**Play Store Review Analysis**

****

**By - Pratiksha More , Somya Gupta**

**Akshay Shinde , Prashant Gour**

**Data science students**

**Almabetter**

**Abstract**

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.Since most PlayStore applications are free, the income model is very obscure and in accessible regarding how the in-application buys.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. We have used a raw data set of Google Play Store from the Almabetter data.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 GooglePlay Store. This journal talks about different classification 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 visualisation done on this data set.

**Introduction**

Mobile applications are one of the fastest-growing segments of downloadable software application markets. Out of all of the markets we choose Google Play store due to its increasing popularity and recent fast growth.One of the main reasons for this popularity is the fact that about 81% of the apps are free of cost.The market has increased to over 845900 Apps and 226,500 unique sellers. This rapid market has, in turn, led to over 500 million users downloading around 40 billion Apps all over the world. Developers and users play key roles in determining the impact that market interactions have on future technology. However, the lack of a clear understanding of the inner working and dynamic of popular app markets impacts both the developers and users. In this article, we seek to shed light on the dynamics of the Google Play Store and how we can use different features from this data set for prediction purposes. In this article, we will provide a longitudinal study of Google Play app metadata which will give unique information that is not available through the standard approach of capturing a single app snapshot. Using feature extraction from a longitudinal app analysis will be used to ﬁnd whether an app will be successful or not. Our Analysis is divided into four phases: data extraction, data cleaning, data visualisation , and applying different models, and it is depicted . First, we collect the data from the almabetter. In the next step, we try to do data cleaning on the data set to reduce the error percentage. After the data set is ready,we try to analyse the data set using different plots and remove the stuff not needed from the data set. The last step includes using different classiﬁcation algorithms on the dataset to see which one gives the highest percentage of accuracy. Finally, we narrate the analysis results to provide a clear vision of the relationship among the areas of interest.We include a detailed discussion of the applicability and future research directions in the last section called Conclusion and future work.

**1 . Problem Statements**

Data science can be summarised into five steps: capture, maintain process, analyse, and communicate. The analysis of Google Play Store applications aided to build more reliable and more interactive applications. This would be very useful for app developers to build an application focused on certain discussed categories in this analysis. This analysis will help in building the application with precise and accurate objectives.

In the initial phase, we focused more on the problem statements and data cleaning, in order to ensure that we give them the best results out of our analysis.Our major challenge was data cleaning, In Data Cleaning, we have performed a few steps to ensure the data quality such as removing NAN values. During the Data Cleaning step we found that 13.60% of reviews were NaN values, and even after merging both the data frames, we could not infer much in order to fill them. Thus, we had to drop them.

The merged data frame of both play store and user reviews, had only 816 common apps. This is just 10% of the cleaned data, we could have given more valuable analysis if we had at least 70% - 80% of the data available in the merged data frames.

User Reviews had 42% of NaN values, which could have been used for developing an understanding of the category wise sentiments, which would help us to fill 13.60% NaN values of the Reviews column.

With the cleaned data, we have performed Exploratory Data Analysis to understand our dataset like number of installations for each category We explore the correlation between the size of the app and the version of Android on the number of installs and so on.

Our motive in the whole project was to analyse the data and find out main components that affect users' decision to download apps. After completion of analysis I concluded that users prefer more free apps. Most of the apps present in play store are more or less of the same size so size doesn’t affect their decision much.

It was found that Most of the apps that are present on the google play store have ratings in between 4 and 5.Also it was observed that Maximum number of applications present in the dataset are of small size.

We found the most popular category of apps on two bases - Number of Installs and Number of reviews. Personalization wins in former criteria whereas Sports wins in later criteria.

In the problem statement we are given with 2 datasetsi.e.play store and User review data set in the user review dataset it was observed that User Reviews had 42% of NaN values, which could have been used for developing an understanding of the category wise sentiments, which would help us to fill 13.60% NaN values of the Reviews column.

Most of the reviews are of Positive Sentiment, while Negative and Neutral have low number of reviews. 8.Sentiment Polarity / Sentiment Subjectivity

Collection of reviews shows a wide range of subjectivity and most of the reviews fall in [-0.50,0.75] polarity scale implying that the extremely negative or positive sentiments are significantly low. Most of the reviews show a mid-range of negative and positive sentiments.

Sentiment subjectivity is not always proportional to sentiment polarity but in the maximum number of cases, shows a proportional behaviour, when variance is too high or low.

Sentiment Polarity is not highly correlated with Sentiment Subjectivity.

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.

From the results and process we have implemented, we can conclude that we have achieved this group project objective which is analysing the Google Play Store apps and determining trends of the Google Play Store and both of our research questions.

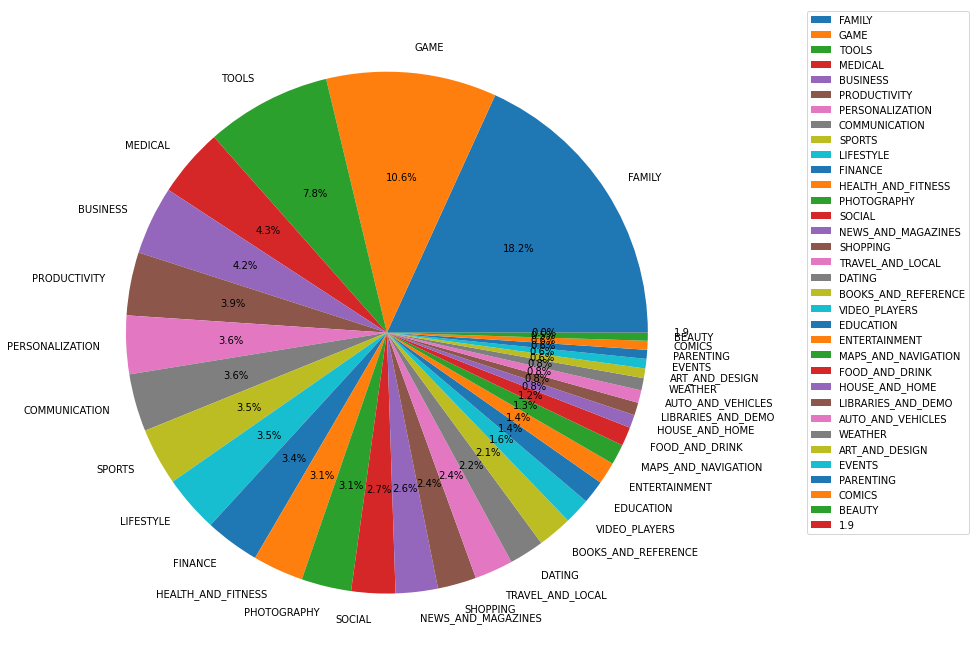
**2. Analysis Methodology**

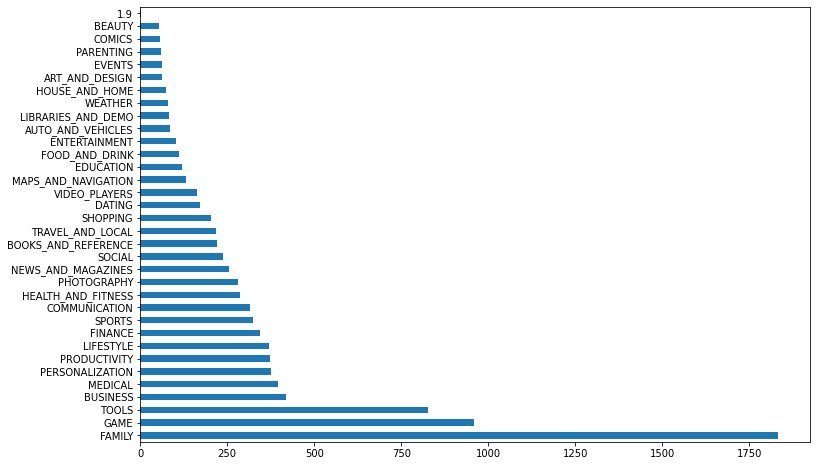
Our analysis approach is divided into three phases: data extraction, data cleansing, and visualisation, and data modelling. In the ﬁrst step, we collected the raw data from almabetter. Then we did basic data cleaning and data visualisation.After visualising the data set, we removed some unnecessary features and made it ready for data modelling.

**2.1. Data Visualisation**

In this data set there are various features that can be used to analyse the data set. In this section we will be analysing different features to ﬁnd which feature determines whether an app will be successful or not. The top ﬁve genres of the google play store include Tools, Entertainment, Education,Business and Medical. In our final ppt we can see that the most number of apps in Google Play store belongs to the categories of Family and game. This shows that apps that belong to the gaming and Family category are more common and apps in this category have high chances of being successful. There are only 7.26% of paid apps in Google play store. The majority of paid apps cost about $1. So the popularity doesn’t depend on this factor.While looking at the popularity feature, you can see that the average rating is quite high, around 4.2 from 5.0. Most ratings are in the range of (4.2 to 4.5). To further deﬁne which category are the highest rated, we will only look at the data for each category that have more than or equal to 4.0 in rating. This will help us in making a more optimal model. To analyse the apps that would produce the most revenue, we will look at the relation table between installs and other parameters. From ppt we can see that installs and reviews have the strongest inverse correlation.This is reasonable because popular apps tend to get more number of reviews.There is no correlation found between installs and other features like size,rating,number of installs and price. There is no correlation between rating and price also. Since the installs parameter is independent and not correlated to any other parameters, we must only use installs to show the popularity of an app. For our analysis we will be only using a number of installs .

* **Category wise percentage and count of no of applications**

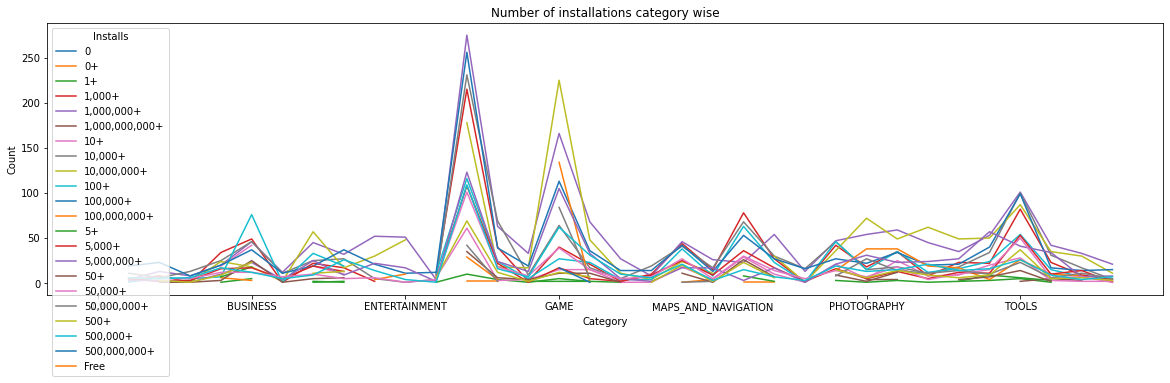




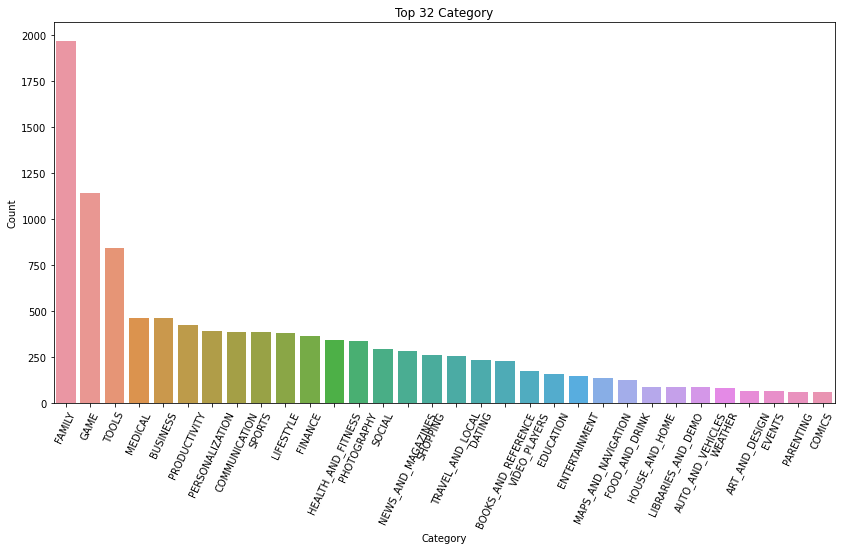
Observation: (Pie Chart) The FAMILY, GAMES and TOOLS are the most dominating applications in the playstore.

* **Comparison top 32 Category present in the google play store as per their installation**

# -**Overall Comparison of Category & install**



This data is much difficult to read and understand so make addition of all the count of install and present comparisons category wise basically it is showing the top category.

Observation : We can see in the above graph that in the top 32 Family, Games and Tools respectively are the most demanded categories with respect to installation

# 

# **Top 32 apps present in the google play store as per their Genres**



Observations: Top Genres like Tools , Entertainment , Education and Medical has the most number of applications.

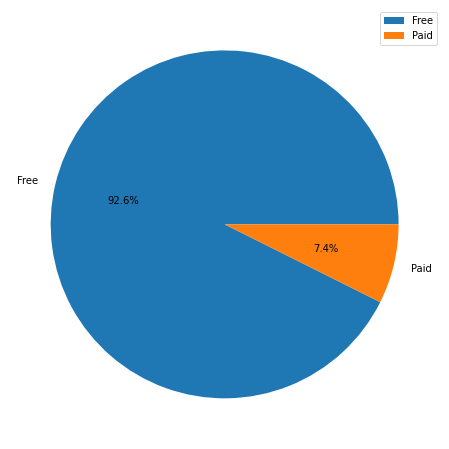
# **Count of Rating in each category differentiated by Customer Ratings**



Observations : For every customer who provides the rating for them we provide the ten coins for each star.

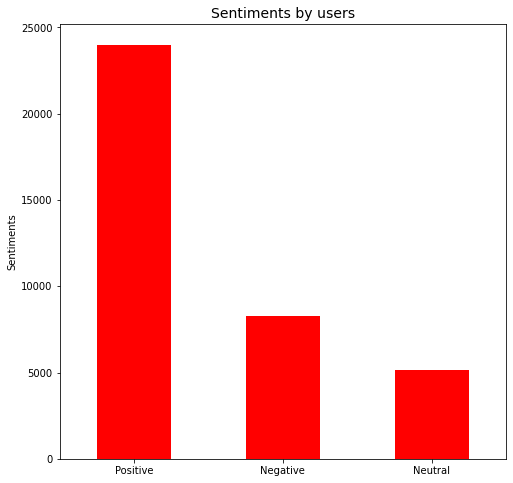
\* From data we find the Total number of coins the play store needs to pay 390803.0 coins.

# **Reviews for paid and unpaid applications**

****

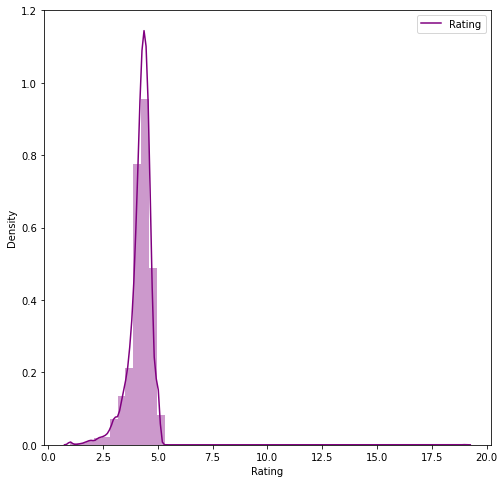
Observations: Mostly Free Apps has installed with well rated .

## **Comparison of positive,negative and neutral sentiments**

****

Observations: Most of the reviews are of Positive Sentiment, while Negative and Neutral have low number of reviews. 8.Sentiment Polarity / Sentiment Subjectivity.

# **Price(Density) V/S Rating comparison**

****

Observations: Ratings according to lies between 4.0 to 5.0 according to categories

**3. Conclusion and future work**

This data set contains a large amount of data that can be used for various purposes. Currently, the data wrangling and visualising this data set can be used for future developers and Google play 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 libraries.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 size of the app and its version of Android, etc on the number of installs .

**Thank You For Reading.**