Customer Retention

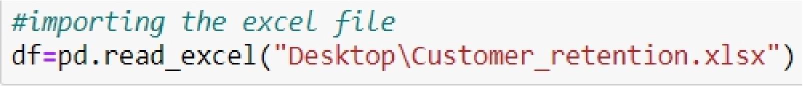
In this project we will be Analyzing a set of Customer Data set for Online and retail shopping.In this we will use different visualization and encoding techniques to do Exploratory Data Analysis on our data set and clean the data.

Problem Statement:

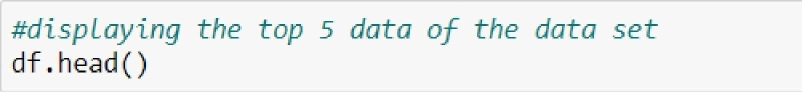
Customer satisfaction has emerged as one of the most important factors that guarantee the success of online store; it has been posited as a key stimulant of purchase, repurchase intentions and customer loyalty. A comprehensive review of the literature, theories and models have been carried out to propose 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.

Exploratory Data Analysis:

In order to Do the Exploratory Data Analysis first we need to import the data set Using read_excel function as the file is in Excel format and then print the top 5 rows of our data set to find whether it was imported correctly or not.

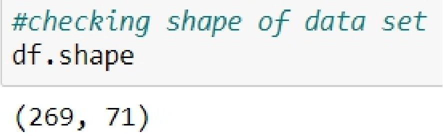


```
#importing the excel file
df=pd.read_excel("Desktop\Customer_retention.xlsx")
```



```
#displaying the top 5 data of the data set
df.head()
```

The above two lines shows how to import the data set and check the top 5 rows of our data set using the head function.After importing our data set we will be finding the shape of data set using the shape function which is shown below.



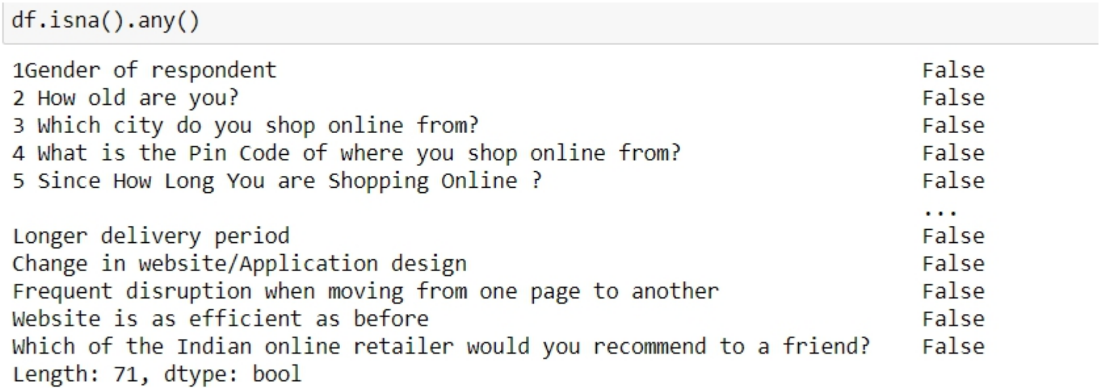
```
#checking shape of data set
df.shape
```

```
(269, 71)
```

So our data set has 269 rows and 71 columns to clean the data and get it into a normal format to predict the model.After checking the shape of our data set we need to check for missing values in our data set for this we will use isna() function.

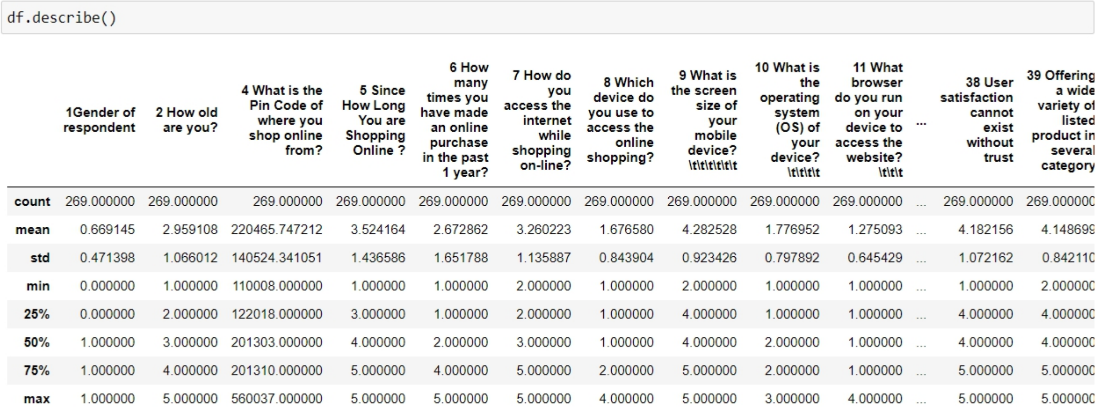
Checking for null values:

In order to Check the null values as explained previously we will use the isna() function to find out the null values in our data set.After using the function we will be able to see that there will be no Missing values or null values in our data set.The output is shown below.



```
df.isna().any()

1Gender of respondent                                                 False
2 How old are you?                                                    False
3 Which city do you shop online from?                                 False
4 What is the Pin Code of where you shop online from?                 False
5 Since How Long You are Shopping Online ?                            False
                                                                       ...
Longer delivery period                                                False
Change in website/Application design                                  False
Frequent disruption when moving from one page to another              False
Website is as efficient as before                                     False
Which of the Indian online retailer would you recommend to a friend?  False
Length: 71, dtype: bool
```

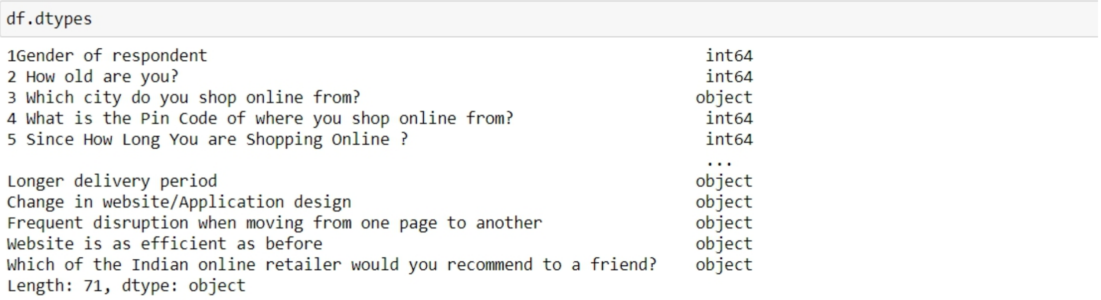
For further clarity we are using the .describe() function in order to check whether we have missed any null values as the .describe() function gives the complete statistics of our data set.The describe() functions the Statistics of all the Numerical columns in our data set as shown below.



```
df.describe()
```

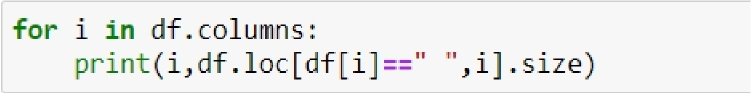
	1Gender of respondent	2 How old are you?	4 What is the Pin Code of where you shop online from?	5 Since How Long You are Shopping Online ?	6 How many times you have made an online purchase in the past 1 year?	7 How do you access the internet while shopping on-line?	8 Which device do you use to access the online shopping?	9 What is the screen size of your mobile device? \t\t\t\t\t\t	10 What is the operating system (OS) of your device? \t\t\t\t	11 What browser do you run on your device to access the website? \t\t\t	...	38 User satisfaction cannot exist without trust	39 Offering a wide variety of listed product in several category
count	269.000000	269.000000	269.000000	269.000000	269.000000	269.000000	269.000000	269.000000	269.000000	269.000000	...	269.000000	269.000000
mean	0.669145	2.959108	220465.747212	3.524164	2.672862	3.260223	1.676580	4.282528	1.776952	1.275093	...	4.182156	4.148699
std	0.471398	1.066012	140524.341051	1.436586	1.651788	1.135887	0.843904	0.923426	0.797892	0.645429	...	1.072162	0.842110
min	0.000000	1.000000	110008.000000	1.000000	1.000000	2.000000	1.000000	2.000000	1.000000	1.000000	...	1.000000	2.000000
25%	0.000000	2.000000	122018.000000	3.000000	1.000000	2.000000	1.000000	4.000000	1.000000	1.000000	...	4.000000	4.000000
50%	1.000000	3.000000	201303.000000	4.000000	2.000000	3.000000	1.000000	4.000000	2.000000	1.000000	...	4.000000	4.000000
75%	1.000000	4.000000	201310.000000	5.000000	4.000000	5.000000	2.000000	5.000000	2.000000	1.000000	...	5.000000	5.000000
max	1.000000	5.000000	560037.000000	5.000000	5.000000	5.000000	4.000000	5.000000	3.000000	4.000000	...	5.000000	5.000000

The above output shows the statistics of all the numerical columns in our data set.Next we will check the data types of all the columns using dtypes function as shown in the below output.



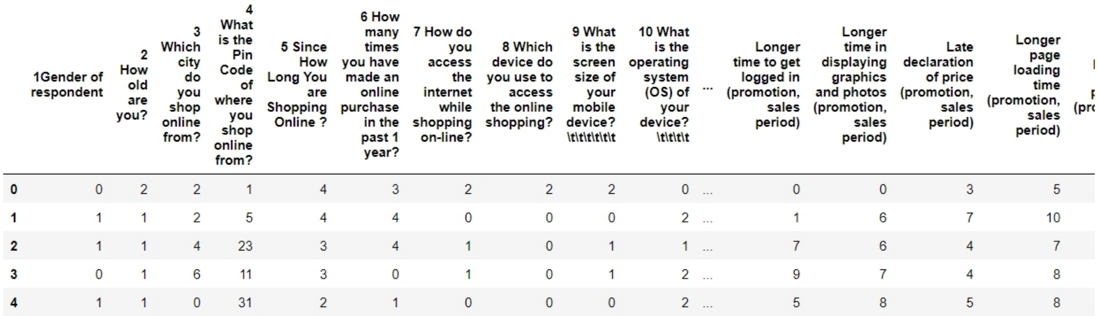
```
df.dtypes

1Gender of respondent                                                  int64
2 How old are you?                                                     int64
3 Which city do you shop online from?                                 object
4 What is the Pin Code of where you shop online from?                  int64
5 Since How Long You are Shopping Online ?                             int64
                                                                        ...
Longer delivery period                                                object
Change in website/Application design                                  object
Frequent disruption when moving from one page to another              object
Website is as efficient as before                                     object
Which of the Indian online retailer would you recommend to a friend?  object
Length: 71, dtype: object
```

As shown in the above output our data set has both numerical and categorical variables.After checking the data types in our data set we will check whether we have any special characters or symbols in our data set which will not be displayed when we used isna() function.So in our data set first let us check whether we have any empty spaces and next we will check for Any ? symbols in our data set.The below codes gives the respective outputs as no special characters and no empty spaces in our data set.



```
for i in df.columns:
    print(i,df.loc[df[i]==" ",i].size)
```

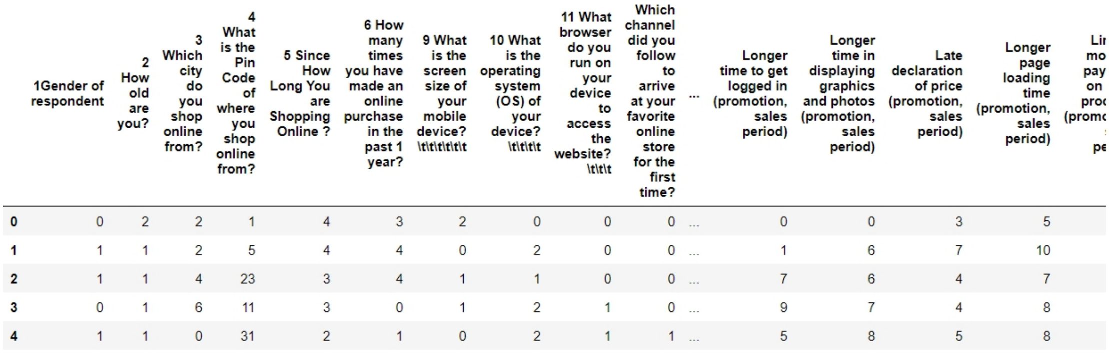
After checking for special characters and empty spaces we will convert all the Categorical columns in to numerical columns By importing Label Encoder.After importing the Label Encoder and fitting our data set all the categorical columns will be converted in to numerical as shown below.



	1Gender of respondent	2 How old are you?	3 Which city do you shop online from?	4 What is the Pin Code of where you shop online from?	5 Since How Long You are Shopping Online ?	6 How many times you have made an online purchase in the past 1 year?	7 How do you access the internet while shopping on-line?	8 Which device do you use to access the online shopping?	9 What is the screen size of your mobile device? \t\t\t\t\t\t	10 What is the operating system (OS) of your device? \t\t\t\t	...	Longer time to get logged in (promotion, sales period)	Longer time in displaying graphics and photos (promotion, sales period)	Late declaration of price (promotion, sales period)	Longer page loading time (promotion, sales period)
0	0	2	2	1	4	3	2	2	2	0	...	0	0	3	5
1	1	1	2	5	4	4	0	0	0	2	...	1	6	7	10
2	1	1	4	23	3	4	1	0	1	1	...	7	6	4	7
3	0	1	6	11	3	0	1	0	1	2	...	9	7	4	8
4	1	1	0	31	2	1	0	0	0	2	...	5	8	5	8

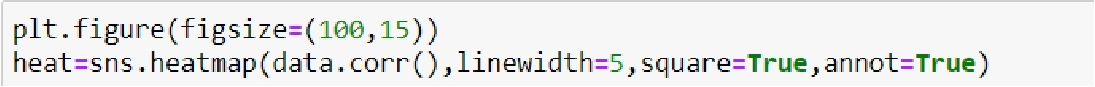
After converting all the columns in to numerical we need to remove the Unnecessary column from our data sets.In our data set the columns How do you access the internet while shopping online?,which device do you use to access the Online shopping and Shopping on the website helps you fulfill certain roles columns does not reflect the major factors of our data set so we can drop them by using the drop function.After dropping the unnecessary columns the below output shows how our data looks

like.



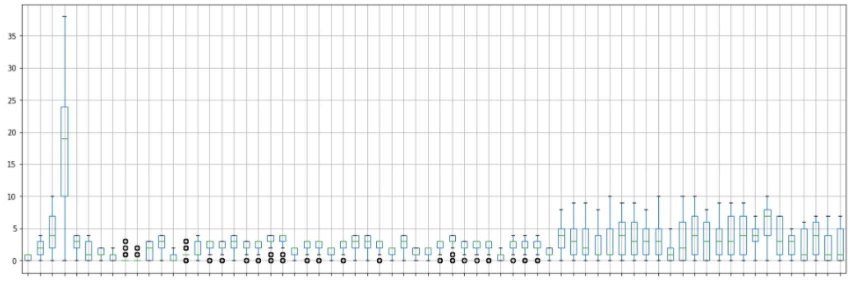
	1Gender of respondent	2 How old are you?	3 Which city do you shop online from?	4 What is the Pin Code of where you shop online from?	5 Since How Long You are Shopping Online ?	6 How many times you have made an online purchase in the past 1 year?	9 What is the screen size of your mobile device? \t\t\t\t\t\t	10 What is the operating system (OS) of your device? \t\t\t\t	11 What browser do you run on your device to access the website? \t\t\t	12 Which channel did you follow to arrive at your favorite online store for the first time?	...	Longer time to get logged in (promotion, sales period)	Longer time in displaying graphics and photos (promotion, sales period)	Late declaration of price (promotion, sales period)	Longer page loading time (promotion, sales period)
0	0	2	2	1	4	3	2	0	0	0	...	0	0	3	5
1	1	1	2	5	4	4	0	2	0	0	...	1	6	7	10
2	1	1	4	23	3	4	1	1	0	0	...	7	6	4	7
3	0	1	6	11	3	0	1	2	1	0	...	9	7	4	8
4	1	1	0	31	2	1	0	2	1	1	...	5	8	5	8

So after dropping the unnecessary columns we will be having 269 rows and 68 columns as we dropped 3 columns from our data set.To find more columns that are not related we will be plotting a heat map for the correlations of the columns.

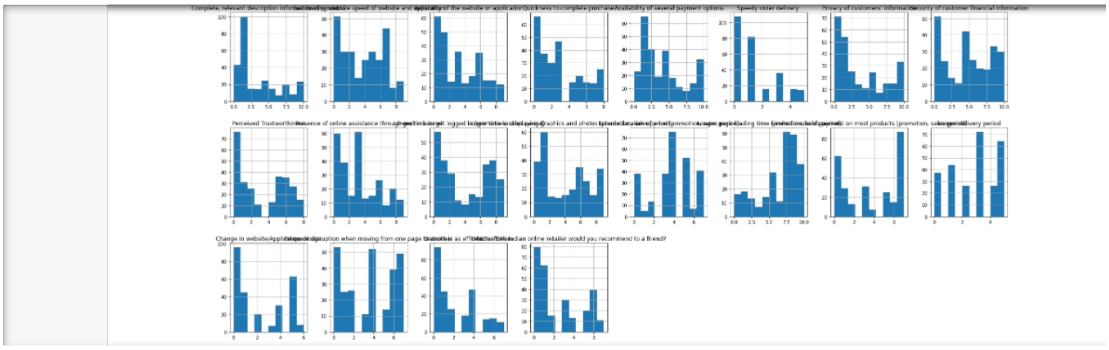


```
plt.figure(figsize=(100,15))
heat=sns.heatmap(data.corr(),linewidth=5,square=True,annot=True)
```

The above code plots the Heat map for all the columns in our data set and after removing the unnecessary columns we will find the Outliers using the Box plots and then we will remove the skewness.The below Output shows the Outliers by using Box plots.



The above output shows that there are outliers in other columns and those are categorical so we wont be removing outliers from them and then we need to check the skewness of our data and for the we will use skew() function we wont remove skewness from our data as we have categorical data. After checking skewness we will plotting a histogram to plot the distribution of data as shown below.



The above output clearly shows that the data is linearly distributed and we successfully Analyzed our data and we can use the above cleaned data that we stored in data frame named “data” to predict the model.