Capstone Project-1 Submission

Play Store App Review Analysis

RAJESH MOHANTY

SAIRINDHRI JENA

MANISHWAR GUPTA

ASHISH MAHARANA

MANISHANKAR KUMAR SAW

Data science trainees,

AlmaBetter, Bangalore

--***--

Rajesh Mohanty – Rmohanty385@gmail.com

Sairindhri jena -- rssairindhri.99@gmail.com

Manishwar gupta -- manishwargupta2109@gmail.com

Ashish maharana --- ashishsid0467@gmail.com

Manishankar kumar saw --- manishankarksaw15@gmail.com

GitHub Link~~

Rajesh mohanty :

https://github.com/Rmohanty385/Playstory-app-review-analysis

Sairindhri jena :

https://github.com/Sairindhrijena/Playstore-app-analysis

Manishwar gupta:

https://github.com/manishwargupta/Play-Store-App-Review-Analysis

Ashish maharana :

https://github.com/ashishsid0467gmailcom/Play-store-app-user-review-analysis

Manishankar kumar saw :

https://github.com/manishankarksaw15/Play-store-App-and--user--review-analysis.git

ABSTRACT-

Software application is vital because specific software is required in almost every industry, in every business ,and for each function. It become more important as time goes on. Mobile app distribution platform such as Google play store gets flooded with millions of new applications uploaded by developers everyday few thousands of new applications are regularly uploaded on Google play store. A huge number of designers working freely on designing the apps and making them successful. With the enormous challenge from everywhere throughout the globe, it is important for a developer to know whether he/she is continuing the correct way or not. Since most Play Store applications are free, the income model is very obscure and inaccessible regarding how the in-application buys, in-application adverts and memberships add to the achievement of an application. In this way, an application's prosperity is normally dictated by the quantity of installation of the application and the client appraisals that it has gotten over its lifetime instead of the income is created. The objective of this experiment is to deliver insights to understand customer demands better and thus help developers to popularize the product. We have tried to discover the relationships among various attributes such as which application is free or paid, what are the user reviews, rating of the application.

***Key Words*:** Google Play Store Apps, Ratings Prediction, Exploratory Data Analysis,

1. PROBLEM STATEMENT

Data is taken from the Google play store dataset. Every row contains various entries regarding a certain app. We will be doing Exploratory data analysis on this data set, which is a very important step in data science cycle, as it not only helps in taking very initial business decisions. Our objective will be to structure the data, clean it and present certain trends that we observe that can help us draw very preliminary conclusions about the probability of success of a newly launched app.

2. INTRODUCTION

In today's scenario we can see that mobile apps playing an important role in any individual's life. In today's era, the Google Play Store is the largest and most popular android app store. It is flooded with millions of applications and it provides wide collection of data on features like ratings, price and number of downloads and apps description. Many apps are being developed as apps are easy to create and its lucrative. But its important for developers to know which apps are loved by customers and are trending in market so that he develop only those apps and also there is a high competition between app providers producing similar applications. Analyzing customer needs is one of the bizarre tasks in the business world today. Hence proposing analyze data to developer that what customer is likely to download, which category got the maximum downloads this all plays a crucial role in app development. With enormous challenge from everywhere throughout the globe, it is important for a designer to realize that he/she is continuing in the right way or not. To hold this income and their place in the market the application designers may need to figure out how to stick into their present position. The dataset with 10k Play Store applications is available to analyze the market of android. It can be examined to analysis the different category such as family, communication, entertainment, tools, music, camera etc.

2.1 GOOGLE PLAY STORE AND USER REVIEW ANALYSIS

In this project we examine the different attributes present in the data set that affect the popularity of the application. We focused on to answer the questions like, what makes an app popular, what should be the price and size of the app, is there some trends in user sentiments. In our data set we have two csv files for data analysis: Play Store data User Reviews At first, we analysis the play store data and in the play store data we have 10841 rows and 13 columns & in the user review data we have 64295 rows and 5 columns of data. We have to take the maximum outcomes from the data which help us to analysis the which type of app is most preferable and comparisons between different insights. Our goal is to filter and make plots accordingly for a better EDA with respect to the final data. We need to explore and analyze the data to discover key factors responsible for app engagement and success. With enormous challenge from The project aims at doing this with the help of a sentimental analysis and machine learning that will analyze customer needs and suggest the developers best app for developing. The analysis is achieved using the survey of the user download behavior on the apps across all the categories on the Google play store. Mobile app stores are becoming extremely lucrative. Android is expanding as an operating system and Mobile app industry is increasing in significantly and thus giving rise to more competitions to the one's that are creating applications. Hence, for a developer to know the recent trends, competition is important so that the value of their app in the store do not degrade. Google play store is a digital distribution service and it allows user to browse and download different apps. It serves as official store of apps for android operating system. Play store is additionally a platform which offers music, digital media store, books, movies and tv programs. Mobile app stores are becoming extremely lucrative. Due to the competition in the market and also expansion in order to help our developer understand what kinds of apps are likely to attract more users and what is the motivating factor for the people to download an app we analyze and research relevant data. They will be getting to know the success rate and they will get to decide what features should be added or modified and what should be maintained according to the current state of their app. Hence we found this topic interesting and convincing for our project work..

We develop Android apps & release on Play Store. As an Developer or say Business Perspective it's very important to know whether users are enjoying the app or facing any issues. To know this Play Store has a Ratings & reviews section for each app released on play store. Users can submit the ratings and has a freedom to write a review for a particular app. This approach is quite a lengthy to rate & review app i.e. navigate to Play store to submit feedback or redirect leaving a current app workflow to open Play Store App link using URI. We never wanted our customers to leave our application, but with this flow, we are forced to redirect the control to Play store app.

2.2 GOOGLE PLAY STORE DATASET

The dataset consists of Google play store application and is taken from Almabetter, which is the world's largest community for data scientists to explore, analyze and share data.

This dataset is for Web scratched information of 10k Play Store applications to analyze the market of android. Here it is a downloaded dataset which a user can use to examine the Android market of different use of classifications music, camera etc. With the assistance of this, client can predict see whether any given application will get lower or higher rating level. This dataset can be moreover used for future references for the proposal of any application. Additionally, the disconnected dataset is picked so as to choose the estimate exactly as online data gets revived all around a great part of the time. With the assistance of this dataset, I will examine various qualities like rating, free or paid and so forth utilizing Hive and after that I will likewise do forecast of various traits like client surveys, rating etc.

The data set contains the following columns:

- **App:** This Column contains the name of the app
- **Category:** This contains the category to which the app belongs. The category column contains 33 unique values.
- **Rating:** This column contains the average value of the individual rating the app has received on the play store. Individual rating values can vary between 0 to 5.
- **Reviews:** This column contains the number of people that have given their feedback for the app.

- **Size:** This column contains the size of the app i.e. The memory space that the app occupies on the device after installation.
- **Installs:** This column indicates the number of time that the app has been downloaded from the play store, these are approximate values and not absolute values.
- **Type:** This column contains only two values- free and paid. They indicate whether the user must pay money to install the app on their device or not.
- **Price:** For paid apps this column contains the price of the app, for free apps it contains the value 0.
- **Content Rating:** It indicates the targeted audience of the app and their age group.
- **Genre:** This column contains to which genre the app belongs to, genre can be considered as a sub division of Category.
- **Last updated:** This column contains the info about the date on which the last update for the app was launched.
- **Current version:** Contains information about the current version of the app available on the play store.
- **Android version:** Contains information about the version of the android OS on which the app can be installed.

2.3 User Review Dataset

- User reviews data frame has 64295 rows and 5 columns. The 5 columns are identified as follows:
- **App:** Contains the name of the app with a short description (optional).
- **Translated Review:** It contains the English translation of the review dropped by the user of the app.
- **Sentiment:** It gives the attitude/emotion of the writer. It can be 'Positive', 'Negative', or 'Neutral'.
- **Sentiment Polarity:** It gives the polarity of the review. Its range is [-1,1], where 1 means 'Positive statement' and -1 means a 'Negative statement'.
- **Sentiment Subjectivity:** This value gives how close a reviewer's opinion is to the opinion of the general public. Its range is [0,1]. Higher the subjectivity, closer is the reviewer's opinion to the opinion of the general public, and lower subjectivity indicates the review is more of a factual information.

2.4 Python

Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation. Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured, object-oriented and functional programming. Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms, and can be freely distributed.

2.5 Data Cleaning and Preparation

Preprocessing is important into transitioning raw data into a more desirable format. Undergoing the preprocessing process can help with completeness and compellability. For instance, you'll see if certain values were recorded or not. Also, you'll see how trustable the info is. It could also help with finding how consistent the values are. We need preprocessing because most real-world data are dirty. Data can be noisy i.e. the data can contain outliers or simply errors generally. Data can also be incomplete i.e. there can be some missing values.

The available data is raw and unusable for Exploratory data analysis, so before we do anything with the data we will have to explore and clean it to prepare it for data analysis.

- **Step1**: We write a function play store info (), that will display 5 attributes about all the columns: Data type, Count of non-null values, Count of null values, number of unique values in that column and percentage of null value in that columns in the play store dataset.

- **Step2**: we start off with the column 'Type' we can see that it has one null value. We checked this row and found out from the play store that it is a free app. We use fillna() function of the pandas library to fill this value.

- **Step 3**: We drop the columns 'Current Ver', 'Android Ver' and 'last updated' from our dataset using the drop() function of the pandas library.

- **Step 4**: We can see that the 'Rating' column has 1474 null values. Due to low variations in the rating values and a lot of repeated values the 'median' would be a suitable statistical indicator to replace the null values with. We calculate the mode of the column using the median () aggregate method, and fill this value in place of null values using the fill na() function.

- **Step 5:** We can see that the 'Reviews' column despite being a numerical indicator is of the 'object' data type, we will convert this to 'int' data type using the as type(int) function.

- **Step 6:** We can see that the size column, which should be numeric, is of the data type 'object', it also has characters 'k' and 'M' in the values which stand for kilobytes and Megabytes, we will replace the 'k' with 1000 and 'M' with 1000000. Some values also have '+' sign in them, which will be removed. Next, we will convert this column into 'int' data type.

- **Step 7:** The 'Installs' column values contain the characters '+' and ',' which are going to prevent us from converting this column into a numeric data type. We will get rid of these using the strip() and replace() functions.

- **Step 8:** The values in the column 'Price' might have the '$' sign in some values and the column is of the data type 'object'. We will first remove the '$' sign using the **strip()** function and then convert the column into 'int' data type.

- **Step 9:** Handling the duplicates in the App column we drop the no of duplicate rows that are present in the App columns.

- **Step 10:** We write a function Ur info(), that will display 5 attributes about all the columns: Data type, Count of non-null values, Count of null values ,number of unique values in that column and percentage of null value in that columns in the User review dataset.

- **Step11:** In the User review dataset the columns are App, Translated Review, Sentiment, Sentiment Polarity, Sentiment Subjectivity in this total 26863 NaN value are present so we drop them using drop na() function.

3. EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis, or EDA, is an important step in any Data Analysis or Data Science project. EDA is the process of investigating the dataset to discover patterns, and anomalies (outliers), and form hypotheses based on our understanding of the dataset.

EDA involves generating summary statistics for numerical data in the dataset and creating various graphical representations to understand the data better. In this article, we will understand EDA with the help of an example dataset. We will use **Python** language (**Pandas** library) for this purpose.

3.1 Free v/s Paid

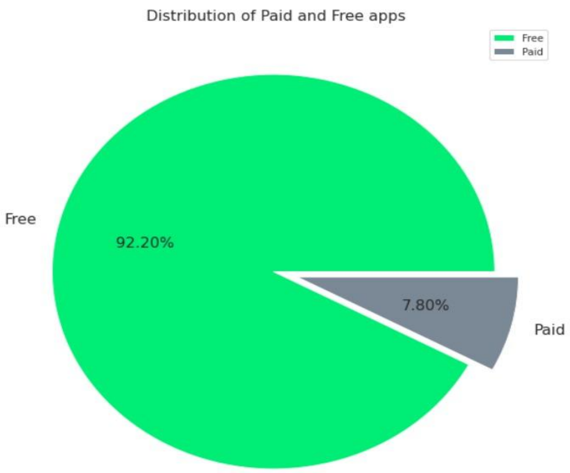

Fig-1: FreevsPaid

Here we can see that 92.2% apps are free, and 7.80%appsarepaidonGooglePlayStore,sowecansaythatMostoftheappsarefreeonGooglePlayStore.

3.2 Top Category of Play store

There are lot of category wise apps are available on playstore so the below curve show hoe the apps are distributed.

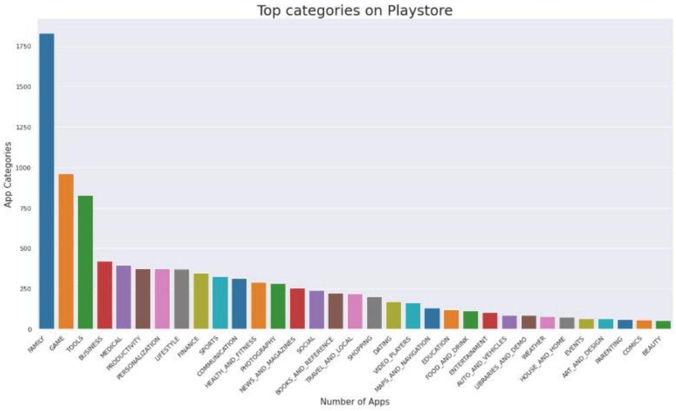

Fig-2:Top Categories on Play store

So, there are all total 33 categories in the dataset From the above output we can come to a conclusion that in play store most of the apps are under FAMILY & GAME category and least are of EVENTS & BEAUTY Category.

3.3 Percentage of User review Sentiments

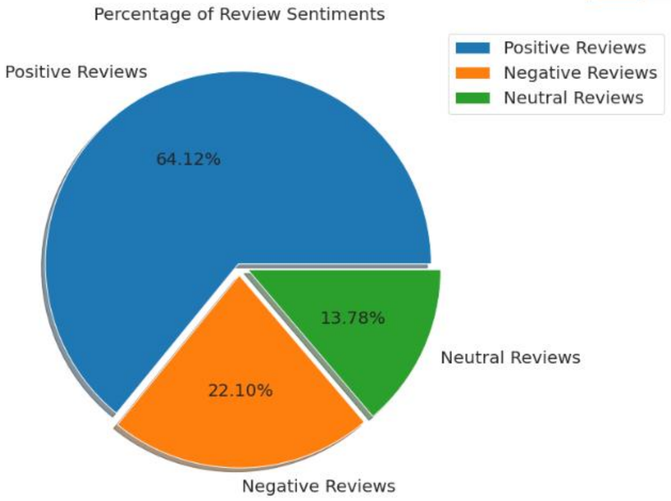

Fig-3: Percentage of User Review Sentiments

From the above pie chart, we can say that most of the apps that are present on the play store has received positive review by the user while there are some apps which have negative reviews as well.

3.4 Top 10 positively reviewed Apps

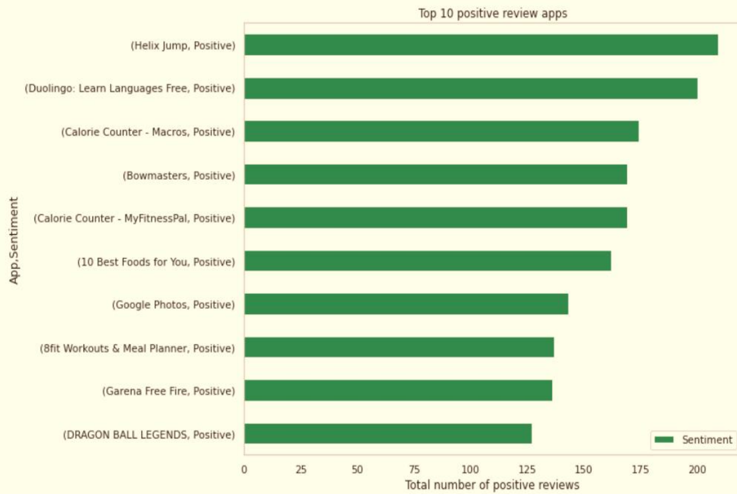

Fig-4: Top 10 Positive Reviewed App

3.5 Top 10 Negative Reviews Apps

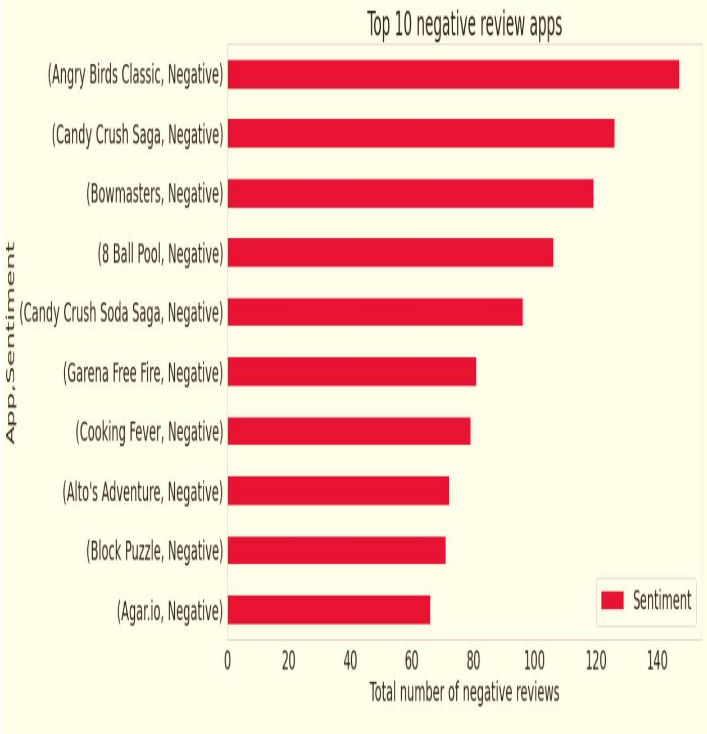

Fig-5: Top 10 Negative Reviewed Apps

3.6 Top 10 Negative Reviews Apps

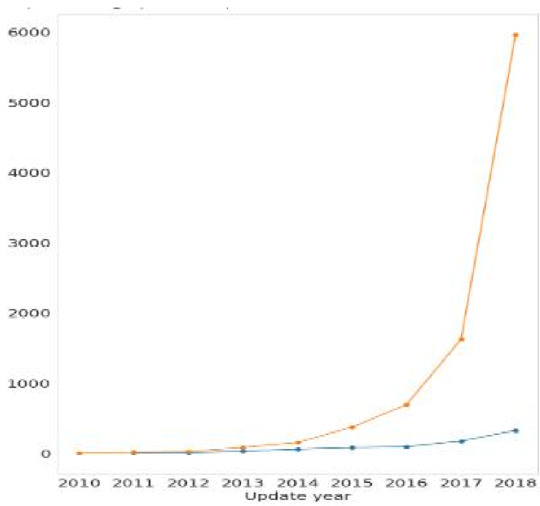

Fig-6: Top 10 Negative Reviewed Apps

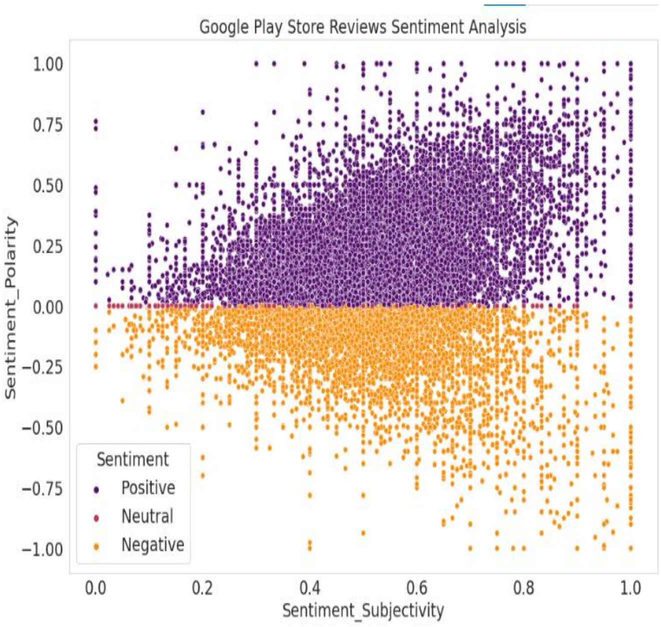

Fig-7: Google play store Reviews Sentiment Analysis

From the above scatter plot it can be concluded that sentiment subjectivity is not always proportional to sentiment polarity but in maximum number of cases, show a proportional behavior, when variance is too high or low.

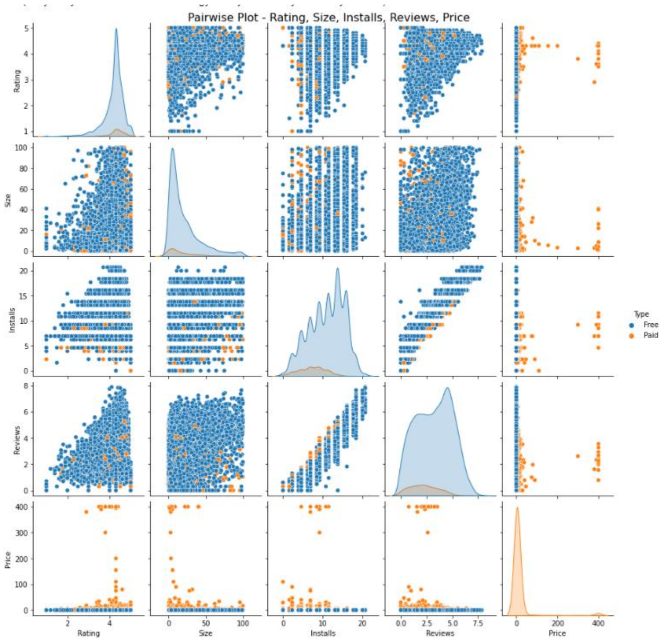

Fig-8: Pair wise plot

Conclusion

Through exploratory data analysis we have observed some trends and have made some assumptions that might lead to app success among the users in the play store.

- Percentage of free apps = 92%
- Percentage of apps with no age restrictions = 82%
- Most competitive category: Family
- Family, Game and Tools are top three categories having 1906, 926 and 829 app count.
- Tools, Entertainment, Education, Business and Medical are top Genres.

References

- GeeksforGeeks
- Wikipedia
- Stackoverflow
- Towards data science
- Python libraries documentation
- Data camp
- 1. Researchgate.net

Thank You