Capstone Project

Play Store App Review Exploratory Data Analysis (EDA)

By,
Prashant Shaw
Data Science Trainee at AlmaBetter

Agenda

- Introduction
- Problem Statement
- Attributes
- Steps Involved in this EDA
- Visualisation
  - Category wise number of Apps
  - Number of App Installed in each Category
  - Box Plot showing Rating distribution for each category
  - Violin plot showing the Distribution of Rating across the whole dataset
  - Number of Reviews against each Category
  - Installation vs Genres
  - Number of Apps vs Content Rating
  - 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
  - Number of Apps Supported in Each Android Ver
  - Pie Chart for Percentage of Sentiment Reviews
  - Distribution of Sentiment Polarity
  - Histogram Plot for Sentiment Subjectivity
  - Number of Apps in each Category with Reviews Sentiment
  - Correlation Heat Map
- Conclusion

Introduction

- Google Play Store formerly known as Android Market, is a digital distribution service operated and developed by Google.
- It is an online store from where android users can find various apps from a wide range of Categories and Genres.
- Users can install the apps from the Google Play Store and also they can give reviews and ratings to the apps based on how good the app is.
- Play Store has lots of data and thus has an enormous potential to drive app-making businesses to success.
- Exploratory Data Analysis (EDA) is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods.
- The main goal or objective of this project is to perform EDA on Play Store Data to discover the key factors responsible for the apps success and to give actionable insights to the developer to work on and capture the android market.

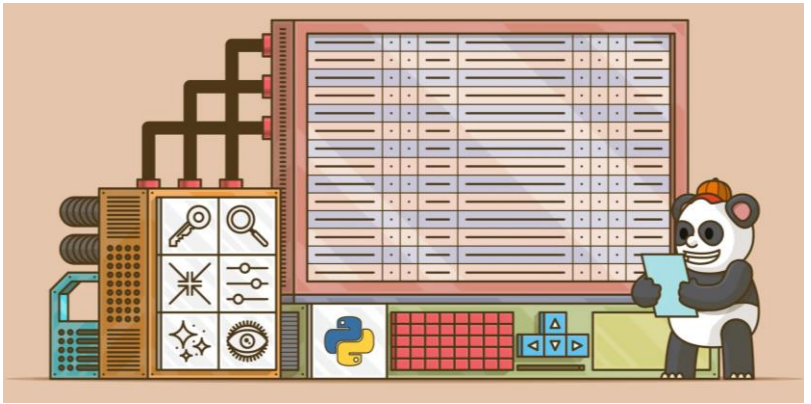
Problem Statement

- Android users have more than a million apps available through the Google Play Store. These apps have come to play a huge role in the way we live our lives today.
- 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.
- Due to the presence of such wide variety of apps and the data associated with it, it has become important to analyze these data and extract meaningful insights which will help developer to capture the android market.
- Various key factors needs to be extracted from the datasets which are responsible for the apps success.

Attributes

Play Store Data –

- Apps – App name.
- Category – The category to which app belongs.
- Rating – Rating of the app.
- Reviews – Number of reviews given to each app.
- Size – Size of the app.
- Installs – Number of installs of each app.
- Type – Free or Paid
- Price – Price of the app in $.
- Content Rating – Age restriction for each app.
- Genres – Genres the app belongs to.
- Last Updated – When the app is last updated.
- Current Ver – Current version of the app.
- Android Ver – Android version on which the app is supported.



User Reviews –

- Apps – App name.
- 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

Exploring Dataset

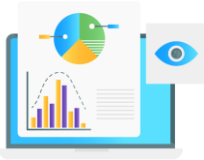
- Checking the summary.
- Converting Reviews, Size, Installs & Price attributes from object type to numeric type.

Cleaning the Dataset

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

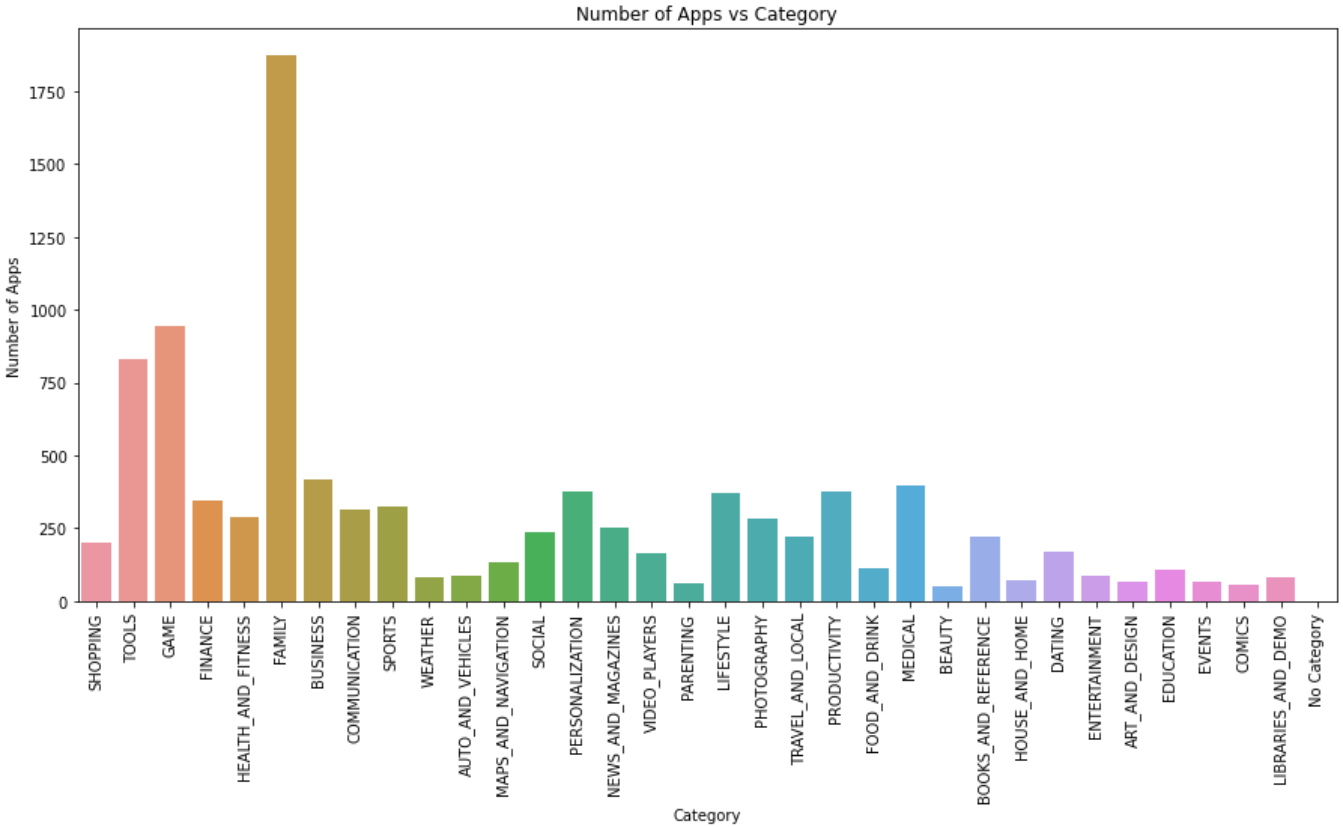
Visualising Data

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



Visualisation

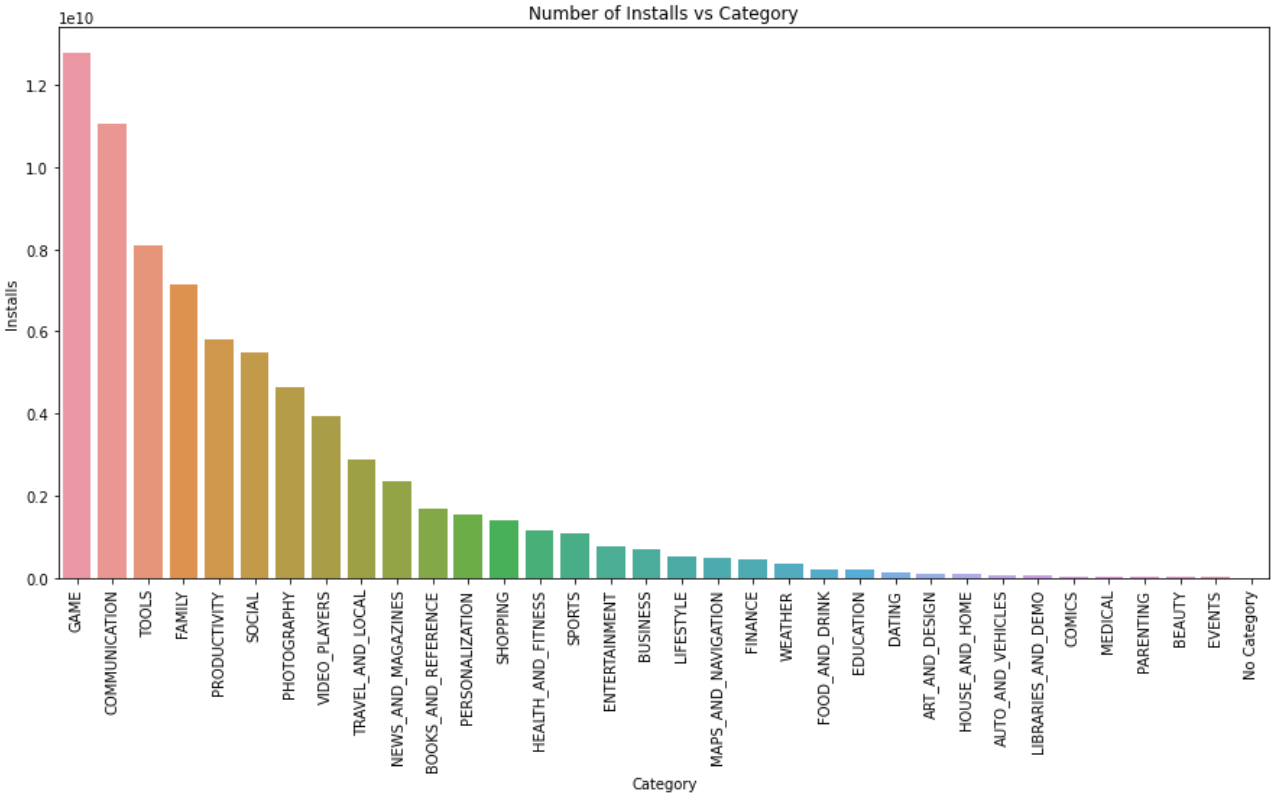
Category wise Number of Apps



Family, Game, Tools, Business & Medical are the top 5 categories which have maximum number of apps.

Visualisation

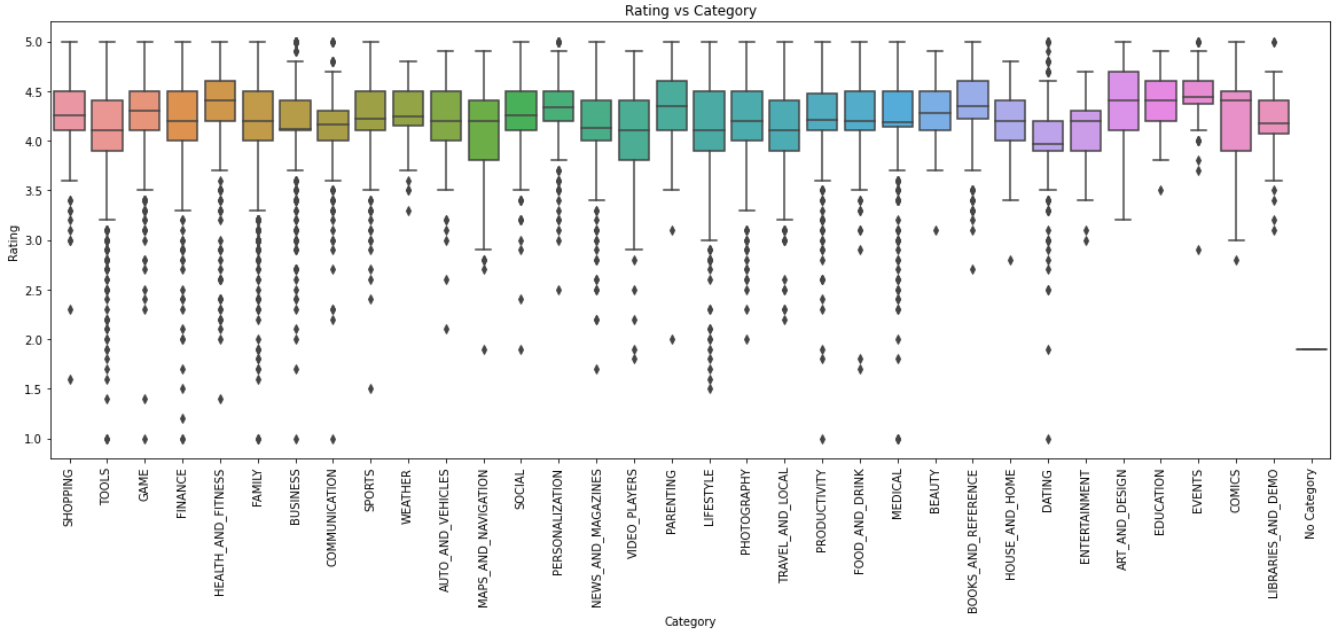
Number of App Installed in each Category



Game, Communication, Tools, Family and Productivity are the top 5 categories which have maximum number of Installs.

Visualisation

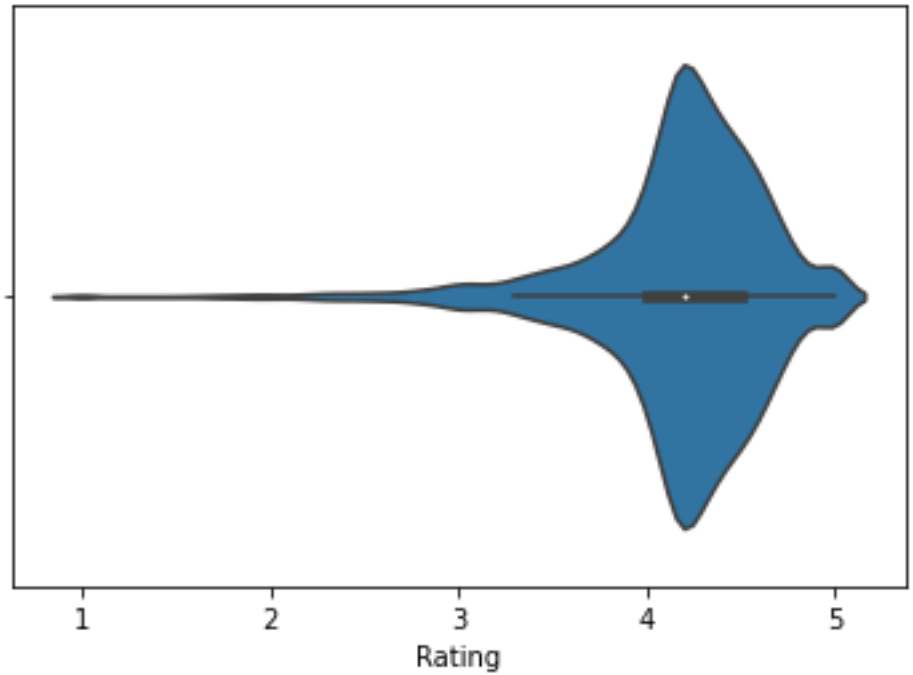
Box Plot showing Rating distribution for each category



Events, Art_And_Design, Comics, Education and Health_And_Fitness are the top 5 highest rated categories based on the median of ratings of these categories.

Visualisation

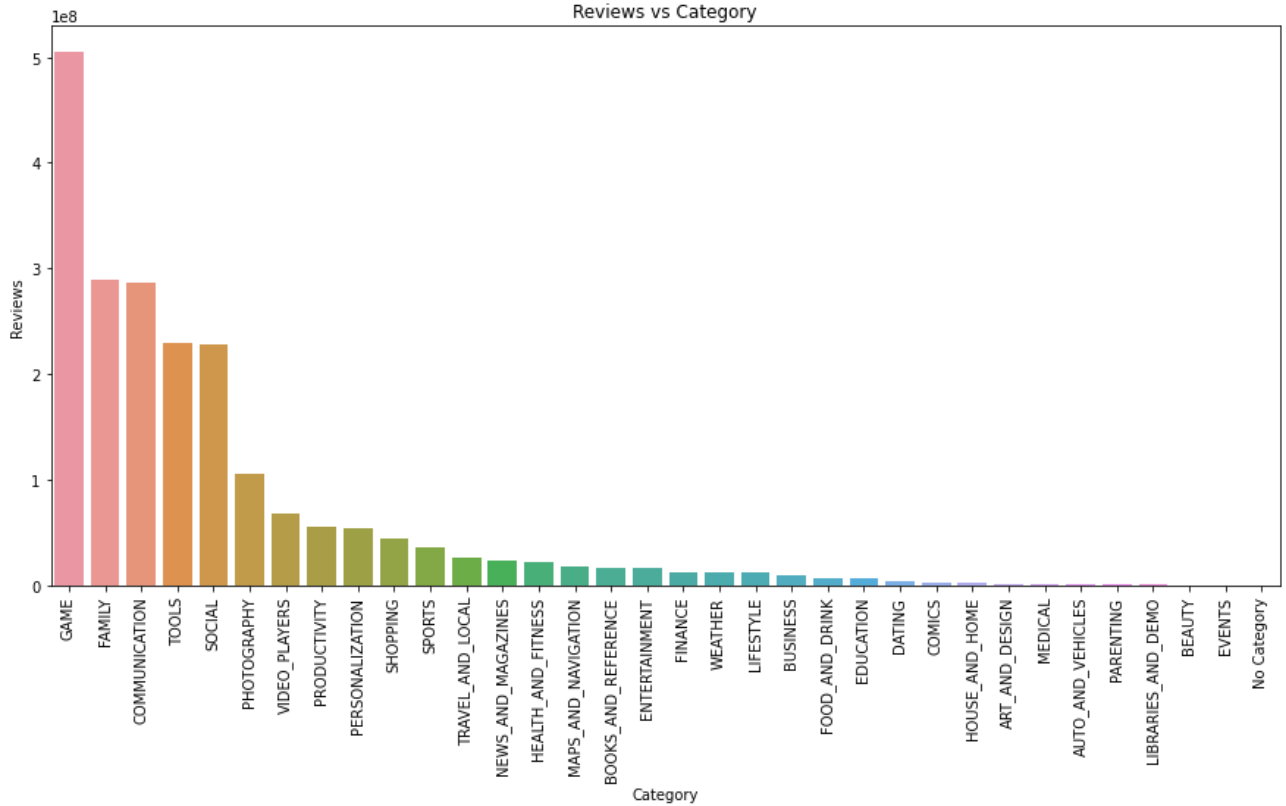
Violin plot showing the Distribution of Rating across the whole dataset



The Violin plot shows that most of the apps are rated between 4 to 5.

Visualisation

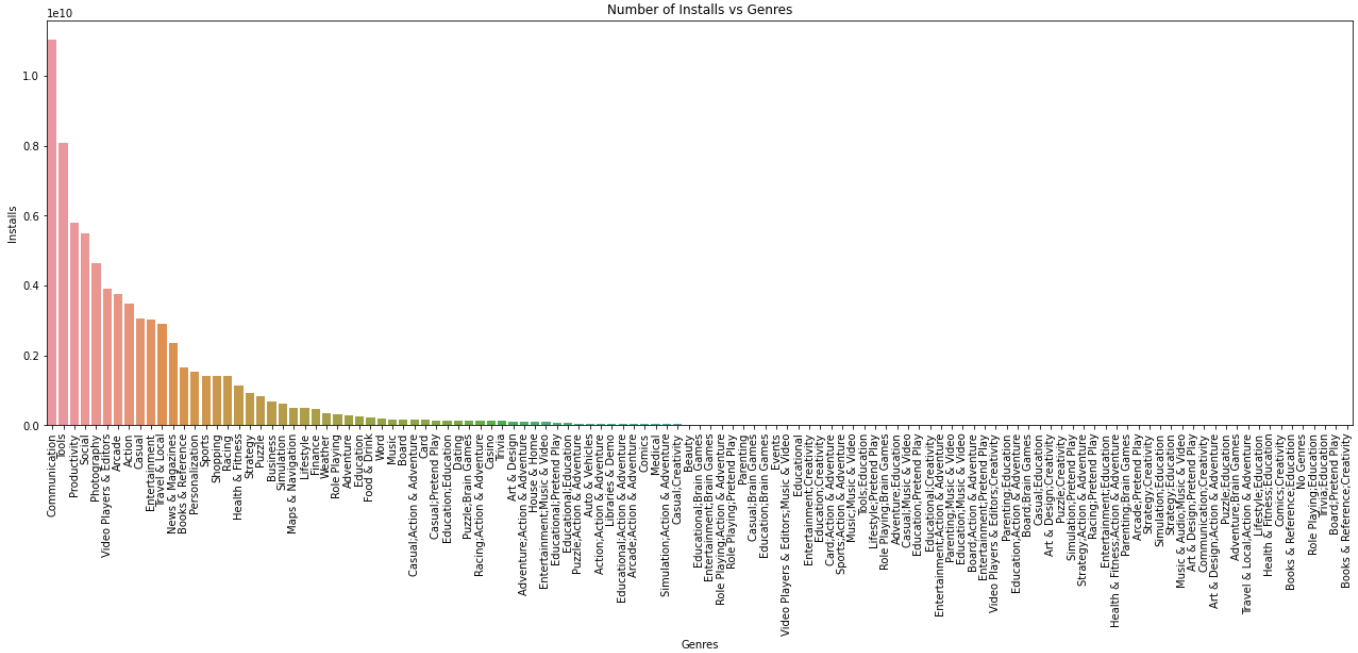
Number of Reviews against each Category



From this Graph we can infer that **Game, Family, Communication, Tools and Social** are the Top 5 categories which are most reviewed.

Visualisation

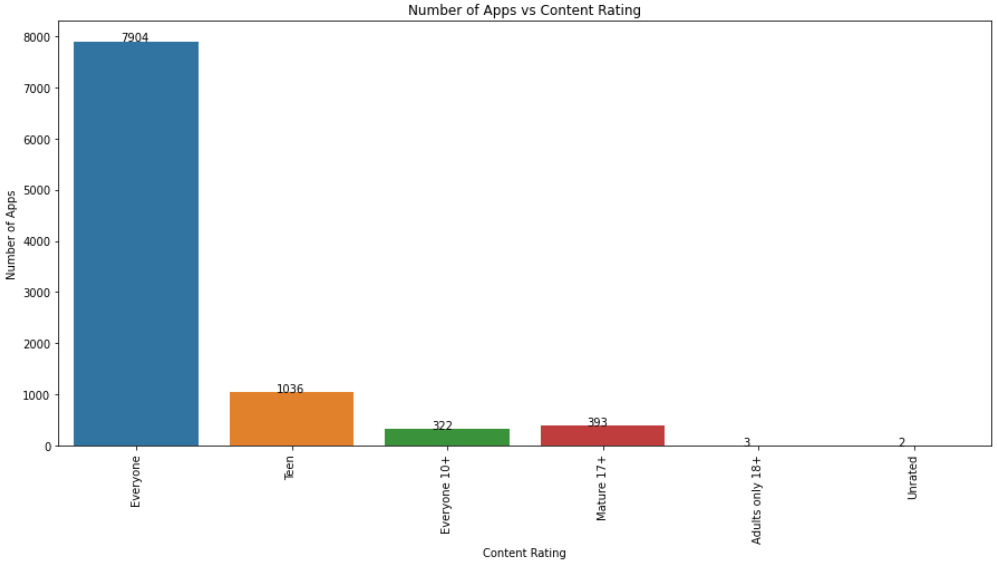
Installation vs Genres

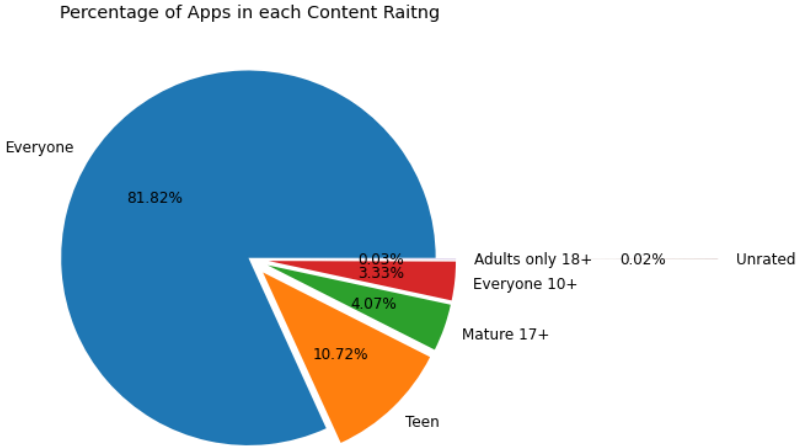


Communication, Tools, Productivity, Social and Photography are the top 5 Genres from where the apps have been installed the most.

Visualisation

Number of Apps vs Content Rating

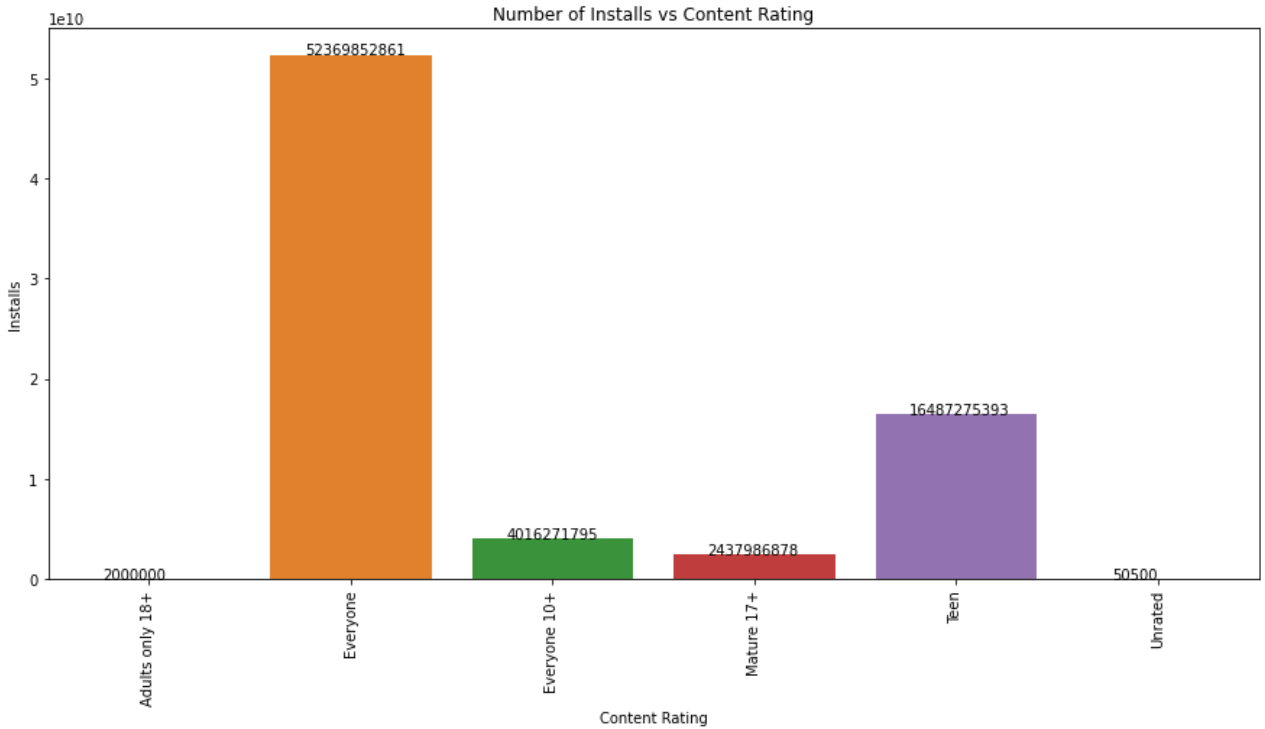




Most of the apps (81.82%) in the Play Store are with content rating everyone. So, anyone can install these apps.

Visualisation

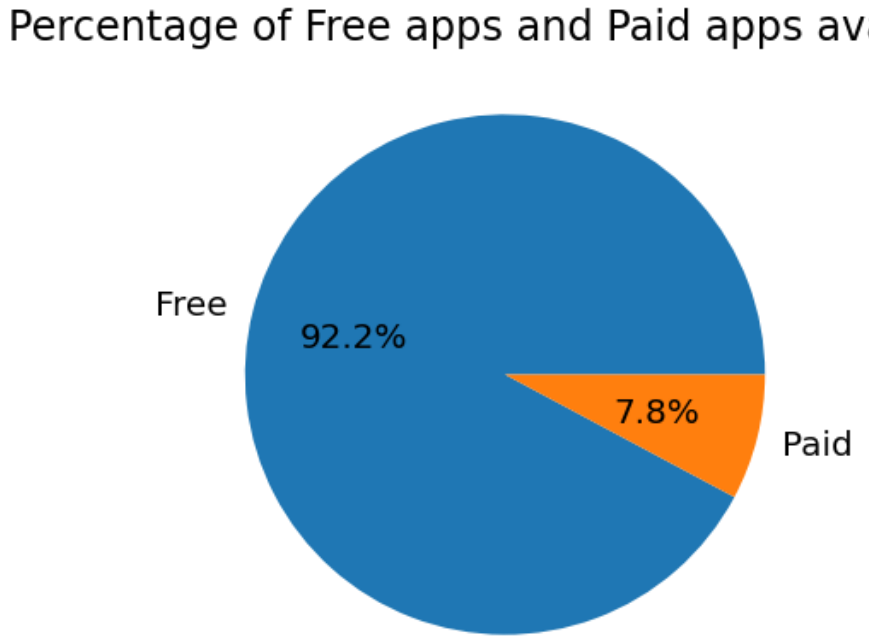
Number of Installs in each Content Rating



As evident since there were more number of Apps with content rating ‘Everyone’ so it has more number of installs.
Also the content rating ‘Teen’ has fairly good amount of installs so, developer can focus more on this section as it is very easy to attract Teens to their apps.

Visualisation

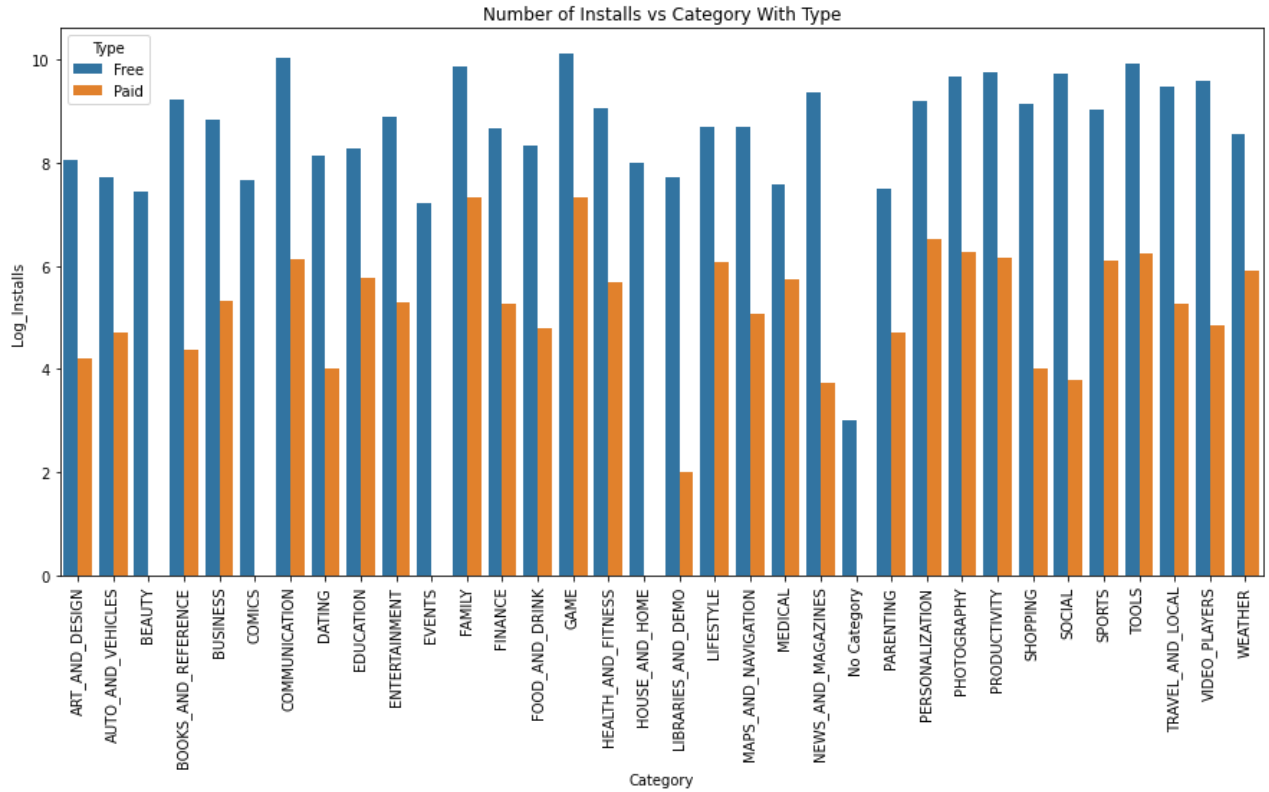
Pie Chart for Percentage of Free Apps and Paid Apps



More Free Apps (92.2%) are present in the play store as compared to the Paid Apps (7.8%).

Visualisation

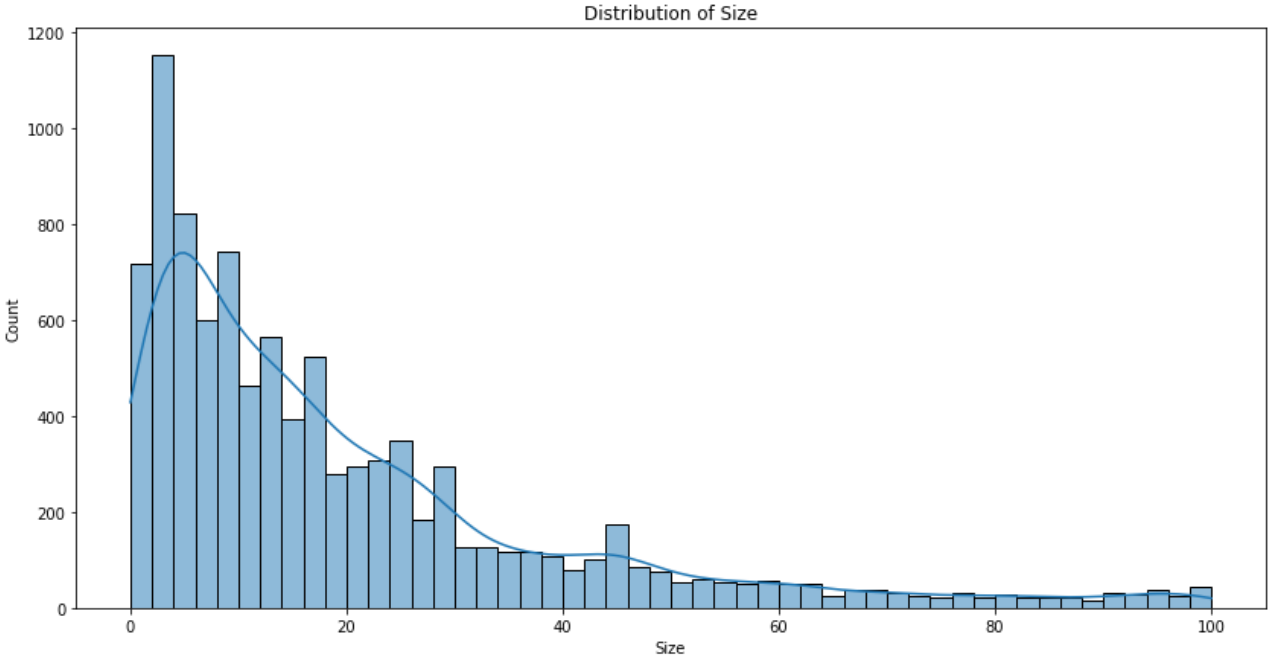
Number of Installs vs Category Based on Type



From this graph it is quite evident that user installs more of Free Apps as compared to the Paid Apps.

Visualisation

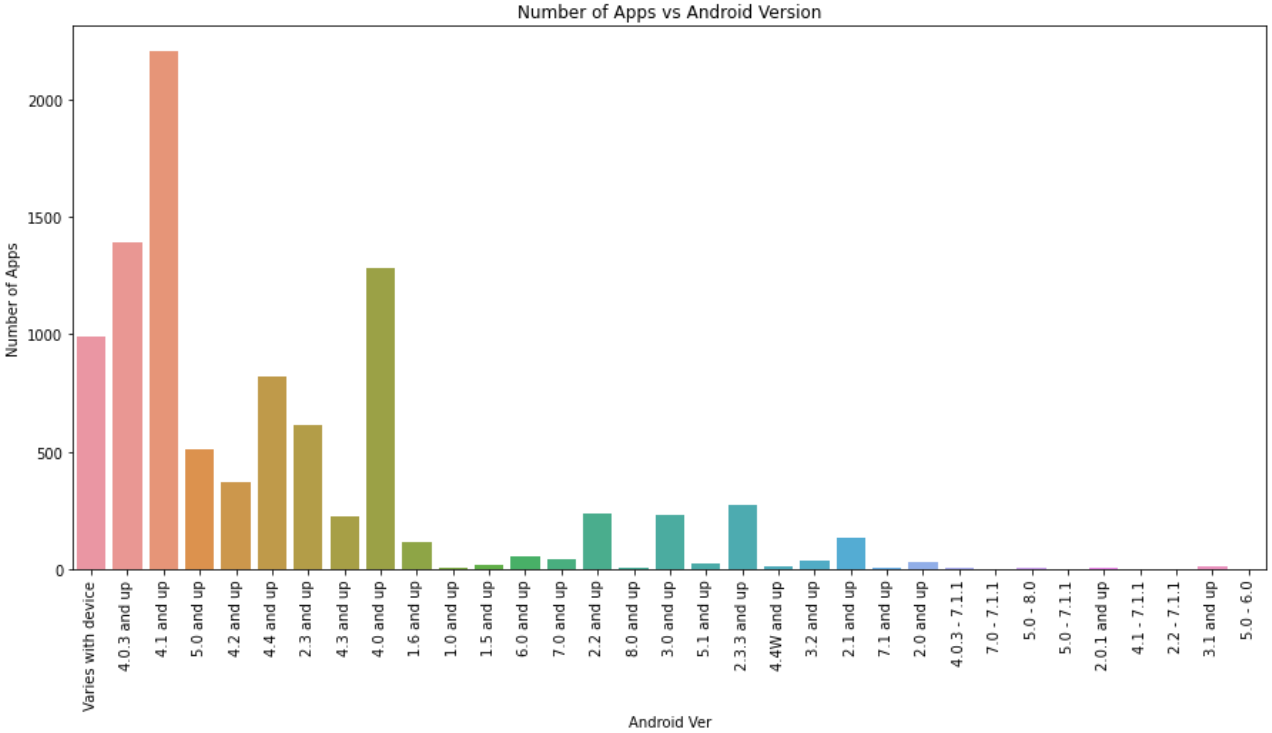
Distribution of Size



This Graph shows that most of the apps in the play store are of smaller sizes. Here the size is in MB.
This encourages the Developer to reduce the size of the app as small as possible.

Visualisation

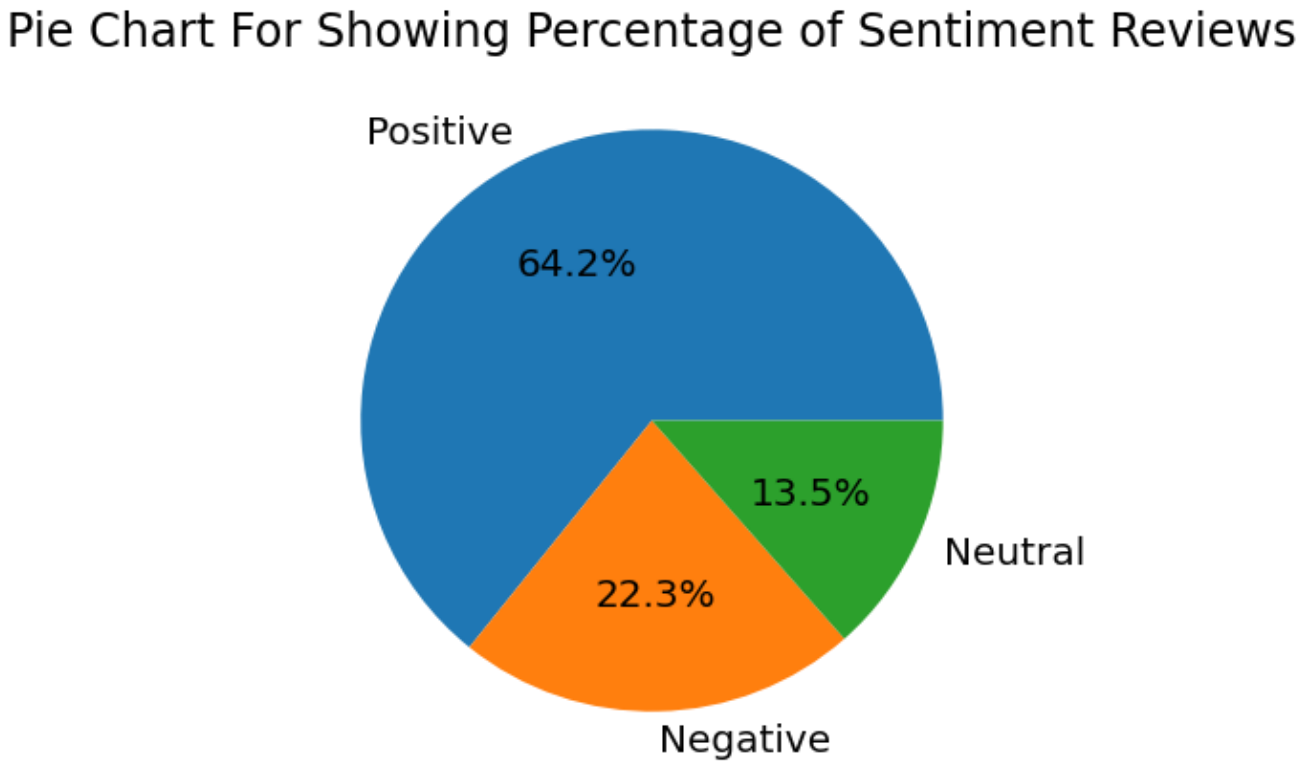
Number of Apps Supported in Each Android Ver



From this Graph we can infer that most of the apps in the play store requires Android Version 4.1 and up to run.

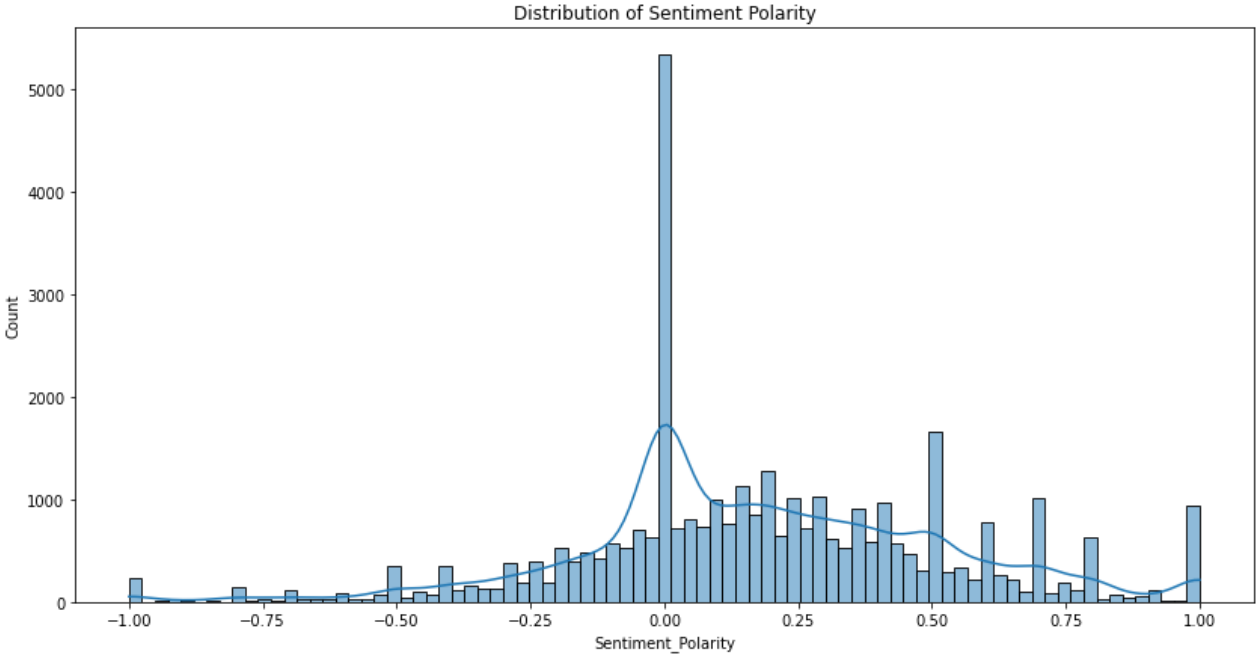
Visualisation

Pie Chart for Percentage of Sentiment Reviews



Visualisation

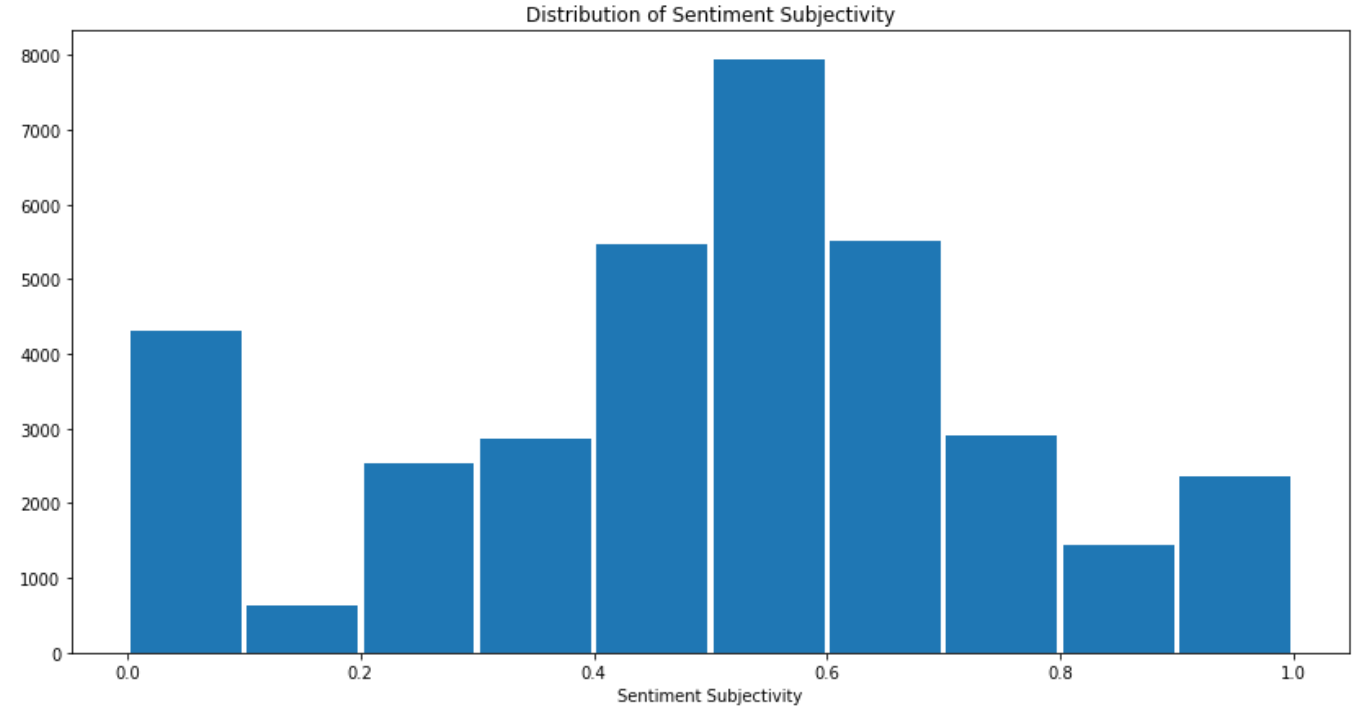
Distribution of Sentiment Polarity



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.

Visualisation

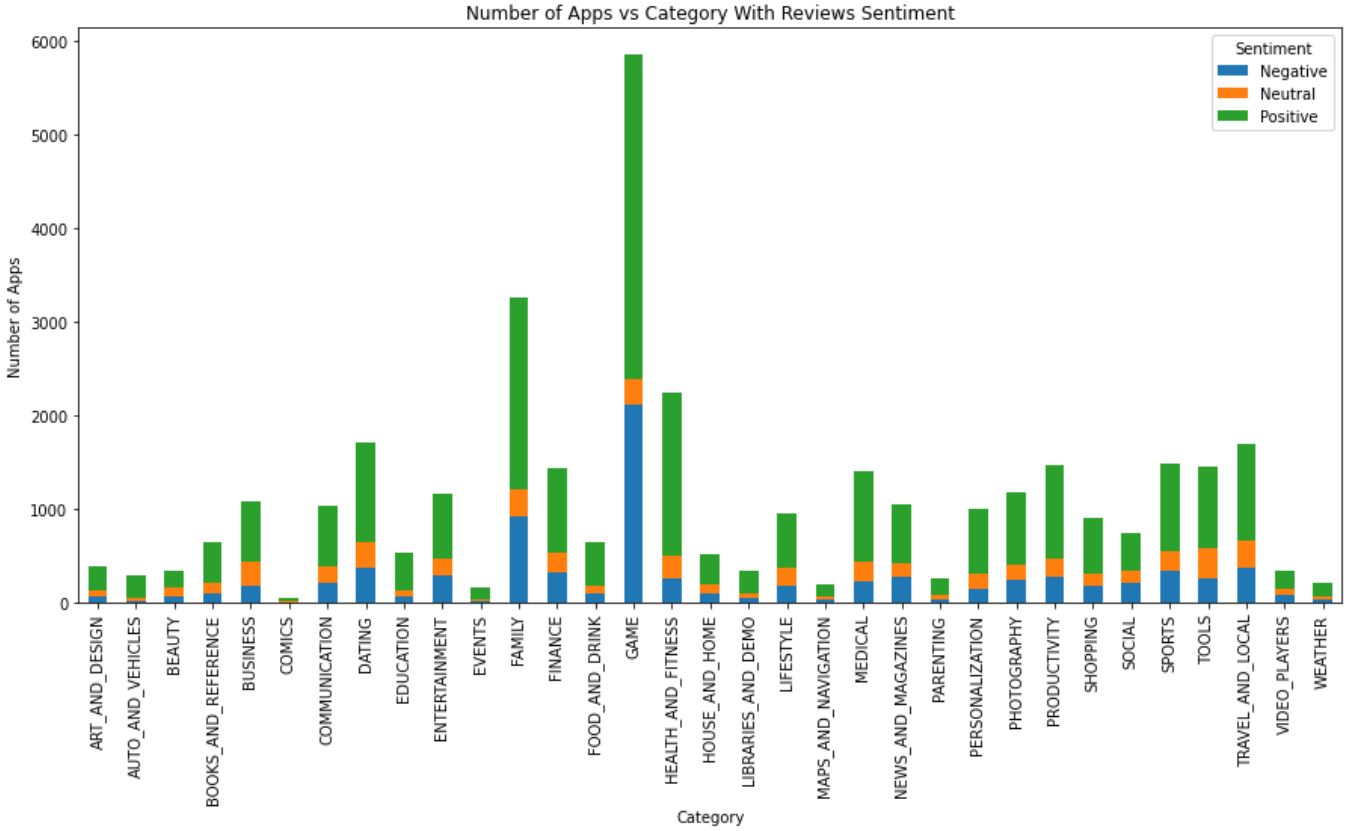
Histogram Plot for Sentiment Subjectivity



From 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.

Visualisation

Number of Apps in each Category with Reviews Sentiment



Visualisation

Correlation Heat Map

- **Installs** is showing fairly good relation with **Reviews.**
- **Size** and **Reviews** are slightly correlated.
- **Sentiment Polarity** and **Sentiment Subjectivity** are slightly correlated.

Conclusion

- **Family** category has the greatest number of apps present in the play store.
- **Game** category is the most installed and reviewed from the play store although the number of apps in Game category is almost half of the number of apps in Family category. This shows that the categories which are more entertaining and interactive will do better instead of having lower number of apps present in it.
- **Events** category has the highest **rating** of around **4.4** based on the median of the ratings given for each category apps.
- **Communication, Tools, Productivity, Social & Photography** are the top 5 **Genres** with most number of **installs**.
- Most of the apps in the Play Store are having **content rating 'Everyone' (81.82%)**.
- **Content Rating Teen** is having a quite good number of **installs** which 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.
- **92.2%** apps are **free** and **7.8%** apps are **paid** apps.
- There are more **free apps** present in the play store than the **paid apps** and also, 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.

Conclusion Contd.

- **Distribution of Size** shows most of the apps are of smaller size. Developer has to focus on reducing the size of the apps.
- Most of the apps are running at **Android Version 4.1** and above.
- **64.2%** of **reviews** are of **positive sentiment**, **22.3%** are of **negative sentiment** and **13.5%** are of **neutral sentiment**.
- Most the **sentiment subjectivity** lies between **0.4** to **0.7**.
- **Installs** is showing fairly good relation with **Reviews. Size** and **Reviews** are slightly correlated. **Sentiment Polarity** and **Sentiment Subjectivity** are slightly correlated.
- I'm Rich - Trump Edition is the most costly app with price tag of $400.0.
- Facebook is the most reviewed app with the review count of 78158306.
- Helix Jump has Positive Review count of 209 which makes it the most positively reviewed app in the dataset.
- Angry Bird classic has negative review count of 147 which makes it the most negatively reviewed app in the dataset.

These are some of the meaningful insights from the dataset. These insights can help the developer to capture the android market more efficiently.


Thank You!