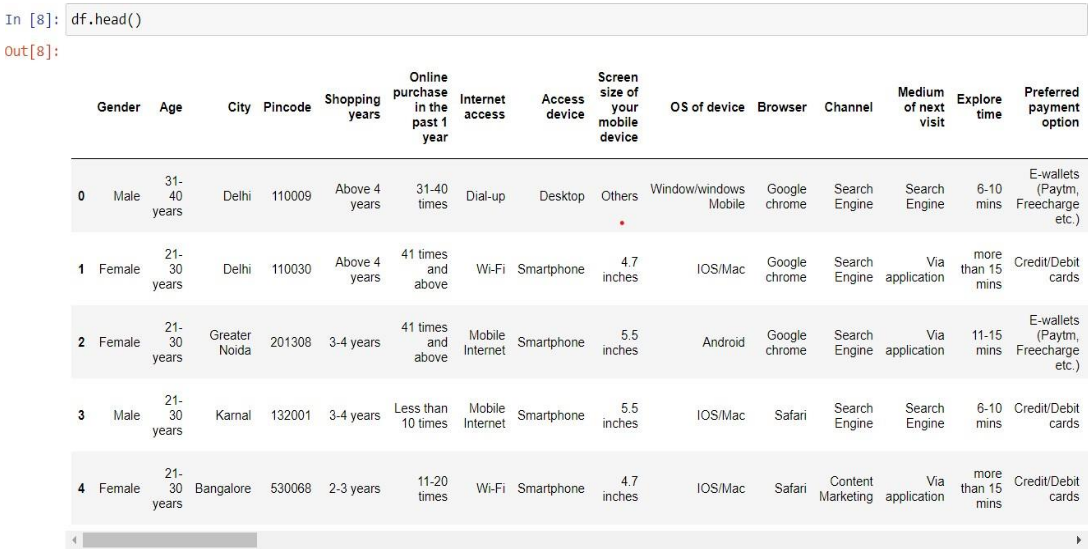
Data consist of 269 rows and 71 columns, because of the large number of columns it is difficult to analyze easily also to find significant columns.
EDA does not give much options as only one column in the dataset is numerical rather than that each columns were categorical columns

Data Sources and their formats

The Sample Data is Highly Confidential and used for the purpose, of customers preferences, loyal customers and trends. The data is in csv format. First I have read the csv file and converted it into a data frame.
The columns names have to change in short names to make the analysis process easy.



```
In [2]: df = pd.read_excel(r"C:\Users\dell\Desktop\cr_dataset.xlsx")

In [2]: df1 = pd.read_csv('customer_retention.csv')

In [20]: data_xls = pd.read_excel('customer_retention_dataset.xlsx', index_col=None)
         data_xls.to_csv('your_csv.csv', encoding='utf-8')

In [3]: pd.set_option('display.max_columns',None)
```



```
In [8]: df.head()
Out[8]:
```

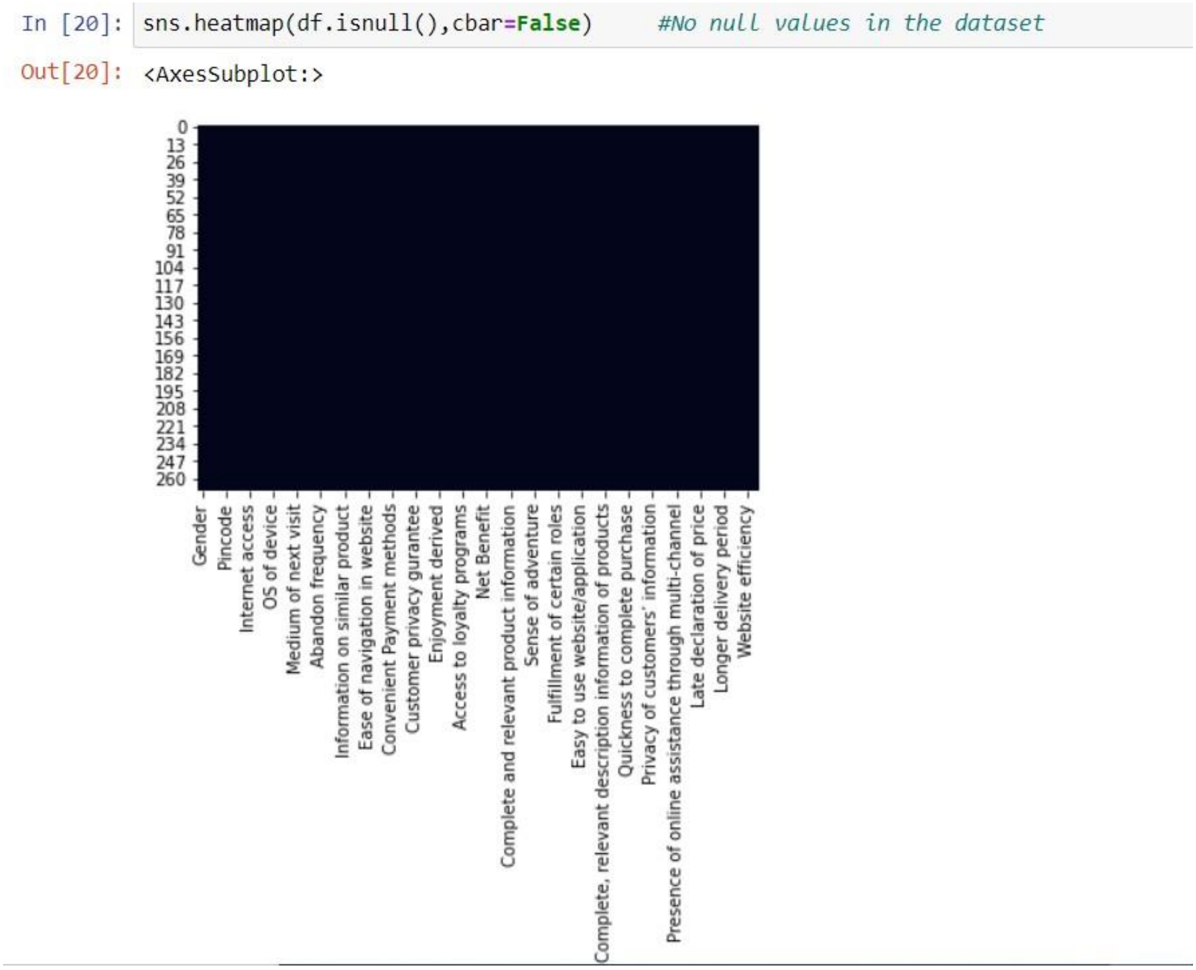
	Gender	Age	City	Pincode	Shopping years	Online purchase in the past 1 year	Internet access	Access device	Screen size of your mobile device	OS of device	Browser	Channel	Medium of next visit	Explore time	Preferred payment option
0	Male	31-40 years	Delhi	110009	Above 4 years	31-40 times	Dial-up	Desktop	Others	Window/windows Mobile	Google chrome	Search Engine	Search Engine	6-10 mins	E-wallets (Paytm, Freecharge etc.)
1	Female	21-30 years	Delhi	110030	Above 4 years	41 times and above	Wi-Fi	Smartphone	4.7 inches	IOS/Mac	Google chrome	Search Engine	Via application	more than 15 mins	Credit/Debit cards
2	Female	21-30 years	Greater Noida	201308	3-4 years	41 times and above	Mobile Internet	Smartphone	5.5 inches	Android	Google chrome	Search Engine	Via application	11-15 mins	E-wallets (Paytm, Freecharge etc.)
3	Male	21-30 years	Karnal	132001	3-4 years	Less than 10 times	Mobile Internet	Smartphone	5.5 inches	IOS/Mac	Safari	Search Engine	Search Engine	6-10 mins	Credit/Debit cards
4	Female	21-30 years	Bangalore	530068	2-3 years	11-20 times	Wi-Fi	Smartphone	4.7 inches	IOS/Mac	Safari	Content Marketing	Via application	more than 15 mins	Credit/Debit cards

➢ Data Pre Processing

Pre-processing Pipe Line is an important step towards the data modelling to make data ready for prediction.
Following Steps I followed for Pre Processing:

➢ Finding the null values

No NULL Values in the Dataset



```
In [20]: sns.heatmap(df.isnull(),cbar=False)     #No null values in the dataset

Out[20]: <AxesSubplot:>
```

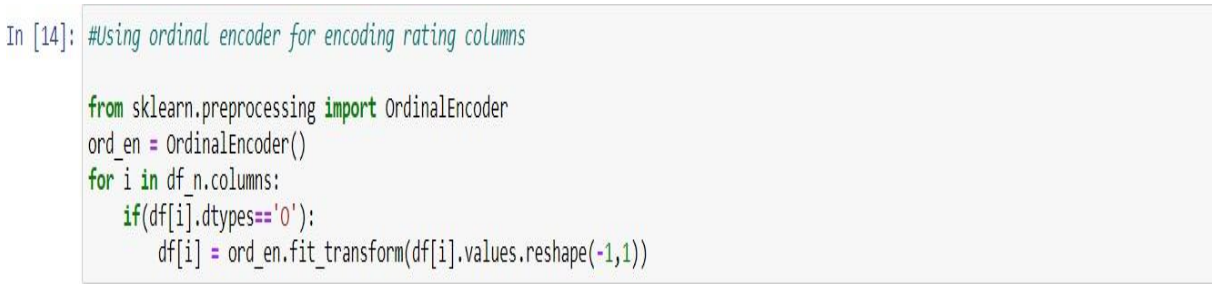
0
13
26
39
52
65
78
91
104
117
130
143
156
169
182
195
208
221
234
247
260
Gender
Pincode
Internet access
OS of device
Medium of next visit
Abandon frequency
Information on similar product
Ease of navigation in website
Convenient Payment methodes
Customer privacy gurantee
Enjoyment derived
Access to loyalty programs
Net Benefit
Complete and relevant product information
Sense of adventure
Fulfillment of certain roles
Easy to use website/application
Complete, relevant description information of products
Quickness to complete purchase
Privacy of customers' information
Presence of online assistance through multi-channel
Late declaration of price
Longer delivery period
Website efficiency

- **Handling outliers and skewness**

  No Need to handle it as all the columns are categorical.
  Only one column was numerical that is of pin code which do not need Preprocessing.

- **Encoding**

  For the encoding of columns which has ratings kind of data, I used Ordinal Encoder.



```
In [14]: #Using ordinal encoder for encoding rating columns

         from sklearn.preprocessing import OrdinalEncoder
         ord_en = OrdinalEncoder()
         for i in df_n.columns:
             if(df[i].dtypes=='O'):
                 df[i] = ord_en.fit_transform(df[i].values.reshape(-1,1))
```

- ## Hardware and Software Requirements and Tools Used

Tool:
Jupyter NoteBook 6.1.4:

- Web-based interactive computing notebook Environment.

Software Requirement:

- The client environment may be Windows, macOS, or Linux.

Hardware Requirement:

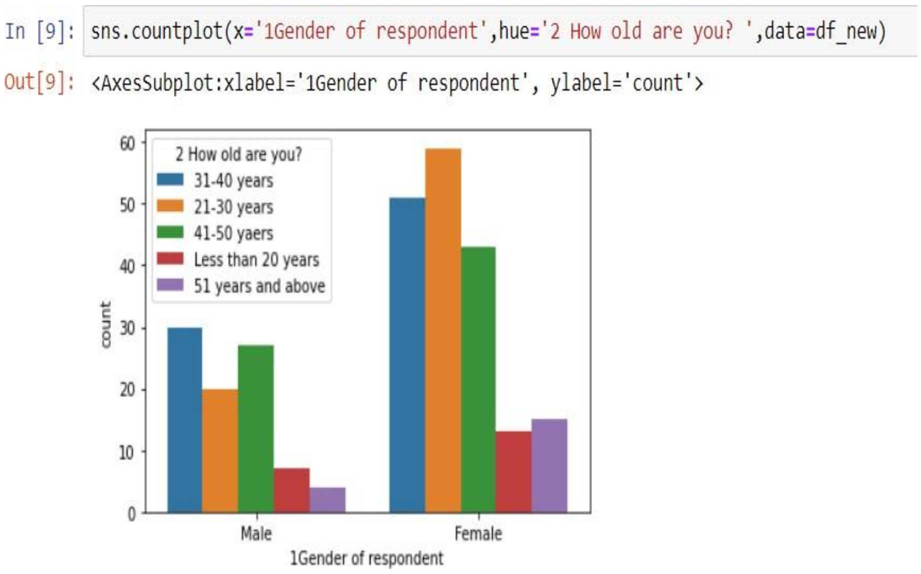
- CPU: 2 x 64-bit, 2.8 GHz, 8.00 GT/s CPUs or better.
- Memory: minimum RAM size of 32 GB, or 16 GB RAM with 1600 MHz DDR3 installed, for a typical installation with 50 regular users.

Libraries:

- Pandas: For reading CSV file, Converting dataset into a data frame, handling date datatype, and more.
- Seaborn and matplotlib: For EDA and Visualization.

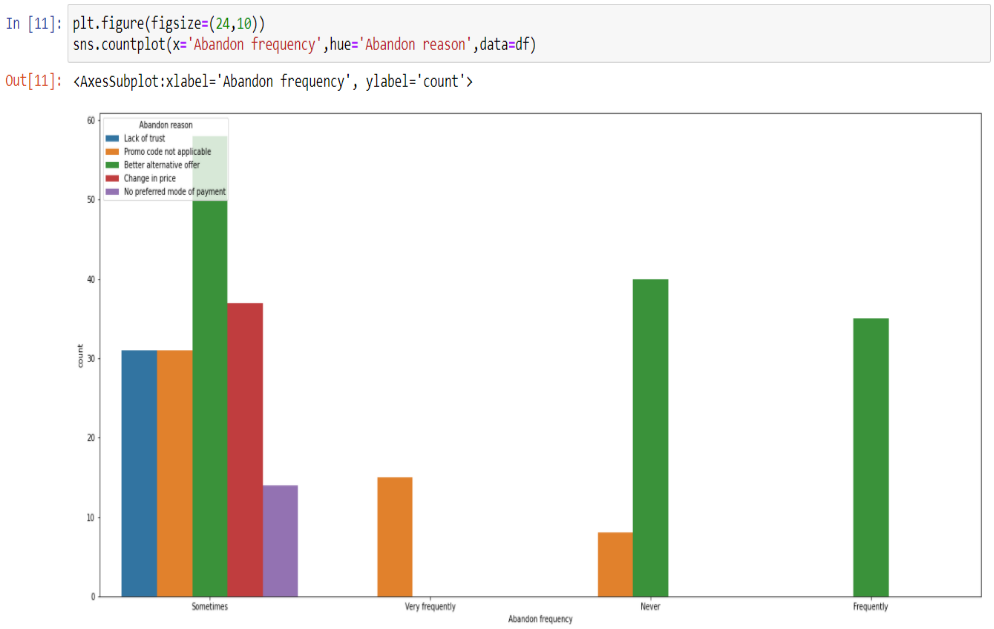
➢ Visualizations

1. Gender with Age



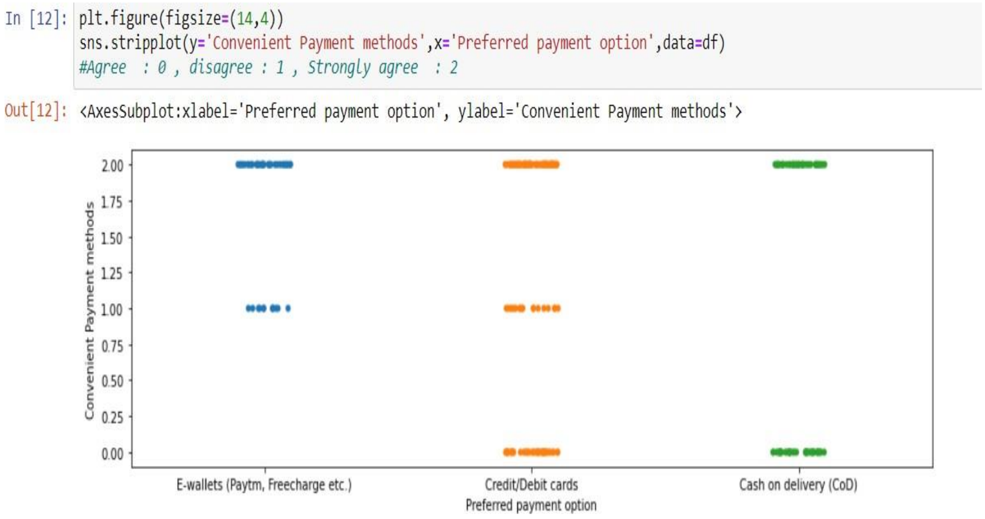
```
In [9]: sns.countplot(x='1Gender of respondent',hue='2 How old are you? ',data=df_new)

Out[9]: <AxesSubplot:xlabel='1Gender of respondent', ylabel='count'>
```

2. Abandon Frequency with its reason



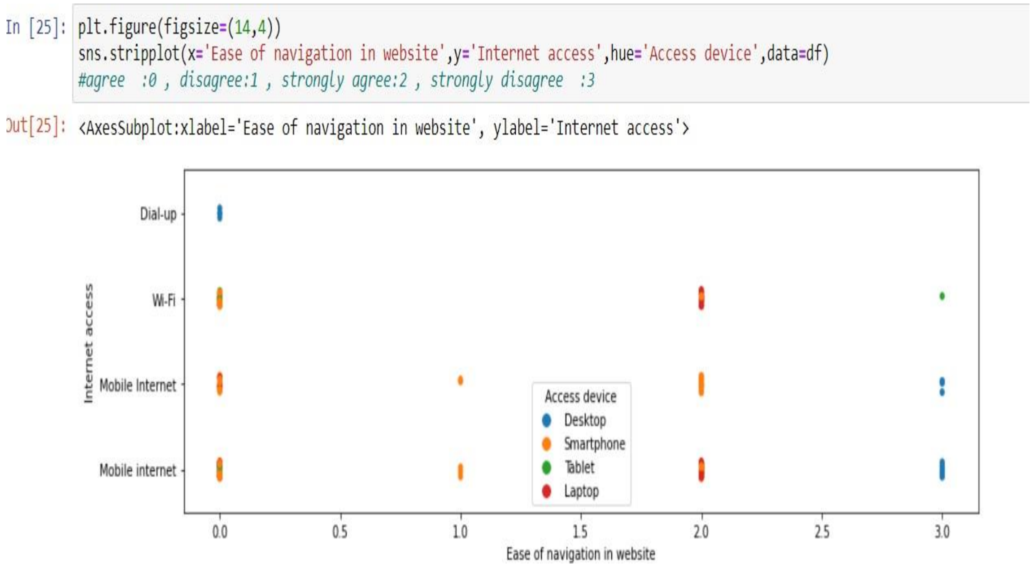
```
In [11]: plt.figure(figsize=(24,10))
         sns.countplot(x='Abandon frequency',hue='Abandon reason',data=df)

Out[11]: <AxesSubplot:xlabel='Abandon frequency', ylabel='count'>
```

3. Convenient payment with preferred payment option



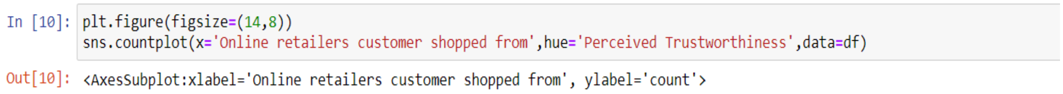
```
In [12]: plt.figure(figsize=(14,4))
         sns.stripplot(y='Convenient Payment methods',x='Preferred payment option',data=df)
         #Agree  : 0 , disagree : 1 , Strongly agree  : 2

Out[12]: <AxesSubplot:xlabel='Preferred payment option', ylabel='Convenient Payment methods'>
```

4. Navigation with internet access

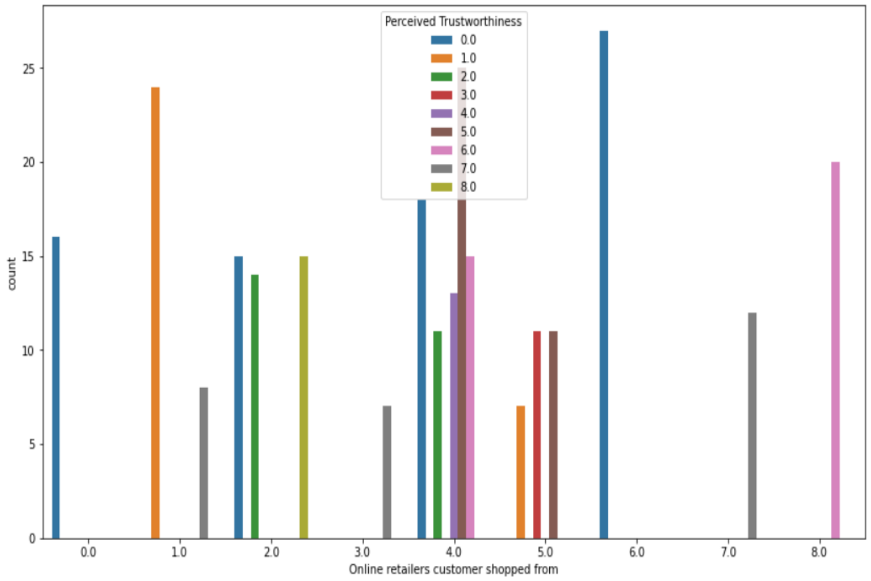


```
In [25]: plt.figure(figsize=(14,4))
         sns.stripplot(x='Ease of navigation in website',y='Internet access',hue='Access device',data=df)
         #agree  :0 , disagree:1 , strongly agree:2 , strongly disagree  :3

Out[25]: <AxesSubplot:xlabel='Ease of navigation in website', ylabel='Internet access'>
```

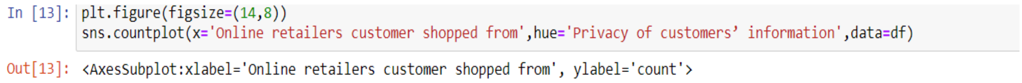
5. Online retailers shopped with trustworthiness



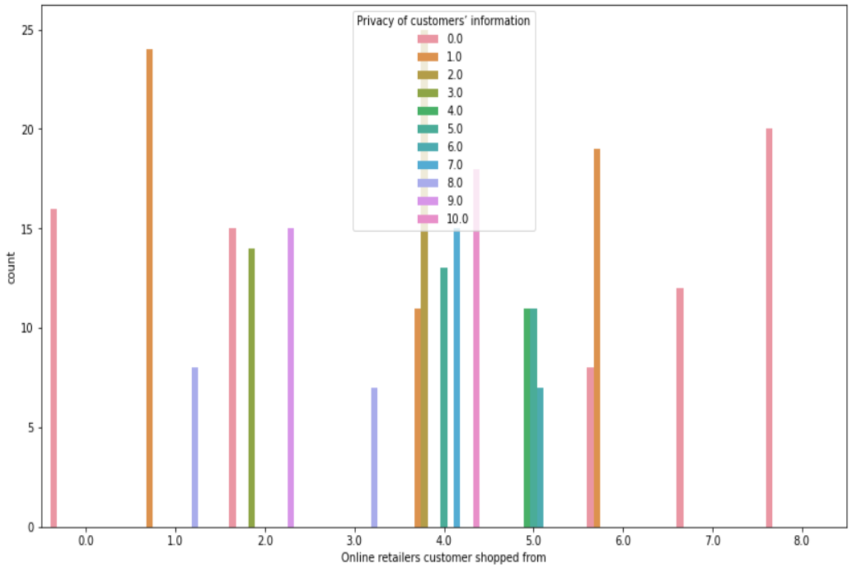
```
In [10]: plt.figure(figsize=(14,8))
         sns.countplot(x='Online retailers customer shopped from',hue='Perceived Trustworthiness',data=df)

Out[10]: <AxesSubplot:xlabel='Online retailers customer shopped from', ylabel='count'>
```



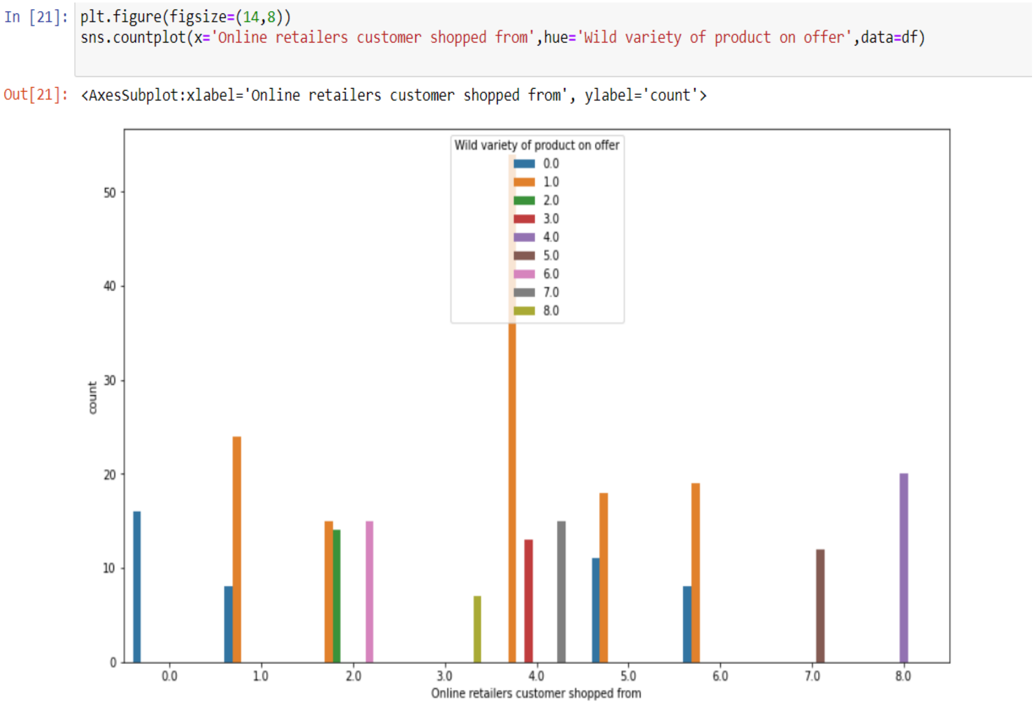
6. Privacy of customer’s information



```
In [13]: plt.figure(figsize=(14,8))
         sns.countplot(x='Online retailers customer shopped from',hue='Privacy of customers’ information',data=df)

Out[13]: <AxesSubplot:xlabel='Online retailers customer shopped from', ylabel='count'>
```



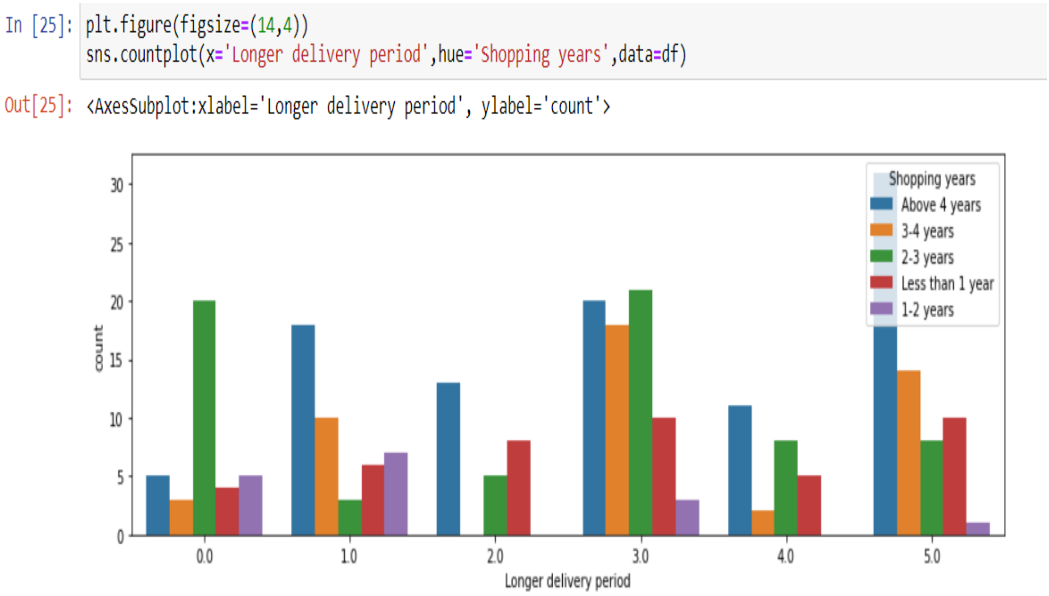
7. Wide variety of product on offer



```
In [21]: plt.figure(figsize=(14,8))
         sns.countplot(x='Online retailers customer shopped from',hue='Wild variety of product on offer',data=df)

Out[21]: <AxesSubplot:xlabel='Online retailers customer shopped from', ylabel='count'>
```

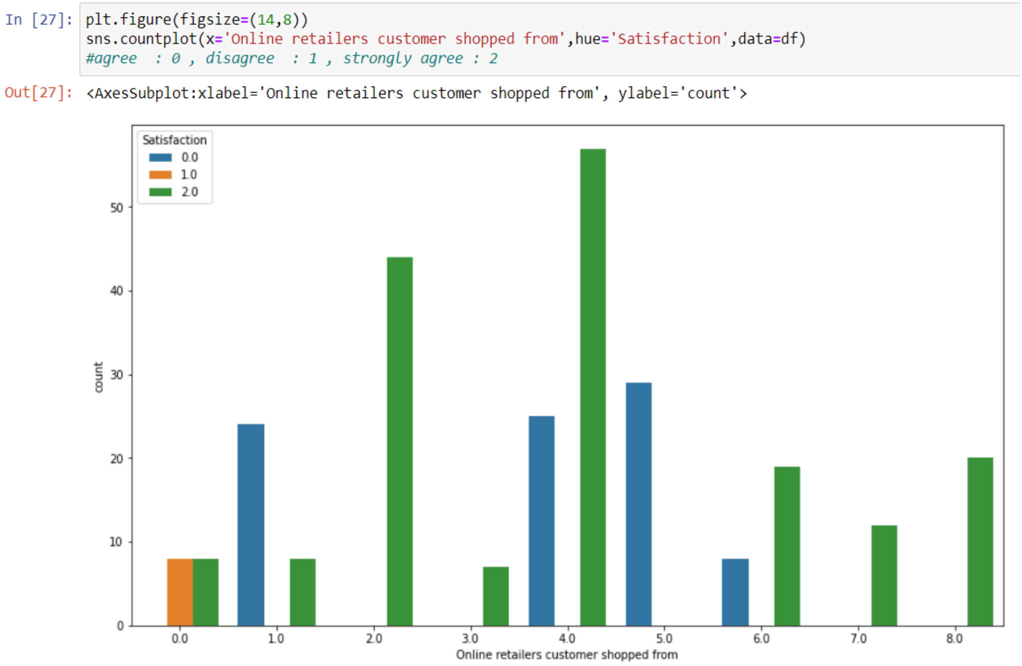
Wild variety of product on offer
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
count
0
10
20
30
40
50
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
Online retailers customer shopped from

8. Delivery period with shopping years



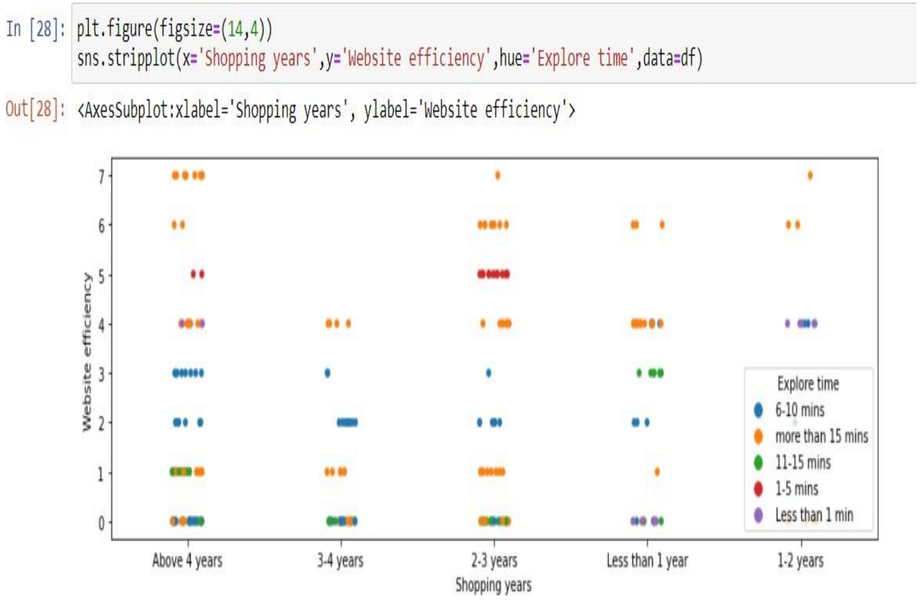
```
In [25]: plt.figure(figsize=(14,4))
         sns.countplot(x='Longer delivery period',hue='Shopping years',data=df)

Out[25]: <AxesSubplot:xlabel='Longer delivery period', ylabel='count'>
```

9. Satisfaction with online retailers customers shopped



```
In [27]: plt.figure(figsize=(14,8))
         sns.countplot(x='Online retailers customer shopped from',hue='Satisfaction',data=df)
         #agree  : 0 , disagree  : 1 , strongly agree : 2

Out[27]: <AxesSubplot:xlabel='Online retailers customer shopped from', ylabel='count'>
```

10. Shopping years with explore time



```
In [28]: plt.figure(figsize=(14,4))
         sns.stripplot(x='Shopping years',y='Website efficiency',hue='Explore time',data=df)

Out[28]: <AxesSubplot:xlabel='Shopping years', ylabel='Website efficiency'>
```

➢ Observations:

- ➢ 21-50 yrs of people shop more.
- ➢ Records contain more data of people shopping from more than 4 years.
- ➢ Source of e-shopping for most people :
- ➢ Smartphone, chrome browser and using mobile data.
- ➢ Source of payment : Credit/Debit cards
- ➢ Prominent reason of not shopped the added products : Got alternative option
- ➢ In each of the columns like: navigation in website, payment option, security, product details, processing speed: Most of the people strongly agree or agree.
- ➢ Assuming that in the columns where ratings are given most of the people given strongly agree or agree.

✓Interpretation of the Results

- ✓ All the columns are categorical. In categorical columns, the categories are somewhat imbalance.
- ✓ Looking at the type of customers there are customers who are of age 21 to 31 who are females and 31 to 41 who are males.
- ✓ Customers use mainly smartphones and mobile internet for online shopping.
- ✓ Main cities are Noida, Delhi and Bangalore from where customers do shopping.
- ✓ Based on the trustworthiness the amazon and flipkart at the top.
- ✓ Snapdeal Lack in the privacy of the customer.
- ✓ Amazon is the only website that is single website used by some customers.
- ✓ Most preferred payment option is cash on delivery.
- ✓ Most of the people do not shop the added product because of the Promo code non applicability or better alternative option.
  That means websites should take care of the alternative option in terms of price may be.
- ✓ Except Amazon every website gives late delivery to the customers using it from long time which is not a good practice.
- ✓ Customers who are using desktop are strongly disagree with ease of navigation of websites reason may be the websites portability or internet issue.

- ✓ Some of the data is non real like customer who are using other websites and given ratings to some other website.

CONCLUSION

➢ Key Findings and Conclusions of the Study

All the columns are categorical and the major drawback of the dataset is the number of columns.
The main reason of customer should not retain is to provide the same service to old and new customers both and cost of the products and promo code applicability and security of the customers data.

➢ Learning Outcomes of the Study in respect of Data Science

As the dataset contains 71 columns and all the columns are categorical in Nature.
If we see it as a Data Science prospective it is a kind of challenging task that how to properly take every column into the consideration and make the proper or accurate prediction.

Secondly in the Categorical columns we cannot find the outliers or the

Dummy data and which will influence the conclusion at the end.
Moreover it is a good case to know more about the categorical datasets.

Also each column has many categories which is also a time consuming task to do.

➢ Limitations of this work and Scope for Future Work

If the data could be of more rows may be result of the analyzing will be better and also some of the columns where kind of similar and without any reason that columns were increasing the number of dimensions of the dataset.

Each column is in categorical form if it would be some numerical columns it will be easy to analysis.