

**Ratings Prediction**



Submitted by

Sachin Gupta

**ACKNOWLEDGMENT**

Working on this project has an incredible experience that will have an impact on my career.

It is pleasant gratification to present Ratings Prediction.

I have completed this project by taking the help from Google, Bing and You tube.

**INTRODUCTION**

Rating prediction is a well-known recommendation task aiming to predict a user's rating for those items which were not rated yet by her.

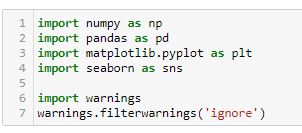
Predictions are computed from users' explicit feedback, i.e. their ratings provided on some items in the past.

In this Project I have done Web Scraping by using Electronics goods that is phones, smart watches,headphones, laptops etc.

After completing Web scrapping and converting all the scrap data into Dataset. From which I understand from this dataset I have to predict Ratings\_of\_the\_product.

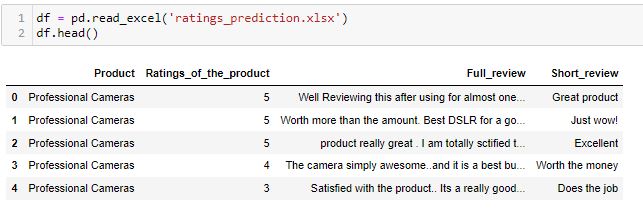
**Importing Libraries**

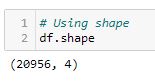
*I am importing all the library which I required for EDA, visualization, prediction and finding all matrices. The reason of doing this is that it become easier to use all the import statement at one go and we do not require to import the statement again at each point***.**



* Data Sources and their formats

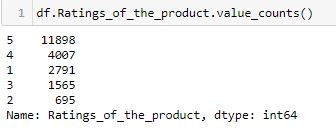
Now I am going to upload or read the files/data-sets using pandas. For this I used read csv method.





There are 20956 rows and 4 columns in the dataset.

Value\_Counts



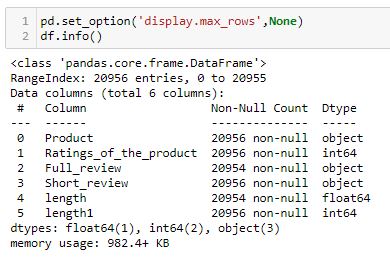
From the above code 5 star reviews are most demanded.



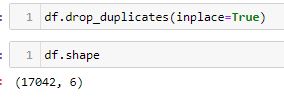
By using len Full\_review & Short\_review shows the length of features columns.



All features are converted into lower case.

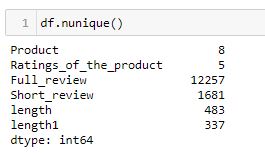


Also, most of the column are object in nature and Ratings\_of\_the\_product,length1 is of int64 data type.



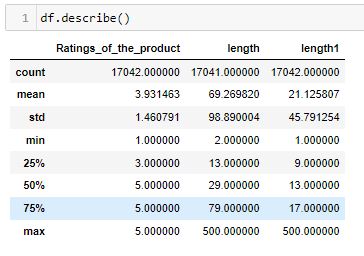
#### After using duplicates I can see there is duplicates value in this dataset.

With using shape I got 17042 rows & 6 columns.



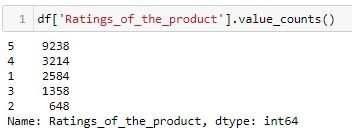
There are few columns which are categorical in nature and few columns are continuous in nature.

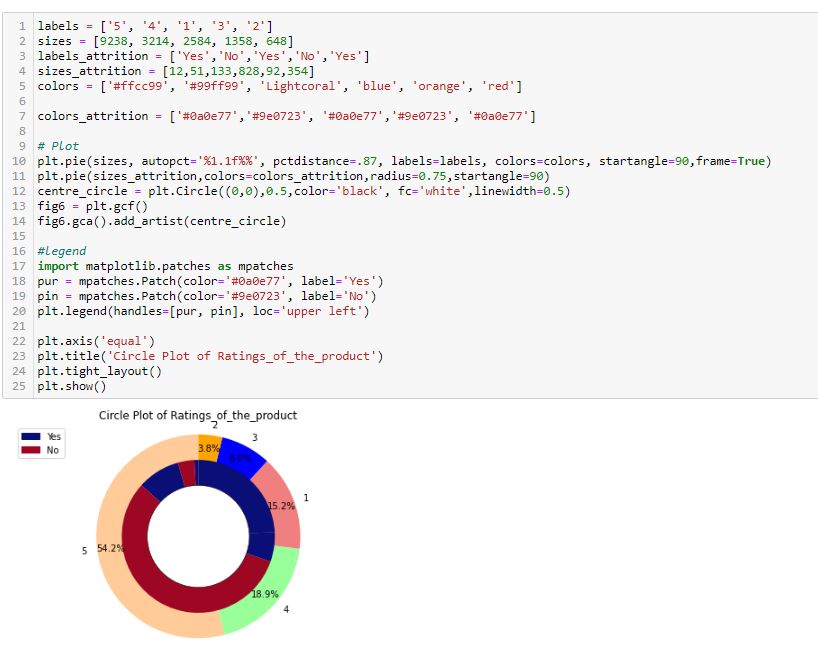
Describe

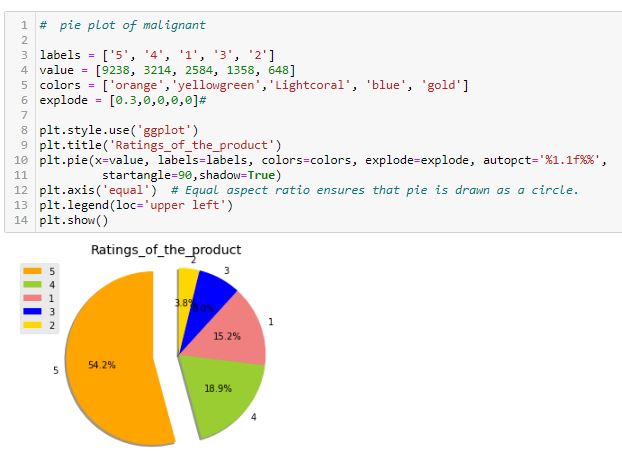


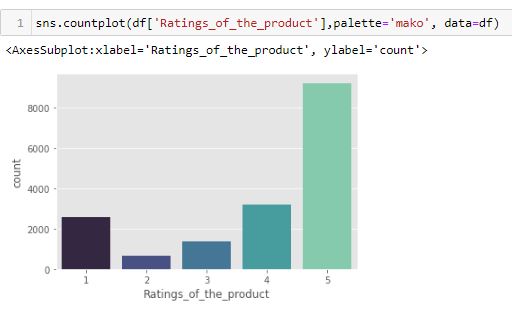
**Data Visualizations**

**Ratings\_of\_the\_product**









From the Above Plots of Ratings\_of\_the\_product it shows 5 ratings reviews are most used for ratings predictions.

Correlation



#### Now I am finding the correlation value of each column, this value is categorized into mainly 2 parts that are:

#### Positive correlated value

#### Negative correlated value The most the value is positive means that column is much co related and vice versa.

# Heat map

#### Heat map is a way to show some sort of matrix plot.

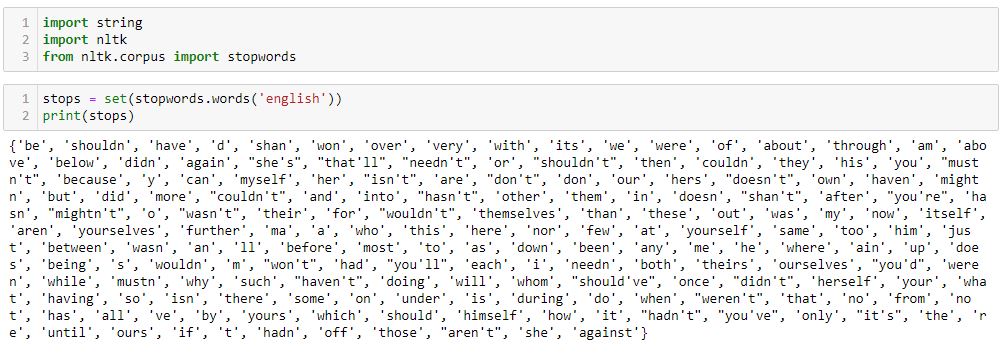
#### To use a heat map the data should be in a matrix form.

#### By matrix we mean that the index name and the column name must match in some way so that the data that we fill inside the cells are relevant.

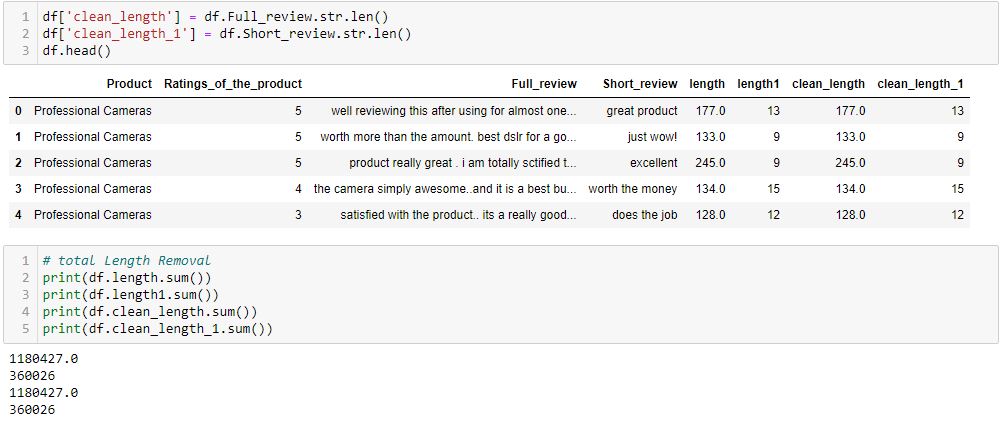
## 

I am using sea born heatmap to plot the correlated matrix and plot the corr value in the heatmap graph.

**Removing Stopwords**

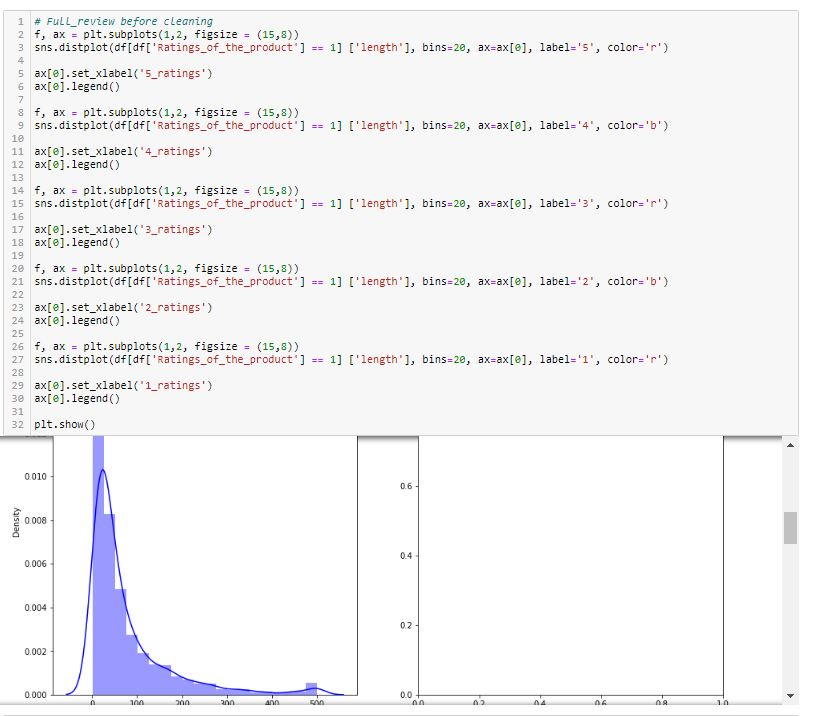


## I have import stopwards that help me to identify all the common words which can remove it form comments columns.



#### After Cleaning and Removal of all the features.

## **Using Distribution Plot after cleaning & before cleaning the features columns.**



## F:\DATA TRAINED\FLIP ROBO TECH\13th Assignment\jpeg\21.JPG

## F:\DATA TRAINED\FLIP ROBO TECH\13th Assignment\jpeg\22.JPG

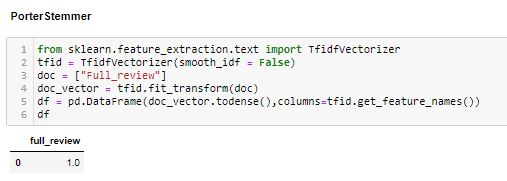


## **WordCloud**



By using Word Cloud it shows numbers of words which are loud.

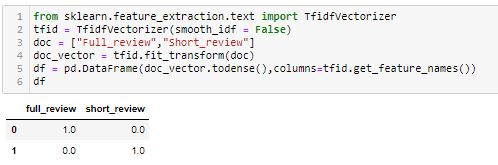
**Porter Stemmer**



I have used the PorterStemmer and WordNetLemmatizer in comment text columns so that i can minimize the numbers of words from the columns and get some meaningful information out of it.

It will also help us to improve the accuracy score.

**TfidfVectorizer**

****