# Play Store Apps Review Analysis

**Rajat Chaudhary, Anukriti Shakyawar,**

**Raman Kumar, Deepmala Srivastava**

**Team: Web Crawlers**

**Abstract:** Google Play Store is a platform for Android device where people can download or buy different apps, games, and other media. Main objective of this project to performed exploratory data analysis to forecast an app's success or failure based on various criteria. Numpy, Pandas, Matplotlib, Seaborn and math libraries are used to analyse the data. Two data sets were provided to us. The initial dataset was playstore data, which comprises 13 different features that can be used The second dataset included user reviews, which were evaluated based on five different criteria to determine which app was most popular with users.

## Introduction

One of the marketplaces for downloadable software programmes with the highest growth is for mobile applications. Playstore is a platform that provides its users with a variety of digital content, not only an app store. Google created and runs the digital distribution service known as Google Play Store, formerly known as Android Market. It is a legitimate app store with a wide selection of media, including apps, books, magazines, music, movies, and television shows. Since 92.2% of the apps are free, this platform has experienced tremendous growth in popularity. A Google survey found that more than 3000 apps are added every other day. The playstore app's dynamics are revealed in this document, which also offers developers useful information they can use to dominate the android market. Given the explosive rise of Android-based gadgets and applications, it would be fascinating to do data analysis on the collected information to gain insightful knowledge from this data.

## 1.Data Description

### DATASET 1: Playstore App data

In this dataset we had the data of apps which had 13 columns and 10841 entries, the following were:

**1.App:** Name of the apps
**2. Category:** Category under which it falls.
**3. Rating:** Applications rating on PlayStore.
**4. Reviews:** Number of reviews of the app.
**5. Size:** Size of the app.
**6. Installs:** Number of Installations of app.
**7. Type:** Whether the app is Free or Paid.
**8. Prize:** Price of the app if it's a Paid app, for Free apps it's zero.
**9. Content Rating:** Appropriate target rating of the app.
**10.Genres:** Genres under which the app falls.
**11. Updated:** The date on which the app was last updated.
**12. Current Version:** Version of the app.
**13. Android Version:** Minimum android version required to support the app.

## DATASET 2: User Reviews data

In this data set we had the customer review who have experienced those apps, In this we have 5 columns and 64295 entries. Here data set is classified by-

**1. Apps**
**2. Translated Review**
**3. Sentiments**
**4. Sentiment Polarity**
**5. Sentiment Subjectivity**

# 1.Analysis Methodology

Integral research, data cleaning and filtering, and data visualization that make up for our three-part analysis strategy. We started by performing some fundamental research on our dataset. When we did this, we found the basic information regarding our data set such as columns, data types and we also found out that we have missing values, data duplication and a few other issues as well, so we had to run a few steps to extract just the data we needed for exploration. Second, we cleaned up our data by eliminating duplicate entries, converting some variables into usable forms(like $ and + signs used in our data), and filtering out any data with null values. Third, we conducted data visualization. For this, we used a variety of tools, including Numpy, Pandas, Matplotlib, Seaborn, and WordCloud, to accomplish data visualization. We performed these steps on both of our DataSets, firstly on Playstore data and then on our User Reviews dataset.

## 3.1. Data Cleaning

While analyzing our dataset the first thing we will do is to examine the null or missing values in our dataset which is very important to remove because it might affect the accuracy and performance of our analysis and can also show false results at the end of our process. This makes our result accurate. There are many missing values in Size & Rating columns and can be seen by plotting graphs. Hence several methods are used to remove these values.

The first step in any data science effort is cleansing the data. The outcomes are better when the data is cleaner. We initially remove duplicate values from the data before we start cleaning it up. Then we purge extraneous characters from our dataset.
Following this, we identify the distinct values for each column and make the appropriate adjustments, such as changing the data types and getting rid of the null and "nan" values.
Lastly, we have done exploratory data analysis of our dataset.

## 3.2. Data Visualization

In our Data Visualization we performed many analysis to find relations in our data set. In which we first performed basic analysis such as Application Type, categories and so on.

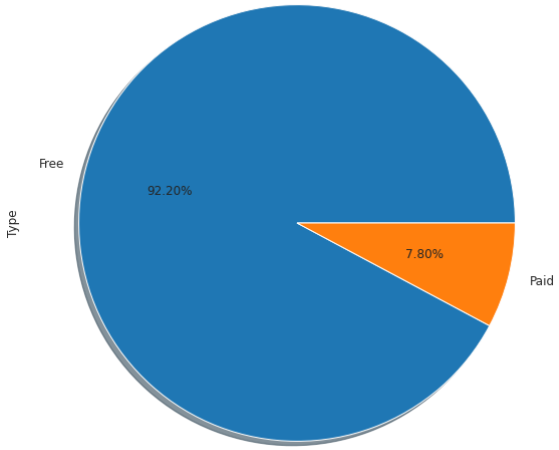



Fig.1: Distribution of Application Type

From the above graph we can conclude that the majority of the apps in the Play Store are Free apps i.e. 92.2% and only 7.80% apps are paid apps.

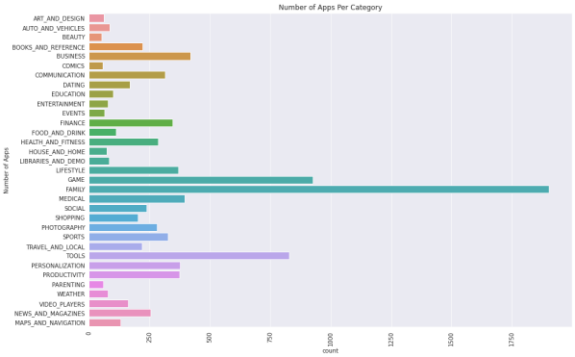



Fig.2: Number of Apps per Category

From this plotting we know that there are many categories in Play Store and most of the apps are from the categories of 'Family', 'Game' and 'Tools' category.

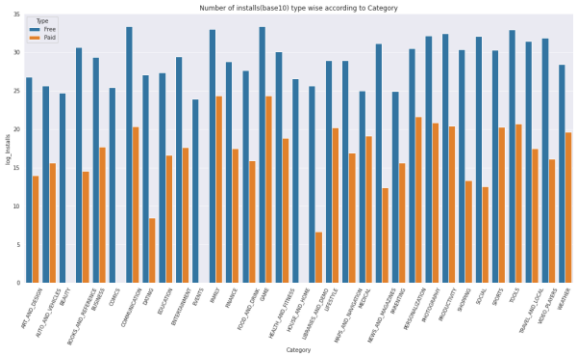



Fig.3: Number of installs type-wise according to categories

This graph shows the number of installations on the basis of Type and Category of apps. In this graph we found out that the app installations have a significantly higher proportion of free software than paid ones. The comparison comparing free and paid apps appears to show no variation because we converted the number of installs to its log.

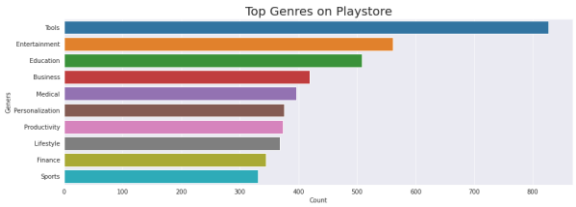



Fig.4: Top genres in playstore

From this plot we can interpret that the top genres in our data set is Tools followed by entertainment, education, business, medical apps and so on.

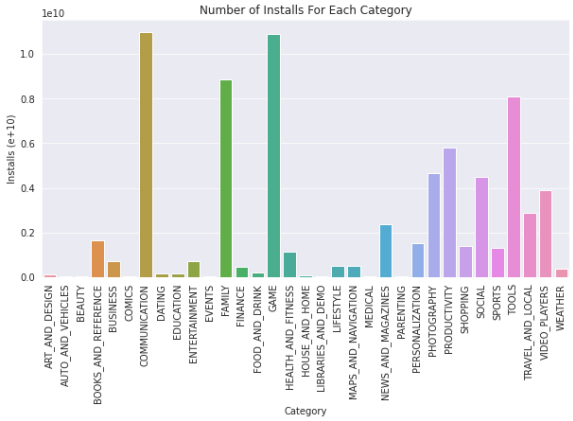



Fig.5: Number of installs for each category

The majority of apps available in the Play Store fall into the Family, Games, and Tools categories, although this is not true according to installations and market demands, as shown by the two plots shown above. Games, communication, and tools have the most installed apps.

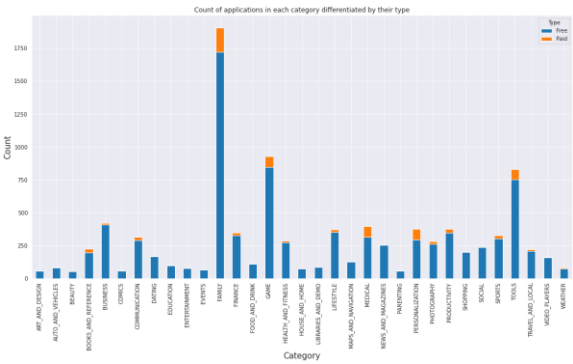



Fig.6: Number of apps per category differentiated by type

This graph shows that some app categories provide more free downloadable apps than others. Most of the apps under the Family, Games, and Tools, as well as Social categories in our dataset, were available for free download. At the same time, the most paid apps were available for download in the Family, Personalization, and Medical categories.

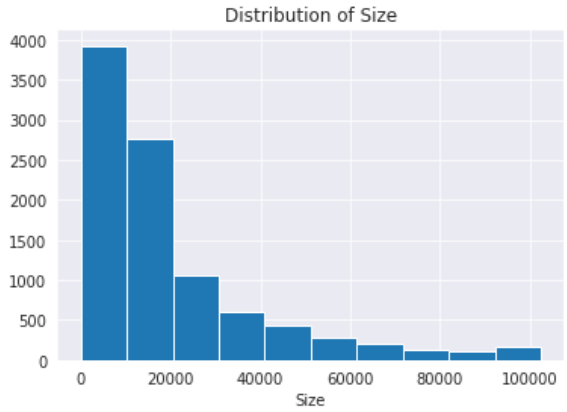



Fig.7: Distribution of Size of app

In this graph we can see that most of the apps present in Play Store are smaller in size and consume less memory and only a handful of apps are larger in size.

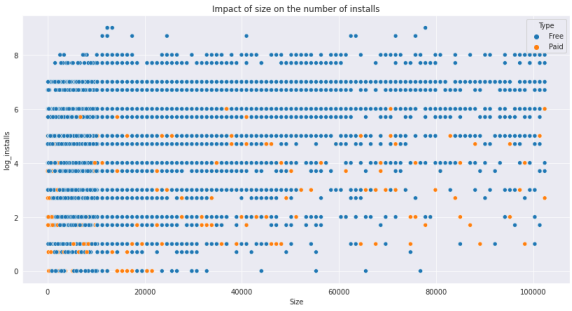



Fig 8. Impact of size on the number of installs

In this graph we can see that the size of application greatly impacts the number of installs by the user. On the other hand we caan see that the bulky applications are less downloaded. We can also derive from this graph that paid applications which are bulky in size are also less installed, which is opposite in the free case, the applications which are bulky are still more downloaded than the paid bulky applications. Hence, we

can say that size affects the installs of apps by the user.

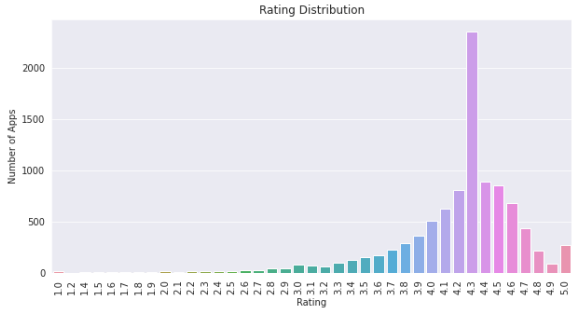



Fig 9. Distribution of App Rating

From this graph we can say that most of the apps in the Play Store are having rating higher than 4 or in the range of 4 to 4.7.

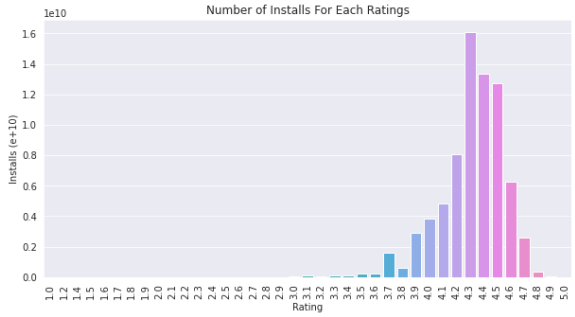



Fig 10. Install per Rating

In this graph we can see that most of the apps downloaded by the customers are of higher rating also which from 4.2 to 4.6 ratings.

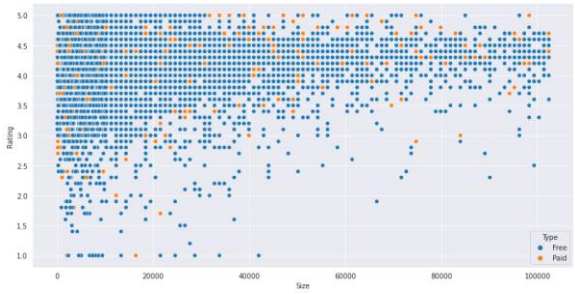



Fig 11. Rating on the basis of Size

In this graph we can see that most of the apps with higher rating are smaller apps in size, we also put paid and free apps in this also but the type of apps are evenly distributed so the type of the app does not affect the rating but the size of the app surely does.

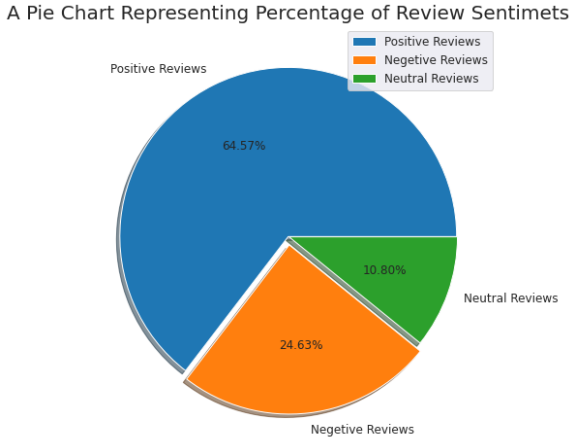
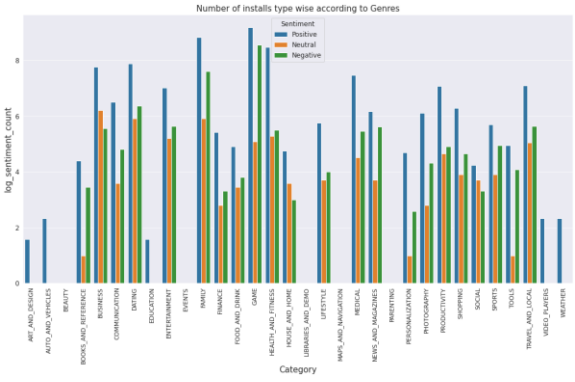



Fig 12. Reviews Sentiment

In this graph we can see that most of the reviews are positive in nature with 64.57% and negative reviews are only 24.63%. There are very less neutral reviews by the users with only 10.8%.




In this we can see that positive reviews are higher in each category of Play Store data, while in categories we can also see that their is very less difference in the positive and negative reviews such as Game and Business.

But there also categories where neutral reviews are also very large as compared to negative reviews such as House and Home and Business. In such conditions the experience of the users also depend for some users it is bad so they provided negative reviews but for users it is not that great so they provided neutral reviews but if such a thing appears in results then it is important of the app to improve their user experience.

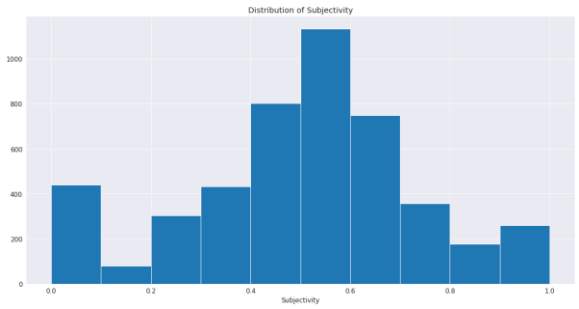



Fig 14. Distribution of Subjectivity

From this graph we can say that the maximum number of sentiment subjectivity lies between 0.4 to 0.7. From this we can conclude that these reviews comes from the experience from the users while using these apps.

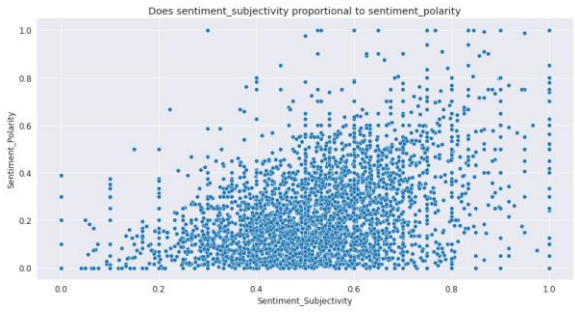



Fig 15. Sentiment subjectivity and Sentiment Polarity

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, it shows a proportional behavior when variance is too high or low.

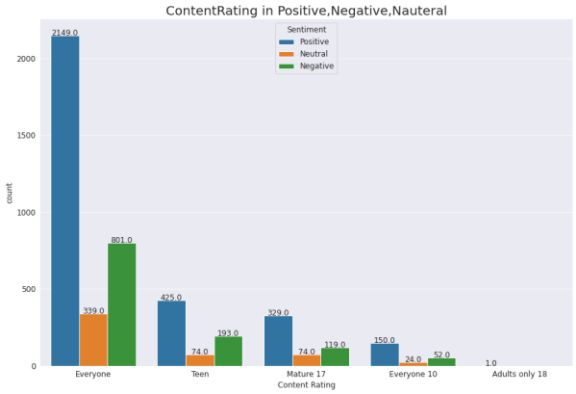



Fig 16. Content Rating on the basis of Age

In this graph we can see that most of the Ratings came from the Everyone category and the ratings gone lesser when we move forward with the age criteria of the apps and gone to even 1 when comes to the adults only category. While talking about the sentiment of the review we can also see that most of the positive ratings came from the everyone category apps with 2149 ratings and 801 negative ratings. In this also we can see that neutral ratings are also low in comparison to positive and negative reviews.

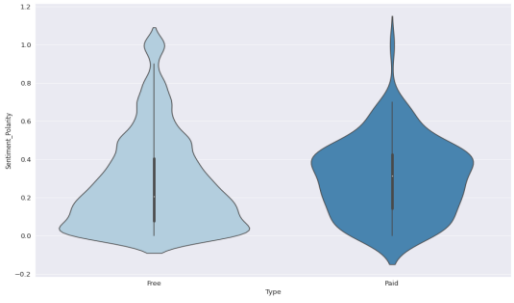



Fig 17. Sentiment polarity relation with type of app

In this graph we can see that in Free apps the sentiment polarity lies majorly in 0.1 and decreases after that which shows that only a handful of users give reviews after experiencing apps, while of the other hand when we the sentiment polarity of paid apps which falls majorly between 0.1 to 0.4 with highest in 0.4, in this we can say that customers first experience these apps and then gave their reviews about those apps.

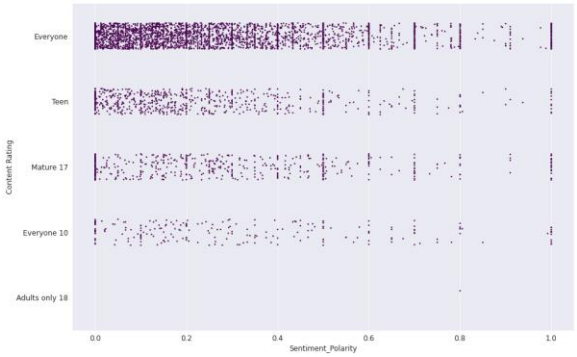



Fig 18. Content Rating relation with Sentiment Polarity

In this we can see that sentiment polarity is low in all categories but in Above 10 category of Content Rating we can see that sentiment polarity is evenly distributed, in this we can say that users have first used the app and then they give their reviews but this can also be seen as less number of reviews in that category. In Everyone category we have highest number of reviews but major part of those reviews lies between 0.1 to 0.4 and not more reviews are going higher than that, in this we can say that customers have not used those apps and gave their reviews beforehand.

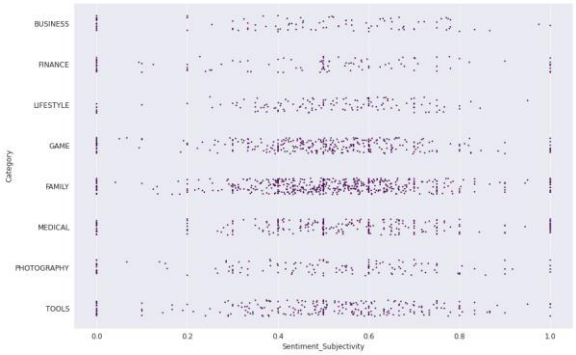



Fig 19. Categories Relation with Sentiment Polarity

In this graph we can see that Sentiment Polarity lies between 0.2 to 0.8 in all Categories, we can also see that the Family Category has the higher number of reviews in which Sentiment Polarity lies between 0.4 to 0.6 which shows that apps in this Category have been reviewed after using them.

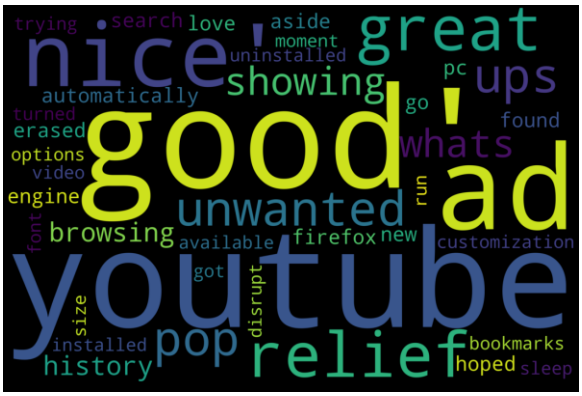



Fig 20. Wordcloud of Reviews

In this Wordcloud we can see what words are mostly used in the reviews.

## Conclusion

The Google Play Store Apps report 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 trending 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.

Other than that, 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 some spikes in term of size and price but it shouldn't reflect that apps with high rating are mostly big in size and pricy as by looking at the graphs they are most probably are due to some minority. Futhermore, 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.

Eventhough 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). As a conclsuion, we learnt that the current trend in the Android market are mostly from these categories which either assisting, communicating or entertaining apps.

## Future Work

We can explore the correlation between the size of the app and the version of Android on the number of installs we can also explore reviews and sentiment of the users as per the the category of the application.

In order to improve the program we could add a system that would create application on its own by using the data set and creating the best user interface by highly rated apps.

## References


1. https://jovian.ml/ritz1602-rs/course project-google-play-store-dataset
2. https://jovian.ai/learn/data-analysis-with-python-zero-to-pandas
3. https://seaborn.pydata.org/examples/index.html
4. https://matplotlib.org/3.1.1/index.html