FLIP ROBO

RATINGS PREDICTION

Submitted by:

ASHUTOSH MISHRA

ACKNOWLEDGMENT

I'd like to extend my gratitude to my mentor Mr. Md. Kashif for giving me this opportunity to work upon this project. Below are all the details of the project which I consumed while preparing and drafting the project –

The references, research papers, data sources, professionals, etc are majorly referred from the website- **Amazon**. Also, some of the other resources like "Research Gate" were explored to gain a deep understanding on the assigned subject.

Above stated details stands correct to the best of my knowledge and I hereby acknowledge the same.

INTRODUCTION

Now a days businesses don’t want to rely only on reviews because sometimes customer’s reviews could be small and meaningless so in this scenario rating system could be more informative to them while studying the shopping behavior of the customers and that’s why almost most of the service based and e-commerce industries are having rating attribute on their websites to analyze these pattern smoothly.

During the past several years incorporating ratings and reviews have become a must for every e-commerce businesses. With the abundance of similar products and services, ratings and reviews can distinguish a business from its competitors. According to research commissioned by Deloitte and Touche LLP, a vast majority (82%) of shoppers research products online before going in-store, and over a third read a product’s reviews on their mobile phones while looking at the same product in-store. This not only helps them feel more confident in their decision but can also change how much they’re willing to spend.

Some industries and services, such as Uber or restaurants, rely on good ratings to make a living and be successful. Look at Uber drivers: Surveys show that half of regular Uber commuters will cancel on a driver if their rating score is below 4.5 (out of 5)! For restaurants, the lowest acceptable rating (LAR) is around 3.8 out of 5. These are pretty high standards. Nevertheless, this is what they are, and that's what both Uber drivers and restaurants have to work with. It's for this reason why for many businesses, good ratings make for a favorable business opportunity; if they choose to see it that way. Building and maintaining an A-rating score will likely result in more customer engagement, more sales, and more growth.

Business Problem Framing

With the Covid-19 impact in the market, we have seen lot of changes across industries. If we talk about e-commerce companies they have grown like never before as people were avoiding physical outings for shopping's and all. Hence covid-19 was like a revolution for the e-commerce industries. As they want to keep adding more customers under their umbrella so it becomes really important for them to study and analyze the buying behavior of existing customers with the help of certain attributes.

One of our client who has a website where people write different reviews for technical products. Now they are adding a new feature to their website i.e. the reviewer will have to add stars (rating) as well with the review. The rating consist of 5 stars and it only has 5 options available as- 1 star, 2 stars, 3 stars, 4 stars, 5 stars. Now they want to predict ratings for the reviews which were written in the past and they don't have a rating. So, we have to build a model which can predict the rating by seeing the reviews.

This Project will be done in two parts. In the very first part I'll be doing web scraping in-order to extract at least 20000 rows of data from amazon having different categories like mobiles, headsets, printers, routers, computer and etcetera. I've to scrape for different attributes like Title, Reviews and at last our Target variable; which is Ratings.

After collecting the data, I'll be building a machine learning model and later on will predict our target variable Ratings with the help of other attributes.

Conceptual Background of the Domain Problem

Data science comes as a vital tool to solve business problems to help companies optimize the standards resulting in overall revenue and profits growth. Moreover, it also improves their marketing strategies and demands focus on changing trends.

Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques globally used for achieving the business goals for any entity or an organization.

Currently the assigned project utilizes **Predictive Modelling** algorithm to solve the business statement.

Review of Literature

As, we know that after Covid pandemic most of the industries are trying to overcome their business challenges; so does our client is. This research project was for a client who has a website where people write different reviews for technical products. Now they are adding a new feature to their website i.e. The reviewer will have to add ratings as well with the review. So, they are looking for new machine learning models in-order to predict ratings for the reviews which were written in the past and they don't have a rating initially.

Being a Data Scientist, I have used the **Classification Model** using multiple algorithms to design and optimize the results as the target variable is in the form of 5 different category (1, 2, 3, 4, 5).

Some of the classes which I explored from Scikit-learn Libraries were – Statistics, Analytical Modelling, Hyper Tuning Method, CV Ratings, Predictive Modelling and Classification Model, etc.

In totality, my project comprises of six different test cases of Classification models where the objective was to train the model and test for accuracy score and different classification reports using Cross Validation Score and the testing accuracy of the different algorithms. Since our target variable 'Ratings' consist of 5 different classes hence I will not be using ROC-AUC curve and ROC-AUC score as it support only two binary classes.

By using Selenium Web-Scraping with the help of python IDE I've extracted out 26840 rows of data from **Amazon** website.

I've scarped for 3 attributes and they are- Title, Review and Ratings.

Also, I have observed that there are 1 numeric column whereas 2 categorical columns are there in the dataset; there are also 706 null values are present in the dataset.

As a result of the above analytical modelling, I have managed to achieve below top model, which are given below as following–

Regression Models	**Testing Accuracy** **(in %)**	**CV Score** **(in %)**
Random Forest Regressor	61	42

Motivation for the Problem Undertaken

See, now a days we are often doing shopping's from e-commerce website and are getting used to with it. As the technology getting advanced day by day one can shop his/her favorite things on a tap of a single click but as a buyer I usually take few minutes and go through by some of the reviews and ratings to decide whether to buy it or not, so reviews and rating are most important aspects for the customer point of view. Thus, by doing this project I came to know how it works behind the scene and how these attributes can really impact the businesses.

Hence, this project guided me to baseline each aspect carefully and be concrete on the decision making process regardless it's an individual or an entity like a company.

In order to cater to the above project, my current knowledge and skill set has aided me a lot which I'd explore on this exponentially further.

Analytical Problem Framing

Mathematical Modeling- I have used the below statistical models to find-out the corresponding mean, median, mode and relationship among the variables.

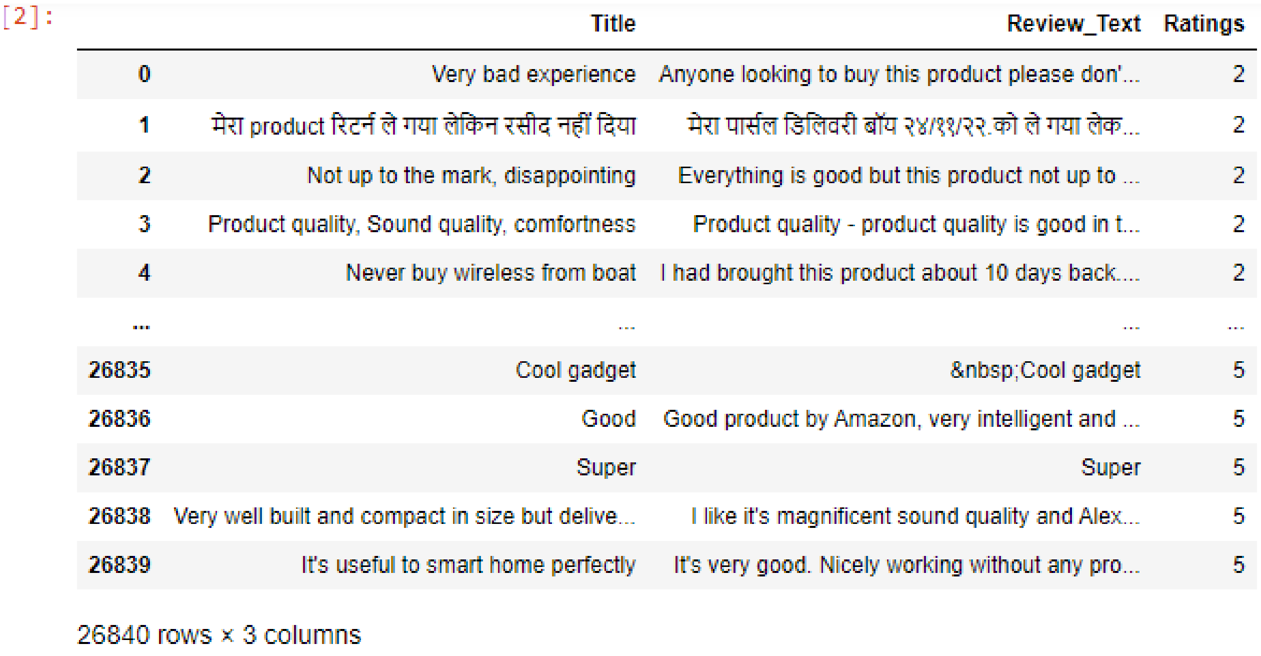
- Descriptive Statistics- To find out the mean, median, mode, percentiles and IQR (**Interquartile Range**)
- Statistical Modeling, Correlation- To find out the relationship among the variables i.e. finding out the positive, negative and zero correlated attributes.
- Multicollinearity- To find out all those variables who are giving similar information to the target variable (‘**Ratings**)
- Skewness- To check whether the attributes are the skewing left hand side or right hand side (Threshold value is=0.5).
- Outliers- To find out variables those are having the value greater than 3.

Analytical Modeling-

- Label Encoder- To convert all the categorical columns into numeric categories
- Simple Imputer-To replace all the null values present in the columns with mean or mode.
- Principle Component Analysis- To remove the curse of dimensionality
- Hyper Parameter Tuning- To find out the best parameter

Data Sources and their formats

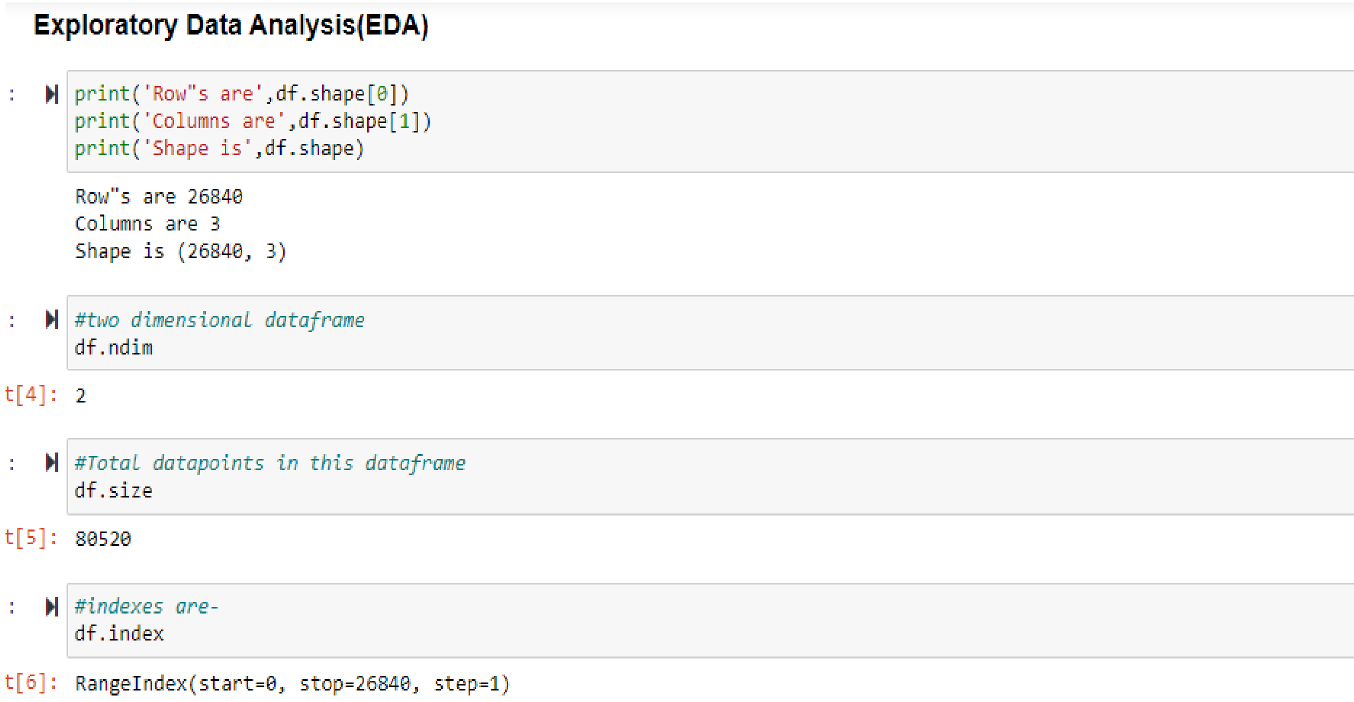
I’ve used selenium web scraping over the **Amazon** website in-order to find out the required attributes data. Below are the attached screenshot for your reference; so that you can check out the insights that were available in the excel file.



[2]:

	Title	Review_Text	Ratings
0	Very bad experience	Anyone looking to buy this product please don'...	2
1	मेरा product रिटर्न ले गया लेकिन रसीद नहीं दिया	मेरा पार्सल डिलिवरी बॉय २४/११/२२.को ले गया लेक...	2
2	Not up to the mark, disappointing	Everything is good but this product not up to ...	2
3	Product quality, Sound quality, comfortness	Product quality - product quality is good in t...	2
4	Never buy wireless from boat	I had brought this product about 10 days back....	2
...	...	...	...
26835	Cool gadget	Cool gadget	5
26836	Good	Good product by Amazon, very intelligent and ...	5
26837	Super	Super	5
26838	Very well built and compact in size but delive...	I like it's magnificent sound quality and Alex...	5
26839	It's useful to smart home perfectly	It's very good. Nicely working without any pro...	5

26840 rows × 3 columns



Exploratory Data Analysis(EDA)

```
print('Row"s are',df.shape[0])
print('Columns are',df.shape[1])
print('Shape is',df.shape)
```

```
Row"s are 26840
Columns are 3
Shape is (26840, 3)
```

```
#two dimensional dataframe
df.ndim
```

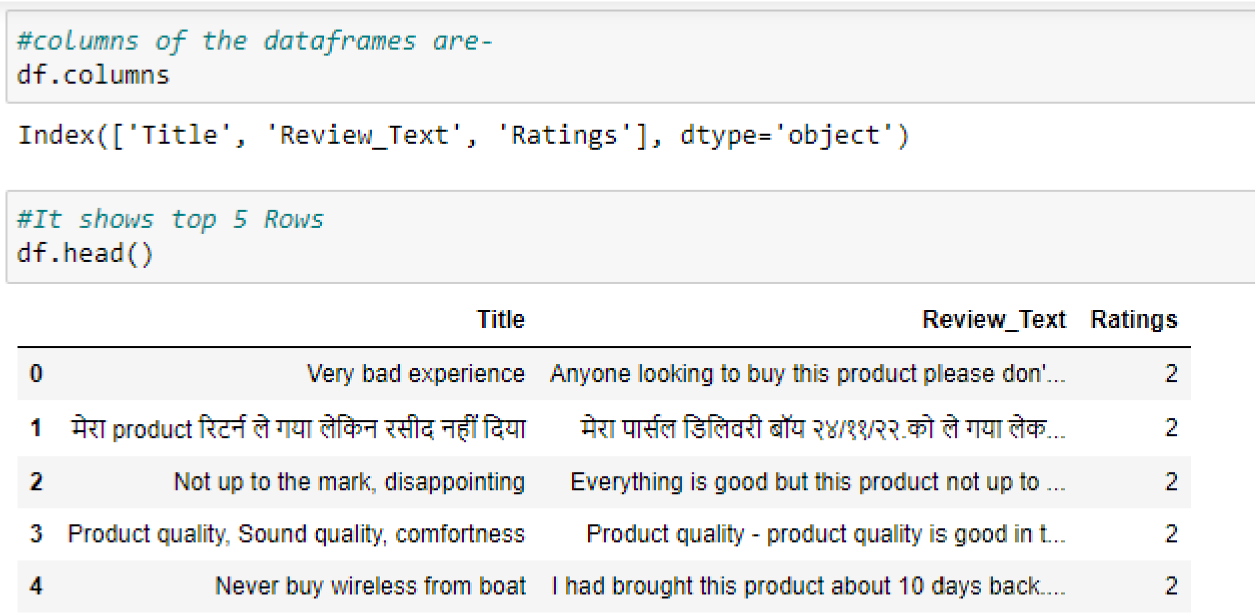
```
t[4]: 2
```

```
#Total datapoints in this dataframe
df.size
```

```
t[5]: 80520
```

```
#indexes are-
df.index
```

```
t[6]: RangeIndex(start=0, stop=26840, step=1)
```

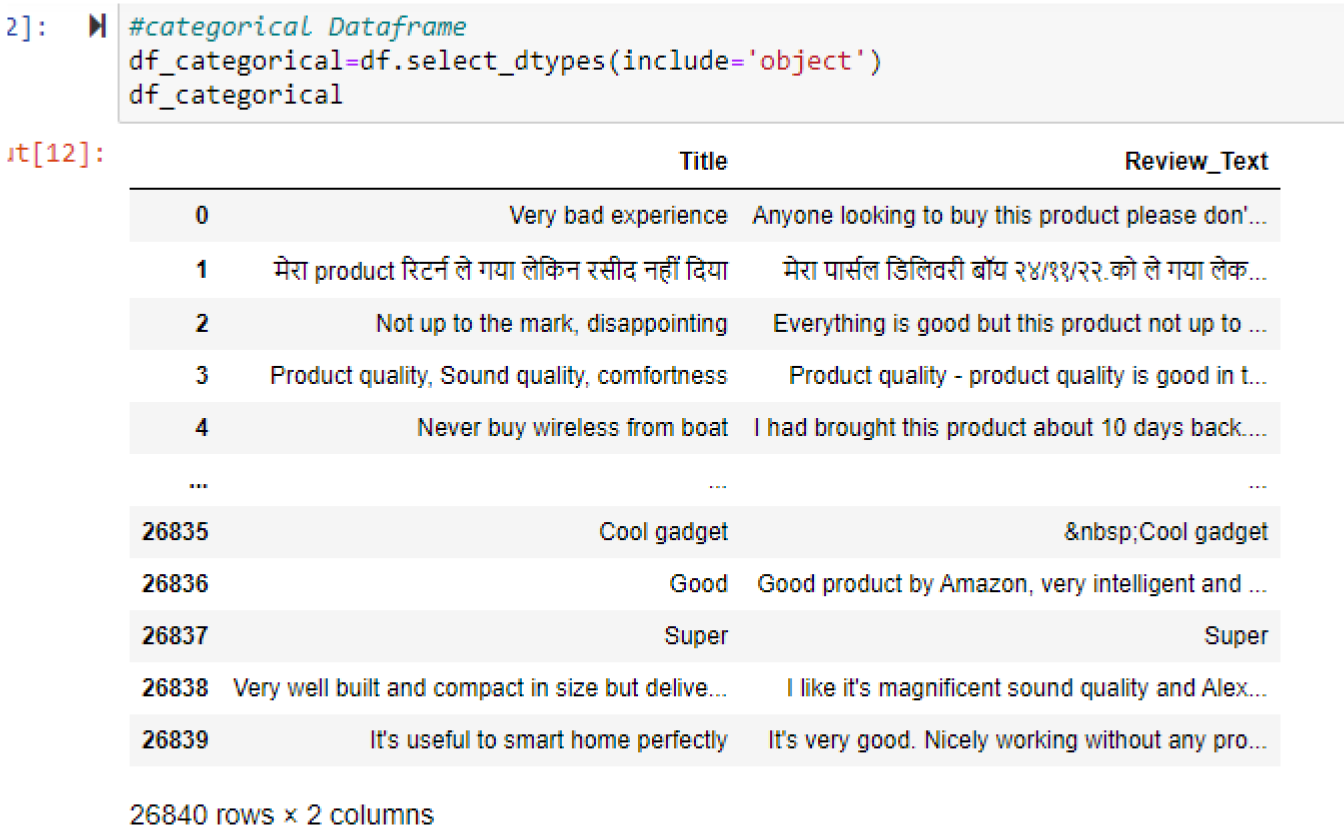


```
#columns of the dataframes are-
df.columns
```

```
Index(['Title', 'Review_Text', 'Ratings'], dtype='object')
```

```
#It shows top 5 Rows
df.head()
```

	Title	Review_Text	Ratings
0	Very bad experience	Anyone looking to buy this product please don'...	2
1	मेरा product रिटर्न ले गया लेकिन रसीद नहीं दिया	मेरा पार्सल डिलिवरी बॉय २४/११/२२.को ले गया लेक...	2
2	Not up to the mark, disappointing	Everything is good but this product not up to ...	2
3	Product quality, Sound quality, comfortness	Product quality - product quality is good in t...	2
4	Never buy wireless from boat	I had brought this product about 10 days back....	2

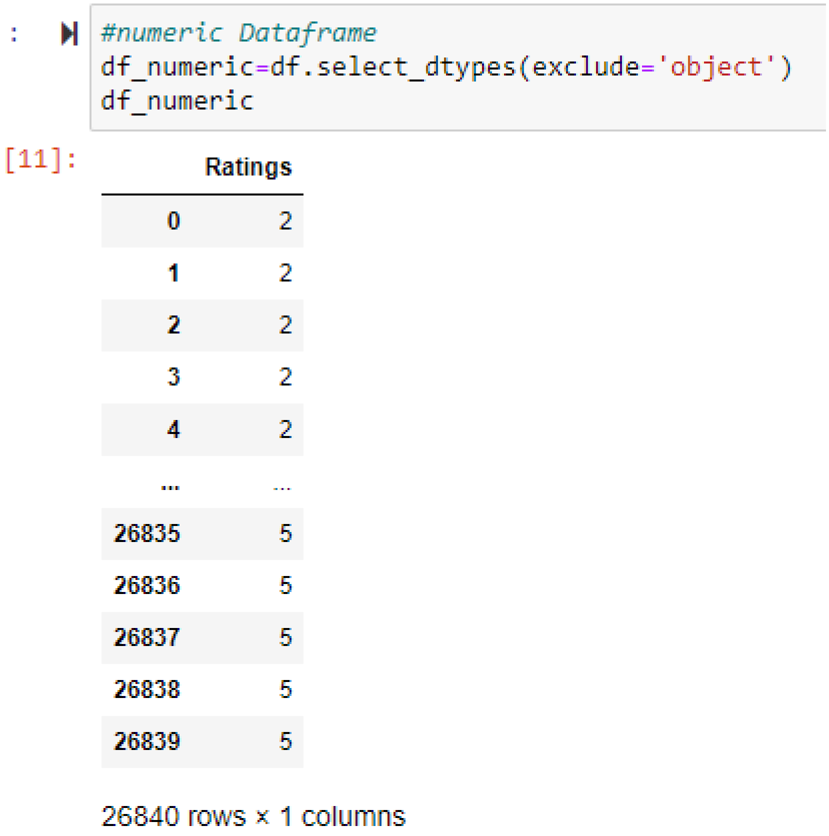


```
2]:  #categorical Dataframe
     df_categorical=df.select_dtypes(include='object')
     df_categorical
```

ıt[12]:

	Title	Review_Text
0	Very bad experience	Anyone looking to buy this product please don'...
1	मेरा product रिटर्न ले गया लेकिन रसीद नहीं दिया	मेरा पार्सल डिलिवरी बॉय २४/११/२२.को ले गया लेक...
2	Not up to the mark, disappointing	Everything is good but this product not up to ...
3	Product quality, Sound quality, comfortness	Product quality - product quality is good in t...
4	Never buy wireless from boat	I had brought this product about 10 days back....
...	...	...
26835	Cool gadget	Cool gadget
26836	Good	Good product by Amazon, very intelligent and ...
26837	Super	Super
26838	Very well built and compact in size but delive...	I like it's magnificent sound quality and Alex...
26839	It's useful to smart home perfectly	It's very good. Nicely working without any pro...

26840 rows × 2 columns

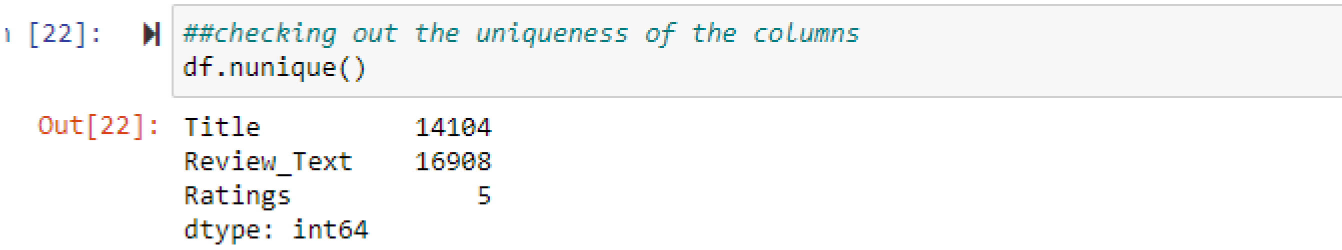


```
:  #numeric Dataframe
   df_numeric=df.select_dtypes(exclude='object')
   df_numeric
```

[11]:

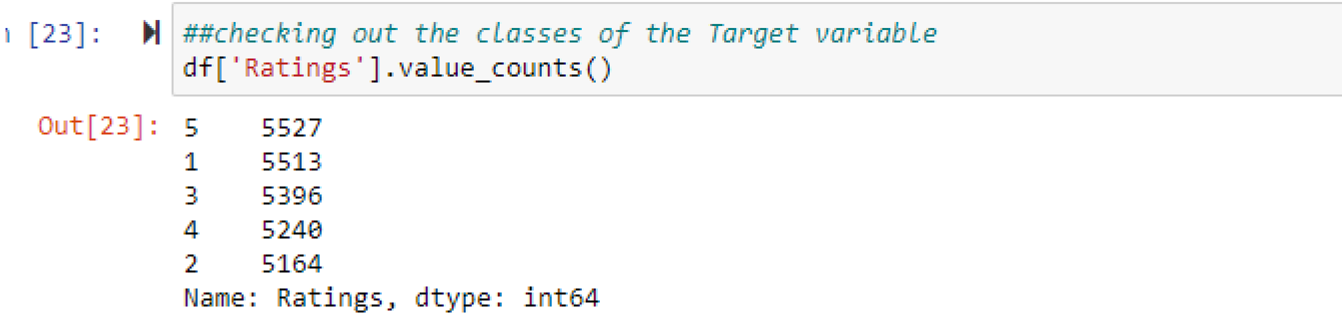
	Ratings
0	2
1	2
2	2
3	2
4	2
...	...
26835	5
26836	5
26837	5
26838	5
26839	5

26840 rows × 1 columns



```
[22]:  ##checking out the uniqueness of the columns
       df.nunique()

Out[22]: Title          14104
         Review_Text    16908
         Ratings            5
         dtype: int64
```



```
[23]:  ##checking out the classes of the Target variable
       df['Ratings'].value_counts()

Out[23]: 5    5527
         1    5513
         3    5396
         4    5240
         2    5164
         Name: Ratings, dtype: int64
```

Data Preprocessing Done

See, both the feature attributes are having very less negative correlation w.r.t 'Ratings' (which is our target variable). I'll not be removing any of the attributes as there is no attribute present having zero correlation. Negative correlation means if input is +ve then output would be -ve and vice-versa whereas, Positive correlation means if input is +ve then output would also be +ve and vice-versa.

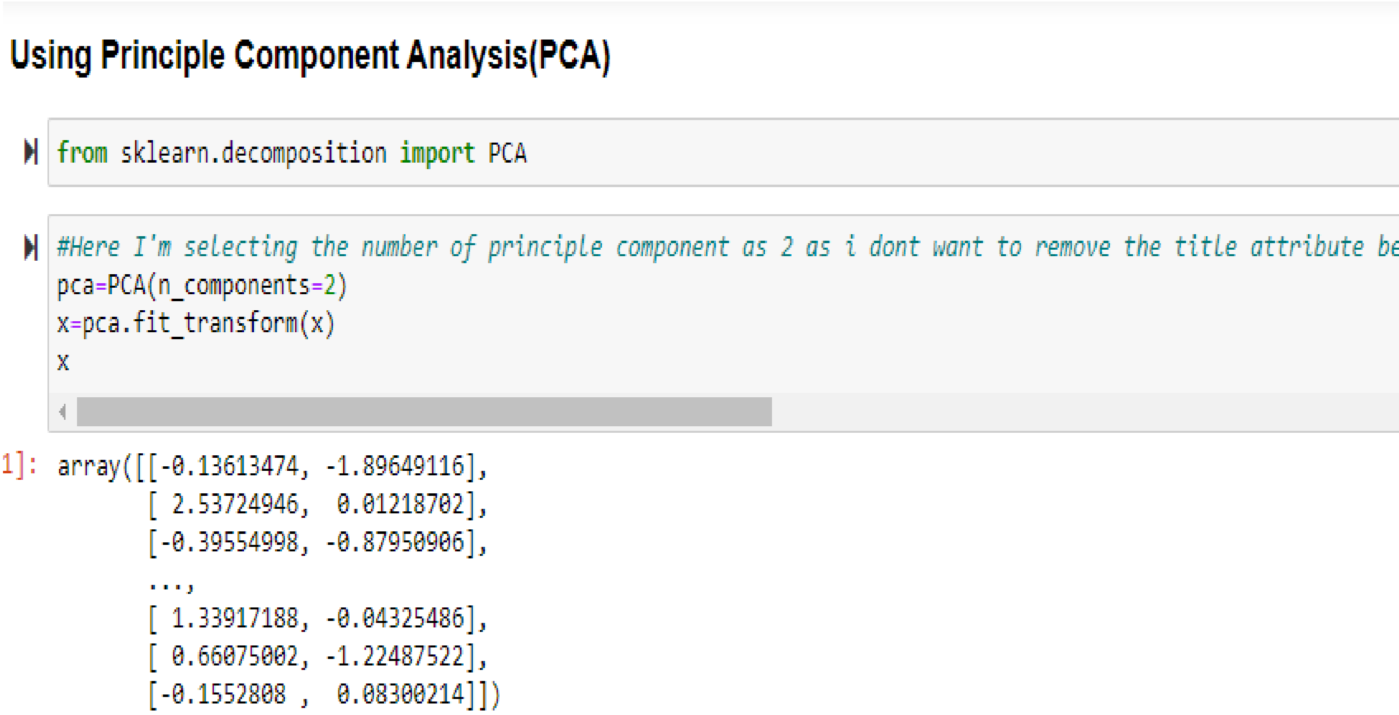
Also, I've used Label Encoder method to convert all the string object type variables into numeric one and performed simple Imputer to replace all the null values present in rating and title columns w.r.t. most frequent value respectively.

The threshold value of Skewness is +/=0.5 and all the attributes are in the range and there is no skewness present in the dataset, also there are no outliers are present as our dataset is almost equally distributed. If we talk about Multicollinearity then we can say that no two variables are highly correlated with each other hence there is no Multicollinearity present in the dataset.

After analyzing above data frames what I can observe is that there are lots of reviews having same meaning but some are written in uppercase and some are in the lowercase, so I'm going to replace the similar words with a single parameter and that would definitely reduce the noise (if present) in the dataset.

I’ve used the Principle Component Analysis (PCA) method in-order to reduce the dimensionality into two principle components. Initially there were two feature variables presented in the data frame and with the help of PCA I’ve converted into two Principle Components for more accuracy.

Kindly see the below screenshots for your reference-

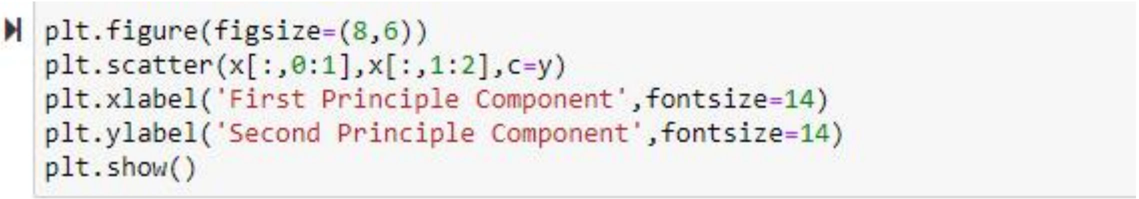


Using Principle Component Analysis(PCA)

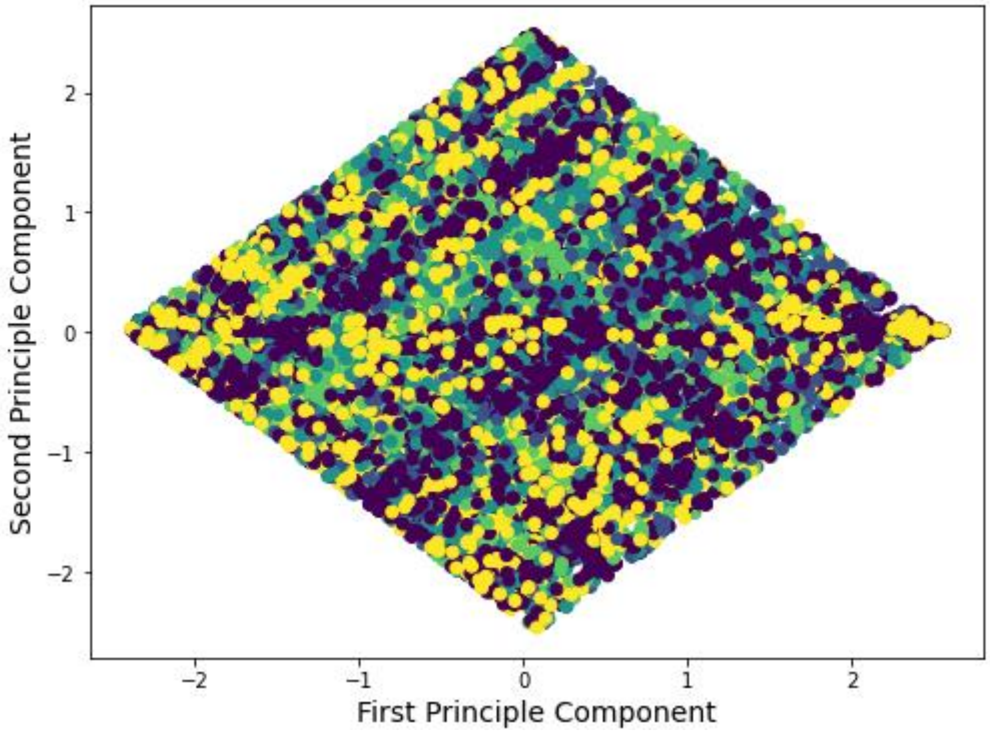
```
from sklearn.decomposition import PCA
```

```
#Here I'm selecting the number of principle component as 2 as i dont want to remove the title attribute be
pca=PCA(n_components=2)
x=pca.fit_transform(x)
x
```

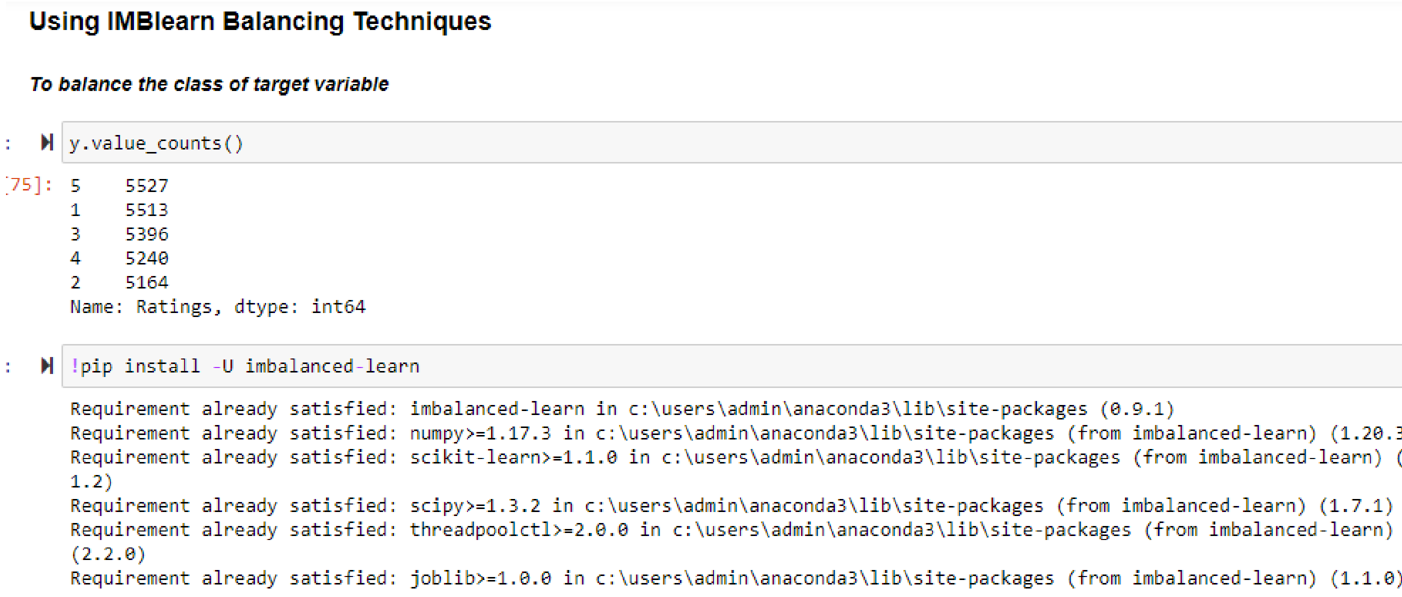
```
1]: array([[-0.13613474, -1.89649116],
           [ 2.53724946,  0.01218702],
           [-0.39554998, -0.87950906],
           ...,
           [ 1.33917188, -0.04325486],
           [ 0.66075002, -1.22487522],
           [-0.1552808 ,  0.08300214]])
```



```
plt.figure(figsize=(8,6))
plt.scatter(x[:,0:1],x[:,1:2],c=y)
plt.xlabel('First Principle Component',fontsize=14)
plt.ylabel('Second Principle Component',fontsize=14)
plt.show()
```



Above graph indicates that our two principle components are almost equally distributed and which is a good sign. At the end I've used Imblearn Balancing technique to convert the unbalanced classes of our target variable 'Ratings' into equal ratio of classes. As you can see in the below attached screenshot the classes are imbalanced in nature.



Using IMBlearn Balancing Techniques

To balance the class of target variable

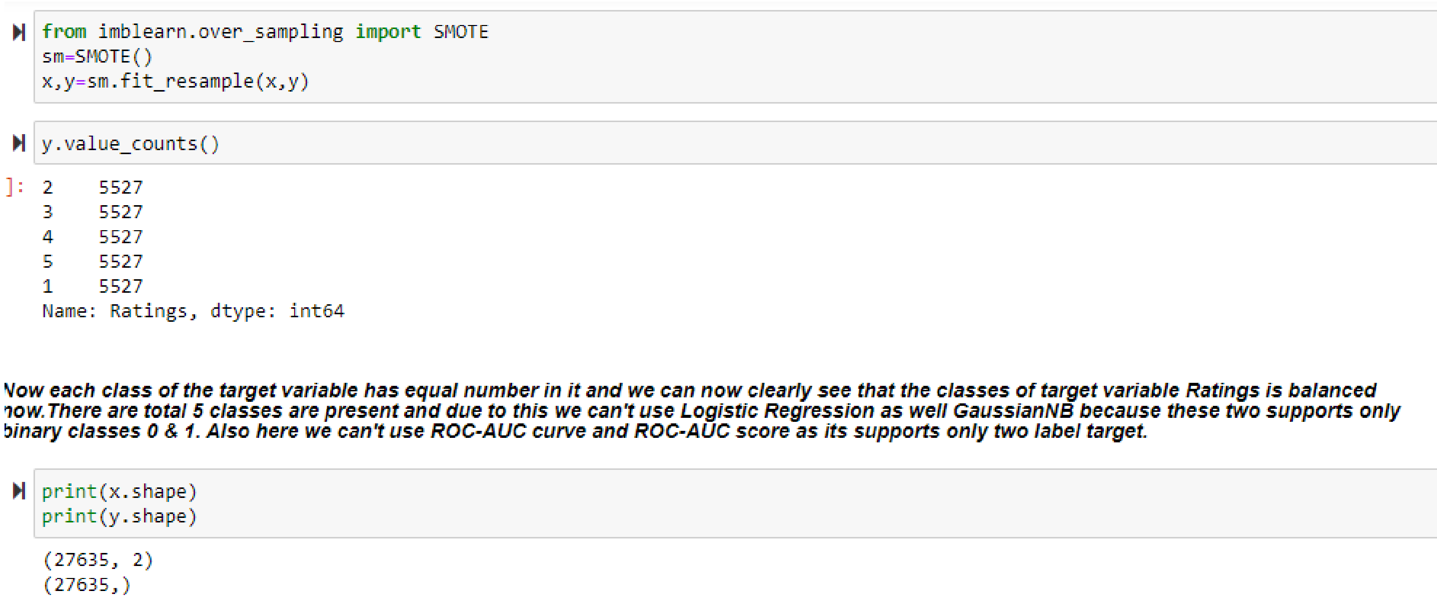
```
y.value_counts()
```

```
5    5527
1    5513
3    5396
4    5240
2    5164
Name: Ratings, dtype: int64
```

```
!pip install -U imbalanced-learn
```

```
Requirement already satisfied: imbalanced-learn in c:\users\admin\anaconda3\lib\site-packages (0.9.1)
Requirement already satisfied: numpy>=1.17.3 in c:\users\admin\anaconda3\lib\site-packages (from imbalanced-learn) (1.20.3
Requirement already satisfied: scikit-learn>=1.1.0 in c:\users\admin\anaconda3\lib\site-packages (from imbalanced-learn) (
1.2)
Requirement already satisfied: scipy>=1.3.2 in c:\users\admin\anaconda3\lib\site-packages (from imbalanced-learn) (1.7.1)
Requirement already satisfied: threadpoolctl>=2.0.0 in c:\users\admin\anaconda3\lib\site-packages (from imbalanced-learn)
(2.2.0)
Requirement already satisfied: joblib>=1.0.0 in c:\users\admin\anaconda3\lib\site-packages (from imbalanced-learn) (1.1.0)
```

After performing the Synthetic Minority Oversampling Technique (SMOTE) we've increased the number of cases in our label in a balanced way as you can see in the below attached screenshot.



```
from imblearn.over_sampling import SMOTE
sm=SMOTE()
x,y=sm.fit_resample(x,y)
```

```
y.value_counts()
```

```
2    5527
3    5527
4    5527
5    5527
1    5527
Name: Ratings, dtype: int64
```

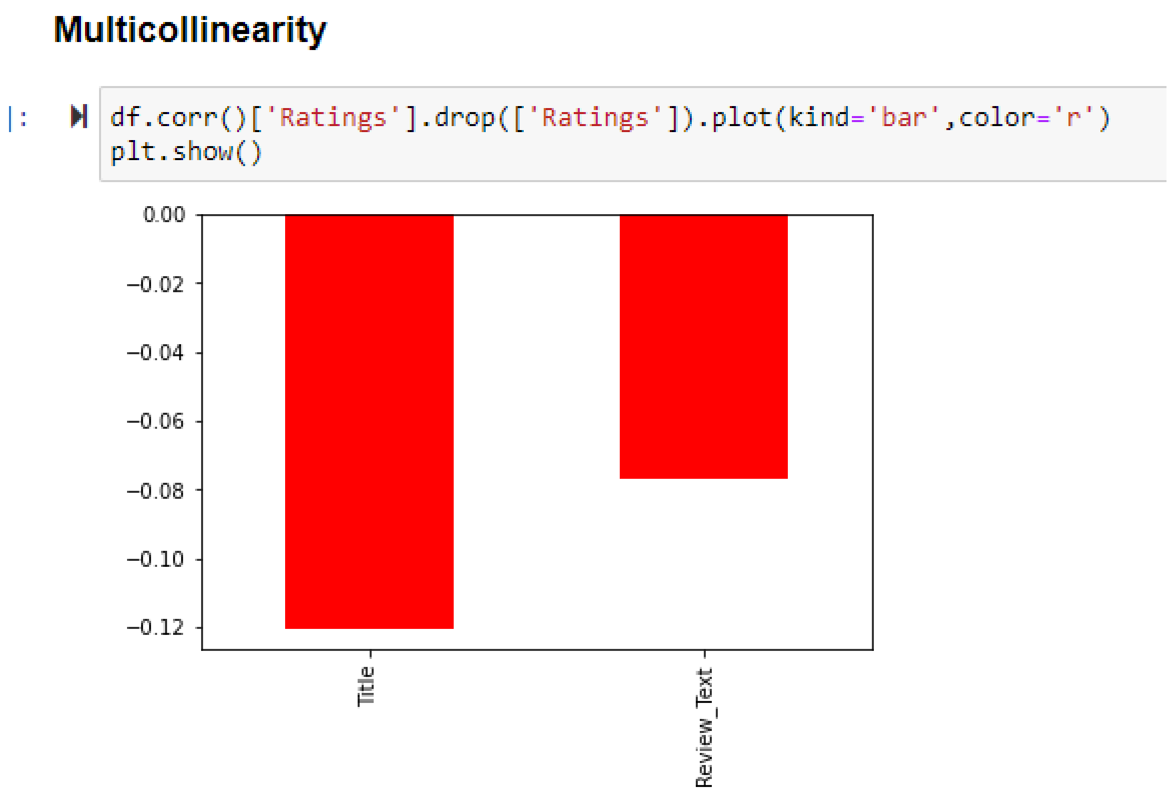
Now each class of the target variable has equal number in it and we can now clearly see that the classes of target variable Ratings is balanced now.There are total 5 classes are present and due to this we can't use Logistic Regression as well GaussianNB because these two supports only binary classes 0 & 1. Also here we can't use ROC-AUC curve and ROC-AUC score as its supports only two label target.

```
print(x.shape)
print(y.shape)
```

```
(27635, 2)
(27635,)
```

Data Inputs- Logic- Output Relationships

Since, as mentioned in our problem statement we've to create model to predict ratings with the help of input variable reviews. So on the basis of output I have checked the correlational input data with my output, as shown in below figure.



Multicollinearity

```
df.corr()['Ratings'].drop(['Ratings']).plot(kind='bar',color='r')
plt.show()
```

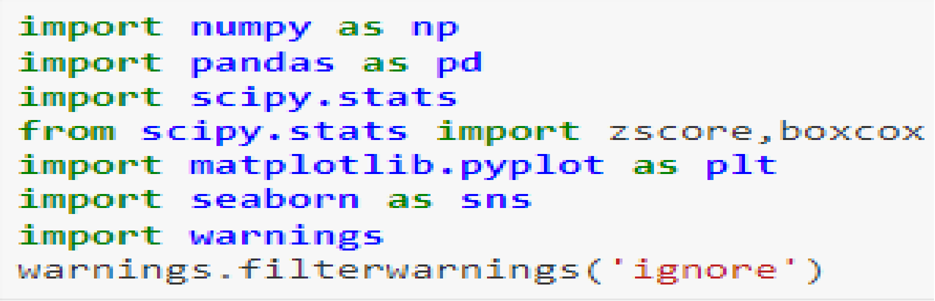
As I checked the input and found that almost all the data is negatively correlated with my output data. Attributes 'Title and Review text' are negatively correlated with my output data.

State the set of assumptions (if any) related to the problem under consideration

No.

Hardware and Software Requirements and Tools Used

I have used **Python IDE** (Integrated Development environment) as a dedicated software throughout solving this project.



```
import numpy as np
import pandas as pd
import scipy.stats
from scipy.stats import zscore,boxcox
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

Python Libraries that I've used throughout the process are-

- Numpy- It is use for linear algebra
- Pandas- data analysis/manipulation library
- Scipy- Utility function for optimization
- Matplotlib-Data visualization and plotting library
- Seaborn- Data visualization and statistical plotting library
- Sklearn- Machine Learning Tool
- Imblearn- Deal with classification problems of Imbalanced classes
- Statsmodels-Deal with advanced statistics

Classes-

- Label Encoder- Encoding the categorical variables into number category
- Simple Imputer- Replacing the null values with mean, median or mode
- Variance_Inflation_Factor- Calculate Multicollinearity
- Power Transformer- Remove skewness
- Standard Scaler- Normalize the feature variables
- Principle Component Analysis- Reduce the dimension of the data frame
- Synthetic Minority Oversampling Technique (SMOTE)
- Cross_Val_Score- CV score
- Grid Search CV- Find out the best parameters for the model

Model/s Development and Evaluation

Identification of possible problem-solving approaches (methods)

Statistical Method-

When I used the describe function then I find out that attribute 'Title' has more median than its mean and so it indicates that there is the possibility of left skewed data in the dataset and the interquartile range for the variables **Titles** and **Reviews** are varying slightly hence it shows that some variable can skew left hand side and it indicates that some of the variables might not be normally distributed.

Also, I have used correlation method to check what are the variables that are giving strong correlation w.r.t Target variable **Ratings**.

Analytical Method-

I've uses Boxplot, Scatter Plots and Distribution Plots to check the outliers and skewness of the variables respectively through the plotting's.

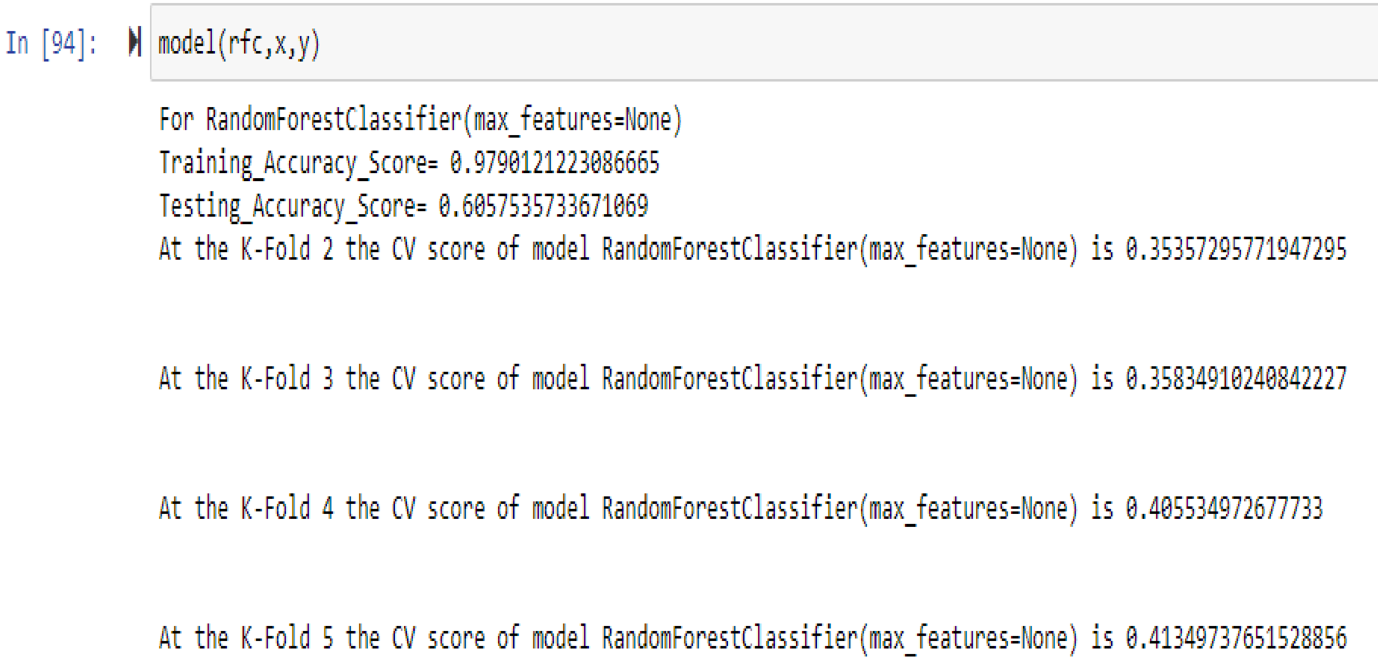
Testing of Identified Approaches (Algorithms)

Listing down all the algorithms used for the training and testing-

- from sklearn.linear_model import KNeighborsClassifier
- from sklearn.model_selection import train_test_split
- from sklearn.metrics import accuracy_score,classification_report,confusion_matrix from sklearn.model_selection import cross_val_score
- from sklearn.model_selection import GridSearchCV
- from sklearn.tree import DecisionTreeClassifier
- from sklearn.ensemble import RandomForestClassifier,AdaBoostClassifier,GradientBoostingClassifier
- from sklearn.linear_model import SGDClassifier
- neighbor=KNeighborsClassifier()
- from sklearn.naive_bayes import MultinomialNB
- dtc=DecisionTreeClassifier()
- rfc=RandomForestClassifier()
- ad=AdaBoostClassifier()
- grd=GradientBoostingClassifier()
- sgd=SGDClassifier()
- mnb=MultinomialNB()

Run and Evaluate selected models

Best Model (Random Forest Classifier)



```
In [94]:  model(rfc,x,y)
```

```
For RandomForestClassifier(max_features=None)
Training_Accuracy_Score= 0.9790121223086665
Testing_Accuracy_Score= 0.6057535733671069
At the K-Fold 2 the CV score of model RandomForestClassifier(max_features=None) is 0.35357295771947295


At the K-Fold 3 the CV score of model RandomForestClassifier(max_features=None) is 0.35834910240842227


At the K-Fold 4 the CV score of model RandomForestClassifier(max_features=None) is 0.405534972677733


At the K-Fold 5 the CV score of model RandomForestClassifier(max_features=None) is 0.41349737651528856
```

Key Metrics for success in solving problem under consideration

Below is the best **Classification Models** where I used below metrics -

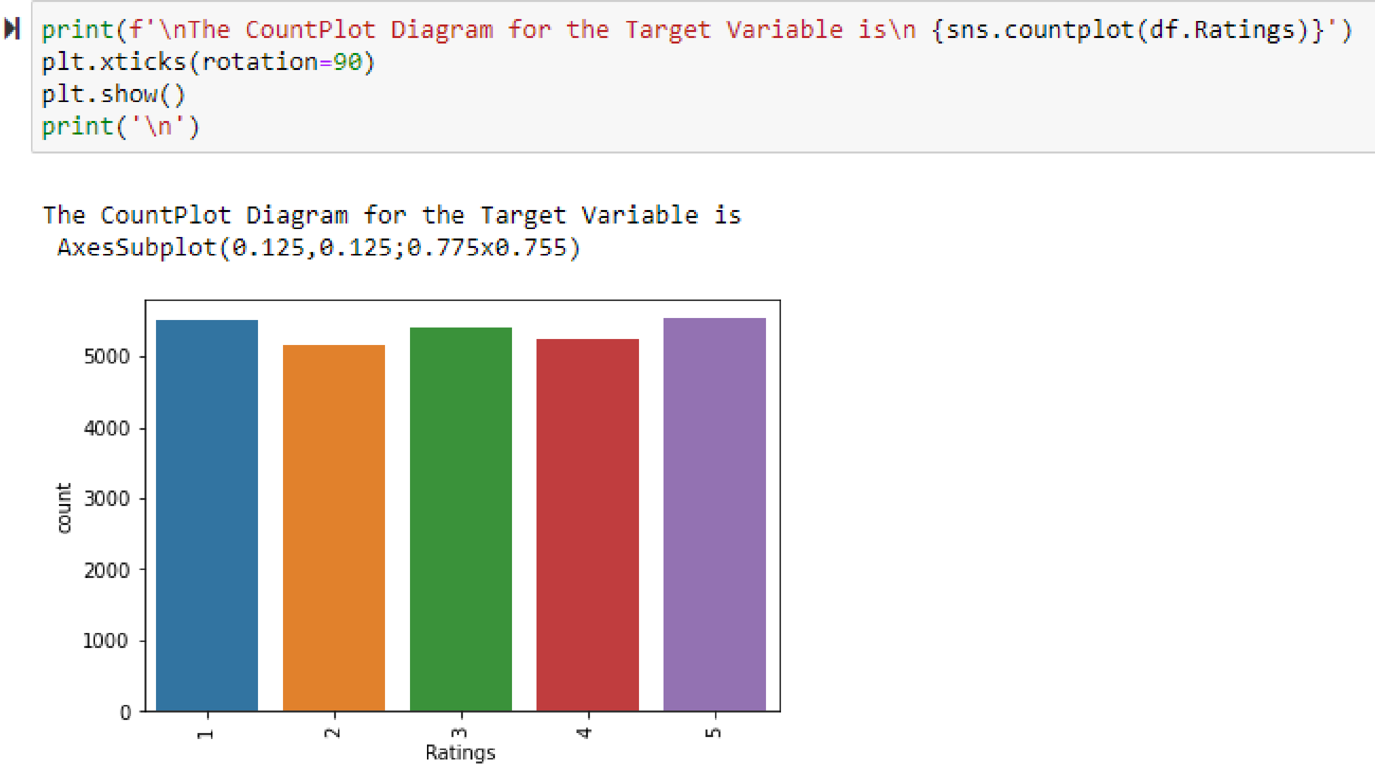
Classification Models	Testing Accuracy (in %)
Random Forest Classifier	61
K Neighbors Classifier	46
Support Vector Classifier	28
Decision Tree Classifier	59
Ada Boost Classifier	25
Gradient Boosting Classifier	33

Visualizations

The plots are –

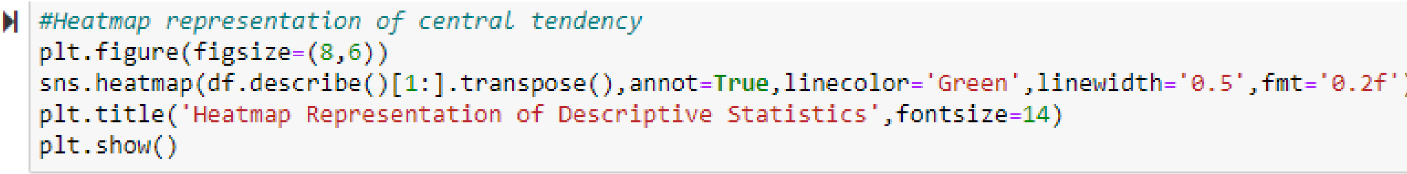
- Count Plot
- Pearson Correlation Heat-map
- Heat-map for Null Values
- Distribution plot, box plot, Violin Plot, Pair-Plot etc.

1.Countplot

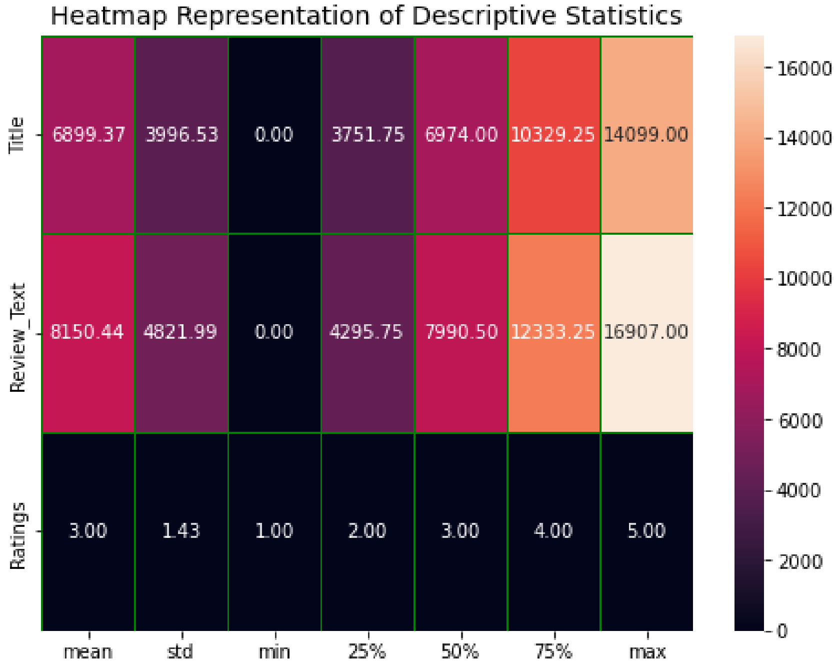


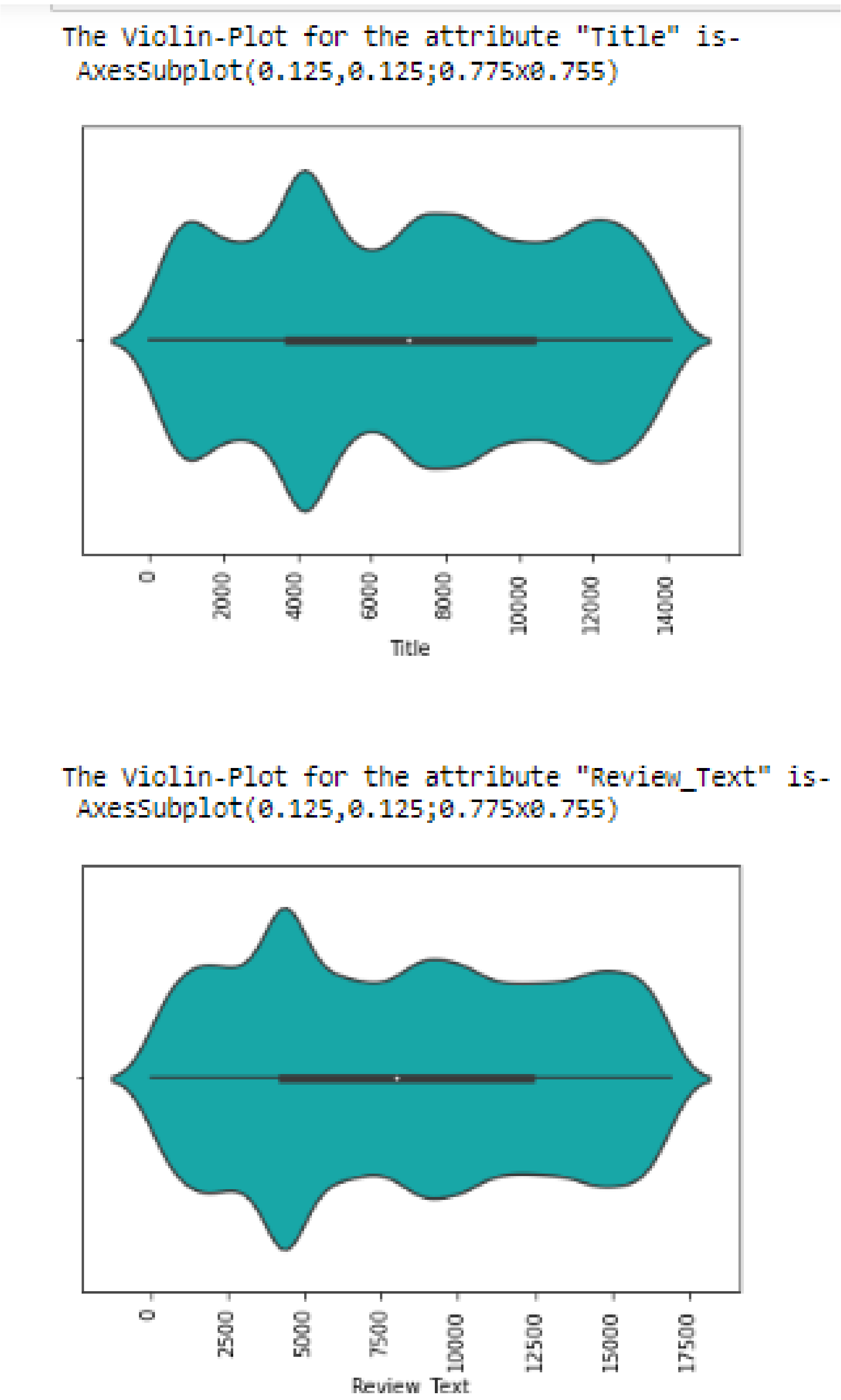
```
print(f'\nThe CountPlot Diagram for the Target Variable is\n {sns.countplot(df.Ratings)}')
plt.xticks(rotation=90)
plt.show()
print('\n')
```

```
The CountPlot Diagram for the Target Variable is
 AxesSubplot(0.125,0.125;0.775x0.755)
```



```
#Heatmap representation of central tendency
plt.figure(figsize=(8,6))
sns.heatmap(df.describe()[1:].transpose(),annot=True,linecolor='Green',linewidth='0.5',fmt='0.2f'
plt.title('Heatmap Representation of Descriptive Statistics',fontsize=14)
plt.show()
```

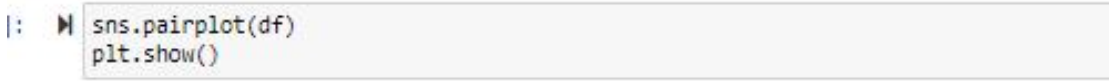




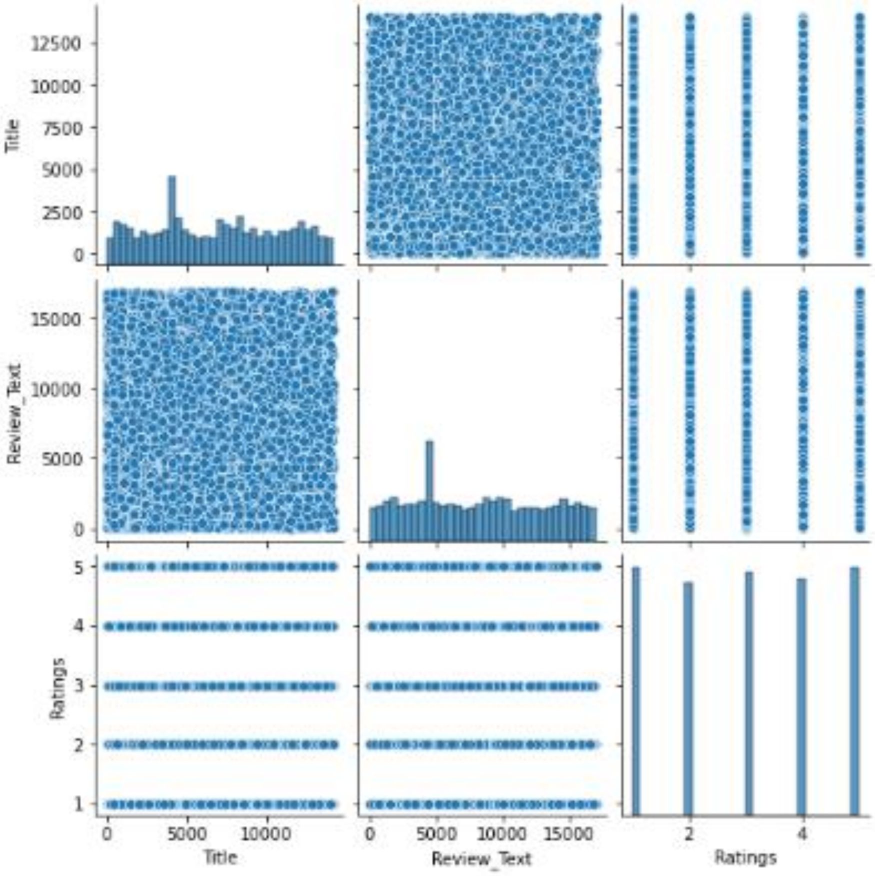
```
The Violin-Plot for the attribute "Title" is-
 AxesSubplot(0.125,0.125;0.775x0.755)
```

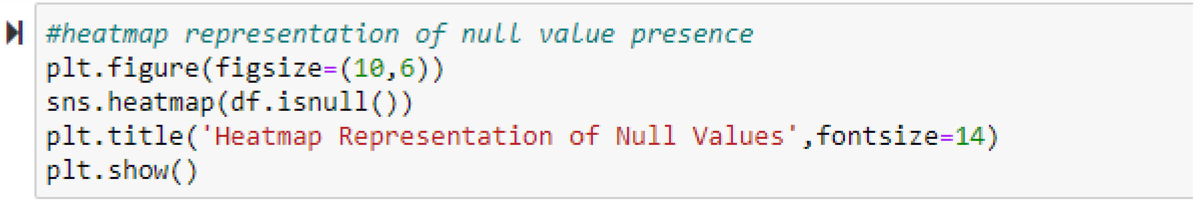
```
The Violin-Plot for the attribute "Review_Text" is-
 AxesSubplot(0.125,0.125;0.775x0.755)
```

3.Pairplot

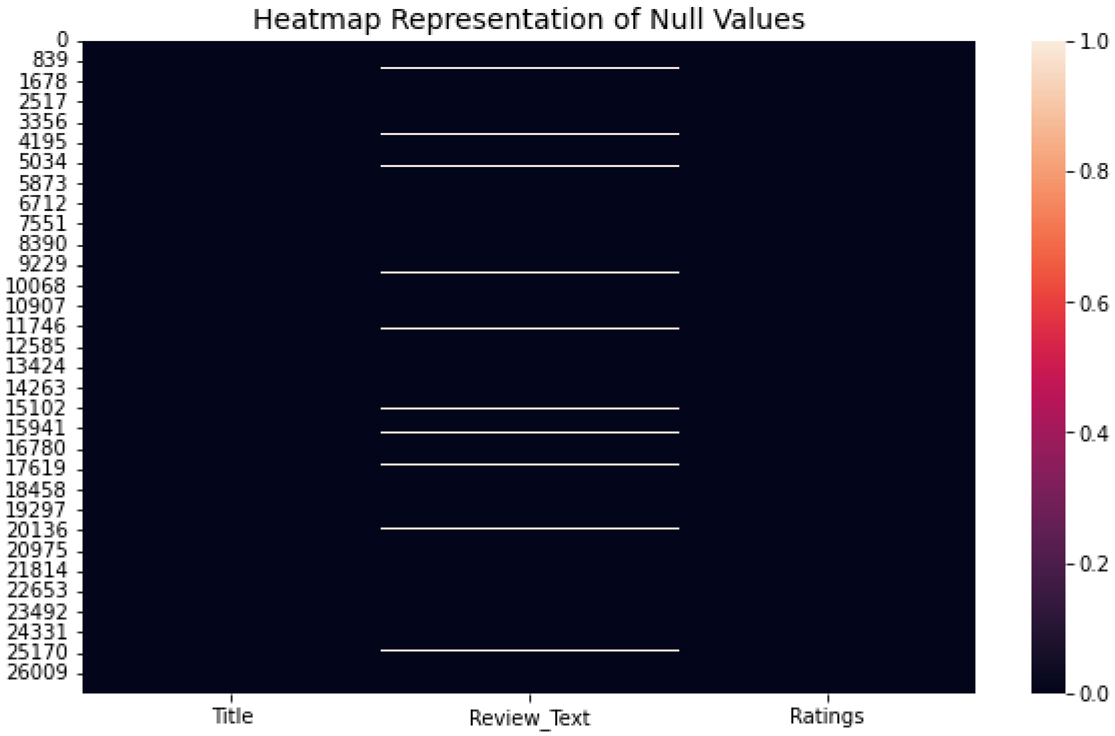


```
sns.pairplot(df)
plt.show()
```

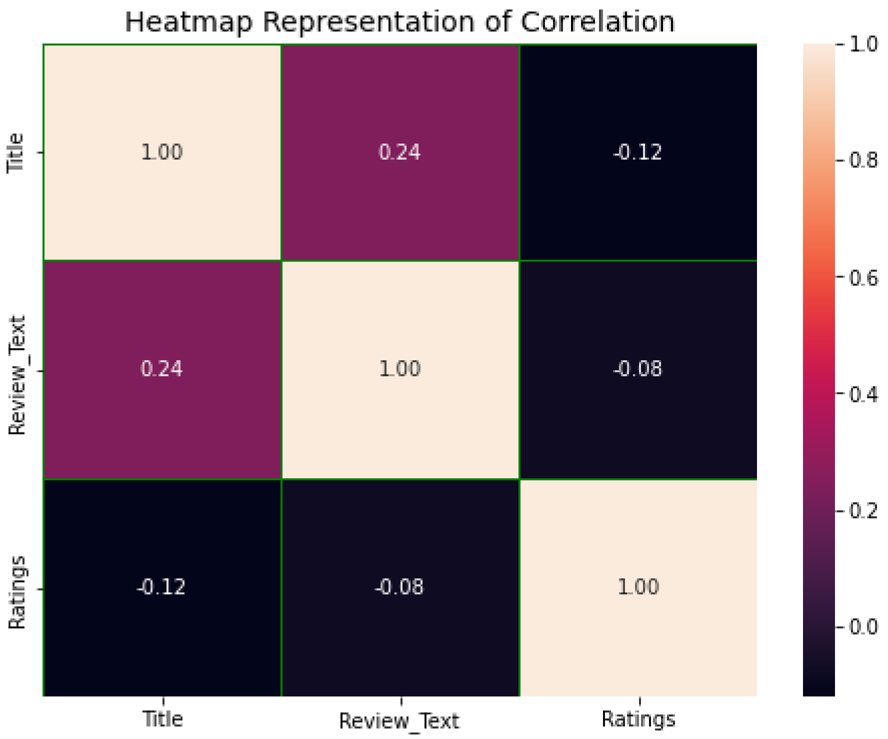




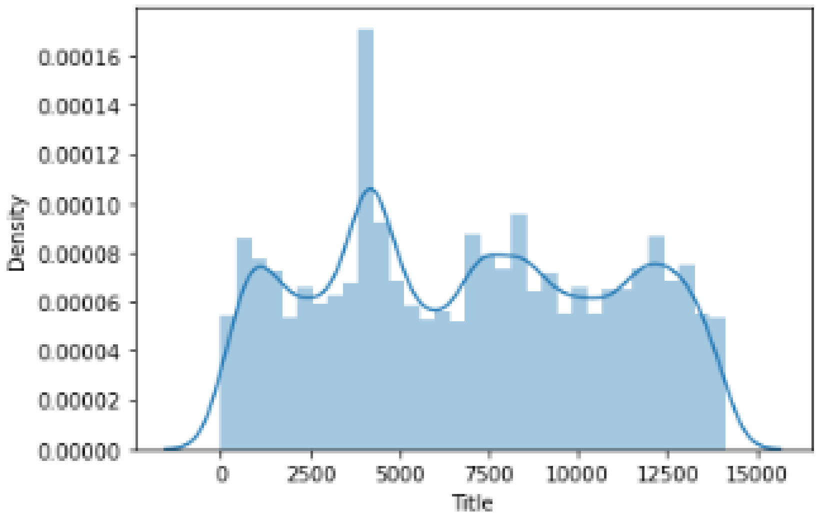
```
#heatmap representation of null value presence
plt.figure(figsize=(10,6))
sns.heatmap(df.isnull())
plt.title('Heatmap Representation of Null Values',fontsize=14)
plt.show()
```



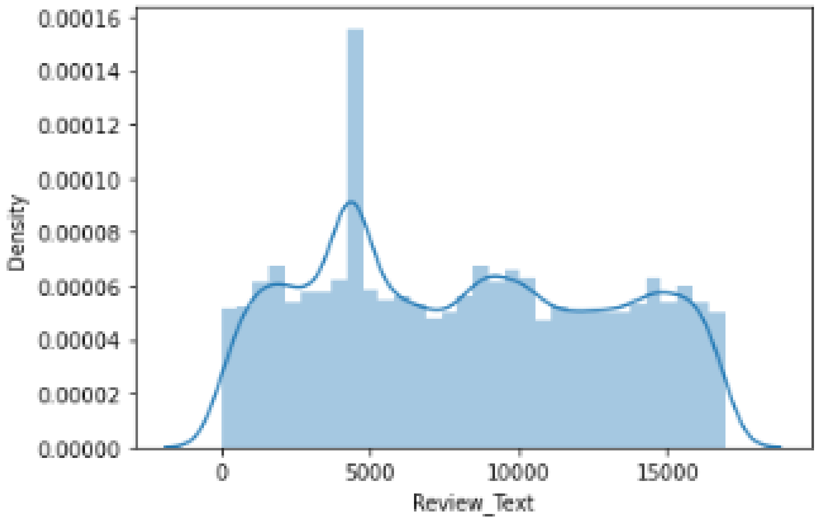
```
plt.figure(figsize=(8,6))
sns.heatmap(df.corr(),annot=True,linecolor='green',linewidth='1',fmt='0.2f')
plt.title('Heatmap Representation of Correlation',fontsize=14)
plt.show()
```

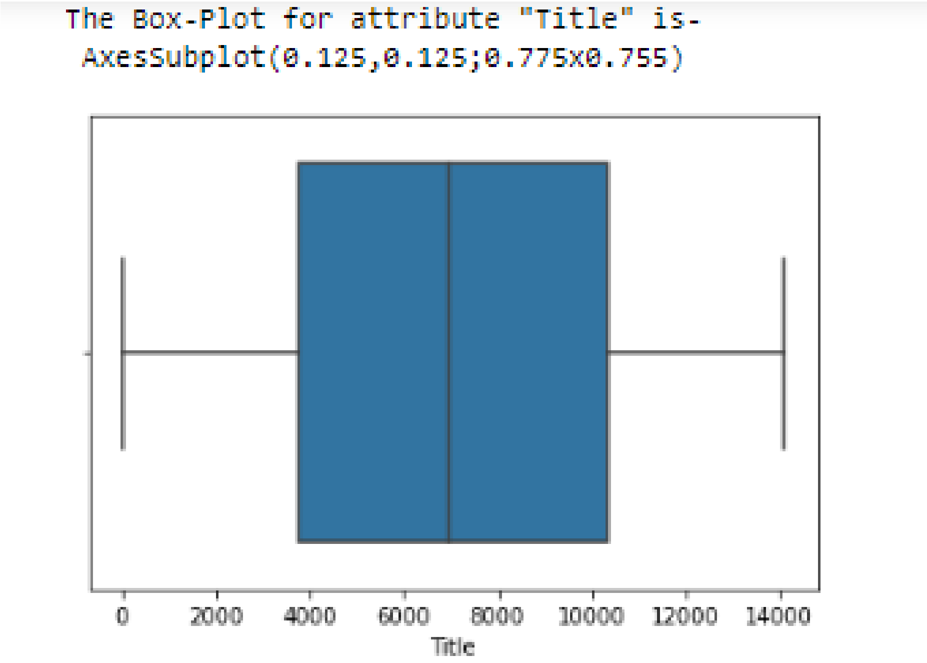


```
The Distribution Plot for attribute "Title" is-
 AxesSubplot(0.125,0.125;0.775x0.755)
```

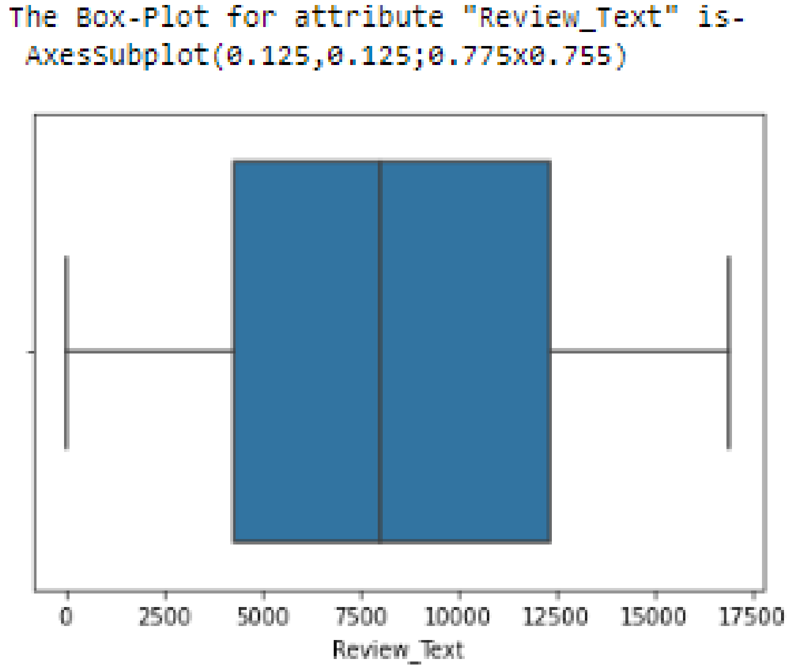


```
The Distribution Plot for attribute "Review_Text" is-
 AxesSubplot(0.125,0.125;0.775x0.755)
```





```
The Box-Plot for attribute "Title" is-
 AxesSubplot(0.125,0.125;0.775x0.755)
```



```
The Box-Plot for attribute "Review_Text" is-
 AxesSubplot(0.125,0.125;0.775x0.755)
```

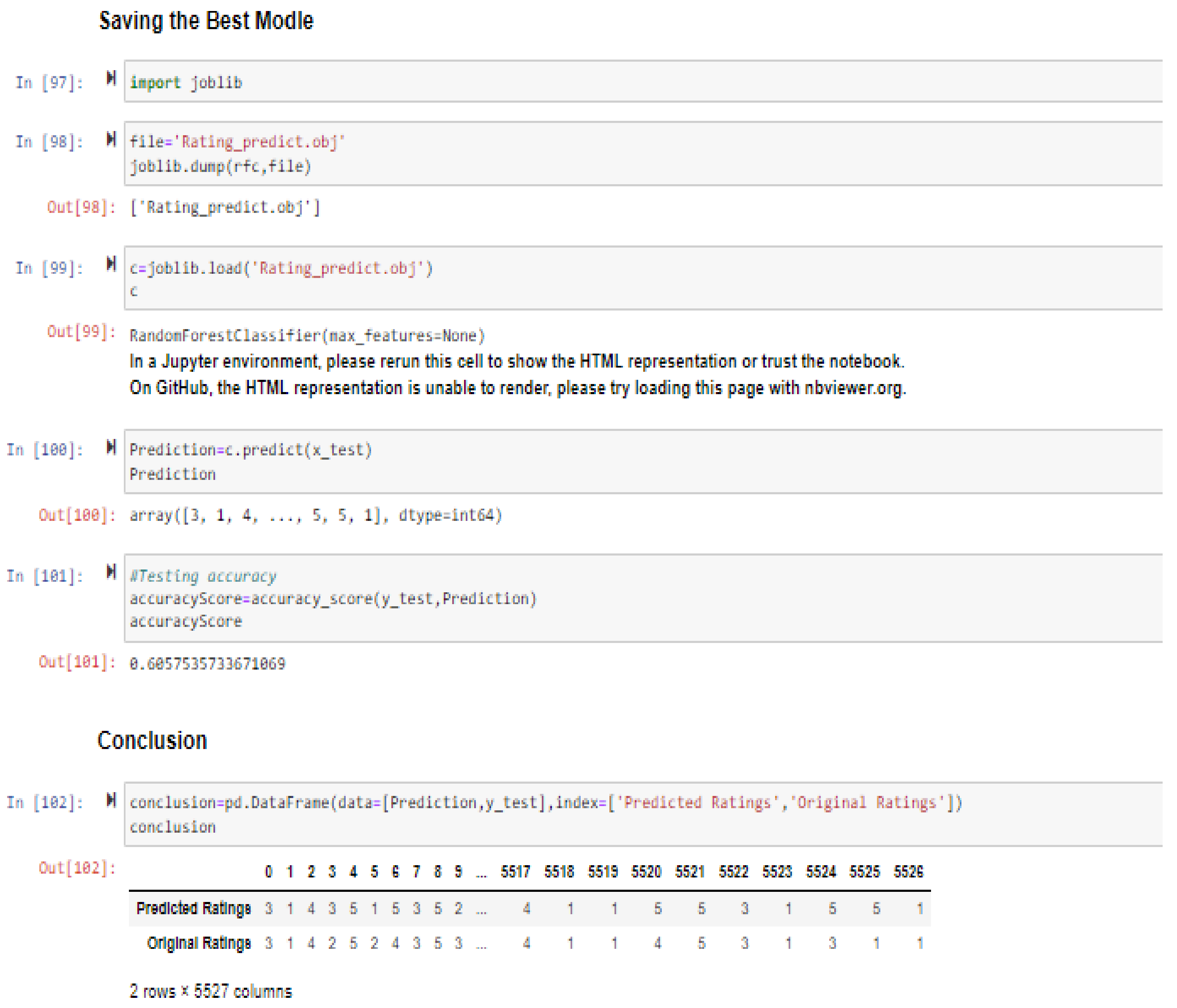
Interpretation of the Results

Below are some of the point’s w.r.t. visualization, pre-processing presented above -

- Null values are present in two of the columns and the null value counts is 706
- In this dataset rating of 5 is on top having 5527 rows out of 26840 rows
- 11040 rows are present having rating 1 or 5 all-together
- 10923 rows are present having rating 3 or 5 all-together
- 10767 rows are present having rating greater than 3 or rating is equal to 5 all-together
- No strong skewness present in the dataset
- No outliers are there in this dataset

CONCLUSION

After comparing all the above model I come to the point that Random Forest Classifier model performs well as compare to other models. This model generates training accuracy as almost 100% and testing accuracy as 61% which is best as compare to others. Also the Cross Validation score for the model is increasing at each instances so we can say that it might come same as that of testing accuracy at some point.



Saving the Best Modle

```
In [97]:  import joblib

In [98]:  file='Rating_predict.obj'
          joblib.dump(rfc,file)

Out[98]:  ['Rating_predict.obj']

In [99]:  c=joblib.load('Rating_predict.obj')
          c

Out[99]:  RandomForestClassifier(max_features=None)
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

```
In [100]: Prediction=c.predict(x_test)
          Prediction

Out[100]: array([3, 1, 4, ..., 5, 5, 1], dtype=int64)

In [101]: #Testing accuracy
          accuracyScore=accuracy_score(y_test,Prediction)
          accuracyScore

Out[101]: 0.6057535733671069
```

Conclusion

```
In [102]: conclusion=pd.DataFrame(data=[Prediction,y_test],index=['Predicted Ratings','Original Ratings'])
          conclusion

Out[102]:
```

	0	1	2	3	4	5	6	7	8	9	...	5517	5518	5519	5520	5521	5522	5523	5524	5525	5526
Predicted Ratings	3	1	4	3	5	1	5	3	5	2	...	4	1	1	5	5	3	1	5	5	1
Original Ratings	3	1	4	2	5	2	4	3	5	3	...	4	1	1	4	5	3	1	3	1	1

2 rows × 5527 columns

Learning Outcomes of the Study in respect of Data Science

I have used the Classification Model using multiple algorithms to design and optimize the results. Some of the classes which I explored from Scikit-learn libraries were – Statistics, Analytical Modelling, Hyper Tuning Method, CV Score, Predictive Modelling and different regression Model, etc.

I condensed to 2 columns keeping in mind the Principle Component Analysis. This resulted in condensed data being trained in a much shorter span of time with utmost accuracy. The visualizations helped in better and quick understanding of the outliers and skewness present in the data sets

With Mathematical Modeling helped me to find-out the corresponding mean, median, mode and relationship among the variables. Whereas Statistical Modeling helped in Correlation for understanding the relationship among the variables

One of the key challenge that I faced was scraping the data from different websites like Flipkart and Amazon as it was very time taking while looking for outputs and several time I was getting exceptions error and it restricted me to go with only one website that’s why I scraped the data from one Amazon only.

Limitations of this work and Scope for Future Work

Since I’ve extracted the dataset only from one website and keeping in the mind that it’s not necessary to have same attributes for all the websites hence output might be different while predicting the target variable ‘Ratings’, if we use different dataset.