**Play Store App Review Analysis - Capstone Project**

**Mohd Azeem Ansari**

**Data science trainee,**

**AlmaBetter, Bangalore**

**Abstract:**

The Google Play Store Apps data analysis provides some useful insights regarding the trending of the apps in the play store. As per the graphs visualizations shown above, most of the trending apps (in terms of users' installs) are from the categories like game, communication, and tool even though the amount of available apps from these categories are twice as much lesser than the category family.

The demands of these apps are most probably due to their nature of being able to entertain or assist the user. Besides, it also shows a good trend where we can see that developers from these categories are focusing on the quality instead of the quantity of the apps as most of them are getting positive sentimental support from the users.

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

**1. Problem Statement**

* Application distribution platform, for example, Play Store gets overwhelmed with millions of new applications being launched on the platform regularly. Lots of designers and developers work on it to make an app successful on the Play Store. In this highly competitive world, it is an immense challenge for a developer to know whether they are focusing on the right path to make their app successful on the platform.
* 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**

Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations. Machine learning approaches are essential for us to take care of numerous issues. In this paper, we present machine learning models and structures in detail. Machine learning has numerous applications in numerous perspectives and has incredible advancement potential. It helps you gather insights and make better sense of the data, and removes irregularities and unnecessary values from data. Helps you prepare your dataset for analysis. It allows a machine learning model to predict our dataset better. Gives you more accurate results.

In future, it is predictable that machine learning could set up ideal speculations to clarify its exhibitions. In the meantime, its capacities of unsupervised learning will be improved since there is much information on the planet however it isn't relevant to add names to every one of them. It is additionally anticipated that neural system structures will turn out to be increasingly unpredictable with the goal that they can separate all the more semantically important highlights. In addition, profound learning will consolidate with support adapting better and we can utilize these points of interest to achieve more assignments.

## **3. Google play store**

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.

## **4. Google play store review dataset**

### 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.

## **5. User Data Review Set**

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.

## **6. Data Cleaning and Preparation**

Data Cleaning is itself a big task and one of the most important part of Exploratory Data Analysis. It is the major pre-processing. Undergoing the preprocessing process can help with completeness and compellability. 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 Play store Review 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.

Different steps involved in the process of data cleaning. They are as follows:

* **Step 1**: First importing all the relevant libraries and mounting the drive. And, then loading. Then we import our playstore review dataset as psdf1and user review dataset as psdf2. We write a function psf1.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.
* **Step 2**: After using that function, we can say that in a data frame :
* There are 13 columns of properties with 10841 rows of data.
* Column 'Reviews', 'Size', 'Installs' and 'Price' are in the type of 'object'.
* Values of column 'Size' are strings representing size in 'M' as Megabytes, 'k' as kilobytes and also 'Varies with devices'.
* Values of column 'Installs' are strings representing install amount with symbols such as ',' and '+'.
* Values of column 'Price' are strings representing price with symbol '$'. Hence, we will need to do some data cleaning.
* **Step 3**: Cleaning the 'Reviews' data using clean\_reviews() function, here ‘M’ i.e. Million will be replaced by 1,000,000 and its datatype converted from 'object' to 'float' datatype using astype() function to perform different operations.
* **Step 4**: Cleaning the 'Size' Column data. There we found one value with '1,000+'  in one of record so we can remove it  from our dataframe as it is not clear  whether it is'M'(Megabyte)  or 'k(kilobyte)' using drop() function of pandas library and using clean\_sizes() function removing ‘M’ and converting kilobytes to megabytes so that all elements have same unit and will replace ‘Varies with device’ with value 0. Its datatype is converted from 'object' to 'float' datatype using astype() function to perform different operations.
* **Step 5:** Cleaning the 'Installs' data, removing ‘+’ and ‘,’ using replace() functions and converting its type from 'object' to 'float'.
* **Step 6**: Cleaning the 'Price' data using clean\_price() removing “$” sign from this column. Its datatype is converted from 'object' to 'float' datatype using astype() function to perform different operations.
* **Step 7:** There are 1474 null values in ‘Rating’ column that is a significant amount of data. . 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 So, we will replace these null values with its median using median() aggregate method and fill this value in place of null values using the fillna() function.
* **Step 8:** Handling the duplicates in the App column we drop the no of duplicate rows that are present in the App columns.
* **Step 9:** We write a function psdf1.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.
* **Step 9:** We write a function ps\_df1.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.
* **Step 10:** 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 dropna() function.

**7. Exploratory Data Analysis**

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. 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.

**7.1 Top 5 apps installed by the user**

The top 5 apps installed by the user are:

* Facebook,
* Google Photos,
* Google+,
* Gmail,
* Google Street View.
  1. **Top 5 apps most reviewed by the users**

The top 5 apps most reviewed by the users are:

* Facebook,
* WhatsApp Messenger,
* Instagram,
* Messenger – Text and Video Chat for Free,
* Clash of Clans.
  1. **Top 5 expensive apps available on the playstore**

The top 5 expensive apps available on the play store are:

* I'm Rich - Trump Edition,
* I'm Rich/Eu sou Rico/أنا غني/我很有錢,
* I am Rich Plus,
* I am rich (Most expensive app),
* I Am Rich Premium.
  1. **The 5 apps from the 'Finance' category having the lowest rating**

The 5 apps from the Finance category having the lowest rating are:

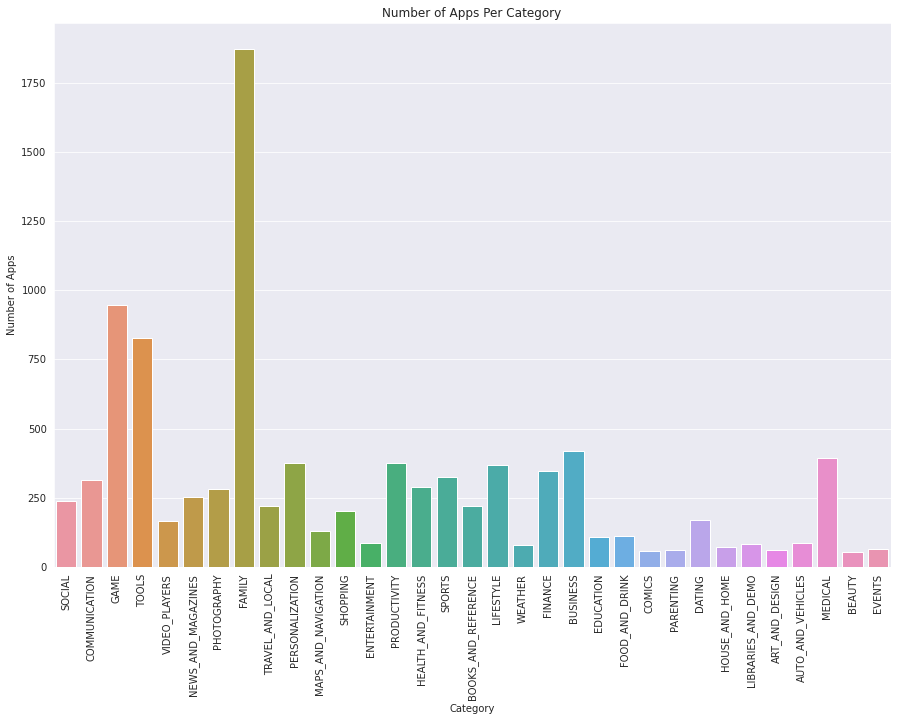
* W CB Mobile Biz,
* Tech CU Card Manager,
* EY TaxChat,
* CB Mobile Access,
* Bitcoin BX Thailand PRO.
  1. **The** **top 5 most installed apps in Game category**

The top 5 most installed apps in Game category are:

* Subway Surfers,
* Candy Crush Saga,
* My Talking Tom,
* Temple Run 2,
* Pou.

**8. Visualization – Different Trends**

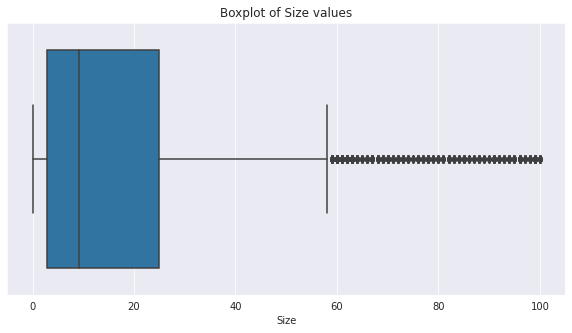
* 1. **Looking for category of apps on a bar graph**

****

**Fig1: Top categories on apps on playstore**

From the above graph we can figure out that most of the apps available on the play store are from the categories of 'Family, 'Game' and also 'Tools’.

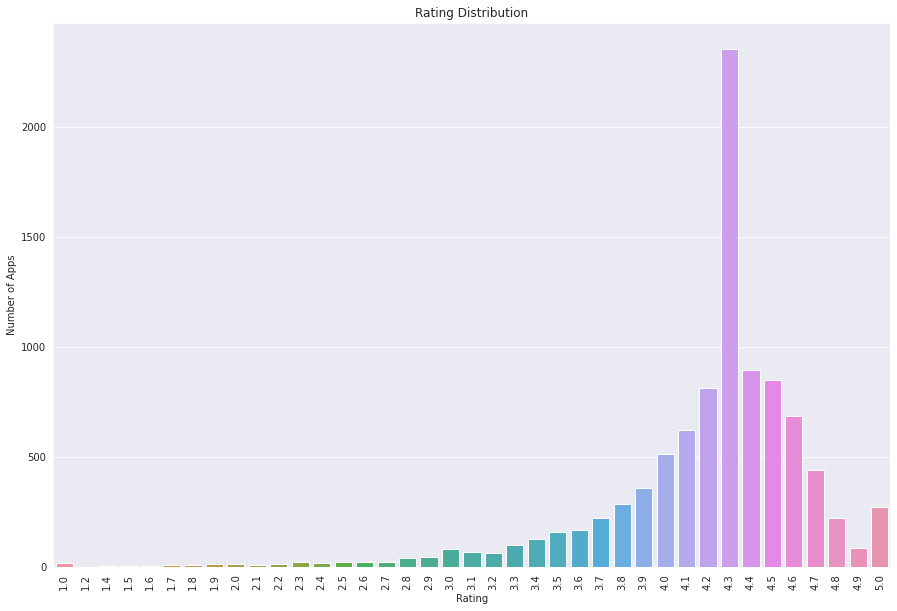
**8.2 Size of Applications**



**Fig 2: Sizes of different apps**

From the above boxplot, it can be concluded that the size of majority of the apps available on the Play Store are within the range of 5Mb to 25Mb and their median size is around 9Mb.

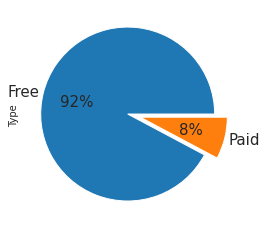
**8.3 Rating**



**Fig 3: Rating Distribution**

From the above plotting, it can be figure out that most of the apps in the Play Store are having rating higher than 4 or in the range of 4 to 4.7.

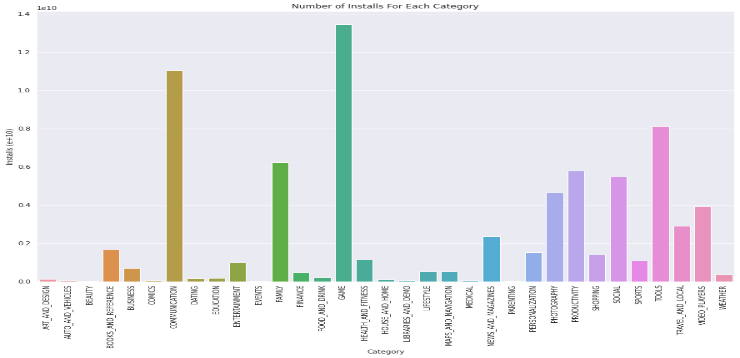
* 1. **Types of Applications**



**Fig 4: Application Types**

From the above pie chart, it can be concluded that majority of the apps available on the Play Store are free apps (i.e. 92%) and only few apps are paid (i.e. 8%).

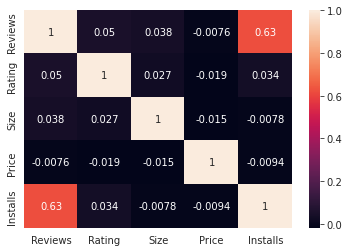
* 1. **Installs trend**



**Fig 5: Installs trend**

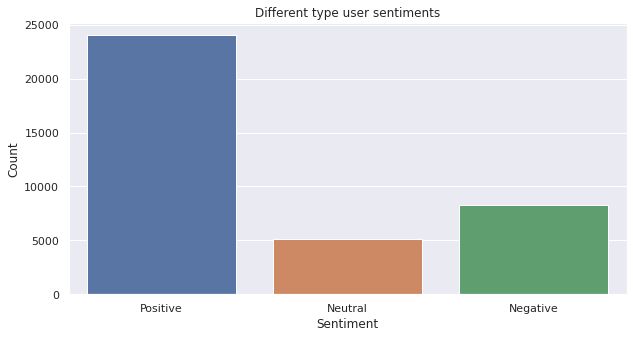
From the above graph, plotting of number of installs for each category, we can conclude that most of the apps being downloaded and installed are from the categories of 'Game' and 'Communication'.

* 1. **Correlation Matrix**



**Fig 6: Correlation Matrix**

Clearly we can see that Reviews and Installs features are more correlated and the value is 0.63.

* 1.  **Sentiments of users towards the apps**

**Fig 7: Sentiments of users**

From the above graph, we can conclude that most of the user have the positive attitude towards the apps available on the app store.

**9. Conclusion**

The Google Play Store Apps data analysis provides some useful insights regarding the trending of the apps in the play store. As per the graphs visualizations shown above, most of the trending apps (in terms of users' installs) are from the categories like Game, Communication, and Tool even though the amount of available apps from these categories are twice as much lesser than the category Family.

The demands of these apps are most probably due to their nature of being able to entertain or assist the user. Besides, it also shows a good trend where we can see that developers from these categories are focusing on the quality instead of the quantity of the apps as most of them are getting positive sentimental support from the users.

On the other hand, the charts shown above actually implies that most of the apps having good ratings of above 4.0 are mostly confirmed to have high amount of reviews and user installs. There are majority of apps which are available free around 92 % and only very few i.e. 8% are paid apps. Furthermore, most of the apps that are having high amount of reviews are from the categories of Social, Communication and Game like Facebook, WhatsApp Messenger, Instagram, Messenger – Text and Video Chat for Free, Clash of Clans etc.

Even though apps from the categories like Game, Social, Communication and Tool of having the highest amount of installs, rating and reviews are reflecting the current trend of Android users, they are not even appearing as category in the top 5 most expensive apps in the store (which are mostly from Finance and Lifestyle).

At the end, we can conclude that the current trend in the Android market are mostly from the categories which are either assisting, communicating or entertaining apps.

**References-**

1. GeeksforGeeks
2. w3schools.com
3. Almabetter study material
4. Python libraries documentation
5. python.org

Thank You