Capstone Project

Google Play Store

Exploratory Data Analysis

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

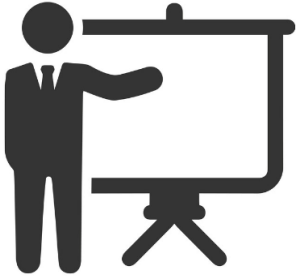
What is Google Play?

- The Play Store is Google’s global online digital content store, where they work to connect users to outstanding digital experiences by creating the best discovery and distribution platform.
- Google Play, also known as the Google Play Store, is where you can download or buy millions of apps, games, and other media onto your Android device. You can find programs for a wide array of interests.
- Many apps will be free, some will have in-app advertisements, some will cost money, while others may offer in-app purchases, or a combination of any of these things.

Content

- Introduction
- Problem Statement
- Attributes
- Steps Involved in this EDA
- Visualisation
  - Number of app based on Category
  - Number of App Installed based on Category
  - Box Plot showing Rating distribution-based Category
  - Distribution of Rating across the whole dataset
  - Number of Reviews against each Category
  - Installation vs Genres
  - Number of Apps vs Content Rating
  - Expensive App Distribution
  - Apps that have made highest Earning
  - Number of Installs in each Content Rating
  - Pie Chart for Percentage of Free Apps and Paid Apps
  - Number of Installs vs Category Based on Type
  - Distribution of Size
  - Pie Chart for Percentage of Sentiment Reviews
  - Distribution of Sentiment Polarity
  - Histogram Plot for Sentiment Subjectivity
  - Correlation Heat Map
- Conclusion

Introduction



- The expansion of smart phones is driving the fast development of mobile app stores. Currently, the two largest global platforms for app distribution are Apple’s App Store (for iOS users) and Google play store (official app store for the Android OS).
- Android is the dominant mobile operating system today more than 87% of all mobile devices running Google’s OS. The Google Play Store is the largest and most popular Android app store.
- 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.
- With immense competition from all over the globe, it is important for a developer to know that he is proceeding in the correct direction. To retain this revenue and their place in the market the app developers might have to find a way to stick into their current position.

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.

- In this problem there are two given datasets –

  - **Play Store data** – This dataset contains the information of the apps like category, genres, price etc.

  - **User Review data** – This dataset contains the reviews given to the various apps and the sentiment of the reviews.

- 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.

- 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.

Attributes Information

Play Store Data –

- **App:** Name of the app
- **Category:** Category of the app.
- **Rating:** Current average rating (out of 5) of the app
- **Reviews:** Number of user reviews given on the app
- **Size:** Size of the app in MB (megabytes)
- **Installs:** Number of times the app was downloaded
- **Type:** Whether the app is paid or free
- **Price:** Price of the app in US$
- **Last Updated**: Date on which the app was last updated

User Reviews –

- **Apps:** Name of the app
- **Translated Review:** Reviews given to each app.
- **Sentiment:** Sentiment of reviews Positive/Negative/Neutral.
- **Sentiment Polarity:** Sentiment polarity score from -1 to 1.
- **Sentiment Subjectivity:** Sentiment subjectivity score.

Steps Involved in this EDA

Data Exploring

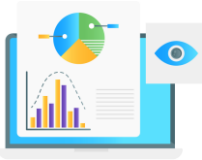
- Checking the summary.
- Converting Reviews, Size, Installs & Price attributes from object type to numeric type.

Data Preprocessing

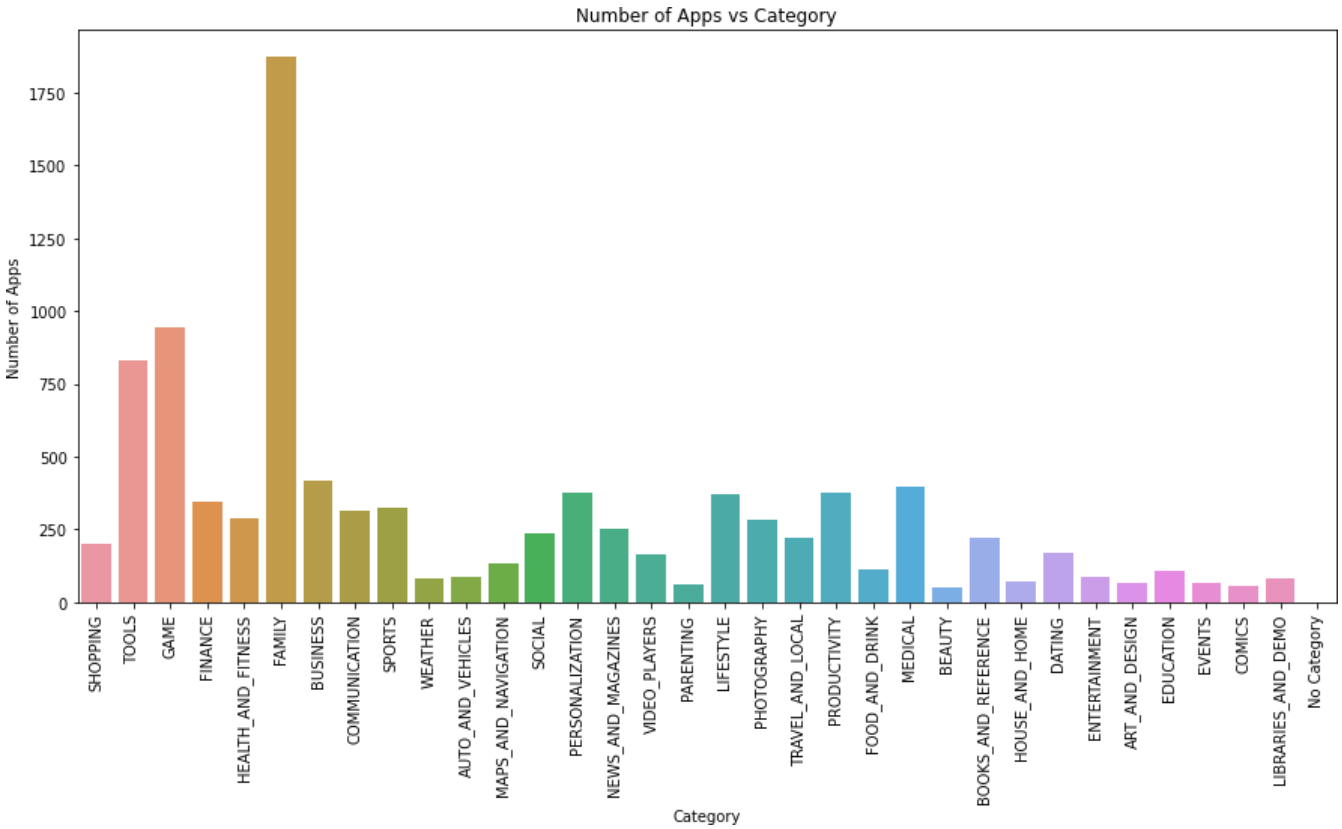
- Filling the null values.
- Removing the duplicate entries.

Data Visualizing

- Plotting various graphs.
- Extracting insights from the dataset.



Number of Apps based on Category



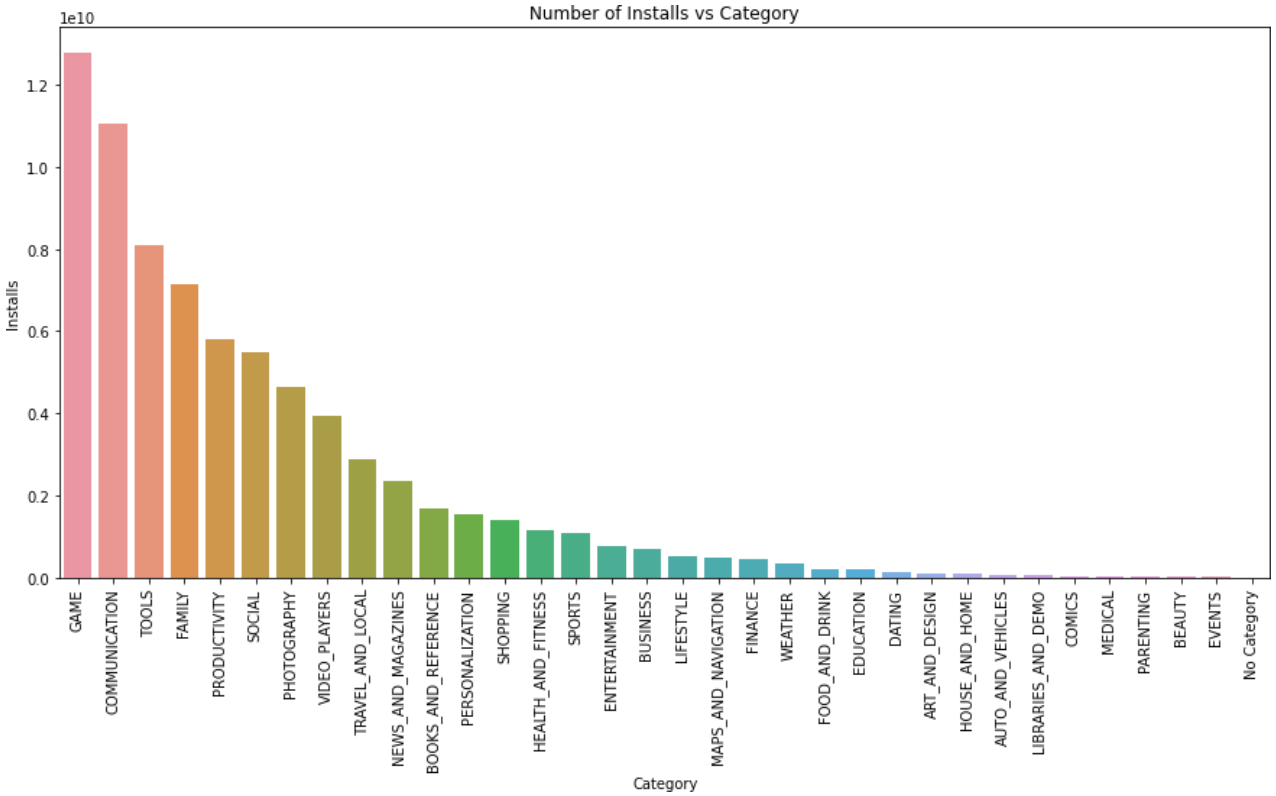
Top categories which have maximum number of apps-

- **Family**
- **Game**
- **Tools**
- **Business**
- **Medical**

Least number of apps category –

- **Beauty**
- **Comics**

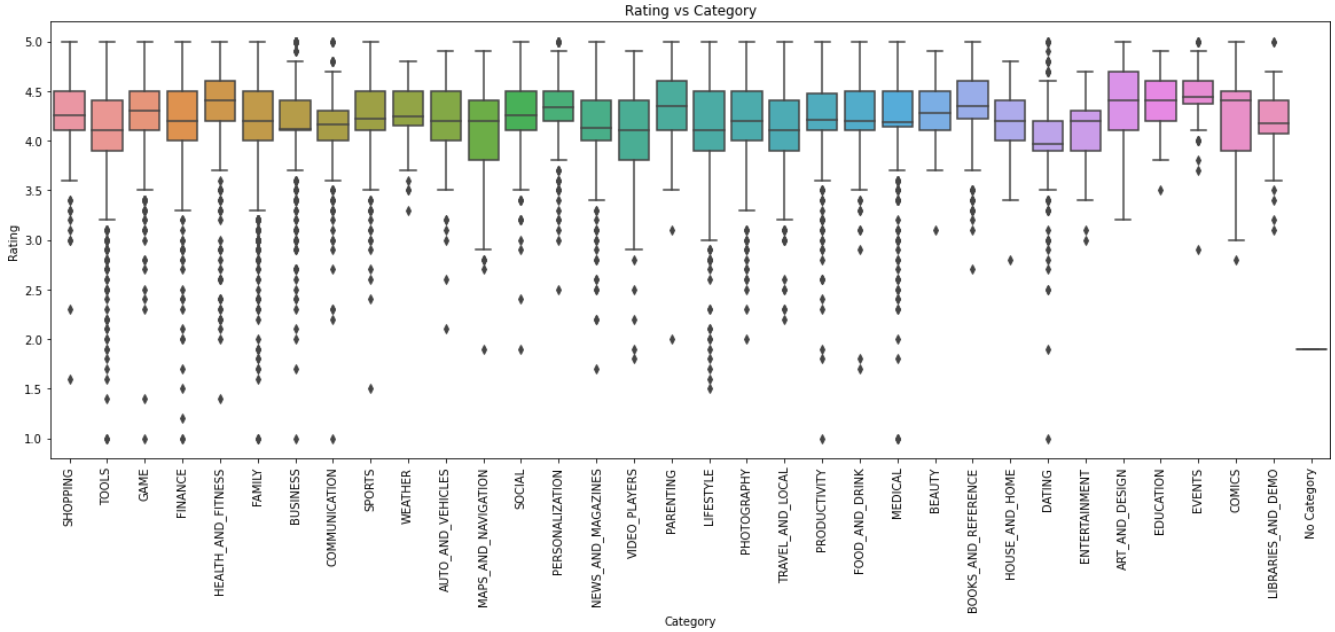
Number of App Installed based on Category



Top categories which have maximum number of Installs-

- **Game**
- **Communication**
- **Tools**
- **Family**
- **Productivity**

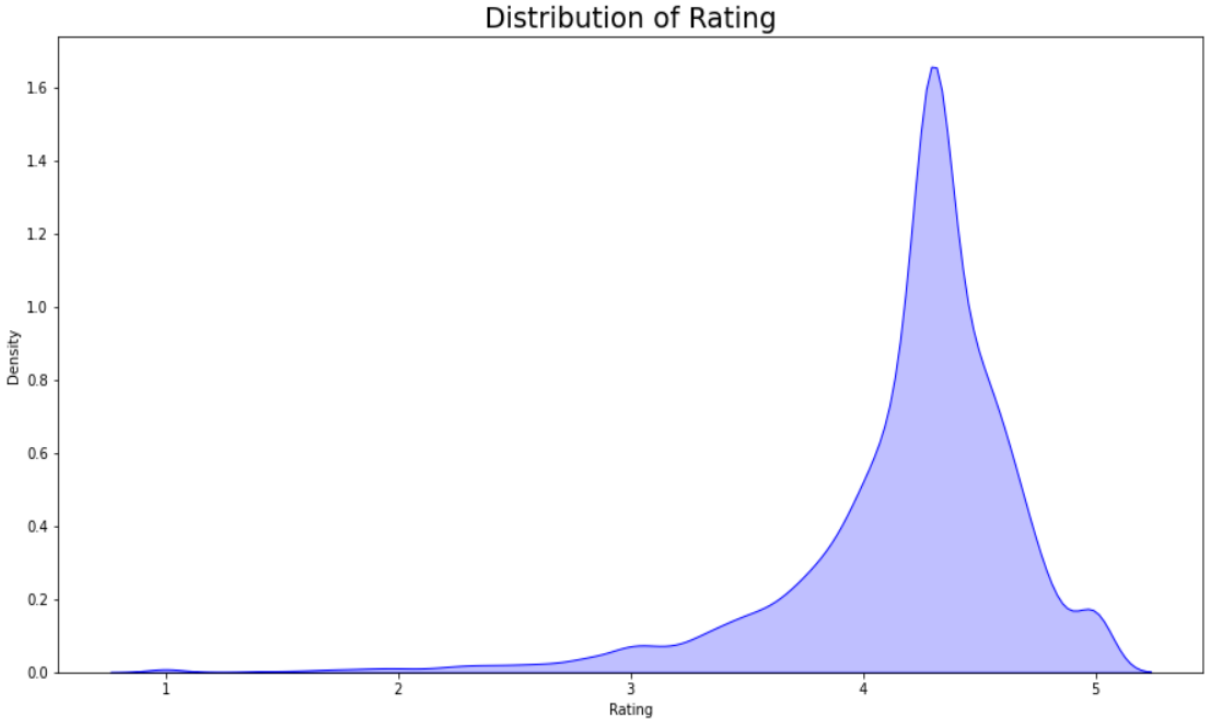
Rating distribution for each category



Highest rated categories based on the median of ratings of these categories-

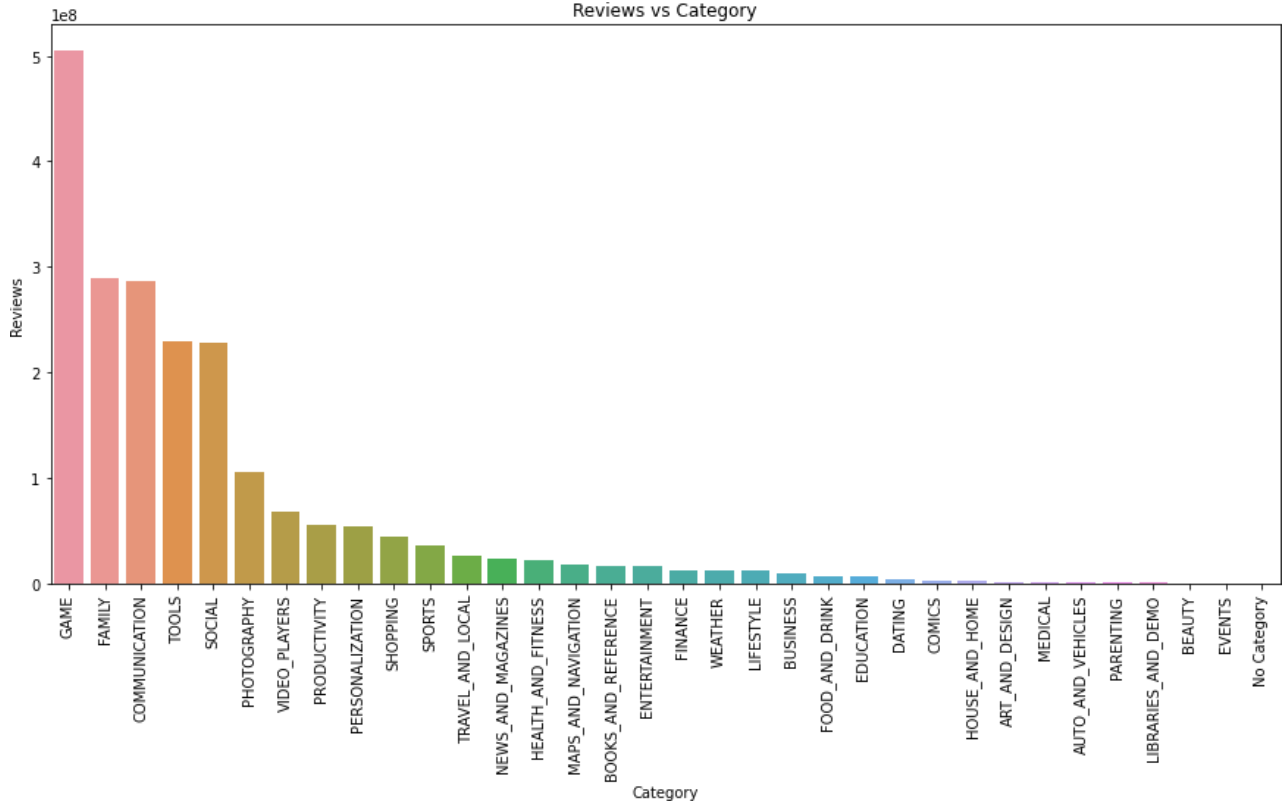
- **Events**
- **Art and Design**
- **Comics**
- **Education**
- **Health and Fitness**

Distribution of Rating across the whole dataset



Most of the apps are rated in between **3.5 to 4.7**

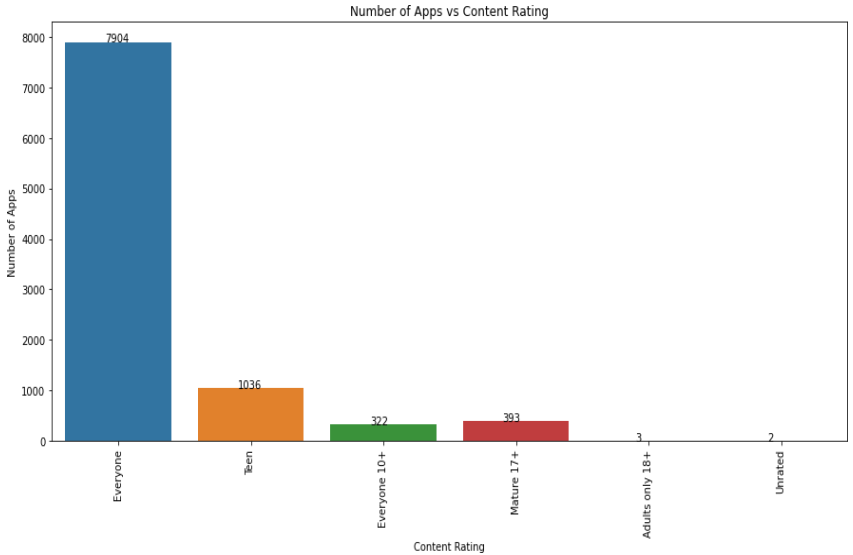
Number of Reviews against bases on Category



Top categories which are most reviewed-

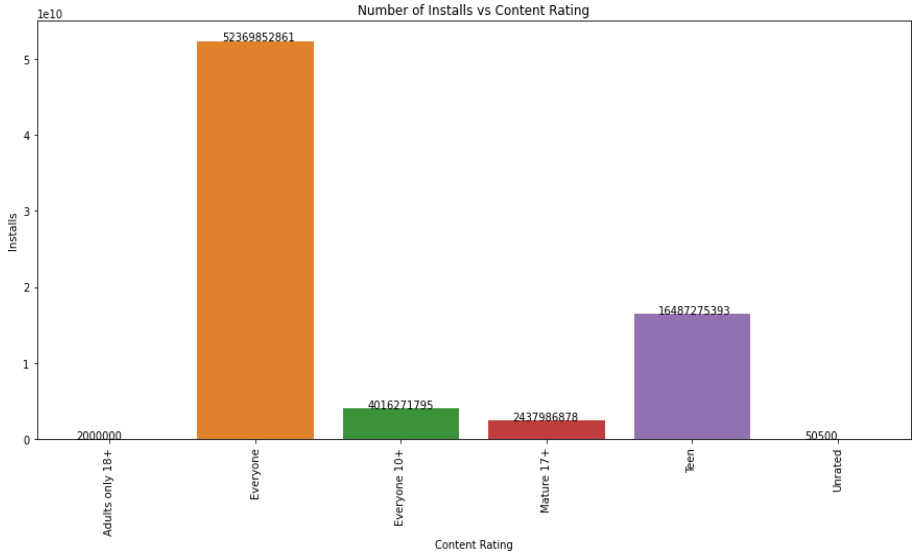
- **Game**
- **Family**
- **Communication**
- **Tools**
- **Social**

Number of Apps vs Content Rating



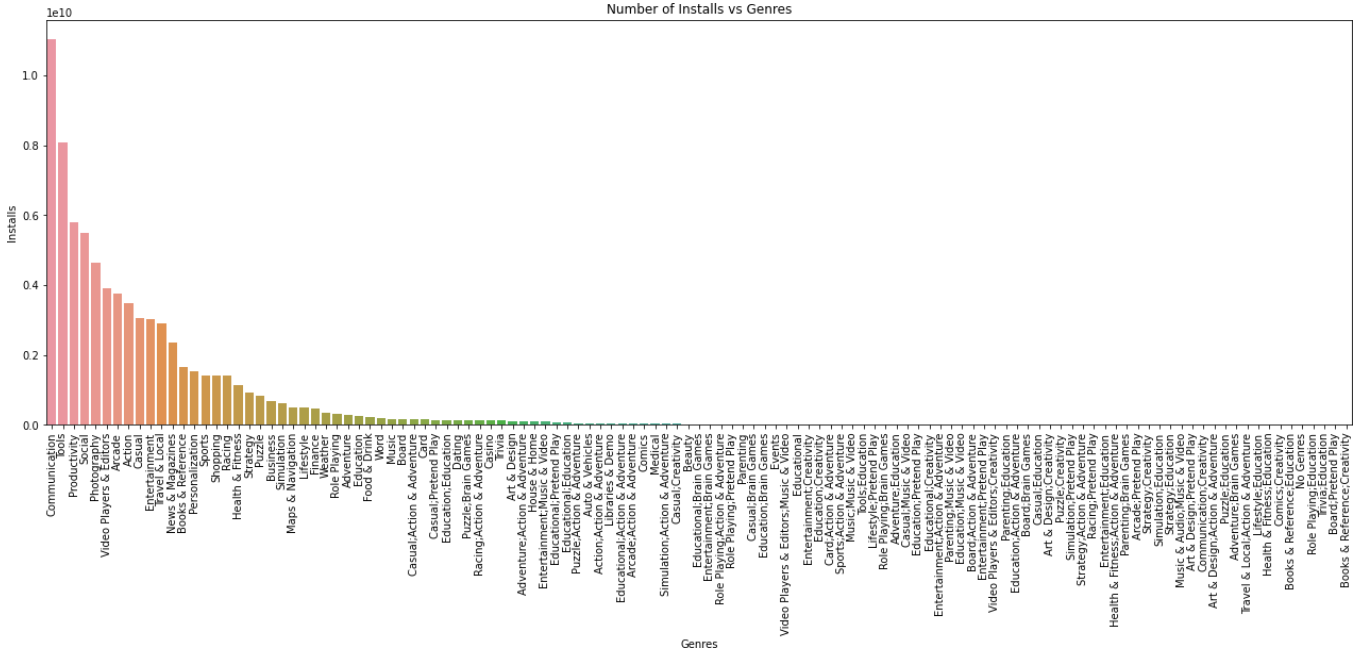
- **Most of the apps** in the Play Store are with content rating **everyone**. So, anyone can install these apps.

Number of Installs based on Content Rating



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

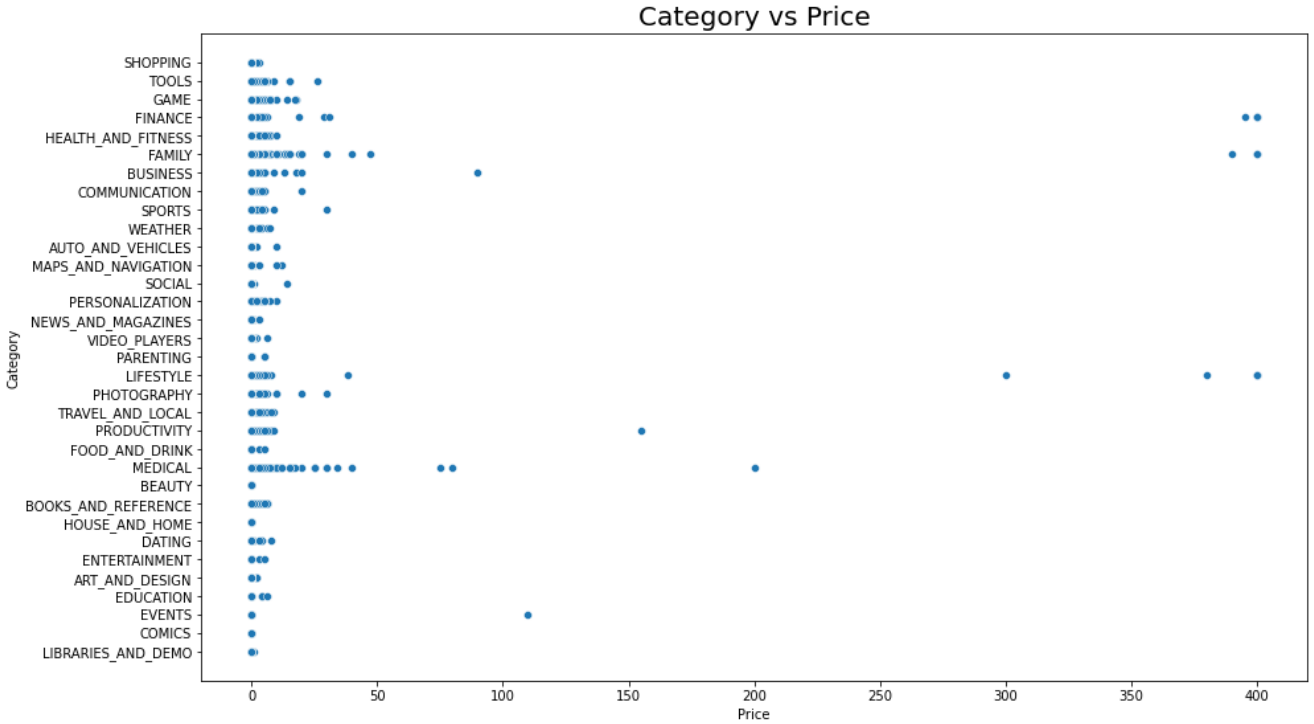
Installation vs Genres



Top Genres from where the apps have been installed the most-

- **Communication**
- **Tools**
- **Productivity**
- **Social**
- **Photography**

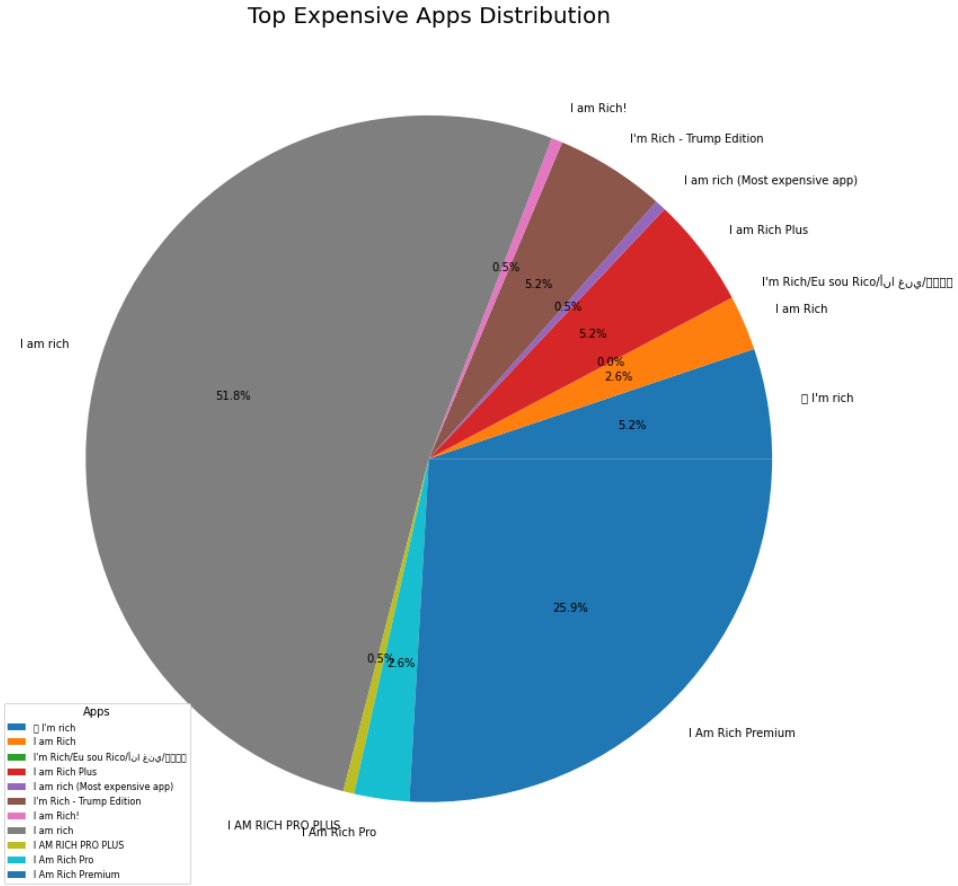
Category vs Price



The highest paid applications are -

- **FINANCE**
- **LIFESTYLE**
- **FAMILY.**

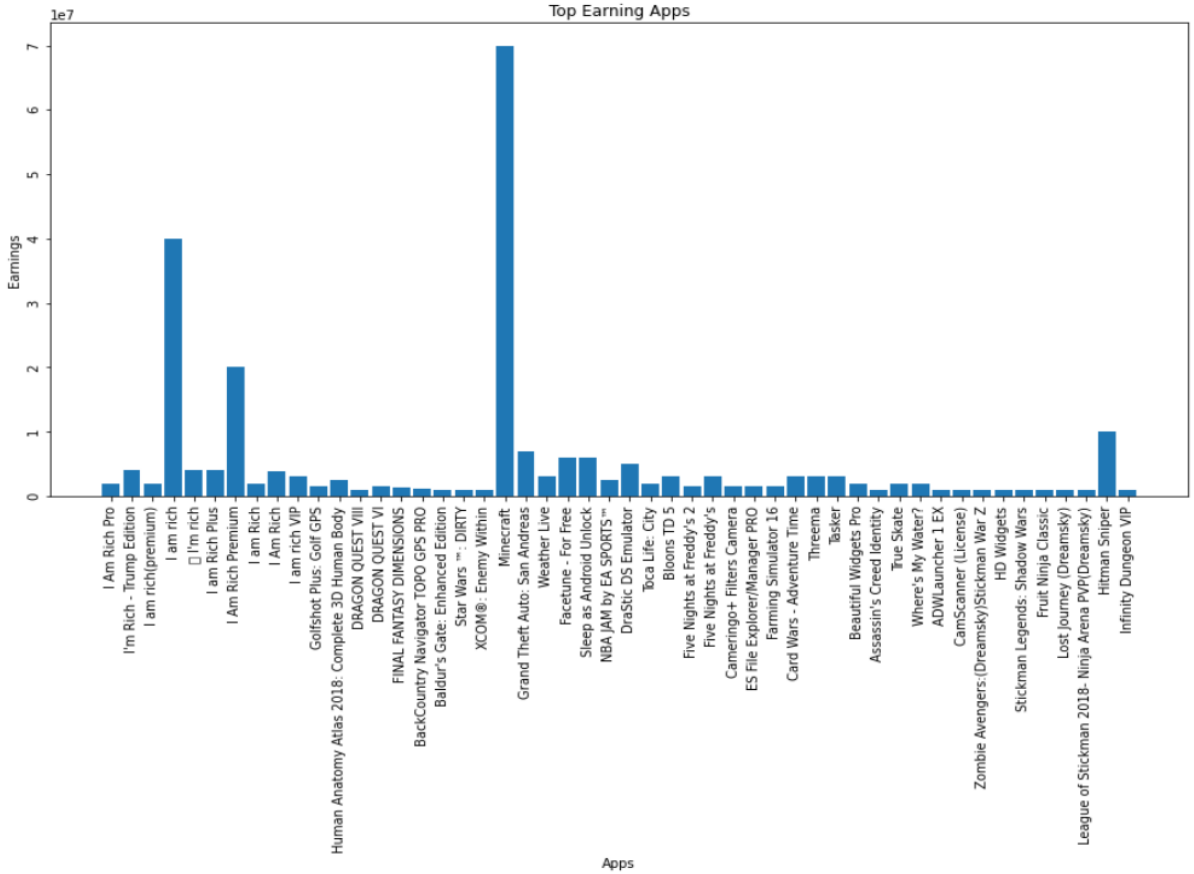
Expensive app distribution



Highest distribution of apps base on price are -

- **I am rich**
- **I Am Rich Premium**
- **I'm rich-Trump Edition**

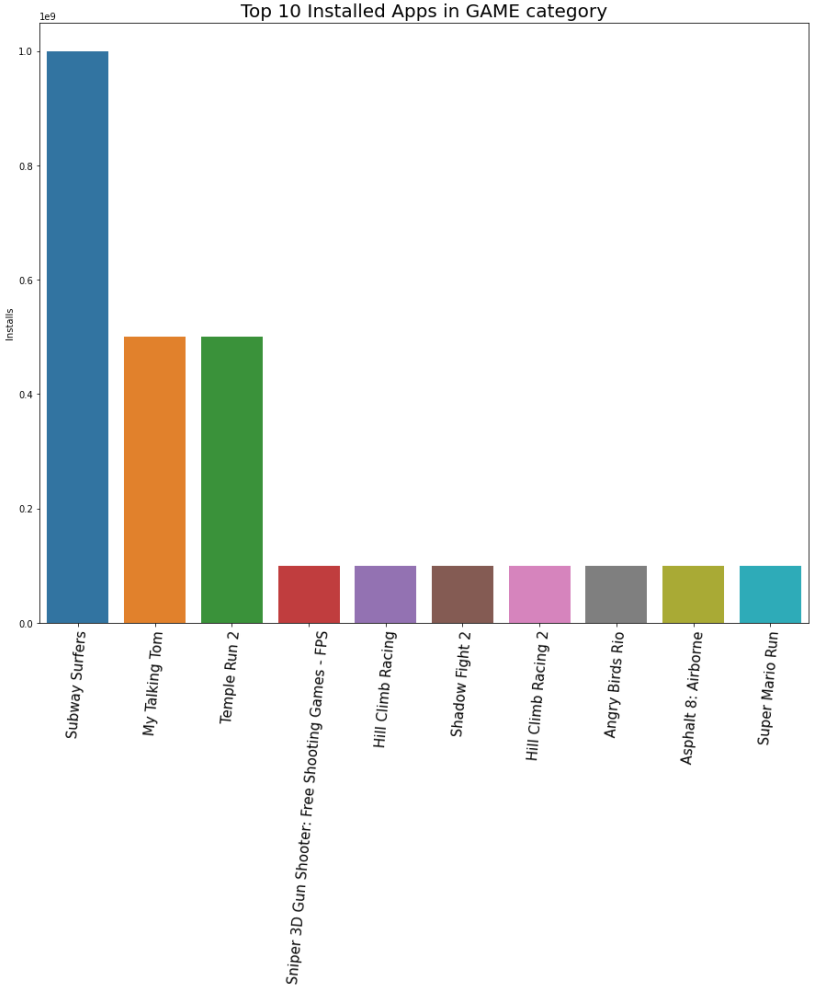
Apps that have made highest earning.

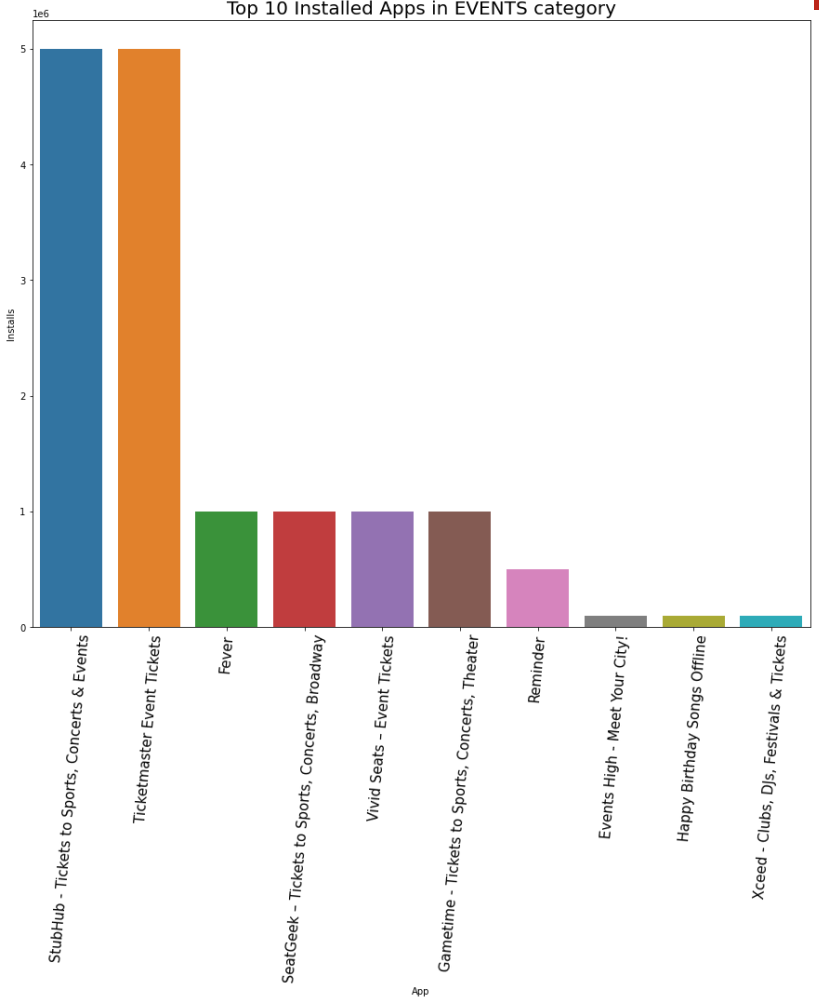


The top five apps with the highest earnings found on google play store are:-

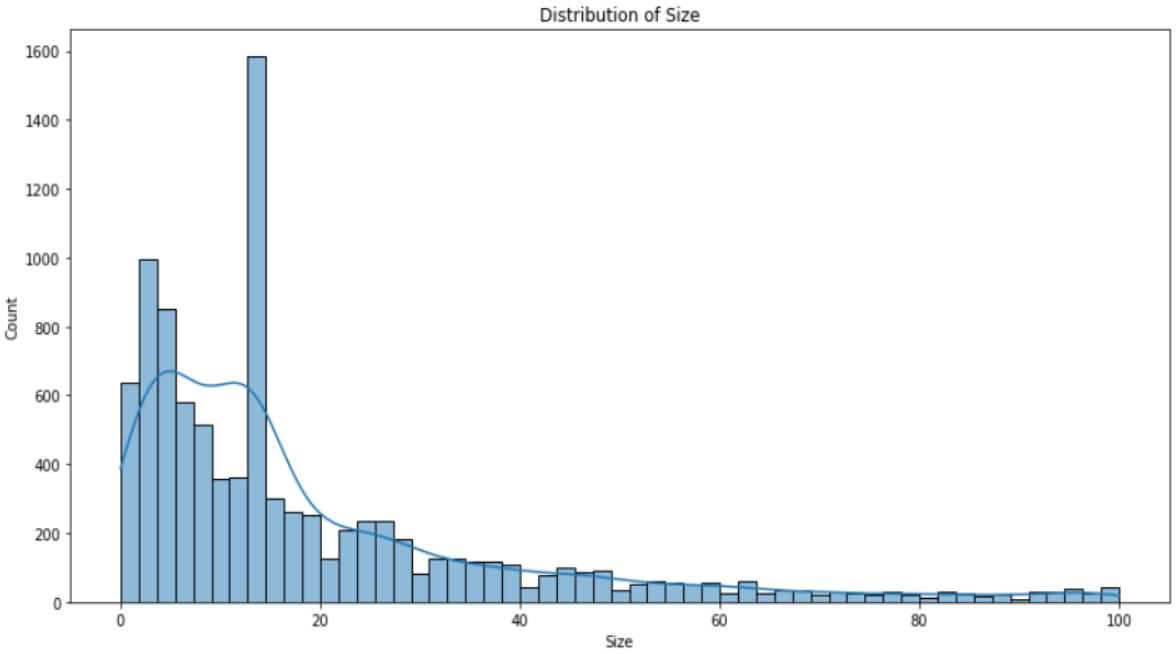
- **Minecraft**
- **I am Rich**
- **I am Rich Premium**
- **Hitman Sniper**
- **Grand Theft Auto: San Andreas**

Top 10 App in specific Category



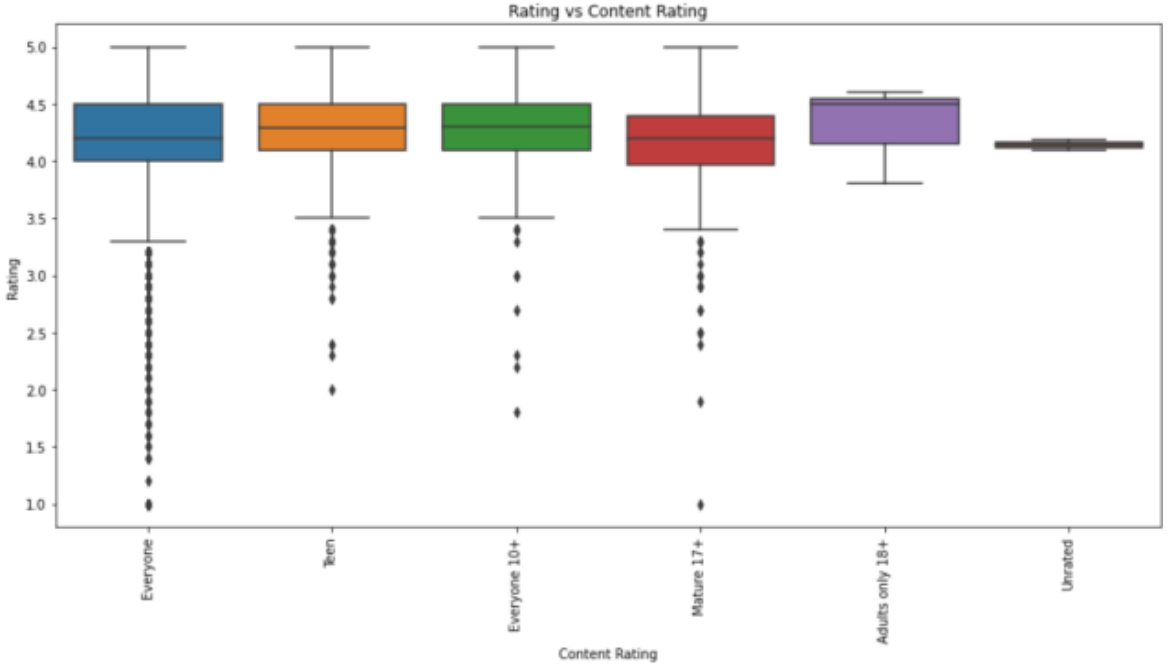


Distribution of Size



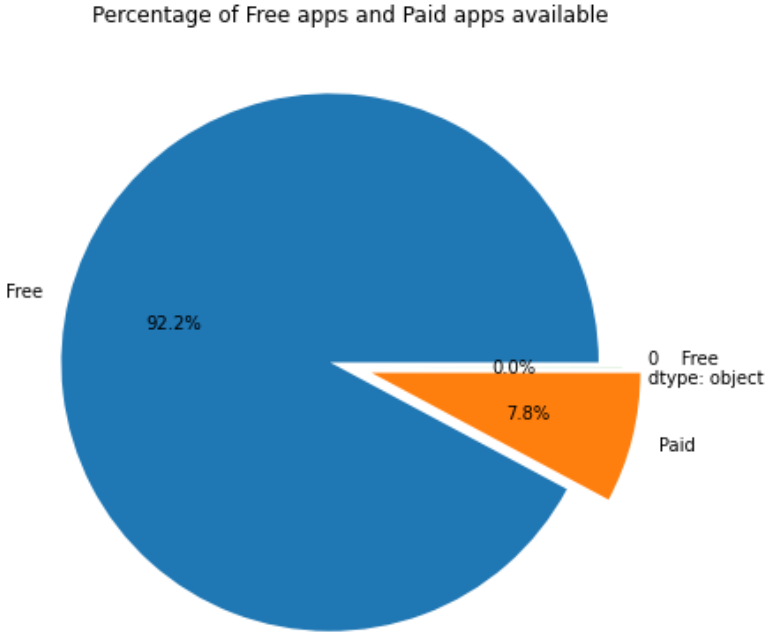
- Most of the apps are present the play store are **smaller sizes** in between **5Mb-20Mb.**
- This encourages the Developer to reduce the size of the app as small as possible.

Rating Based on Content



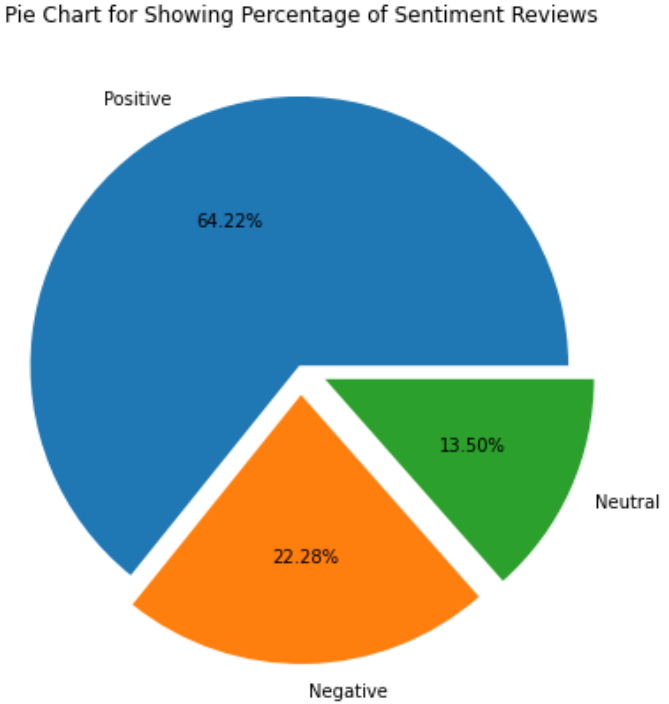
- **Adults Only 18+** contents are having the **highest ratings of 4.5**
- Followed by **Everyone 10+** with rating **4.3**
- **Teen** with rating **4.2**
- **Everyone** with rating **4.2**
- **Mature 17+** with rating **4.2**
- Unrated with rating 4.1

Free Apps vs Paid Apps



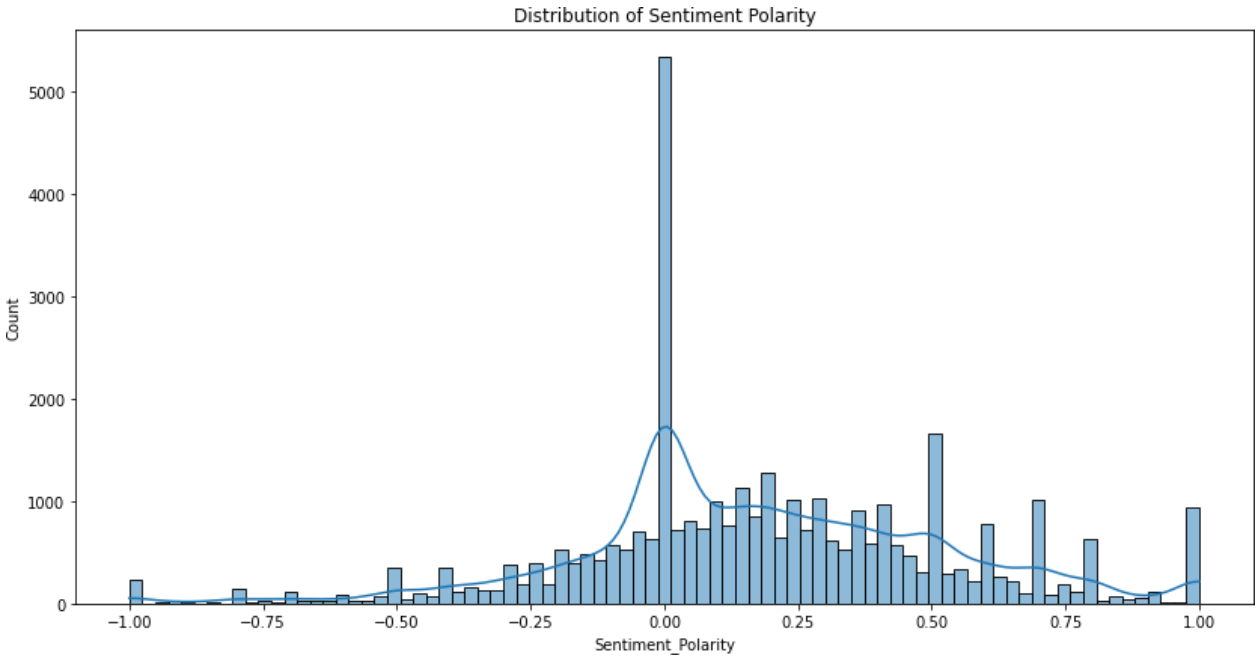
Mostly of the apps are **Free of cost (92.2%)** are present in the play store as compared to the **Paid Apps (7.8%)**

Percentage criteria for Sentiment Reviews



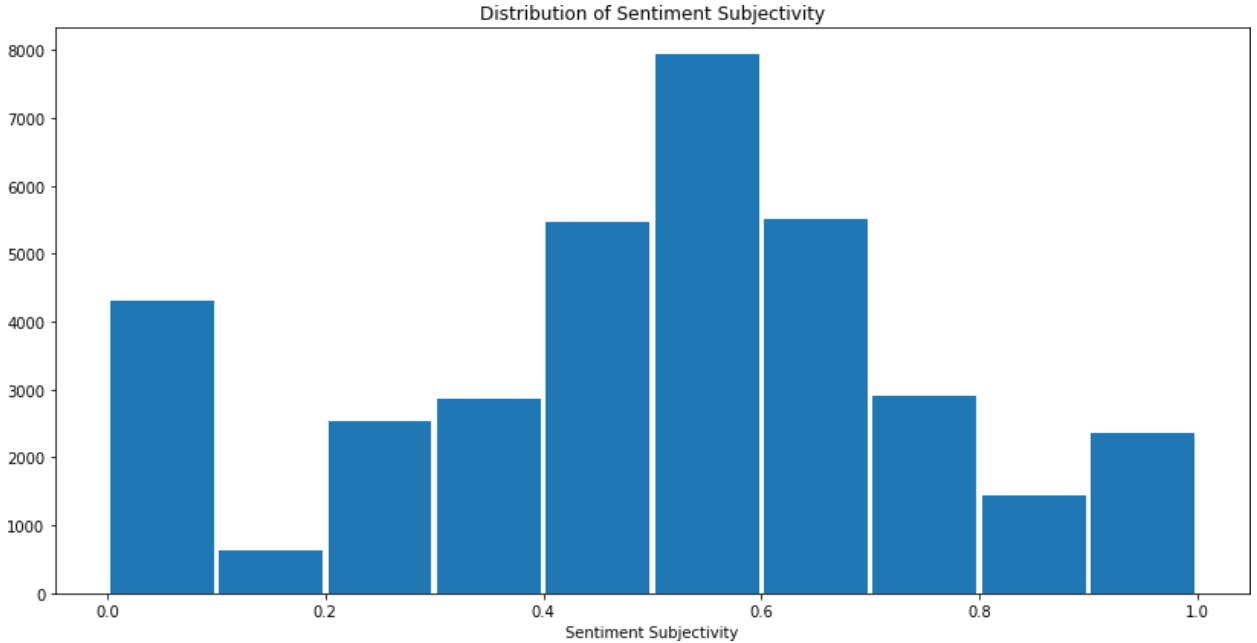
Most of the app's present with Positive review, then Negative reviews and Neutral review each, respectively.

Distribution of Sentiment Polarity



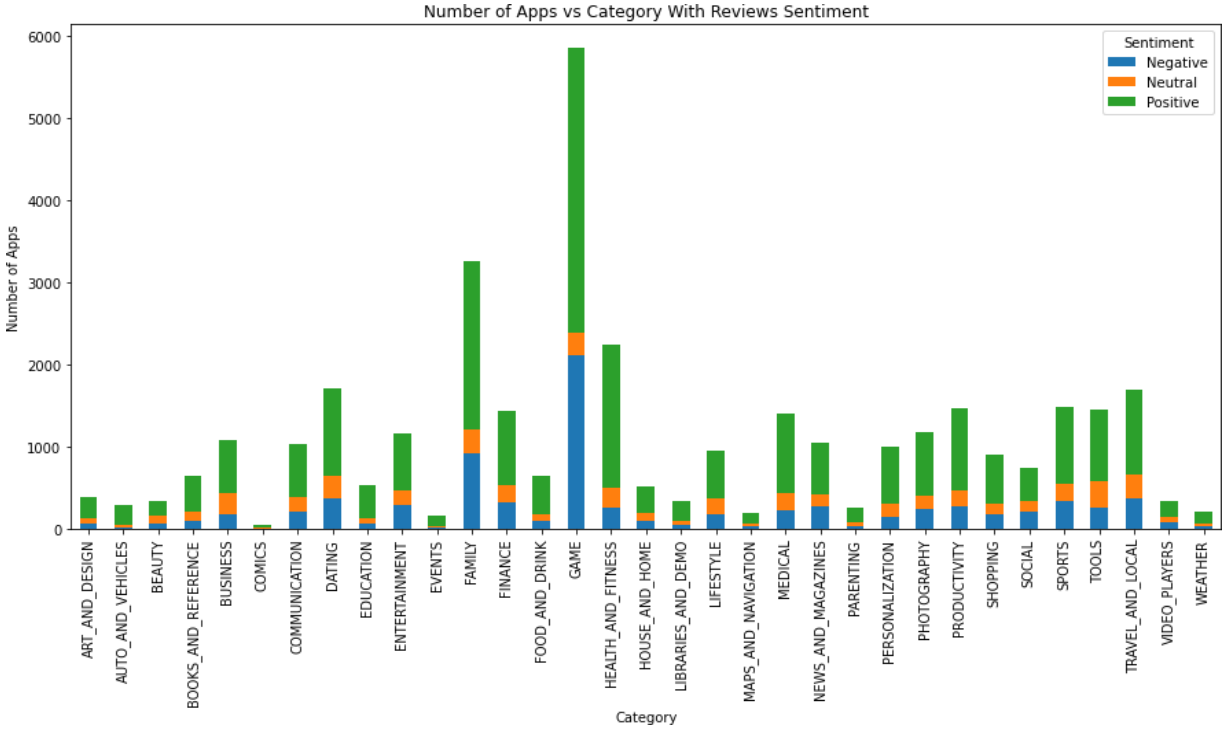
- The width of the distribution is **more towards the left** of the graph which makes it **left skewed.**
- So, the **Polarity of most of the users is towards the positive side** as we already saw in the pie chart.

Histogram Plot for Sentiment Subjectivity



- Most the sentiment subjectivity **lies in between 0.4 to 0.7**
- Which shows **that most of the reviews are towards subjective** point of view of the users.

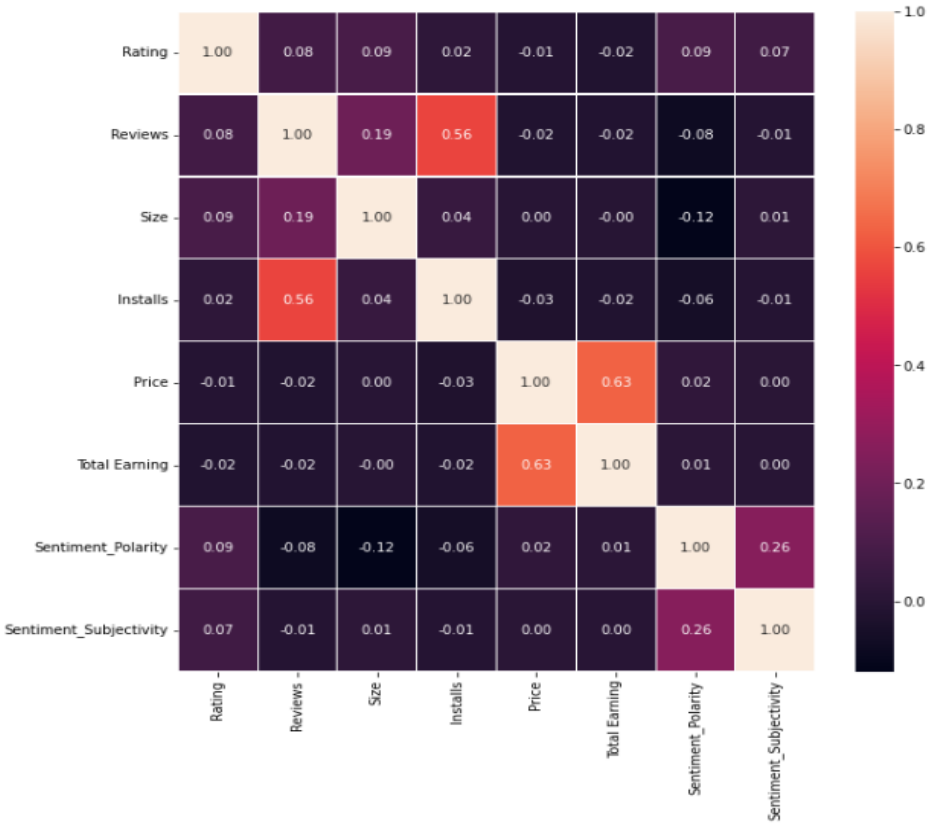
Number of Apps in each Category with Sentiment Reviews



The Top 5 Categories with most positive reviews are –

- **Game**
- **Family**
- **Health Fitness**
- **Dating**
- **Travel and Local**

Correlation Heat Map



- **Strong positive correlation** between **Price** and **Total Earning column.**
- **Strong positive correlation** between the **Reviews** and **Installs.**
- **Size** and **Reviews** are **slightly correlated.**
- Sentiment **Polarity** and Sentiment **Subjectivity** are **slightly correlated.**

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.
- Distribution of Size shows most of the app’s present in the play store are of smaller size.
- 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.
- Helix Jump has the highest number of positive reviews and Angry Birds Classic has the highest number of negative reviews.

Inferences and Conclusion

- 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.

- Game category has a greater number of positive reviews as well as negative reviews since there is more installs from the Game category.

- Content Rating Teen is having highest number of installs.
  - It shows that the present youths are quite good at operating apps and thus developers can develop more apps which suits to the interest of the teens.

- Number of free apps present in the play store are higher than paid apps.
  - Its quite evident users prefer to install free apps more as compared to the paid apps.
  - This gives direction that the developers can launch more of the free apps and for earning money.
  - They can use other means such as through advertisements in the apps or monetizing certain section of the app which serves certain special purpose or any other means.

- From the correlation matrix we can infer that reviews and installs are having a strong correlation.
  - It's quite evident as the greater number of installs more will be the number of reviews.

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.


THANK YOU