# Google Play Store Analysis

**Hark Pun**
**Data Science Trainee,**
**AlmaBetter, Bangalore.**

## Abstract:

Mobile app distribution platform such as Google play store gets flooded with several thousands of new apps everyday with many more thousands of developers working independently or in a team to make them successful. With immense competition from all over the globe, it is imperative for a developer to know whether he is proceeding in the right direction.

Thus, an app's success is usually determined by the number of installs and the user ratings that it has received over its lifetime rather than the revenue it generated. In this thesis, on a smaller scale, we have tried to perform exploratory data analysis to dive deeper into the Google Play Store data that we collected, discovering relationships with specific features such as how the number of words in an app name for instance, affect installs, in order to use them to find out which apps are more likely to succeed.

***Keywords: exploratory data analysis, Play store app, user review***

## 1. Problem Statement:

The expansion of smart phones is driving the fast development of mobile app stores. We have picked Google play store and did a thorough analysis of its features that were available to us for predicting the success of a particular app.

i. As the mobile industry is growing rapidly it is increasing the level of competition however, increased competition also leads to increased chances of failure. So, the developers need to do enough research as an enormous amount of time, effort and the money are invested into the process, so business cannot afford an app failure.

ii. Play store works on a 70:30 payment distribution ratio Building an app is therefore not an easy task to deal with, as there are a bunch of developers creating app every now and then. But to reach the level of success that what makes an app stand out in the crowd.

iii. The lower the number of downloads the less it has a chance that it will do great business ahead in future in the android market.

This is one big problem that we tried to solve in our EDA is a thing that does not come to one so easily and for that, we analysed the features of Google play store and came to a conclusion that will help developer to understand their app success rate using our proposed success parameter.

## 2. Introduction:

It has been observed that the significant growth of the mobile application market has a great impact on digital technology. Having said that, with the ever-growing mobile app market there is also a notable rise of mobile app developers; eventually resulting in sky-high revenue by the global mobile app industry.

We used two types of datasets to perform EDA one is 'Play Store Data' and another one is 'User Review Data'. The dataset that we used had 9659 unique apps and contained over 4814617393 reviews. For the EDA part, we analysed our dataset and created several charts to visualize the relationship between each attribute.

After performed the EDA on play store data, we observed the importance of each feature and the correlations between each of them. Hence, from the analysis, it can be seen that determining the success rate of an app will play a very important role for developers and can bring certain changes that might affect the lifetime of their app

Attributes Information:

**User Review Data**

- **App -** Application name
- **Translated Review** – User review
- **Sentiment -** Positive/Negative/Neutral
- **Sentiment Polarity** - Sentiment polarity score
- **Sentiment Subjectivity** – Sentiment subjectivity score

**Play Store Data**

- **App -** Application name
- **Category** - Category for which app belong
- **Rating** - Overall user rating
- **Reviews** - Number of reviews received
- **Size** - Amount of memory that app can occupy
- **Install** - Number of user downloads
- **Type** -Whether app is free or paid
- **Price** - Price for app if app is paid
- **Content Rating** - For user usage based on age criteria
- **Genres** - From which genres that app belong
- **Last Update** - Latest update release date

# 3. Steps Involved:

To extract the insights from the dataset we are doing EDA on the given dataset.

EDA consists of –

- **Data Exploration** - Analyse and explore the data.
- **Type Casting** - Converting datatype of 'Reviews', 'Install', 'Rating', 'Price' columns into numerical format from object and also converting 'Last Update' column into datetime format from object datatype.
- **Data Cleaning** - Dropping irrelevant columns, removing null values and deleting the duplicate row which are present in dataframe.
- **Data Imputation** - Imputing null values with mean value for numerical data and imputing null values with mode value for categorical data.
- **Data Visualization** - Plotted several graphs and extracted more insights from our dataset which will help the developers in rolling out the app in the play store.

# 4. Exploratory Data Analysis

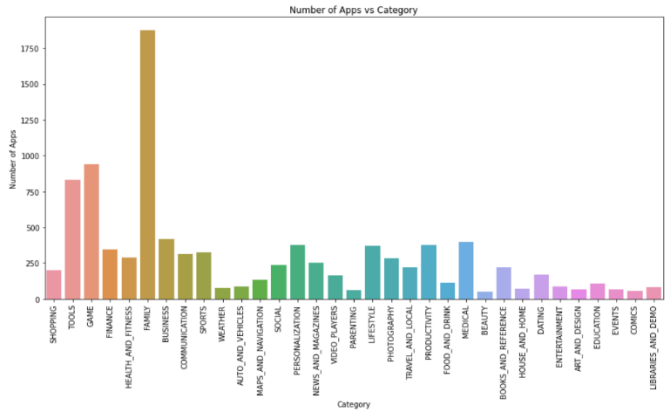
In this chapter we analyze our dataset to summarize their main features with visual representations to see what the data can tell us beyond the formal modelling

### 4.1 App Information

After analyzing first dataset, we found out an information that it contains total 9659 unique app present in play store dataset.

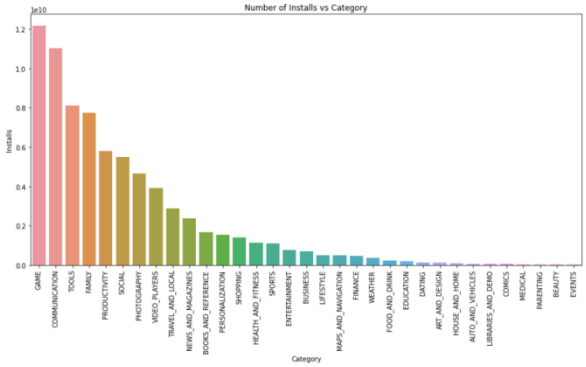
### 4.2 Category vs App

The Category column in our dataset has 33 different types of categories. Figure shows the bar chart of the number of categories. Top 5 category in the play store which have the greatest number of apps are – FAMILY, GAME, TOOLS, BUSINESS & MEDICAL.



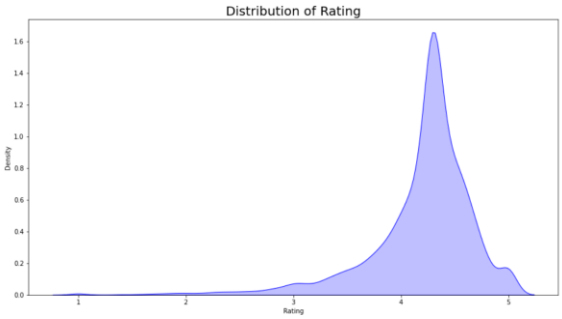

### 4.3 Category vs Installs

Figure shows that while Games still rule the Play Store, FAMILY, previously the 2nd most dominant by app numbers, has been dethroned by COMMUNICATION category. Similarly, the TOOLS category which previously occupied the 3rd spot still have on third places. Hence, looking at categories by apps might be misleading for a developer. A developer wanting to attract a large user base should pick a category based on the number of installs and not by the number of apps in the Play Store.



### 4.4 Distribution of Ratings

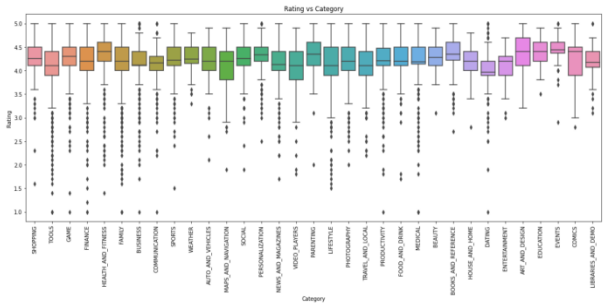
Figure show that most of the number of app are rated between 3.5 to 4.7.




### 4.5 Ratings vs Category

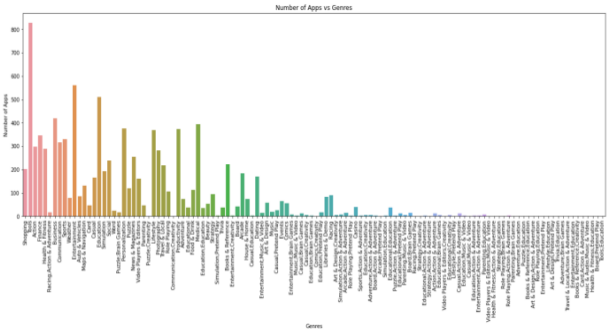
Figure shows box where top 5 categories with highest number of ratings are EVENTS, ART_AND_DESIGN, COMICS, EDUCATION, HEALTH_AND_FITNESS.

App with lowest ratings are- DATING, ENTERTAINMENT, MAPS_AND_NAVIGATION, WEATHER, BEAUTY



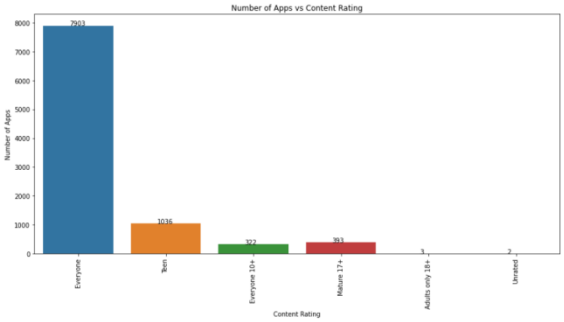

### 4.6 Apps vs Genres

Top 5 Genres which are having the greatest number of apps in play store are - TOOLS, ENTERTAINMENT, EDUCATION, BUSINESS & MEDICAL.



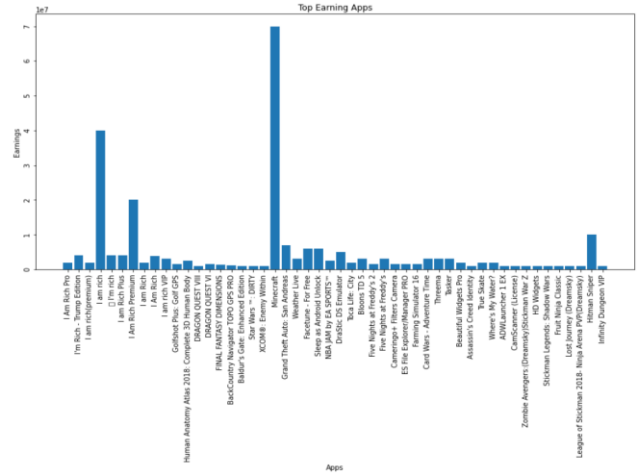

### 4.7 Apps vs Content Rating

Most of the apps are present in play store can anyone use with no age restriction.



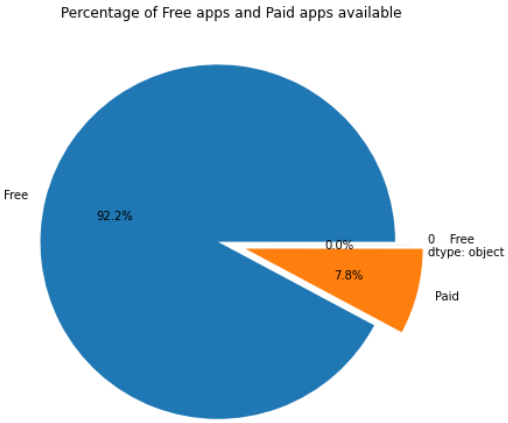

### 4.8 Apps that have made highest earning.

The top five apps with the highest earnings found on google play store are:- Minecraft, I am Rich, I am Rich Premium, Hitman Sniper, Grand Theft Auto: San Andreas



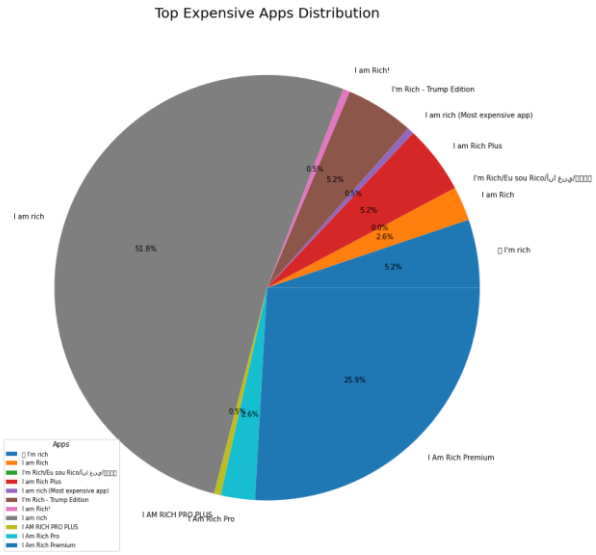
### 4.9 Free app vs Paid app

Most of the apps present in play store is Free.



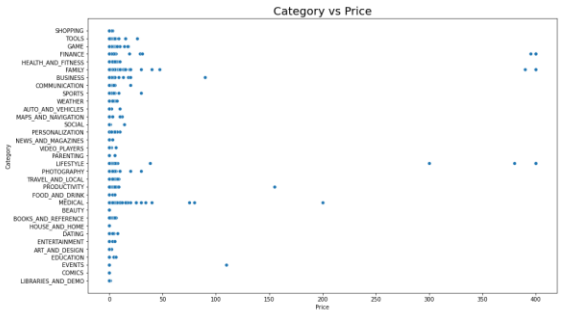

### 4.10 Category vs Price

Highest distribution of app is 'I am rich', 'I Am Rich Premium', "I'm rich-Trump Edition".



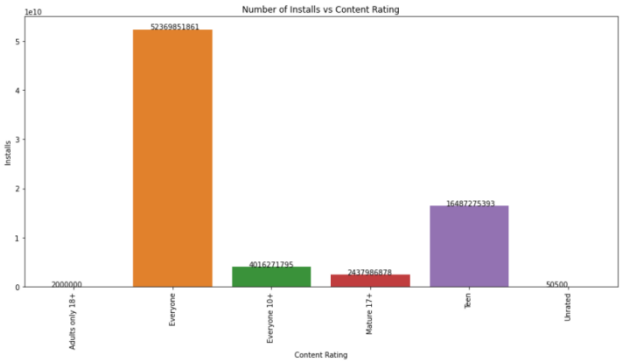

### 4.11 Category vs Price

The highest paid applications are FINANCE, LIFESTYLE, and FAMILY.



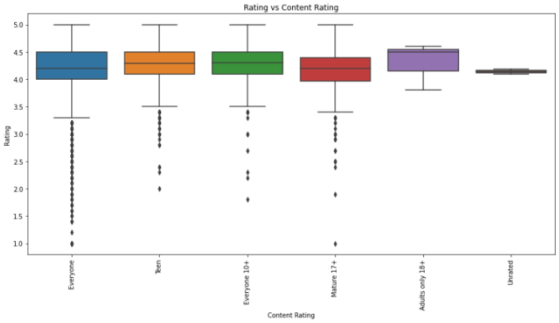

### 4.12 Installs vs Content Rating

Since there is huge number of apps with content rating Everyone in the play store compared to other content ratings therefore the number of installs is also much higher for apps with content rating everyone.

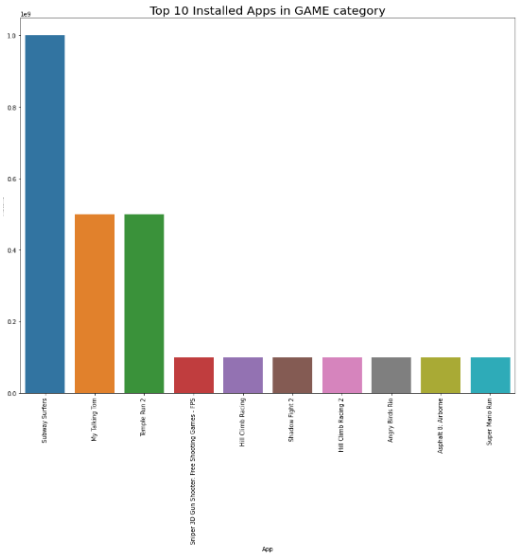


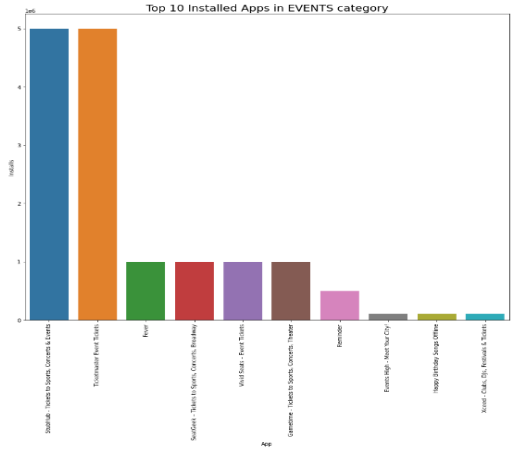

### 4.13 Rating vs Content

Adults Only 18+ contents are having the highest ratings of 4.5 followed by Everyone 10+ with 4.3 then follows Teen with rating 4.2, Everyone with rating 4.2, Mature 17+ with rating 4.2 and Unrated with rating 4.1



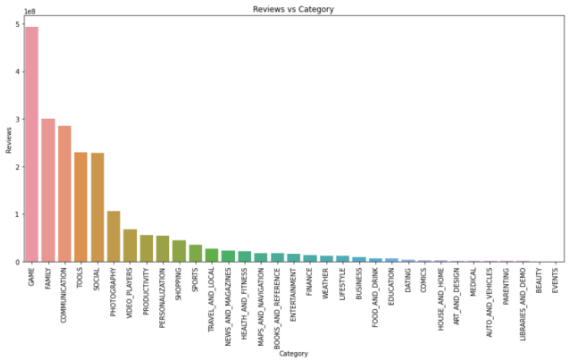
### 4.14 Top 10 App in specific Category



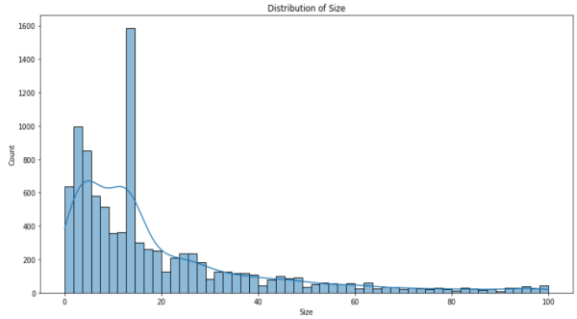

### 4.15 Reviews vs Category

Top categories having most number of reviews in the play store are - GAME, FAMILY, COMMUNICATION, TOOLS & SOCIAL




### 4.16 Distribution of Size

From the above graph we can infer that most of the apps are of smaller size.

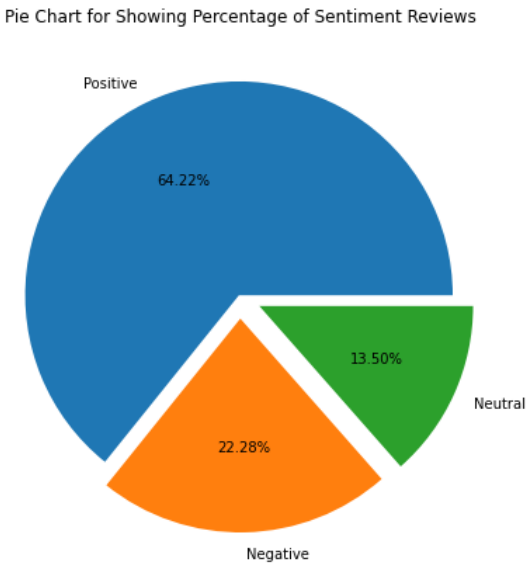



## Merging the User Review dataset with the Play Store dataset

There are three main columns in the User Review dataset - Sentiment, Sentiment Polarity and Sentiment Subjectivity.

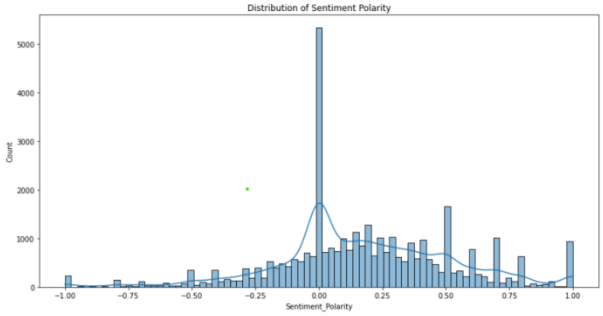
- **Sentiment** – Sentiment is the view or opinion of the user about the app. So, the view/opinion may be Positive, Negative or Neutral.
- **Sentiment Polarity** - Sentiment polarity column contains values from -1 to 1. Where -1 is the most negative polarity and 1 is the most positive polarity. This column can also contain 0 which means neutral polarity.
- **Sentiment Subjectivity:** Sentiment subjectivity contains values ranging from 0 to 1. Where 0 being the very much objective sentence and 1 is very much subjective. Subjectivity refers to the degree to which a person is personally involved in an object.

### 4.17 Pie chart for Sentiment



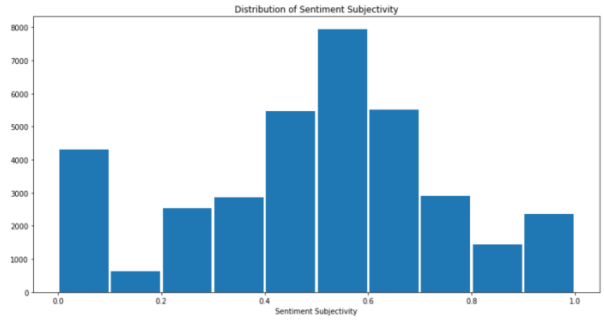

### 4.18 Distributions of Sentiment Polarity

The above graph shows that the width of the distribution is more towards the left of the graph which makes it left skewed. So, the Polarity of most of the users is towards the positive side as we already saw in the pie chart. Also, most of the reviews are having 0 polarity.



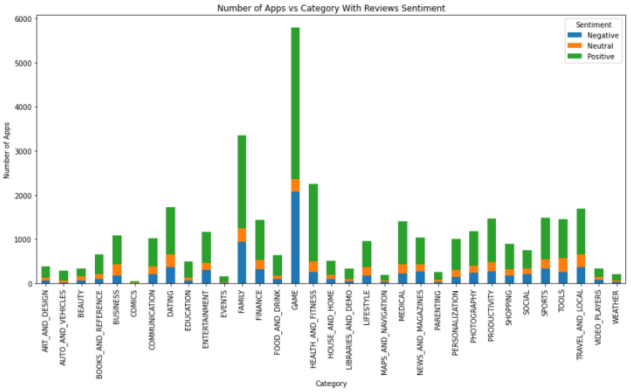

### 4.19 Histogram plot of Sentiment Subjectivity

From the above histogram plot we can infer that most the sentiment subjectivity lies between 0.4 to 0.7 which shows that most of the reviews are towards subjective point of view of the users.



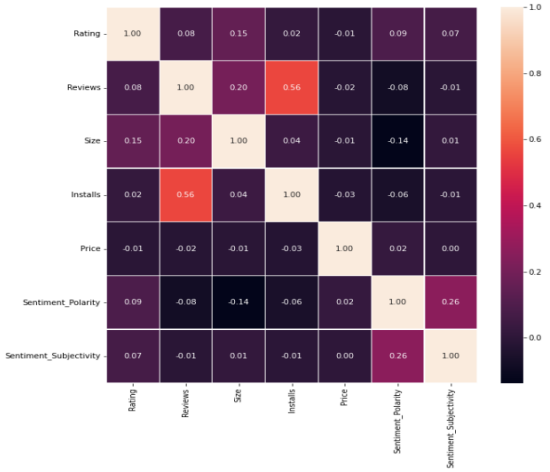

### 4.20 Apps vs Category with Reviews Sentiment Stacking

The Top 5 Categories with most positive reviews are – GAME, FAMILY, HEALTH_FITNESS, DATTING, TRAVEL_AND_LOCAL




### 4.21 Correlation heat map for the final play store dataset

Relationship between Install and Reviews is 0.56 and Relationship between Total Earning and Price is 0.63



## 5. Result

In this EDA the given datasets are analysed, and several graphs has been plotted which can be used to give more insights to the dataset.

- Most competitive category: Family
- Category with the highest number of installs: Game
- Category with the highest average app installs: Communication
- Percentage of apps that are top rated = ~80%
- Percentage of apps with no age restrictions = ~82%
- 92.2% apps are free, and 7.8% apps are paid apps.
- Most the sentiment subjectivity lies between 0.4 to 0.7.
- I'm Rich - Trump Edition is the costliest app with price tag of $400.0.
- Communication, Tools, Productivity, Social & Photography are the top 5 Genres with the greatest number of installs.
- Distribution of Size shows most of the app's present in the play store are of smaller size.
- 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
- Most popular app in the Play Store based on the number of reviews: Facebook
- 64.2% of reviews are of positive sentiment, 22.3% are of negative sentiment and 13.5% are of neutral sentiment.
- Installs is showing fairly good relation with Reviews. Size and Reviews are slightly correlated. Sentiment Polarity and Sentiment Subjectivity are slightly correlated.
- The apps whose size is greater than 90 MB has the highest number of average user reviews, i.e., they are more popular than the rest.
- Helix Jump has the highest number of positive reviews and Angry Birds Classic has the highest number of negative reviews.

## 6. Conclusion

The dataset contains possibilities to deliver insights to understand customer demands better and thus help developers to popularize the product. Dataset can also be used to look whether the original rating of the app matches the predicted rating to know whether the app is performing better or worse compared to other apps on the Play Store.

## 7. Reference


- https://www.kaggle.com/datasets/lava18/google-play-store-apps
- https://datahack.analyticsvidhya.com/blogathon/
- https://www.geeksforgeeks.org/
- https://medium.com/