“CAPSTONE PROJECT”

Exploratory Data Analysis

GOOGLE PLAY STORE APP DATA-SET & REVIEW ANALYSIS

By,

Tushar Khairnar

Shreyash Sarode

Taha Nakedar

Pradnya Pagar

Data science trainees,

AlmaBetter, Bangalore.

----------***----------

Abstract - *The google play store is one of the largest and most popular Android app stores. It has an enormous amount of data that can be used to make an optimal model. Google play store is engulfed with a few thousands of new applications regularly with a progressively huge number of designers working freely or on the other hand in a group to make them successful, with the enormous challenge from everywhere throughout the globe. We have used a raw data set of Google Play Store. This data set contains 13 different features that can be used for predicting whether an app will be successful or not using different features. This data set is scraped from the Google Play Store. This journal talks about different classifier models that we used for prediction purposes and finding which one gives the highest accuracy. This journal also gives detailed information on feature extraction and the complete Data visualization done on this data set. Most Play Store applications are free, the income model is very obscure and inaccessible regarding how the in-application buys, 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. Application (App) ratings are feedback provided voluntarily by users and function important evaluation criteria for apps. However, these ratings can often be biased due to insufficient or missing votes. Additionally, significant differences are observed between numeric ratings and user reviews.*

This Study aims to predict the ratings of Google Play Store apps using machine learning Algorithms. We have tried to perform Data Analysis and prediction into the Google Play store application dataset . Using Python library, we have tried to discover the relationships among various attributes present in our dataset such as which application is free or paid, about the user reviews, rating of the application.

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

1. PROBLEM STATEMENT

The dataset contains immense possibilities to improve business values and have a positive impact. It is not limited to the problem taken into consideration for this project. Many other interesting possibilities can be explored using this dataset.

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 but also in preparing the data for further modelling for use in machine learning algorithms. 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

A Python library is a collection of related modules. It contains bundles of code that can be used repeatedly in different programs. It makes Python Programming simpler and convenient for the programmer. As we don't need to write the same code again and again for different programs. Python libraries play a very vital role in fields of Machine Learning, Data Science, Data Visualization, etc

The Python Standard Library contains the exact syntax, semantics, and tokens of Python. It contains built-in modules that provide access to basic system functionality like I/O and some other core modules. Most of the Python Libraries are written in the C programming language. The Python standard library consists of more than 200 core modules. All these work together to make Python a high-level programming language. Python Standard Library plays a very important role. Without it, the programmers can't have access to the functionalities of Python. But other than this, there are several other libraries in Python that make a programmer's life easier.

2.1 GOOGLE PLAY STORE AND USER REVIEW ANALYSIS

In today's scenario we can see that mobile apps playing an important role in any individual's life. It has been seen that the development of the mobile application advertise has an incredible effect on advanced innovation. Having said that, with the

consistently developing versatile application showcase there is additionally an eminent ascent of portable application designers inevitably bringing about high as can be income by the worldwide portable application industry.

With enormous challenge from everywhere throughout the globe, it is basic for a designer to realize that he is continuing in the right heading. 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 Google Play Store is observed to be the biggest application platform. It has been seen that although it creates more than two-fold the downloads than the Apple App Store yet makes just a large portion of the cash contrasted with the App Store. In this way, I scratched information from the Play Store to direct our examination on it.

With the fast development of advanced cells, portable applications (Mobile Apps) have turned out to be basic pieces of our lives. Be that as it may, it is troublesome for us to follow along the fact and to understand everything about the apps as new applications are entering market each day. It is accounted for that Android market achieved a large portion of a million applications in September 2011. Starting at now, 0.675 million Android applications are accessible on Google Play App Store. Such a lot of applications are by all accounts an extraordinary open door for clients to purchase from a wide determination extend. We trust versatile application clients consider online application surveys as a noteworthy impact for paid applications. It is trying for a potential client to peruse all the literary remarks and rating to settle on a choice. Additionally, application engineers experience issues in discovering how to improve the application execution dependent on generally speaking evaluations alone and would profit by understanding a huge number of printed remarks.

We develop Android apps & release on Play Store. As a 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

Most of the info scientist use python due to the good built-in library functions and therefore the decent community. Python now has 70,000 libraries. Python is simplest programing language to select up compared to other language. That is one among the most reasons to use python. Specifically, for data scientist the foremost popular data inbuilt open-source library is named panda. The Python standard library consists of more than 200 core modules. All these work together to make Python a high-level programming language. We used following libraries,

-**Numpy :** The name "Numpy" stands for "Numerical Python". It is the commonly used library. It is a popular machine learning library that supports large matrices and multi-dimensional data

-**Pandas :** Pandas are an important library for data scientists. It is an open-source machine learning library that provides flexible high-level data structures and a variety of analysis tools. It eases data analysis, data manipulation, and cleaning of data. Pandas support operations like Sorting, Re-indexing, Iteration, Concatenation, Conversion of data, Visualizations, Aggregations, etc.

-**Matplotlib :** This library is responsible for plotting numerical data. And that's why it is used in data analysis. It is also an open-source library and plots high-defined figures like pie charts, histograms, scatterplots, graphs, etc

-**Seaborn :** is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics. For a brief introduction to the ideas behind the library, we can read the introductory notes or the paper.

2.5 STRUCTURES

- Import Packages
- Dataset Information
- Data Cleaning
- Exploring app categories
- Distribution of app ratings
- Size and price of an app
- Relation between app category and app price
- Filter out "junk" apps
- Popularity of paid apps vs free apps
- Sentiment analysis of user review

2.6 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 use describe() function to get information about statistical inference from dataset

- **Step3**: we start with the checking the null values in the data so that we get rough overview how much missing values in the data. Then we think what can we do with this null values, there are two option in front of us. 1. Drop the null values and 2. Fill null values with mode or median of that column.

- **Step 4:** A box plot is a method for graphically depicting groups of numerical data through their quartiles. This represents the average of this numerical data. As we plot boxplot we can see there are values which are greater than 5. The rating of every app has to be max 5 and not greater than it so we had to remove this row of outlier.

- **Step 5**: 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 median of that corresponding column. We calculated the mode of the column using the median () aggregate method, and fill this value in place of null values using the fillna() function.

- **Step 6:** Next we were gone through about other columns which are 'Type, Android Ver, Current Ver. Then we dropped NAN values found in such columns.

- **Step 7:** 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 8:** 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' datatype.

- **Step 9:** The 'Installs' column values contain the characters '+' and ',' which are going to prevent us from converting this column into a numeric datatype. We will get rid of these using the replace() functions.

- **Step 10:** The values in the column 'Price' might have the '$' sign in some values and the column is of the datatype 'object'. We will first remove the '$' sign and then convert the column into 'int' datatype.

- **Step 11:** Handling the duplicates in the App column we drop the no of duplicate rows that are present in the App columns.

- **Step 12:** In the User review dataset the columns are App, Translated Review, Sentiment, Sentiment Polarity, Sentiment Subjectivity in this total **26868** NaN value are present so we drop them using dropna() 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 CATEGORY WISE APPS SHARE IN PLAY STORE

After looking at the data category wise we see that "Family" accounts for 18.97% (1829) of all of them which highest among all, while "Games" and "Tools" account for 9.93%(959) and 8.56%(825) respectively.

When it comes to least share then we can see that "Beauty" is on top with 0.55% and after that "Comics" with 0.58% "Parenting" with 0.62% and "Art and Design" with 0.65%.

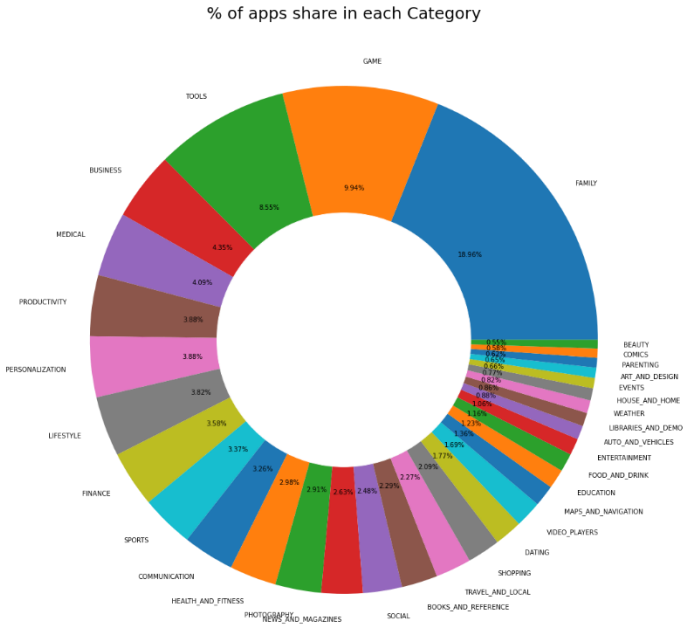


Fig -3.1: Percentage of apps share in each category

3.2 NUMBER OF INSTALLS FOR EACH CATEGORY

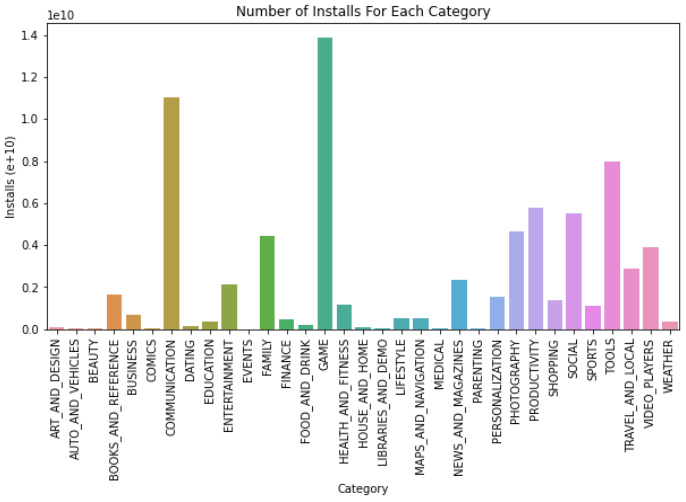


Fig -3.2: Number of installs for each category

This tells us the category of apps that has the maximum number of installs.
The Game, Communication and Tools categories has the highest number of installs compared to other categories of apps

3.3 TOP 10 INSTALL APPS ACCORDING TO CATEGORY

Function are more useful and time saving while analyzing the data. So we have created a function which takes category name and gives the top 10 install app from that particular category.

Let's see which are the top 10 install apps in the "communication" category

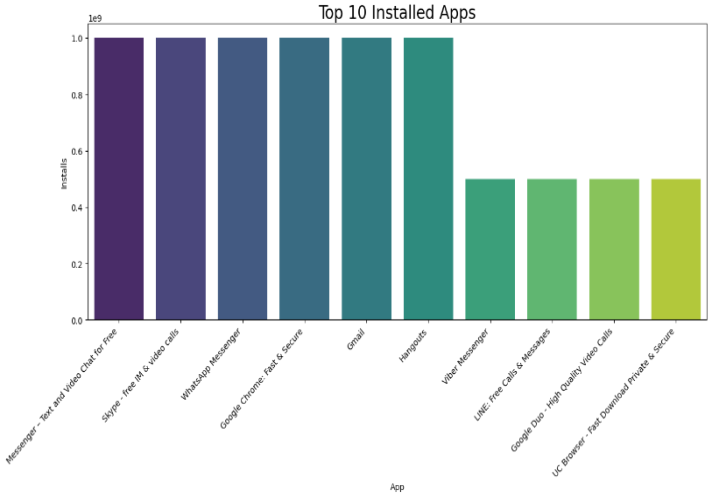


Fig -3.3: Top 10 installed apps

In "communication category 'messenger', 'Skype, and 'WhatsApp' are the top 3 install apps.

3.4 CATEGORIES WITH HIGHEST RATINGS

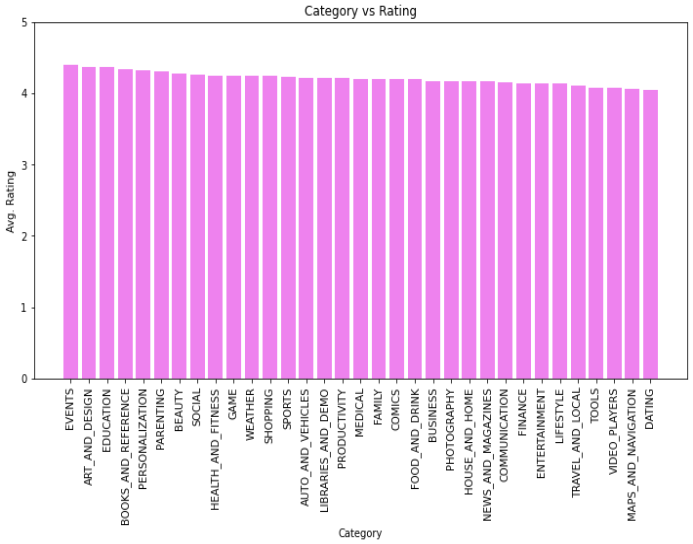


Fig -3.4: Category Vs ratings

In the above plot we plotted categories with highest ratings. By seeing Plot we can say that 'events', 'art and design' and 'education' have slightly high rating.

3.5 AVERAGE RATING OF APPS

After saw the plot We can conclude that most of the people gave rating between 3.5 to 4.8 or we can say that most of the apps have ratting in between 3.5 to 4.8.

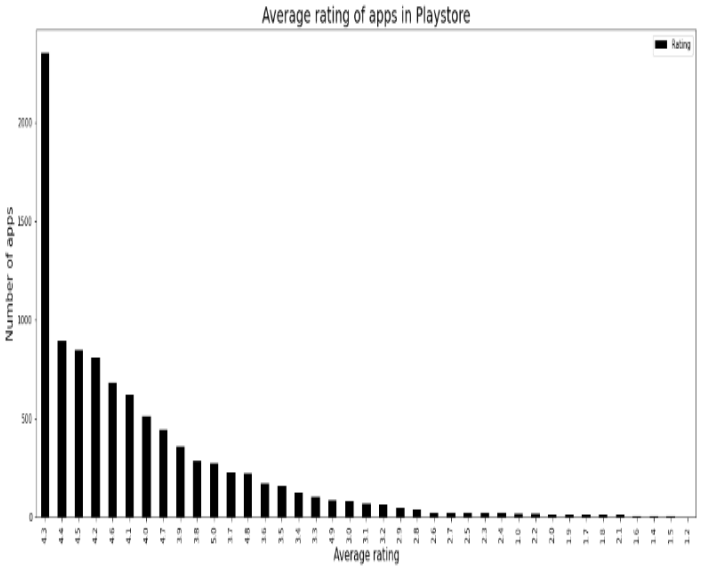


Fig -3.5: Average ratings of apps

3.6 PERCENTAGE OF FREE AND PAID APPS IN STORE

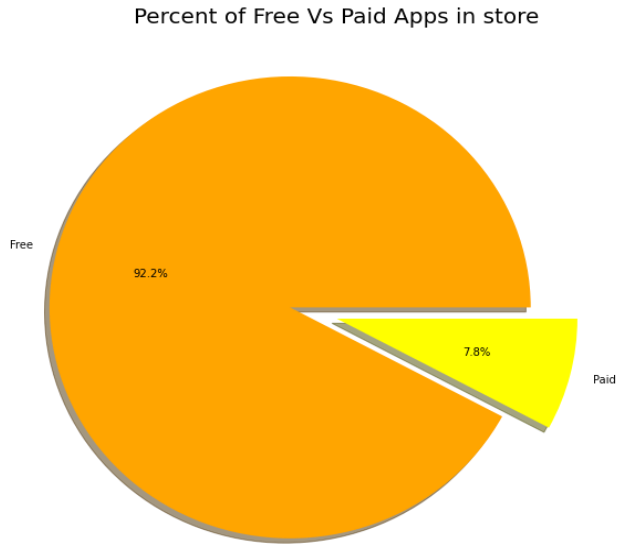


Fig -3.6: Free Vs Paid

Here we can see that 92.2% apps are free, and 7.80% apps are paid on Google Play Store, so we can say that Most of the apps are free on Google Play Store.

3.7 FREE VS PAID CATEGORY WISE

In below plot light blue color shows free apps and red color shows paid apps. By comparing free and paid apps on play store we can say that most of the apps are free. Less no of paid apps shows that people prefer free apps over paid. Every category has more than 75 % of apps that are free for users.

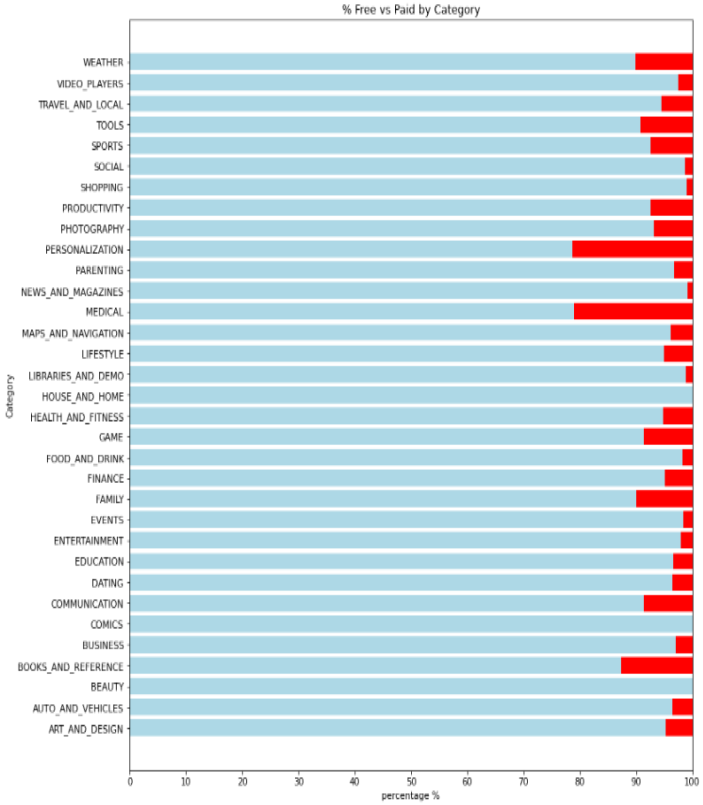


Fig -3.7: Percentage of Free Vs Paid with category

3.8 A. NUMBER OF APPS THAT CAN BE INSTALLED AT A PARTICULAR PRICE

The paid apps charge the users a certain amount to download and install the app. This amount varies from one app to another.

There are a lot of apps that charge a small amount whereas some apps charge a larger amount. In this case the price to download an app varies from USD 0.99 to USD 400.

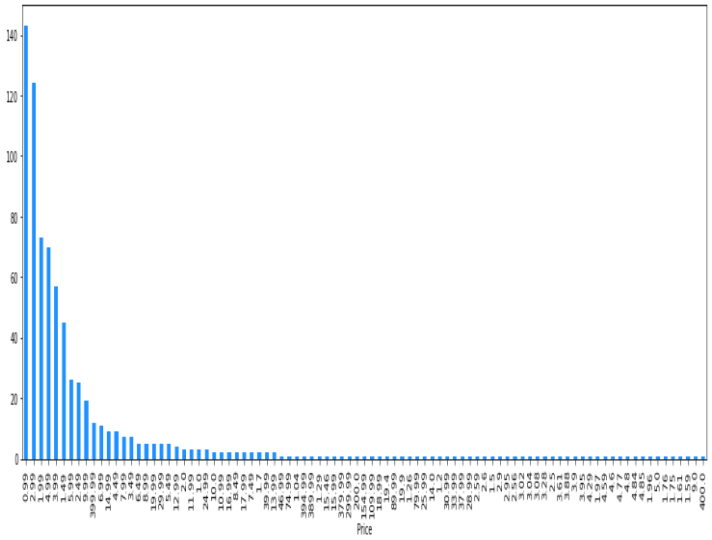
In order to select the top paid apps, it won't be fair to look just into the numer of installs. This is because the apps that charge a lower installation fee will be installed by more number of people in general.

Here a better way to determine the top apps in the paid category is by finding the revenue it generated through app installs.

This is given by:

Revenue generated through installs = (Number of installs)x(Price to install the app)

Fig -3.8 A. Number of apps that can be installed at a particular price



The paid apps charge the users a certain amount to download and install the app. This amount varies from one app to another.

There are a lot of apps that charge a small amount whereas some apps charge a larger amount. In this case the price to download an app varies from USD 0.99 to USD 400.

In order to select the top paid apps, it won't be fair to look just into the numer of installs. This is because the apps that charge a lower installation fee will be installed by more number of people in general.

Here a better way to determine the top apps in the paid category is by finding the revenue it generated through app installs.

This is given by:

Revenue generated through installs = (Number of installs)x(Price to install the app)

3.8 B. CATEGORIES IN WHICH THE TOP 10 PAID APPS BELONG TO

As we can see above fig. shows distribution of category with respect to number of apps in which categories of top 10 paid apps are lies. So that ‘LIFESTYLE’ and ‘GAME’ category are belongs to most paid features.

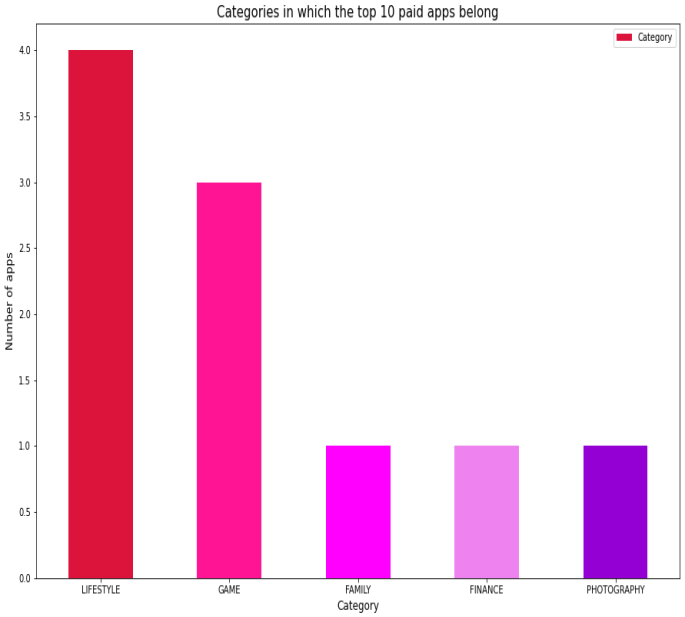


Fig -3.8 B. Category in which the top 10 paid apps belongs to

3.8 C. Top 10 apps based on revenue generated through installation charges

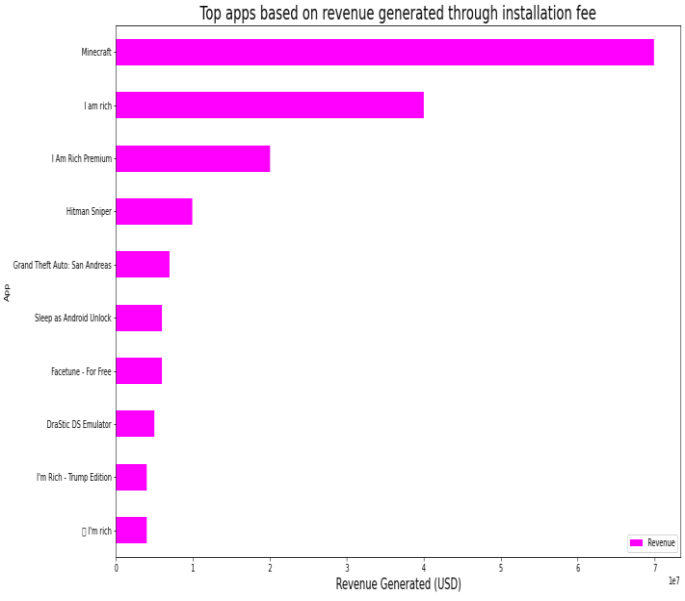


Fig -3.8 C. Top 10 apps based on revenue generated through installation charges

So in given fig. we can determined which top paid apps with revenue generated. Finally, we can plot the graph and find out which are the apps with the highest number of earnings.

Inference:

Top 10 Earning Apps are below in Google -

1) Minecraft

2) I am rich

3) I Am Rich Premium

4) Hitman Sniper

5) Grand Theft Auto: San Andreas

6) Facetune - For Free

7) Sleep as Android Unlock

8) DraStic DS Emulator

9) I'm Rich - Trump Edition

10) 💎 I'm rich

3.9 Apps from the 'Content Rating' column is found more on the play store

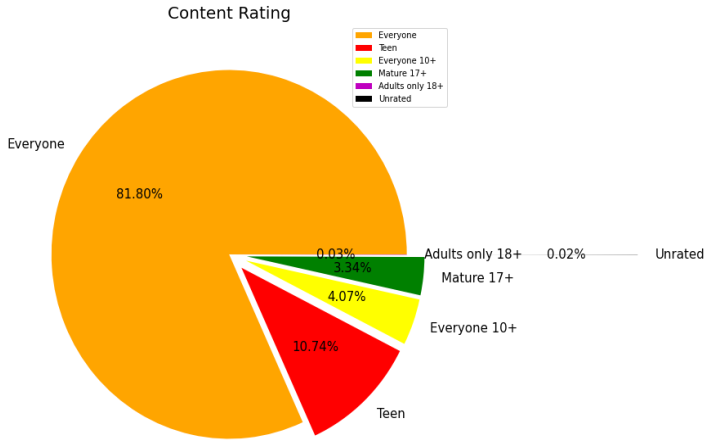


Fig -3.9 Percentage of content rating

A majority of the apps (82%) in the play store are can be used by everyone. The remaining apps have various age restrictions to use it.

We can conclude from above plot that most of the apps on play store are contain the content which is suitable for everyone. And that's because the creater wants to catch more user for their app

3.10 DISTRIBUTION OF APPS IN TERM OF THEIR RATING, SIZE AND TYPE

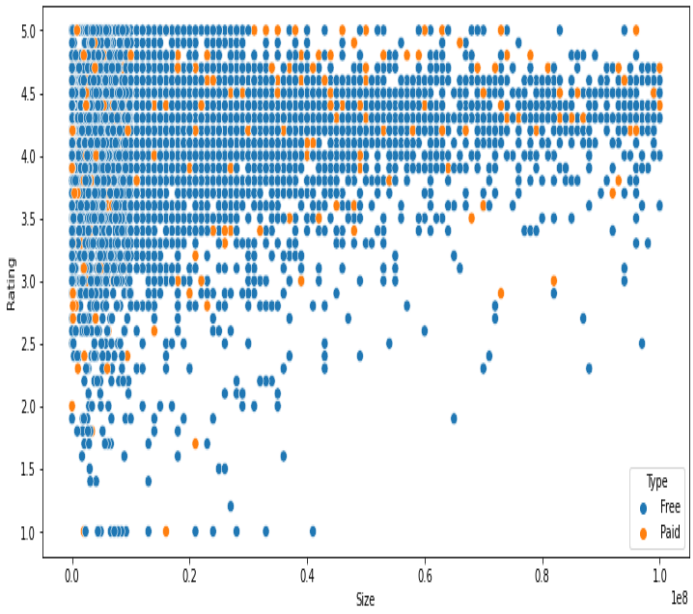


Fig -3.10 distribution of apps in term of their rating, size and type

From this scatter plot, we can imply that majority of the free apps are small in size and having high rating. While for paid apps, we have quite equal distribution in term on size and rating.

3.11 TOP 20 APPS WITH THE HIGHEST NUMBER OF USER REVIEWS

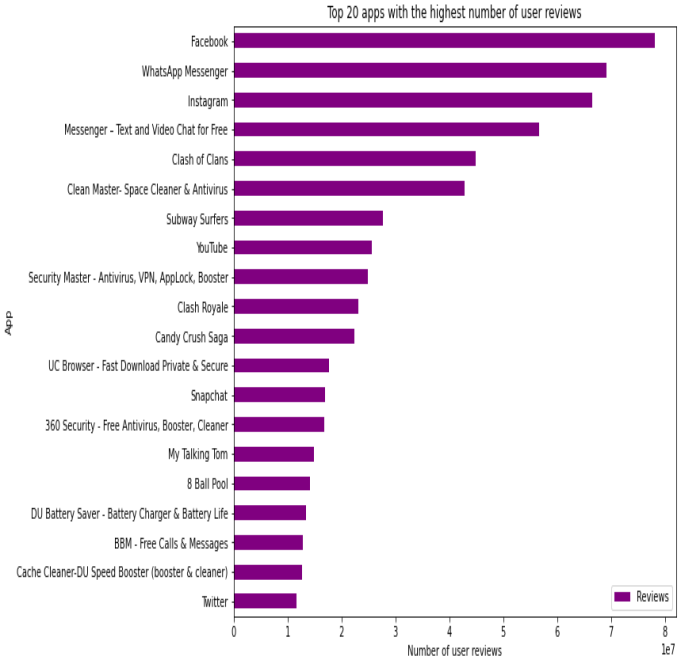


Fig -3.11 Top 20 apps with the highest number of user reviews

We can conclude that "Facebook", "WhatsApp Messenger", "Instagram" are the top three apps have highest reviews.

As we are familiar with this thought that social media is necessary for each and every one that's why people love this applications.

3.12 CORRELATION BETWEEN ALL THE COLUMNS OF THE DATASET

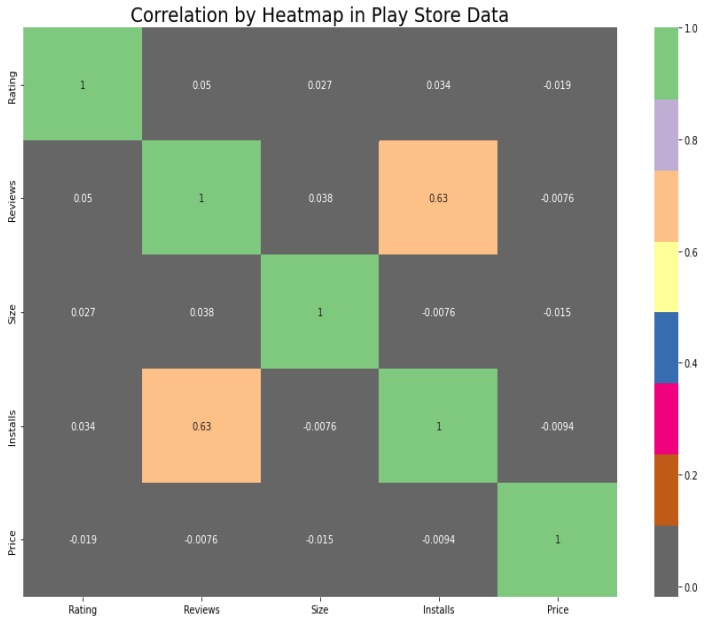


Fig -3.12 Correlation between all the columns of the dataset

Clearly, we saw that reviews and installs are more correlated and the value is 0.64. It is much more obvious that a higher number of installs has a higher number of reviews.

There is a negative correlation between price and install apps, with the price of the app influencing the number of installation of the app.

3.13 PERCENTAGE OF SENTIMENT REVIEWS

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.

Percentage of Review Sentiments-:

Positive Reviews = 64.12%

Negative Reviews = 22.10%

Neutral Reviews = 13.78%

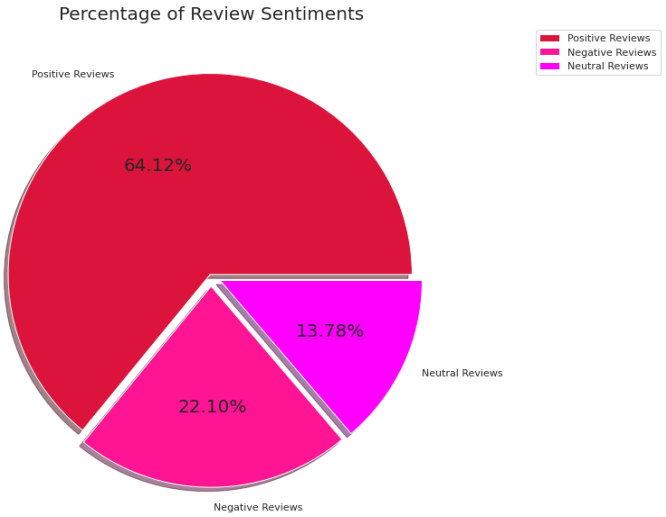


Fig -3.13 Percentage of Sentiment Reviews

3.14 SENTIMENT ANALYSIS BASED ON CATEGORY

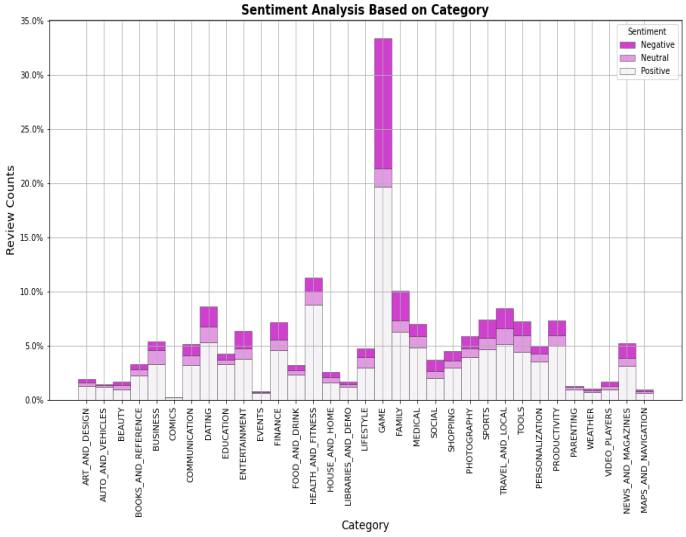


Fig -3.14 Sentiment Analysis Based on Category

Here we notice that most of the reaction are from 'Game' catagory and very less from 'Comics', 'Events', 'Maps_And_Navigation' and 'Weather'.

-This shows that people take much interes in Game catagory app as compare to other apps.

3.15 IS SENTIMENT SUBJECTIVITY PROPORTIONAL TO SENTIMENT POLARITY?

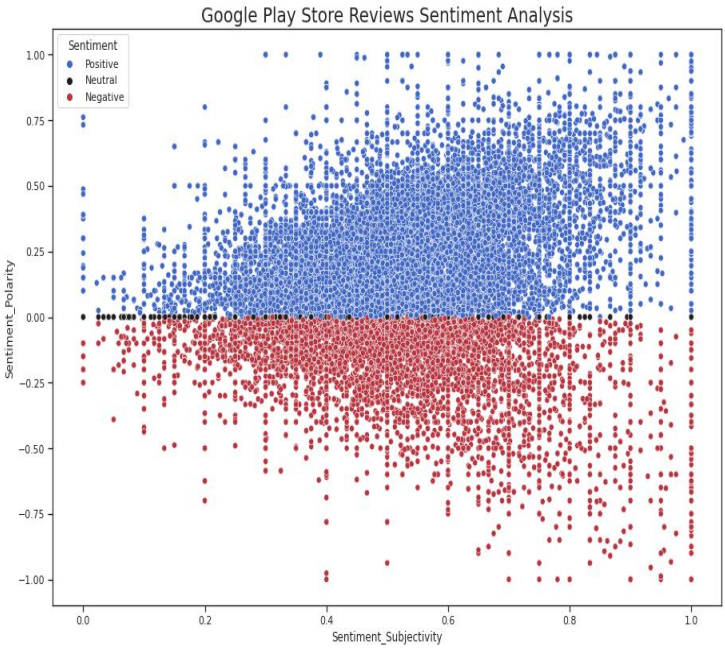


Fig -3.15 Sentiment subjectivity proportional to sentiment polarity

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.

Summary~

- ✓ 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.
- ✓ There are 20 free apps that have been installed over a billion time
- ✓ There are 20 free apps that have been installed over a billion time
- ✓ Minecraft is the only app in the paid category with over 10M installs. This app has also produced the most revenue only from the installation fee
- ✓ Category in which the paid apps have the highest average installation fee: Finance
- ✓ The apps whose size varies with device has the highest number average app installs
- ✓ The apps whose size is greater than 90 MB has the highest number of average user reviews, ie, they are more popular than the rest.
- ✓ Overall sentiment count of merged dataset in which Positive sentiment count is 64%, Negative 22% and Neutral 13%.
- ✓ Sentiment Polarity is not highly correlated with Sentiment Subject

Conclusion~

That's it! We reached the end of our exercise.

After undergoing these algorithms and process, we concluded that our hypothesis is true. Meaning you can predict the app ratings, however significant preprocessing must be done before you start the classification and regression processes.

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.

To deal with those data was really fun with many different analysis. Data cleaning and dealing with duplicate value was also most important to deal with correct methodology. Understand the dataset and predicting the solution of the problem was also great big task

This data set contains a large amount of data that can be used for various purposes. Currently, the decision tree model made using this data set can be used for future developers and Google plays store team to glance at the google play store market and what categories of the apps should be made to keep google play store popular in the future. It can be used to improve business values and google play store in general. It is not just limited to the problem we solved. Using this data set, we applied various classification algorithms and found that Decision tress fits best for our problem statement. We also discovered how different algorithms work in different cases. We found that the Decision tree is easy to visualize and explain the model implementation and it also saves computational power. Using this data set the future work includes the prediction of other parameters such as the number of reviews and installs based on the regression model, identifying the categories and statistics of the most installed apps, exploring the correlation between the size of the app and its version of Android, etc on the number of installs.

The dataset contains possibilities to deliver insights to understand customer demands better and thus help developers to popularize the product.

After analyzing the dataset I got answers to some of the serious and interesting questions which any of the android users would love to know.

Thank You