Capstone Project 1

Play store app review analysis

Done by:

Neha Jadhav
Akriti Panchbudhe
Akanksha Jadhav

Content

WHY WE ANALYZE THE PLAY STORE?

Mobile App Market is set to grow 20% by 2023

Android apps comprise 75% of the Market Share. 85% sharein Brazil, India, Turkey

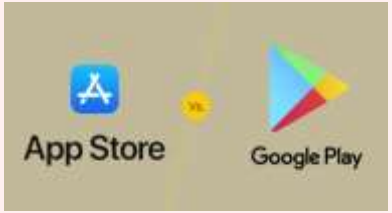
What makes an App popular? Can we predict how popular it's going to be?

What are some interesting patterns in user behavior related to app usage & feedback

Introduction

- Google Play was launched on March 6, 2012.
- Google Play is an online store where people go to find their favorite apps, games, movies, TV shows, books, and more.
- It provides 2 million apps & games to billions of people around the world, generating over $120 billion in earnings for developers to date.
- The main objective of the Project is that to understand the Users demand what they expecting from using their apps and thus it helps to update and develop the product by Developers.

Google Play Store is found to be the largest app market in the world. It has been observed that although it generates more than double the downloads than the Apple App Store but makes only half the money compared to the App Store.



Problem statement

- Two datasets are provided, one with basic information and the other with user reviews for the respective app.
- We must examine and evaluate the data in both datasets in order to identify the important characteristics that influence app engagement and success.
- For this project we analyze Play store data of 2017-2018 ,Google playstore is mostly use app store worldwide also top global market share.
- Our main objective is to find key factor responsible for app success and engagement of users.
- Thousands of new app regularly update playstore of different category.
- We found distribution of every app based on their size, installs, reviews and much more.

Attributes in Google Playstore Data

This contains data on the Google Play applications. It has 10,841 rows of data which has following columns.

- **App Category**: Category of the app. This could be beauty, business, medical,art and design entertainment, education, family...etc.
- **Rating**: It has How users rate the app out of 5, with 1 being the lowest rating and 5 being the highest.
- **Reviews**: number of user reviews each app has received.
- **Size**: The memory size needed to install the application.
- **Installs**: The number of times each application has been installed by users.
- **Type**: Whether the app is free or a paid app.
- **Price**: The price of the app.
- **Content Rating**: This column specifies the intended audience for the app. Can be for teens, mature audience, or everyone.
- **Genres**: The sub-category for each app. Example: for the Education category, this could be Education: Pretend Play, for example.
- **Last Updated**: Release date of the most recent update for the app.
- **Current Ver**: The app's current version.
- **Android Ver**: The oldest version of Android OS supported by the app.

Attributes in User Reviews

- **Sentiment Analysis**: This file contains the result of the sentiment analysis conducted by the dataset creator. It has 64,295 rows of data with the following columns:
- **App** : Name of the app.
- **Translated_Review**: Either the original review in English, or a translated version if the orignal review is in another language.
- **Sentiment**: The result of the sentiment analysis conducted on a review. The value is either Positive, Neutral or Negative.
- **Sentiment_Polarity**: A value indicating the positivity or negativity of the sentiment, values range from -1 (most negative) to 1 (most positive). Sentiment polarity for an element defines the orientation of the expressed sentiment, i.e. it determines ifthe text expresses the positive, negative or neutral sentiment of the user about the entity in consideration.
- **Sentiment_Subjectivity**: A value from 0 to 1 indicating the subjectivity of the review. Lower values indicate the review is based on factual information, and higher values indicate the review is based on personal or public opinions or judgements.

Data Cleaning 

- Google Play store dataset has 10,841 observation of data with fields.
- User reviews dataset has 64295 observation of data with fields.
- Two data set -1) play store data 2) user reviews

List of fields:

Playstore data

- ❑ App
- ❑ Category
- ❑ Rating
- ❑ Reviews
- ❑ Size
- ❑ Installs
- ❑ Type
- ❑ Price
- ❑ Content rating
- ❑ Genres
- ❑ Last updated
- ❑ Current version
- ❑ Android version

User reviews

- ❑ App
- ❑ Translated review
- ❑ Sentiment
- ❑ Sentiment polarity
- ❑ Sentiment subjectivity

Cleanse and validate data

Understand the structure of the dataset and clean data before analysis

Removing extraneous data

Finding Missing value in dataset.

- Correct data type(INT, FLOAT, DATE).
- Replace null value with aggregate function (mean, mode , median).
- Checking outliers.
- Conforming data to a standardized pattern.

Transform data

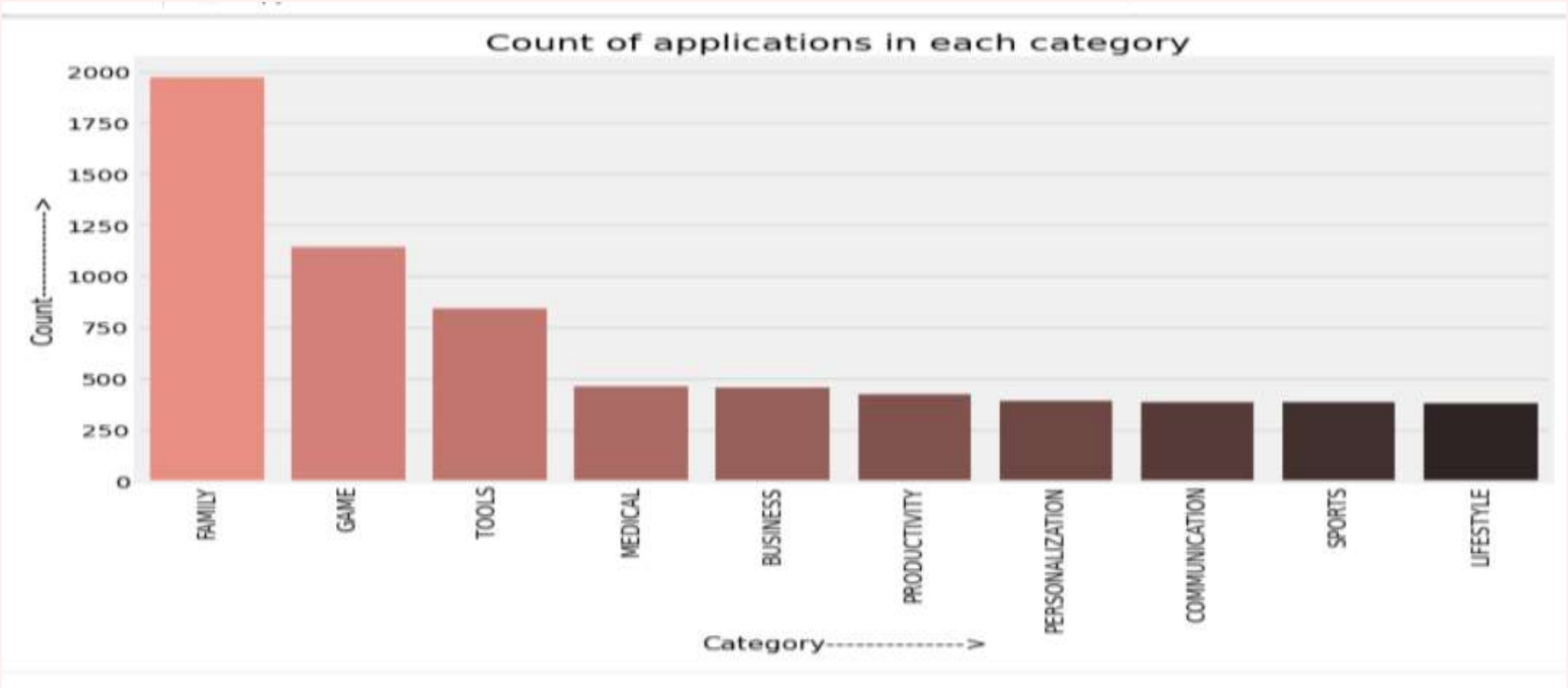
- Transforming data is the process of updating the format or value entries in order to reach a well-defined outcome, or to make the data more easily understood by a wider audience.
- The dataset collected from the Play store is semi structured or unstructured and contains significant superfluous data (defined as not contributing significant meaning).Some data type needs to change in required format as int, float, date.
- we will be working with most frequently henceforth are Installs, Size, Reviews, Rating and Price.

Exploratory data analysis

- Exploratory data visualizations (EDVs) are the type of visualizations we assemble when we do not have a clue about what information lies within our dataset.
- With more than 1 billion active users in 190 countries around the world, Google Play continues to be an important distribution platform to build a global audience. For businesses to get their apps in front of users, it's important to make them more quickly and easily discoverable on Google Play. To improve the overall search experience, Google has introduced the concept of grouping apps into categories.

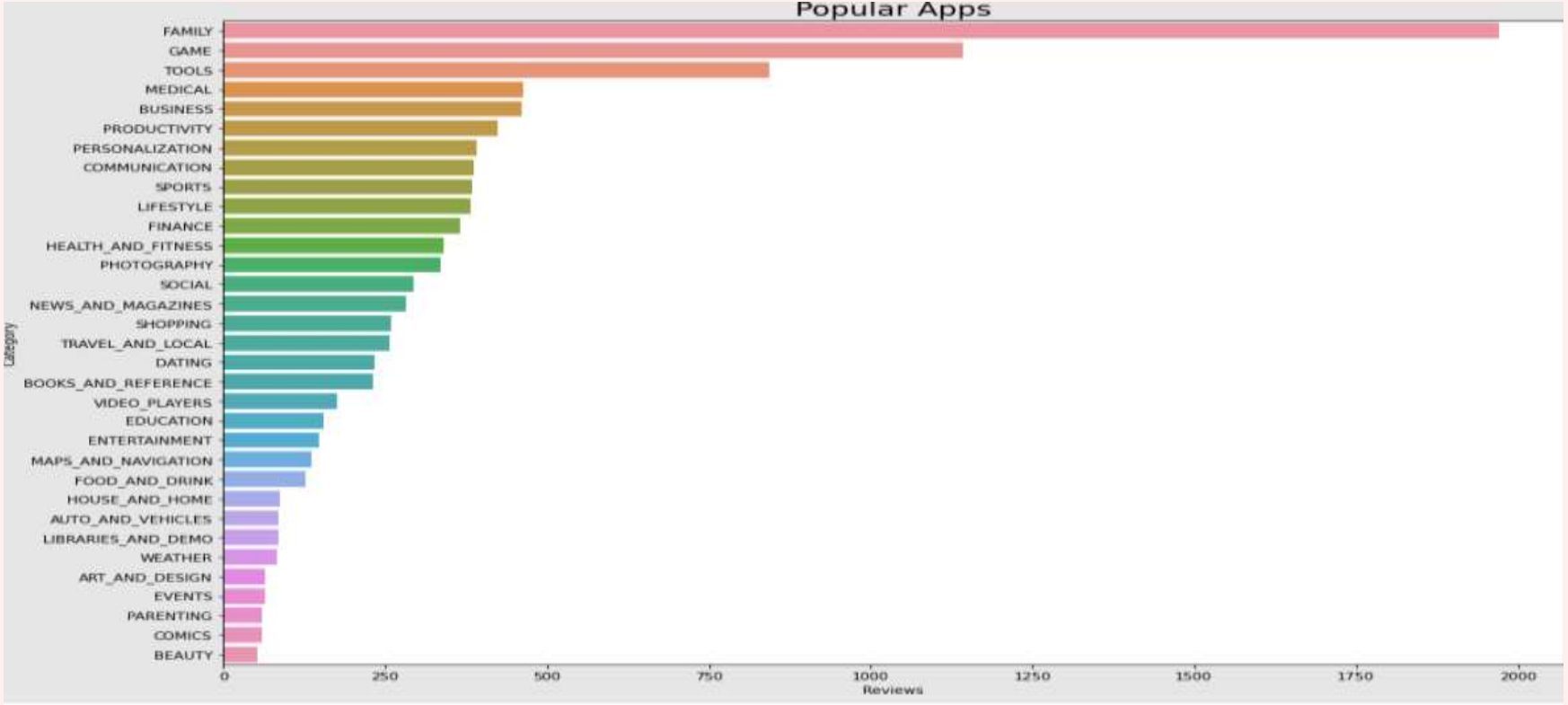
Let's move to the visualization part-

Count of applications in each category



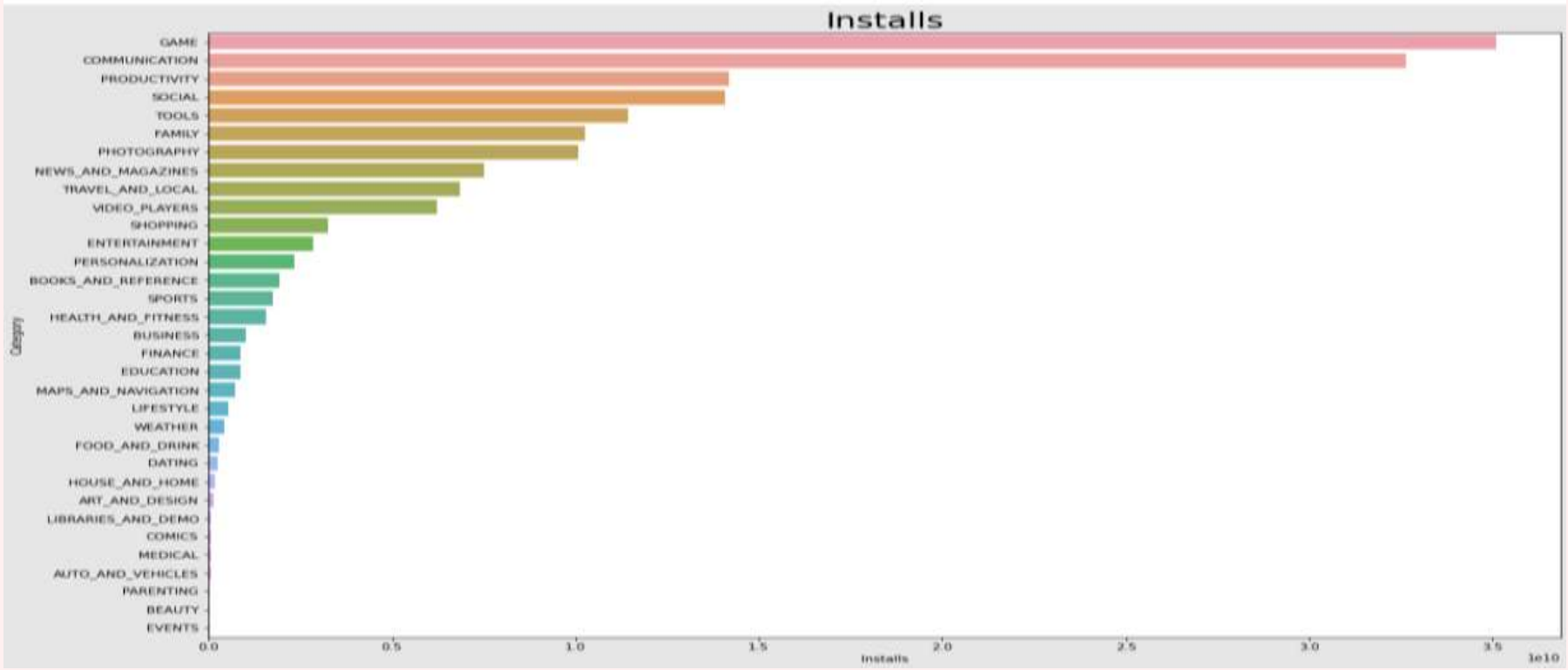
Now we know that the 'Family' and 'Game' category rules the play store market, followed by Tools, Medical, and Business. Okay Cool ..... Data Analytics will understand our daily requirements and fill the market with similar apps.

Popular Apps with respect to reviews



- From above barplot we can see that there are total 33 different categories in the dataset.
- The most popular apps are from 'Family' and 'Games' category. The least reviews are for the apps from 'Beauty' and 'Comics' Category.

Top Installed categories with respect to categories

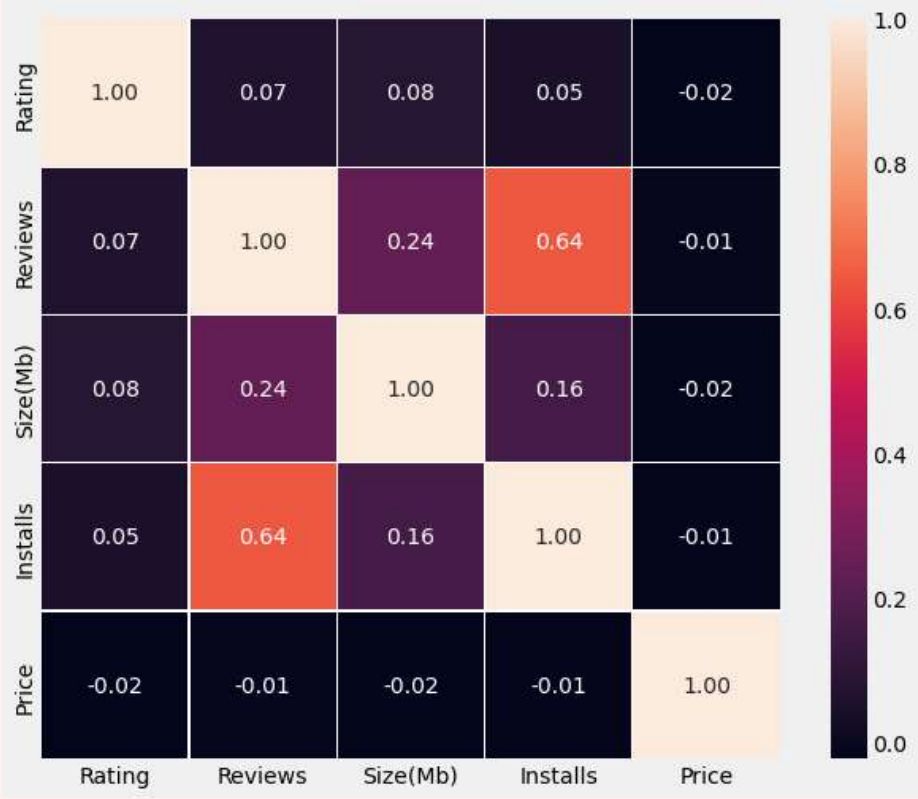


We can see that most of the apps has been installed from 'Game' and 'Communication' category.
Let's see which apps are most installed Category wise.

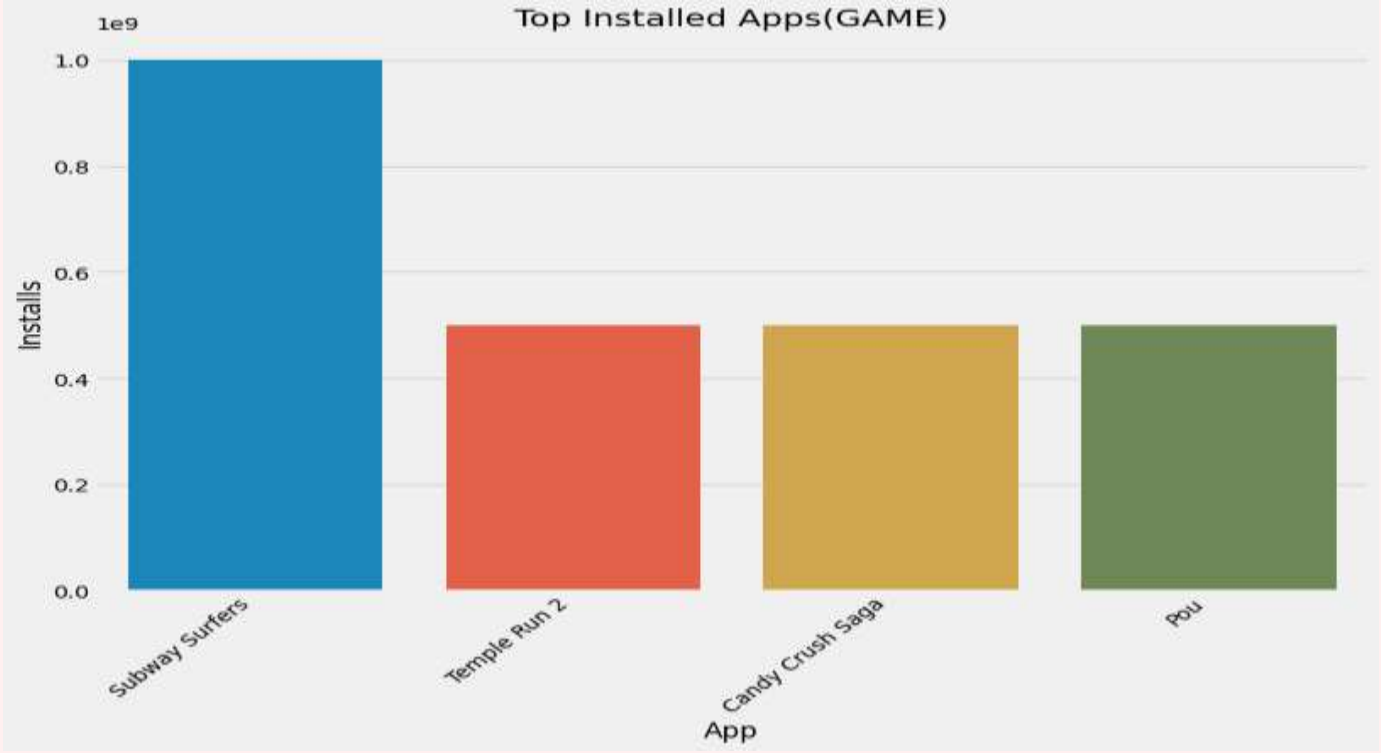
Correlation Heatmap between the columns

From the above heatmap we can see that the 'Reviews' and 'Installs' columns has most correlation.

It is much more obvious that a higher number of installs has a higher number of reviews. There is a negative correlation between price and install apps, with the price of the app influencing the number of installation of the app.

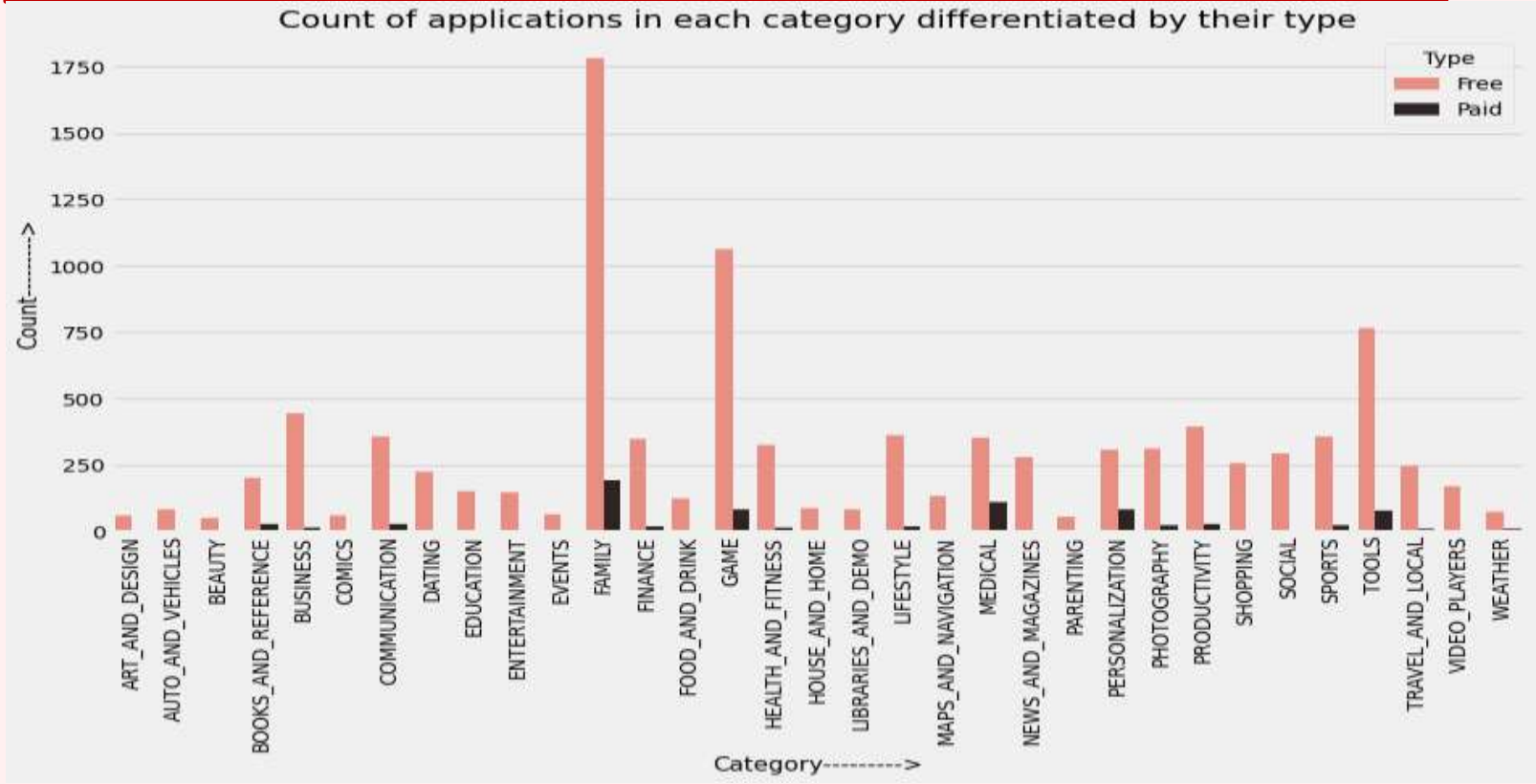


Top Installed gaming Apps



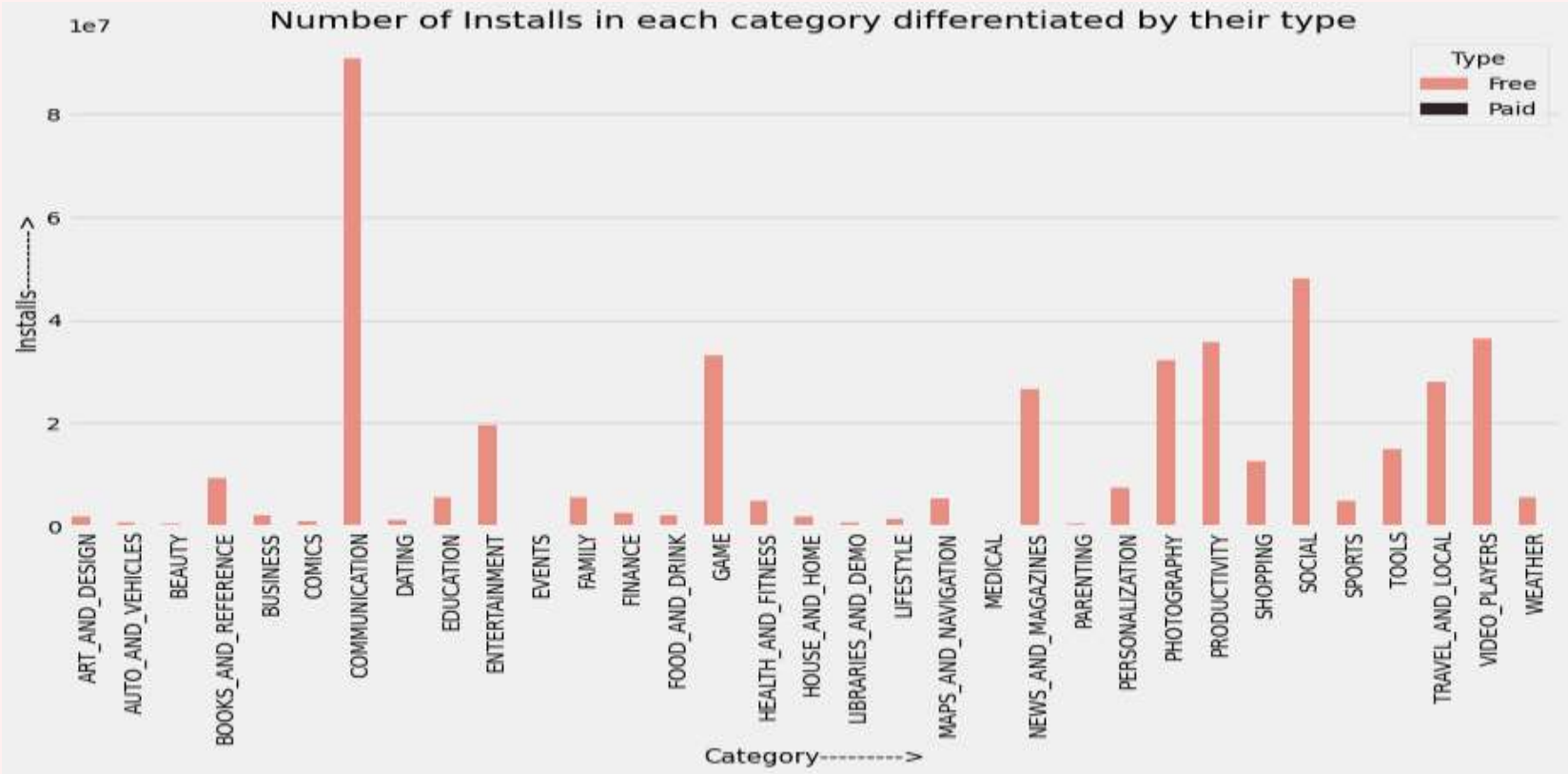
From the above plot we can see that the most installed games from the 'Game' Category are 'Subway Suffers', 'Temple Run 2', 'Candy Crush Saga' and 'Pou'.

Count of Applications in each category differentiated by their type



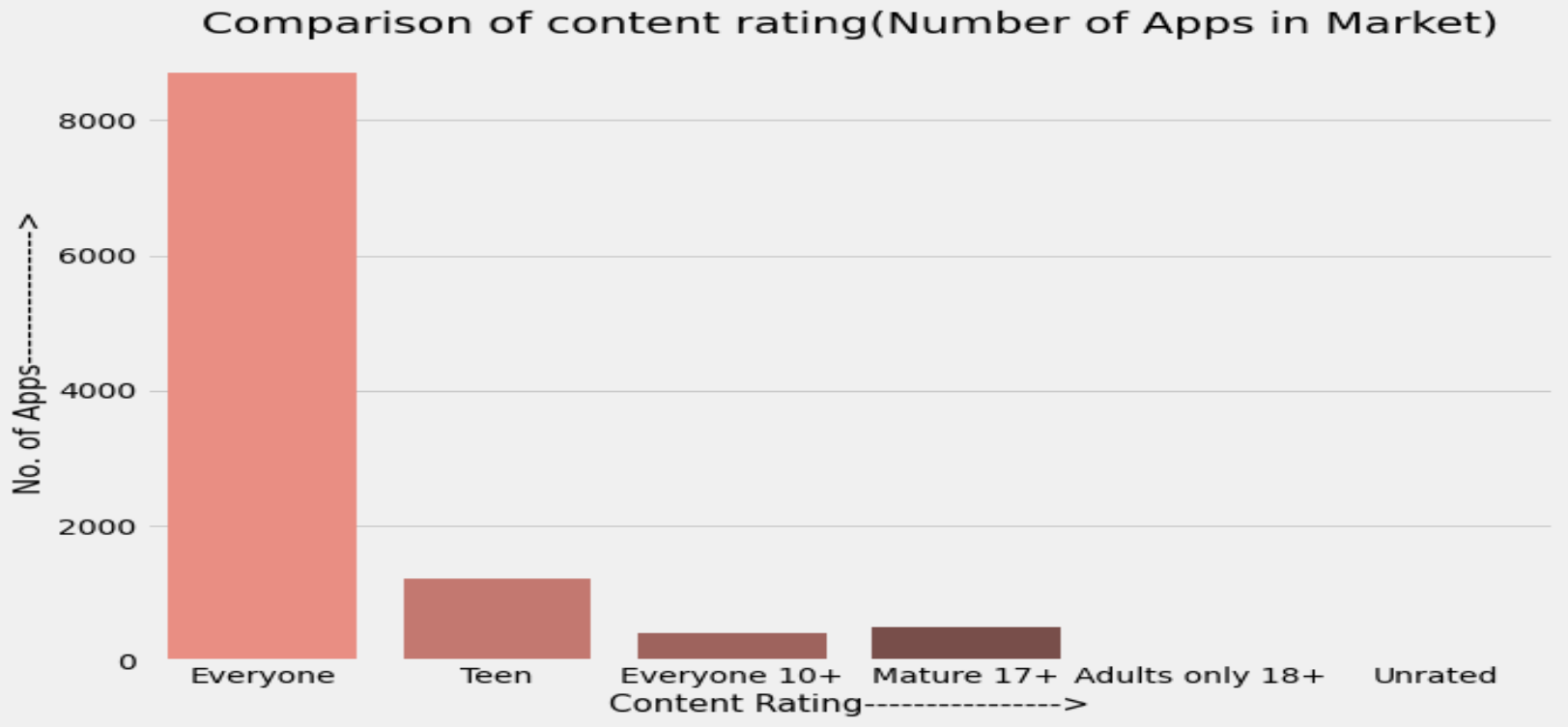
- It looks like certain app categories have more free apps available for download than others. In our story, the majority of apps in the Family, Game, and Tools categories were free to install.
- At the same time Family, Medical, Tools, and Game categories had the biggest number of paid apps available for download.

Number of installs in each category differentiated by their type



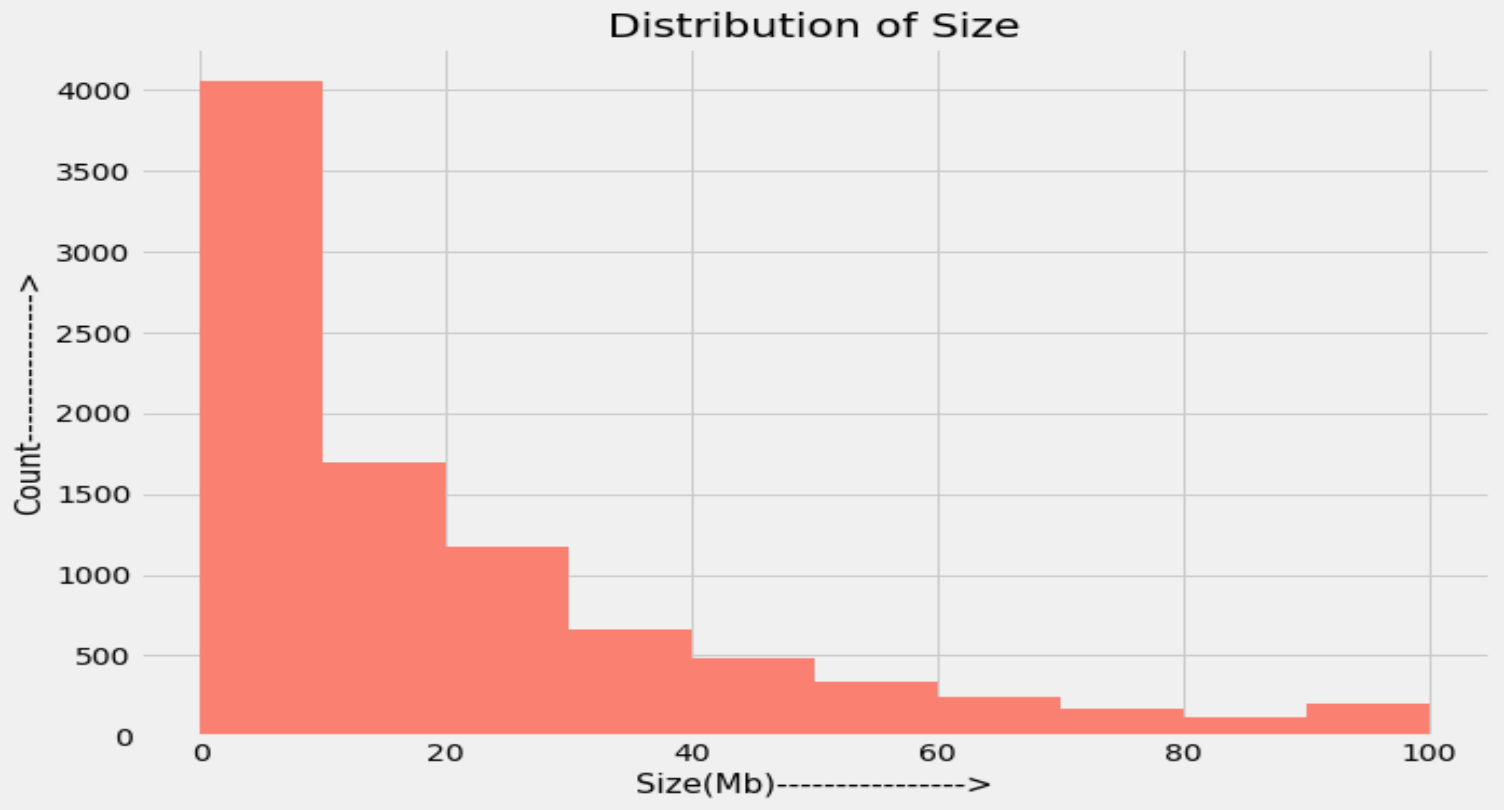
It can be concluded that the number of free applications installed by the user is high when compared with the paid ones.

Comparison of content rating (Number of Apps in Market)



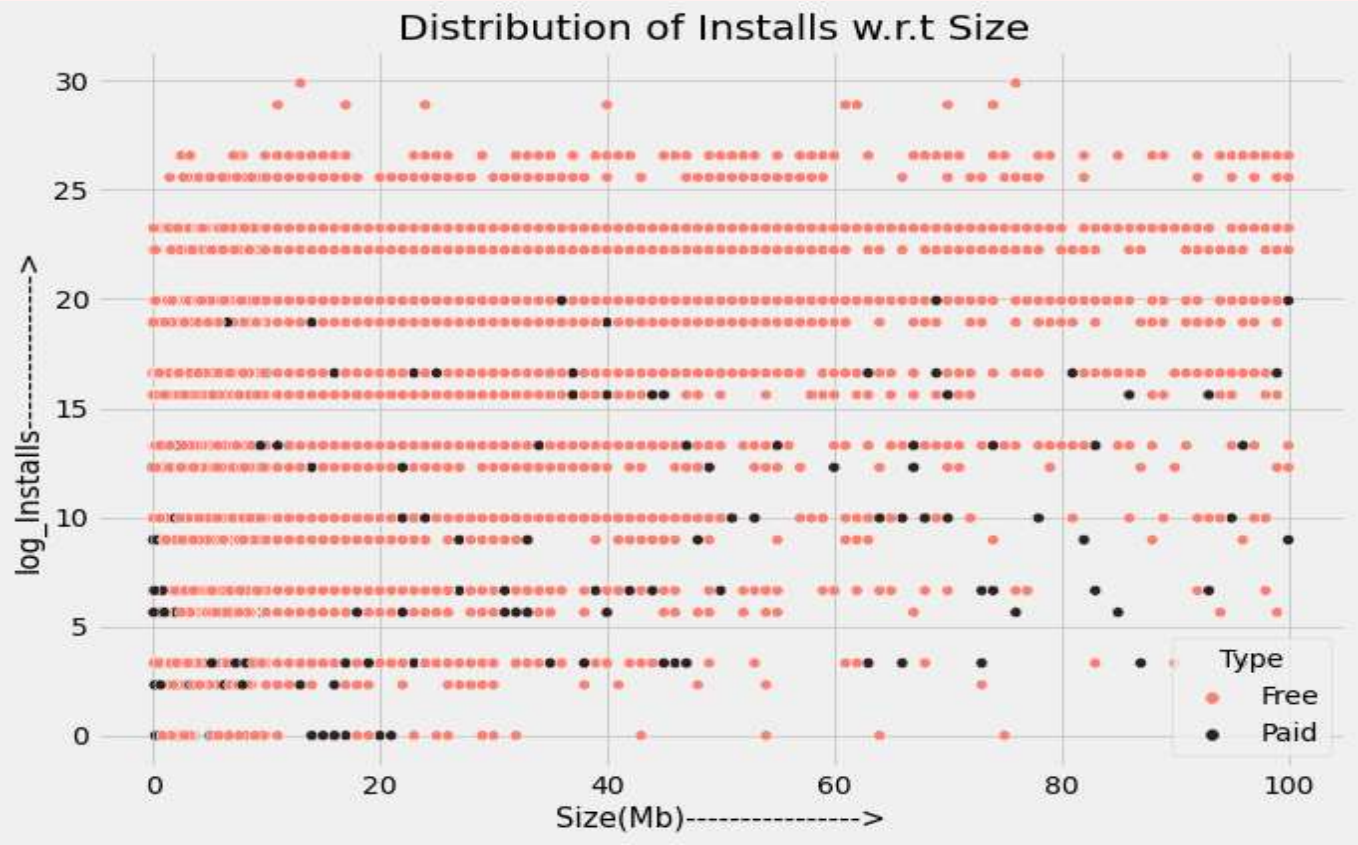
From the above plot we can see that most of the content ratings are from 'Everyone' Category.

Distribution of Apps according to size



As we can see as the size of the app goes on increasing the number of counts goes on decreasing.

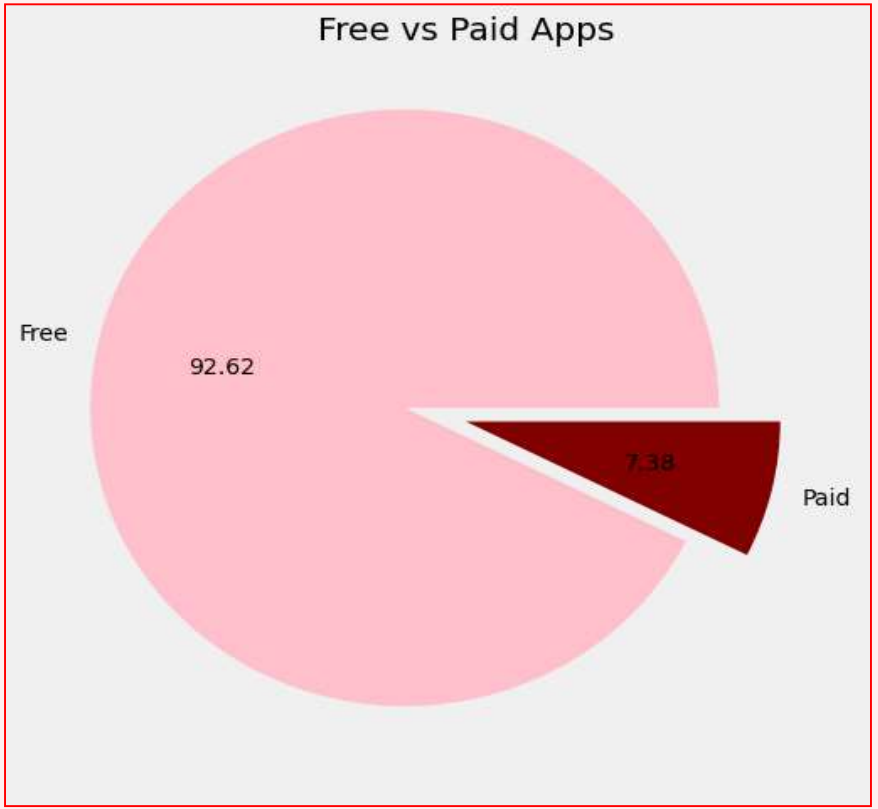
Distribution of Installs according to size



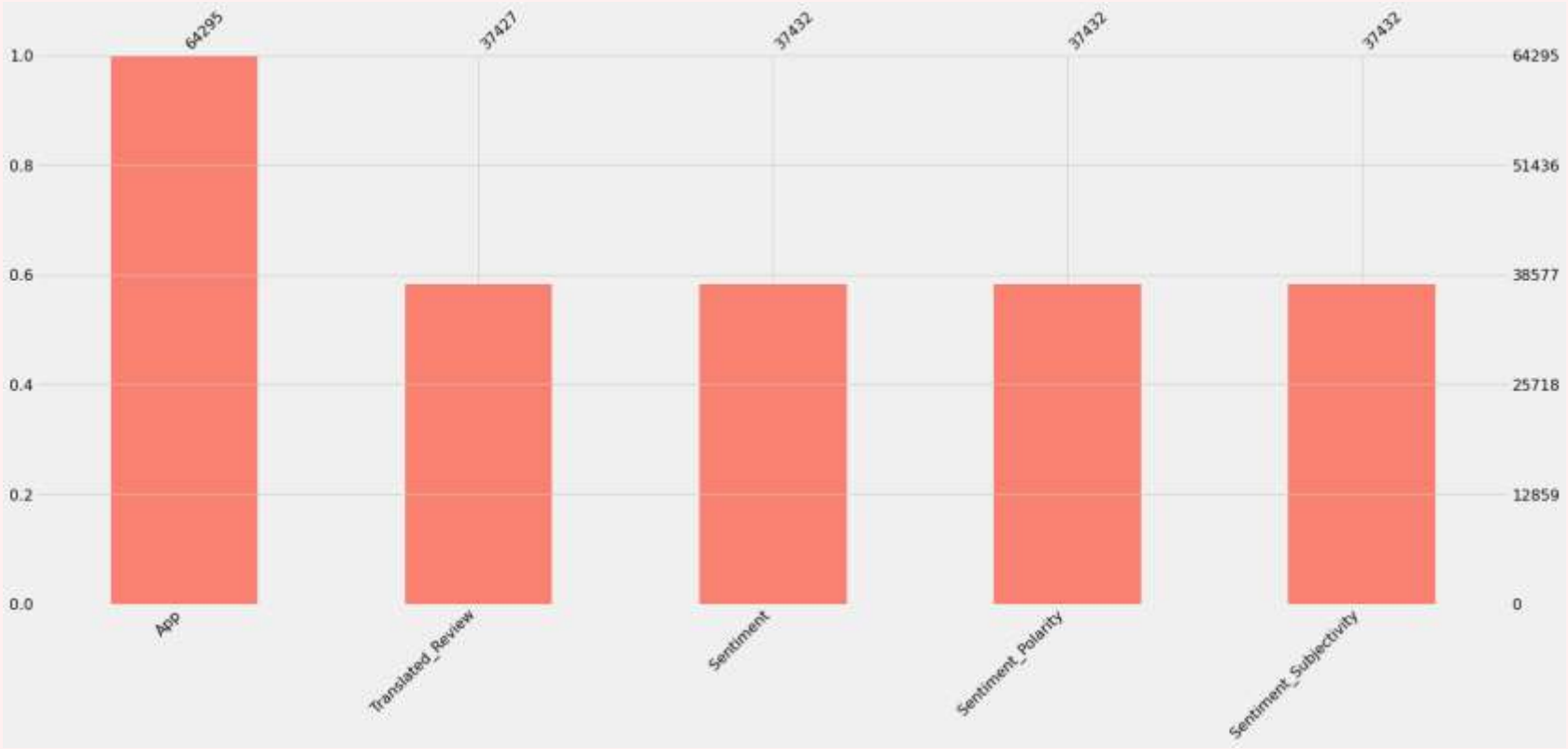
From the above scatter plot we can see that as the size increases the number of installs is less

Free Vs Paid Applications

From the piechart we can see that 92.62% of apps in google play store are free and 7.38% are paid.

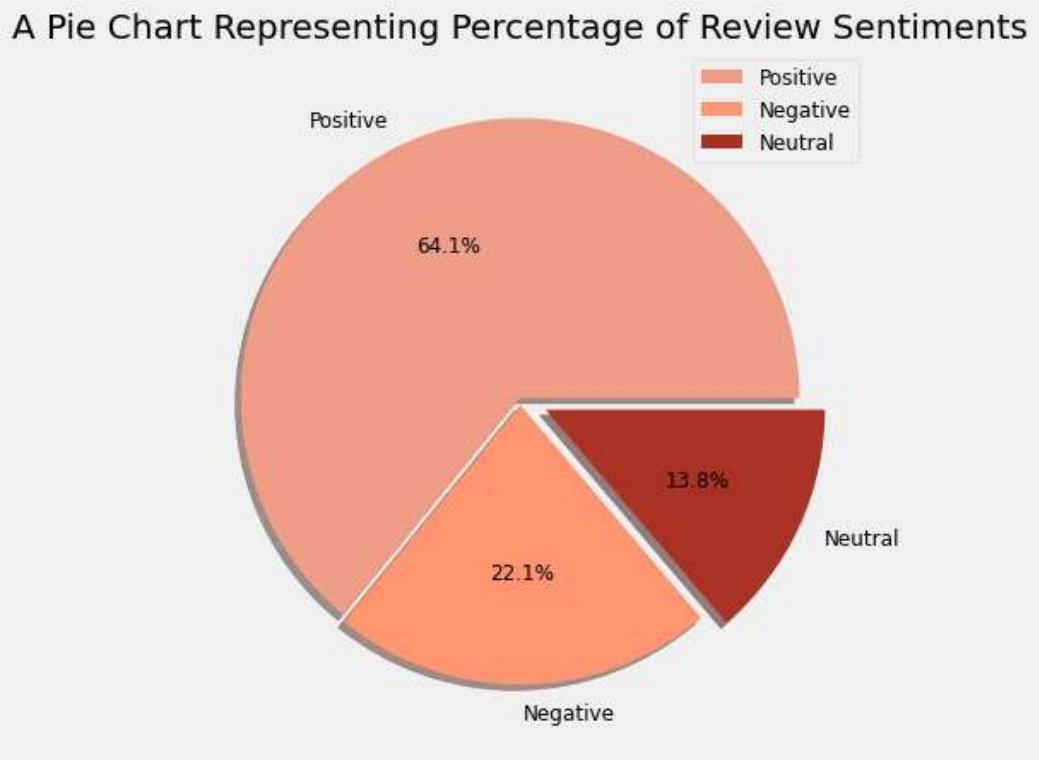


Sentiment analysis



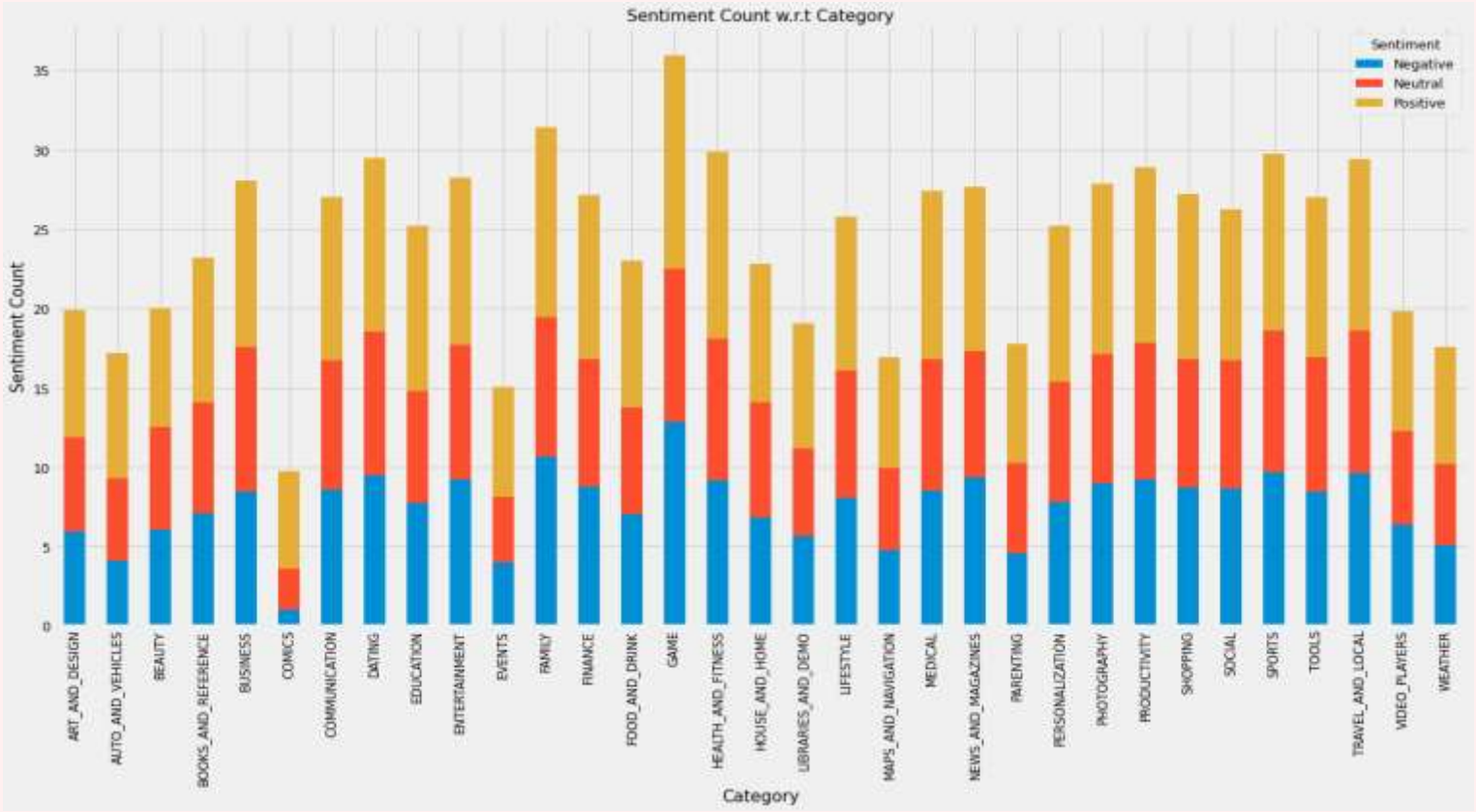
From the above chart we can conclude that there are null values present in the Translated_Review, Sentiment,Sentiment_polarity and Sentiment_subjectivity columns.

Percentage of Review Sentiments



It can be seen from above chart that positive reviews are more than negative reviews.

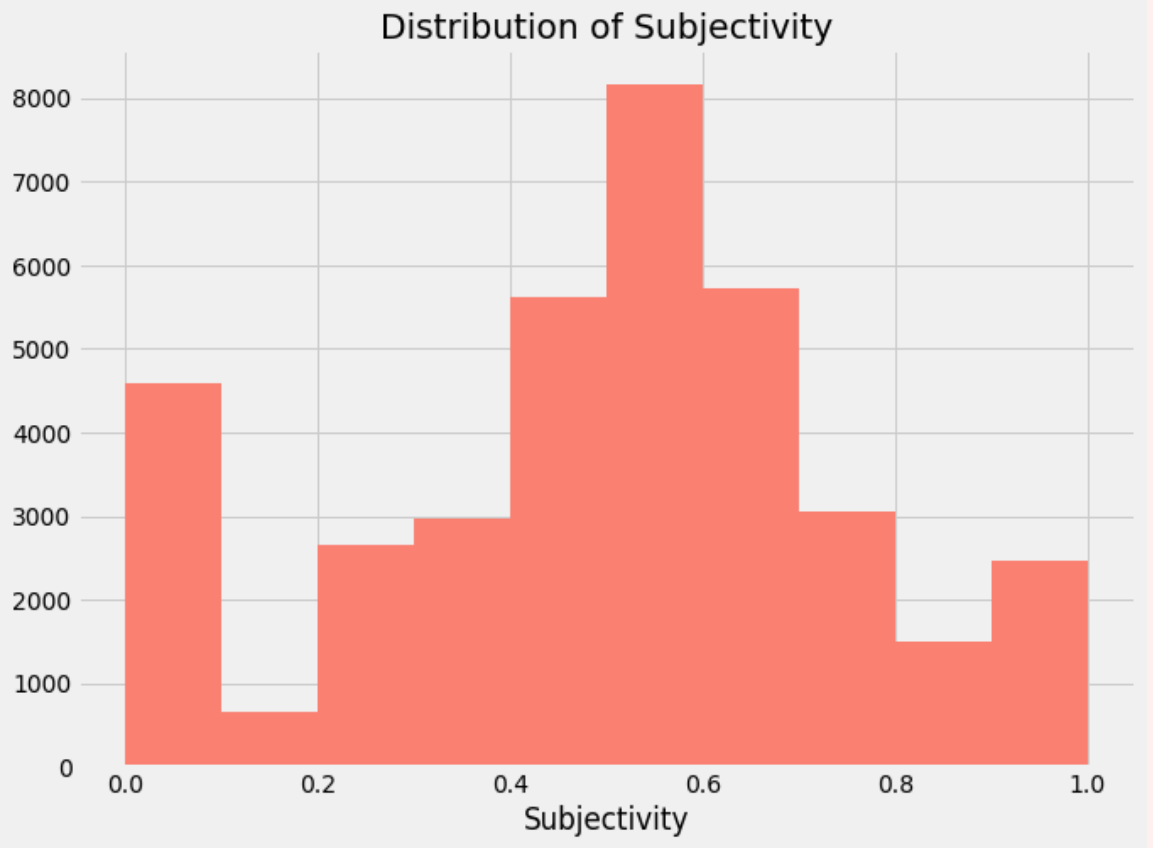
Sentiment Count w.r.t. category



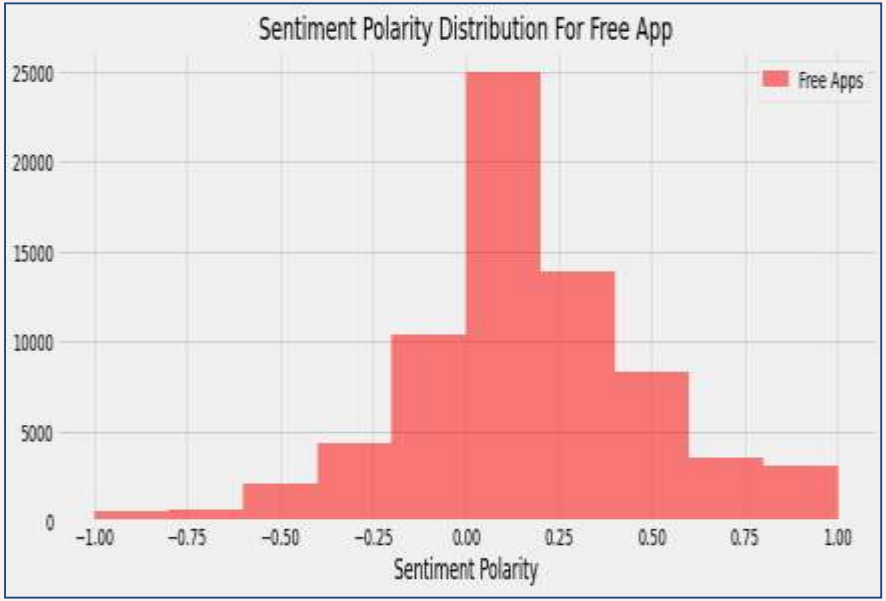
Positive sentiment count is more for many apps than negative count.

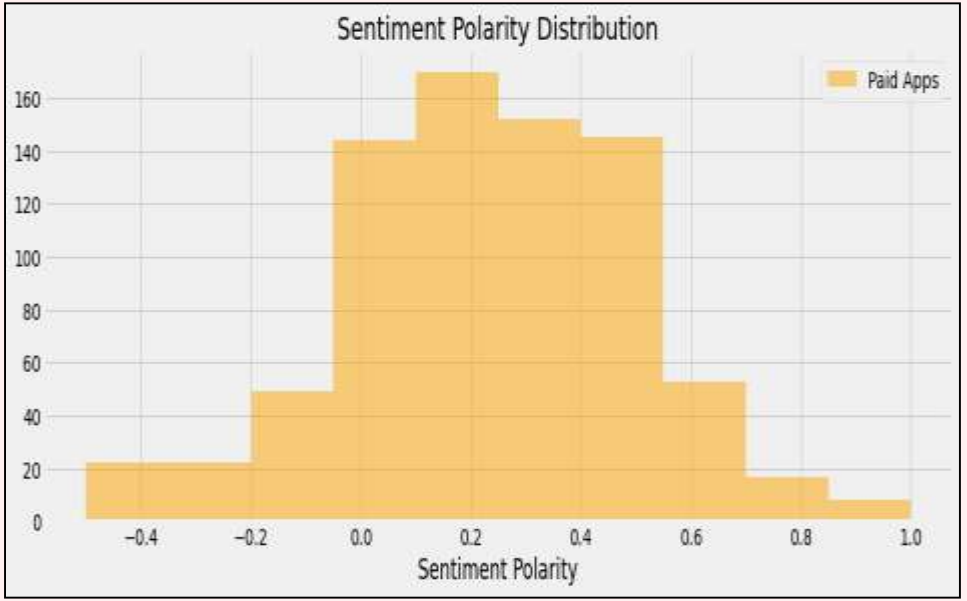
Distribution of Subjectivity

- It can be seen that the maximum number of sentiment subjectivity lies between 0.4 to 0.7.
- From this, we can conclude that the maximum number of the audience give reviews to the applications, according to their experience.
- The higher subjectivity means that the text contains personal opinion rather than factual information.
- Sentiment_Subjectivity > 0.5(refers to that mostly it is public opinion and not a piece of factual information)



Sentiment Polarity Distribution for free and paid Apps





- paid apps have positive sentiment polarity as compared to free apps.
- While free apps have neutral sentiment polarity.

Insights from data

- 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 ratings 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.

Sentiment Polarity

- The polarity of a sentiment measures how negative or positive the context is.
- In the data that we have, the polarity ranges from -1 (most negative) to +1 (most positive).

Future Work

- Exploring the correlation between the size of the app and the version of Android on the number of installs.
- Exploring reviews and sentiment of the users as per the the category of the application.

Conclusion

- Most reviews are for category of 'Family' and 'Game'.
- Most installed apps are from 'GAME' and 'COMMUNICATION' Category.
- Top installed games are 'Subway Surfers', ' Candy Crush Saga' , 'Temple Run 2'and 'Pou'.
- Free apps have more number of Ratings than the paid apps.
- Paid apps have more positive user reviews than the free apps while free apps have more number of negative reviews than paid apps.
- Positive sentiment percentage for paid apps is more than the free apps.

References

- ✓ GeeksforGeeks
- ✓ Stackoverflow
- ✓ W3schools

Thank You