AlmaBetter

Project-EDA on Play Store data

Sumit Jadhav, Rushikesh Shinde,
Aakash Jadhav, Aniket Okate

—----------

Abstract:

There are millions of applications uploaded by the developers on the daily basis. Without any check and balance millions of users download these applications. Theses duplicated applications damage the users trust on Google play store and can grab the confidential information of user. There is no more information provided by developers on the front end of the application that can define the legitimacy of the application.

The categories of these games' applications use respectively are Word, Trivia, Simulation, Sports, Strategy, Racing, Role_Playing, Puzzle, Music, Educational, Card, Casino, Casual, Board, Action, Adventure, and Arcade.

This visualization is more helpful for game developers in the development phases, also for the users of the game's application for the selection of the game that they want to play.

With analysis we find out behaviour of data,which application/category of application is more popular on play store.As well as find out advantages and disadvantages of the application.

Keywords:

EDA, Data analysis, Visualization,, Behaviour of dataset, Data preprocessing, Data cleaning.

Introduction:

It has been observed that the significant growth of the mobile application market has a great impact on digital technology. Having said that, with the ever-growing mobile app market there is also a notable rise of mobile app developers; eventually resulting in sky-high revenue by the global mobile app industry.

There are about 2.8 million applications uploaded on playstore.By google's survey almost 3739 applications get released on it daily.Some of them we used to analysis.Play store become the very popular organization related to Android version because of their wide range of application categories.These applications provide lots of data that can be analyzed and help to Entrepreneurs and a person who launch their own application.They can use this analysis to launch application relative to trending category and understand reviews of application,overcome burgs and provide a useful and complete application.

The dataset is around 10800 rows and 13 columns in Dataframe1 and around 64000 row and 5 columns in Dataframe2, which contains numerical values,text,date values and mix type of values.

Objectives:

- Data Summery.
- Data Cleaning.
- Visualization to find out behavior of data.
- Other visualization by various features.
- Conclusion.

Actual Work:

1.Data summary

Find brief information about dataset..Checking data types of each feature.If there is any kind of problem in data type then change data type. Separate out numerical and categorical features.

Also find mean,mode,max,min,SD values of each features.

2.Data Cleaning

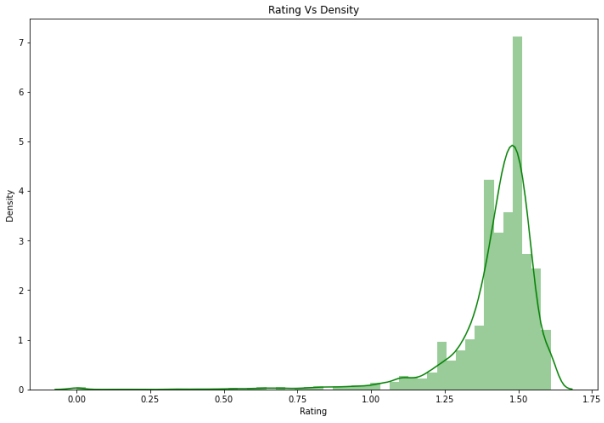
Checking null values are present in the dataset. If present then null value cells are replaced by mean,mode value or if this is a categorical variable then replaced with another word.

Both the datasets contains null values so we drop those rows.

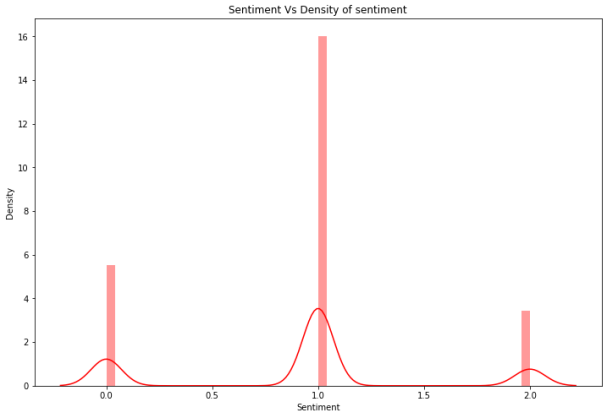
3.Data visualization

1.Check dataset is balanced or imbalance: We consider “Rating” and “Sentiment” as target variable for dataset1 and dataset2 respectively.

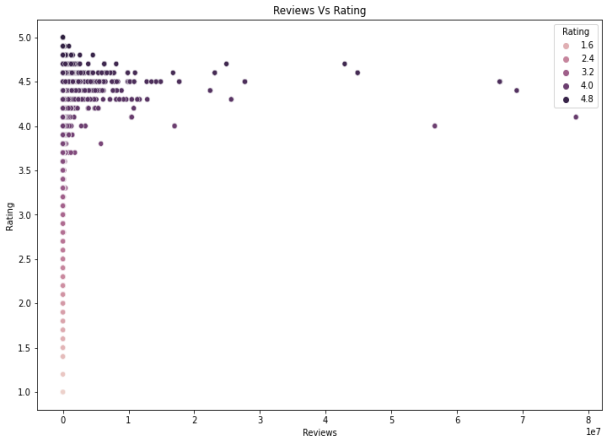
-Dataset 1:



-Dataset 2:



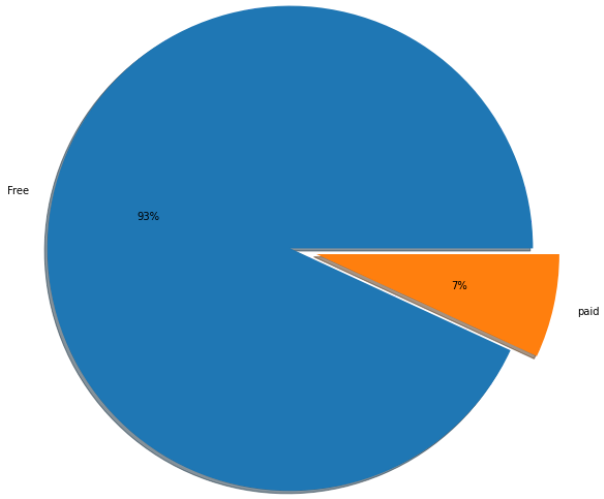
2.Check linearity of dataset 1.



3.Find correlation with each feature.

4.Feature wise count of application.

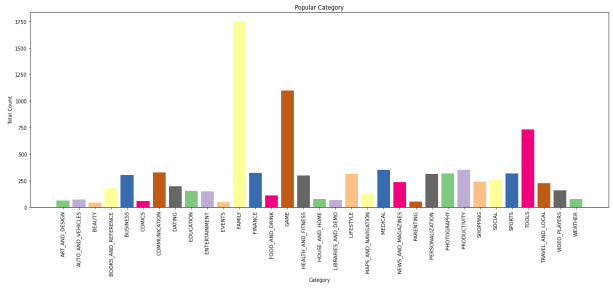
5.To find the percentage of "Free" and "Paid" applications.



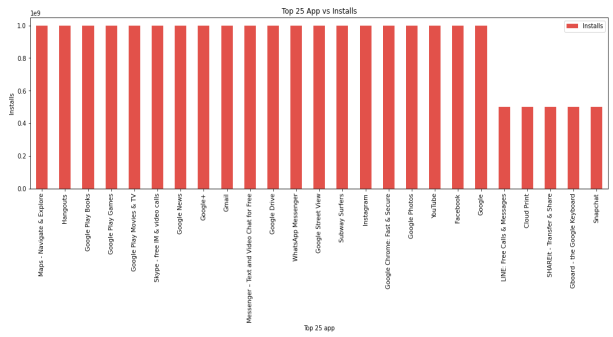
-Free applications=93%

-Paid application=7%

6.Analyse data to found which category is most popular on basis of number of applications are present on play store.

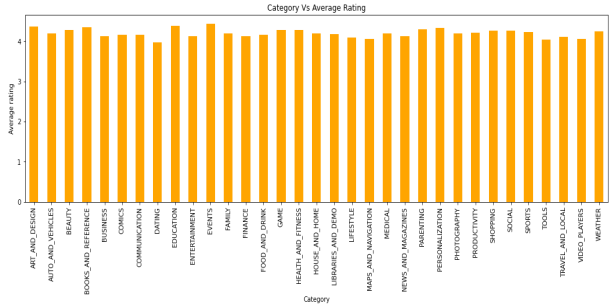


7.To find popular Application on the basis of installs.

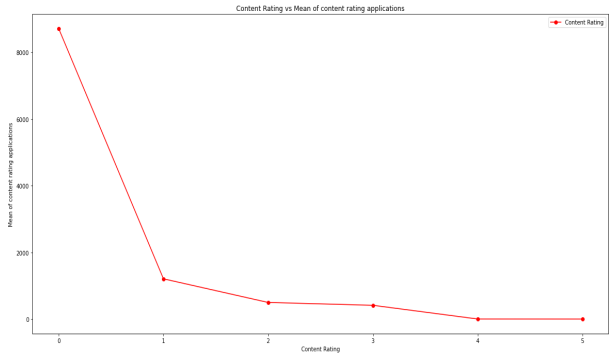


8.Find reviews of the perticular application

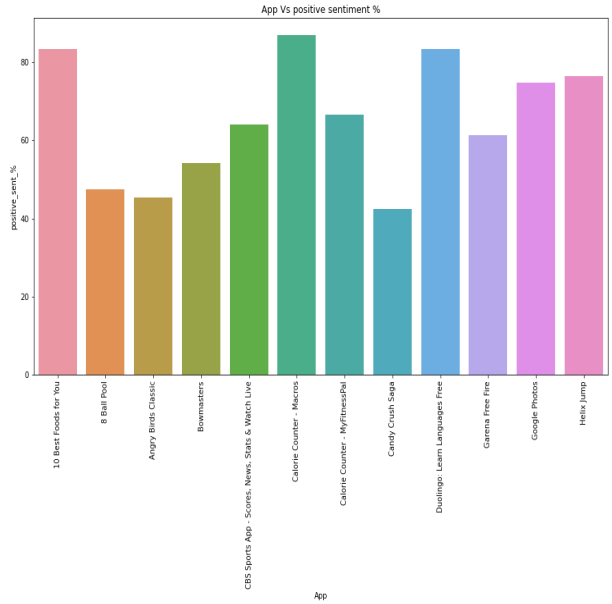
9.To analyze which category is most popular on the basis of Rating



10.To analyze type of applications basis of content ratings



11.Application have more number of positive reviews.



Conclusion:

1.Imbalanced datasets.

2.These two datasets are not normal distribution.

3.Dataset number 1 have non linear data.

4.Other visualization:

a)On the basis of free and paid applications

free=93%

paid=7%

b)On the basis of above analysis we can observe which category is trending on Play Store and the application which have maximum downloads.

c)Then we find reviews of the application with help of Developers to update their applications and remove problems.

d)Find average rating and seen that most of the application categories have rating near to "4"

e)content rating means generations with which applications can be used,so we found that "Everyone" category is widely used.

f)With using "Sentiment" column we can find most loved applications on playstore by means of their positive reviews count.