Capstone Project

Play Store App Review Analysis

Presented By

Tushar Yuvraj Khairnar

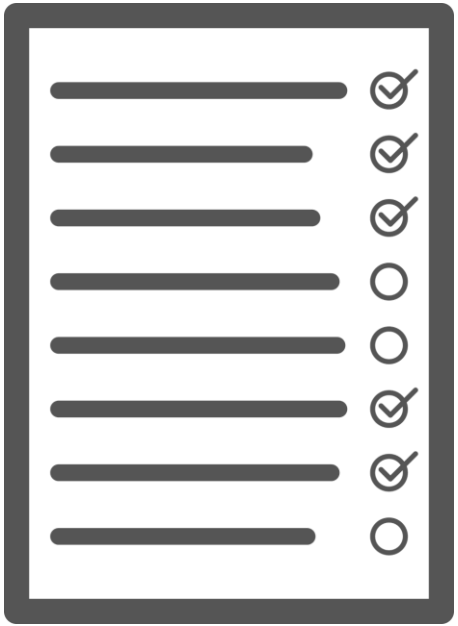
Shreyash Sharad Sarode

Taha Yusuf Nakedar

Pradnya Kewal Pagar

Data Science Trainee, AlmaBetter

Agenda



- Introduction
- why analyze the google play app store?
- Problem Statement
- Category wise play store apps installs
- Category wise most popular apps
- Top 10 apps in play store considering all the parameters
- Average installs, category wise
- Most installed apps in communication category
- Average sizes of apps in each category
- Category wise percentage of paid apps
- Category wise top installed paid apps
- Average rating of paid apps
- Correlation between Rating ,Installs and Price
- Category wise installed apps with content rating
- Percentage reviews sentiment distribution

WHY ANALYZE THE GOOGLE PLAY APP STORE?

- Google Play Store is an online App store like Games, Movies, Books etc. where people can find their favorite Apps. And it provides more than 10 million Apps and Games to billions of people over the world and it’s generating more than 120$ billion earning to developers.

- 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.

Introduction

- Today's Day to day scenario we can see that mobile apps playing an important role in any individual’s life. With enormous challenge from everywhere throughout the globe, it is important for a designer to realize that he/she is continuing in the right way or not. To hold this income and their place in the market the application designers may need to figure out how to stick into their present position.

- The dataset with 10k Play Store applications is available to analyze the market of android. It can be examined to analysis the different category such as family, communication, entertainment, tools, music, camera etc. In this project we examine the different attributes present in the data set that affect the popularity of the application. We focused on to answer the questions like, what makes an app popular, what should be the price and size of the app, is there some trends in user sentiments. In our data set we have two csv files for data analysis

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.

Play Store need to focus more on ⬇

1. Focusing more on content available for Everyone will increase the chances of getting the highest installs.
2. Most of the apps are Free, so focusing on free app is more important.

3. Developing apps related to the least categories as they are not explored much. Like events and beauty.

4. They need to keep in mind that the sentiments of the user keep varying as they keep using the app, so they should focus more on users needs and features.

Dataset Preparation

- **Loading the data sets:** Two datasets, First Play store app
  dataset and User Reviews dataset.
- **Import Libraries:** NumPy, Pandas, Seaborn and Matplotlib
- **Data cleaning:** Null values, Finding and removing Outliers, Removing duplicate data.
- **Data Imputation:** Filling the missing categorical values with mode and numerical values with median. Conversion of price, installs, reviews into numerical values.
- **Exploratory Data Analysis:** Analyzing the data sets to summarize their main characteristics using statistical graphics and data visualizations method.

Attributes in Google Play store Data

1.APP : This column contains the name of the app for each observation.

2.CATEGORY : This column contains category to which the app belongs.

3.RATING : This column contains the average rating for the app.

4.REVIEWS : This column contains the number of reviews that the app has received on the play store.

5.SIZE : This column contains the amount of memory the app occupies on the device.

6.INSTALLS : This column contains the number of times that the app has been downloaded and installed from the play store.

7.TYPE : This column contains the information whether the app is free or paid.

8.PRICE: If the app is a paid app, this column contains the data about its price.

9.CONTENT RATING: This column contains the maturity rating of the app i.e. the age group of the audience for which it is suitable.

10.GENRES: This column contains the data about to which genre the app belongs. genres can be considered as a further division of the group of category.

11.LAST UPDATED: Contains the date on which the latest update of the app was released.

12.CURRENT VERSION: Contains information on the current version of the app available on the play store.

13.ANDROID VERSION: Contains information about the android versions on which the android version is supported.

Attributes in User reviews

1. **App-** Application name
2. **Translated Review-** User review
3. **Sentiment-** Positive/Negative/Neutral
4. **Sentiment Polarity-** Sentiment polarity score
5. **Sentiment Subjectivity-** Sentiment subjectivity score

OVERVIEW OF ANALYSIS

Data Cleaning

Understand the structure of the dataset and clean data before analysis

Data Exploration

Uncover initial patterns, characteristics, and points of interest using visual exploration

Predictive Modeling

Formulate a statistical model to forecast an outcome using relevant predictors

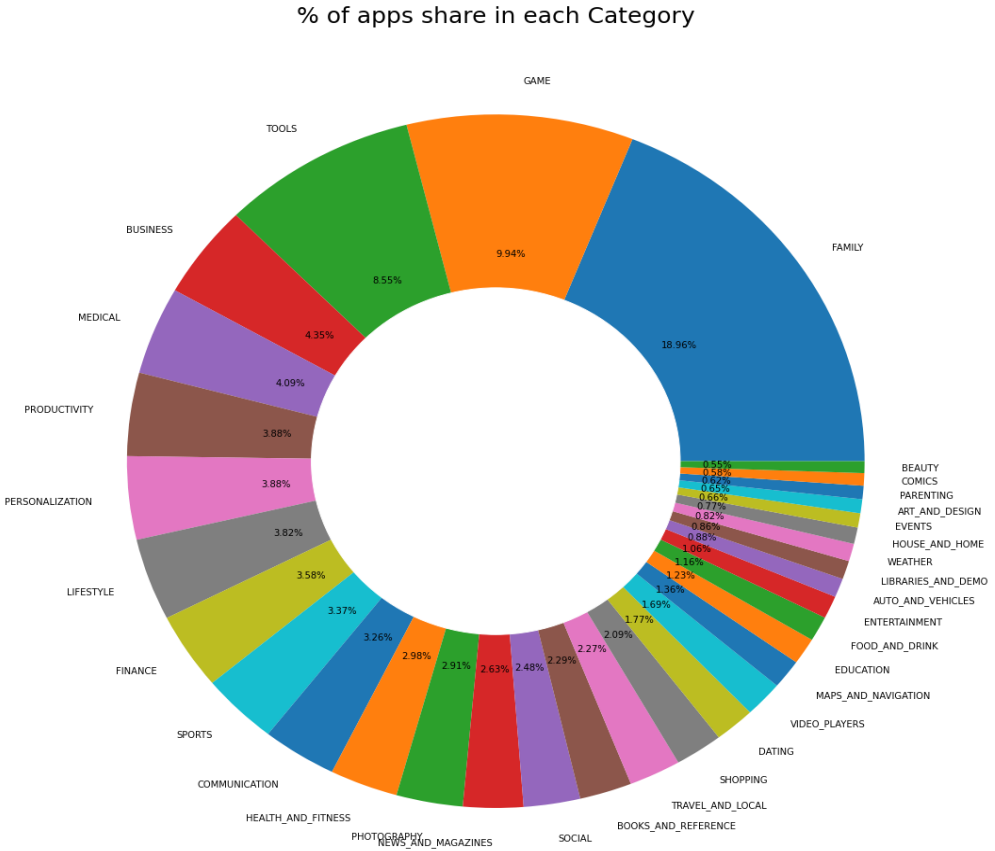
Handling the NaN Values

- Missing Data can occur when no information is provided for one or more items or for a whole unit. Missing Data is a very big problem in a real-life scenarios. Missing Data can also refer to as NA(Not Available) or NaN(Not a Number) values in pandas. In DataFrame sometimes many datasets simply arrive with missing data, either because it exists and was not collected or it never existed.
- Columns which we have handles the NaN Values-:
- The Rating column has 1474 NaN values.
- The Type column has 1 NaN value.
- The current Var column contains 8 NaN values.
- The Android Var Column contains 2 NaN values.

- Above, we handled all the NaN values in a particular column.

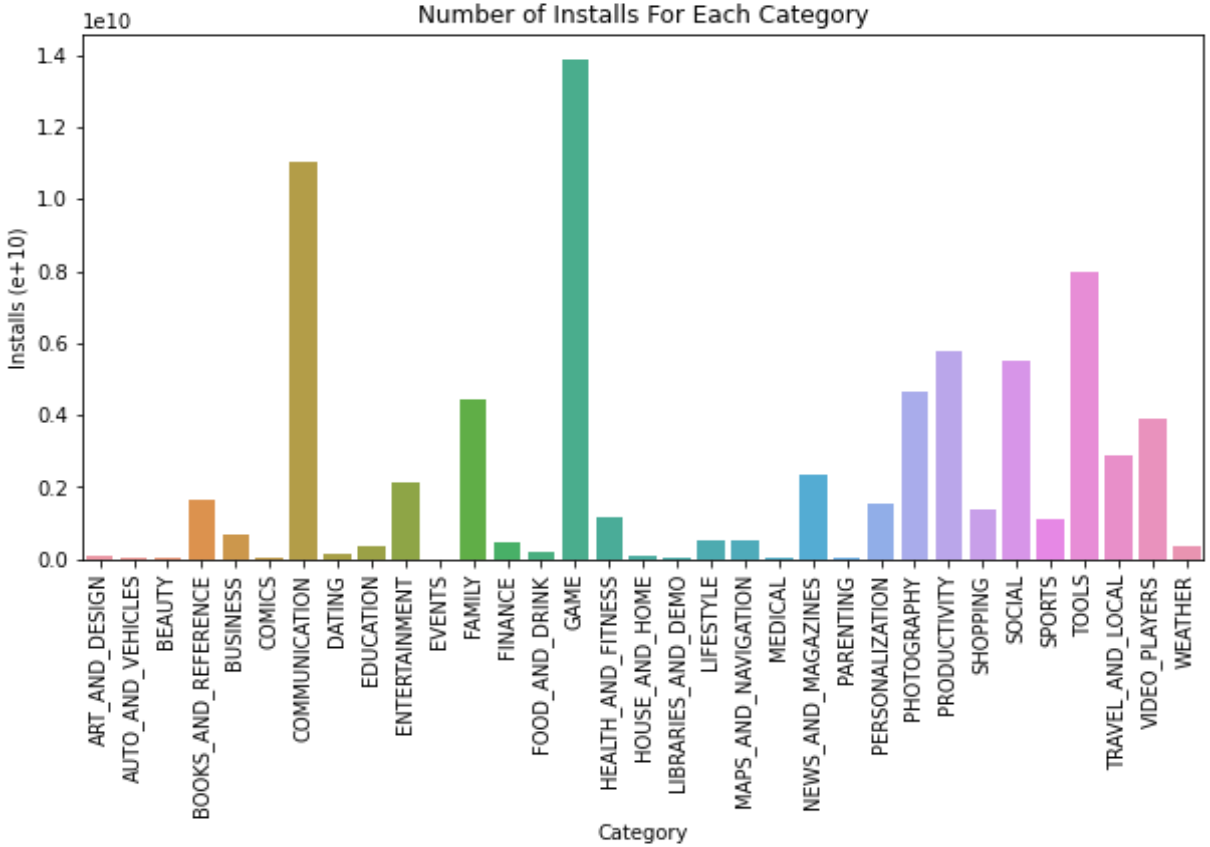
Share in the PlayStore App Category wise

From the Plot we can see that Play store has almost 33 categories. In this plot topmost apps share are from Family and Games. And least are from Comics and Beauty.



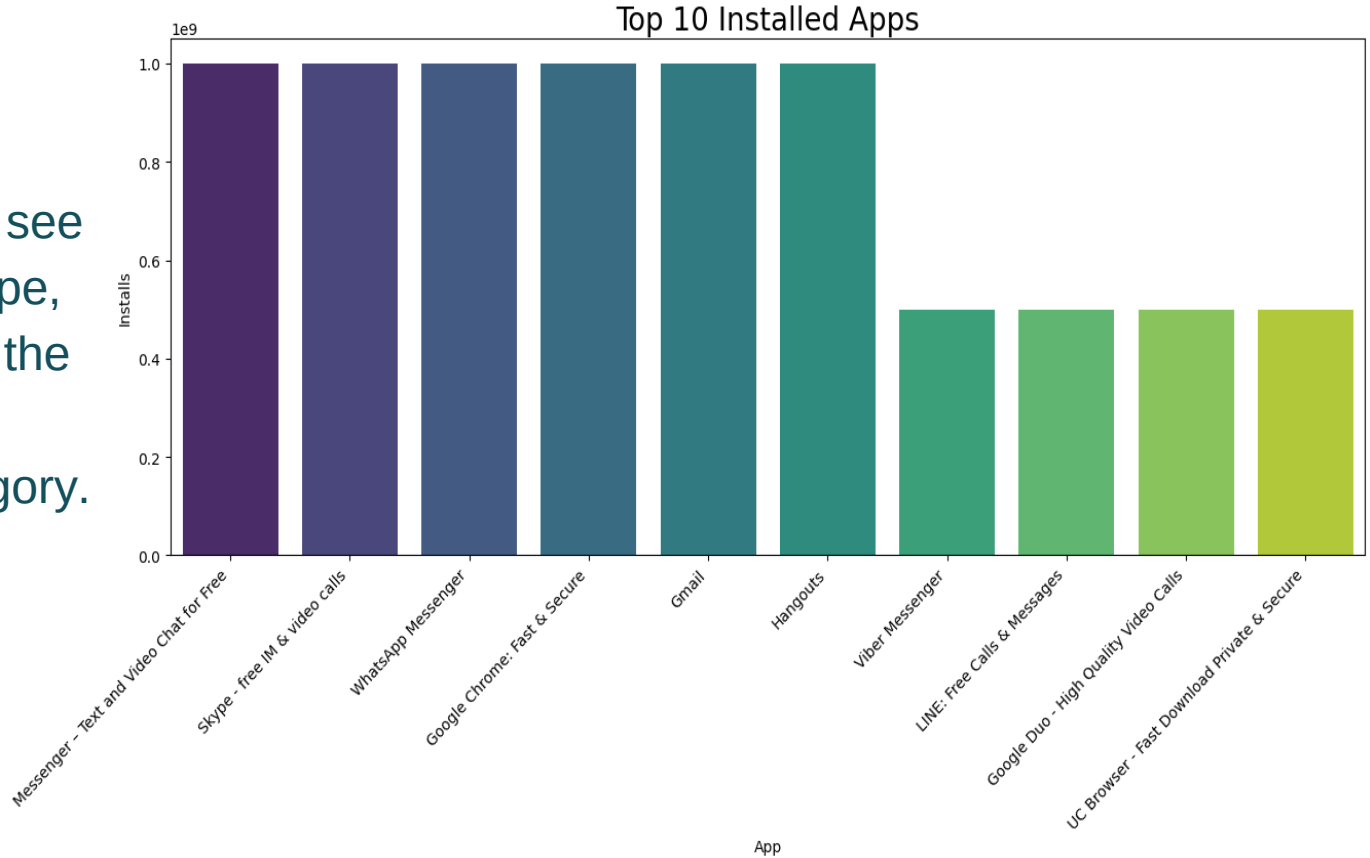
Category of App most installed by user -:

- The weightage of games and communication is much higher than all other categories.



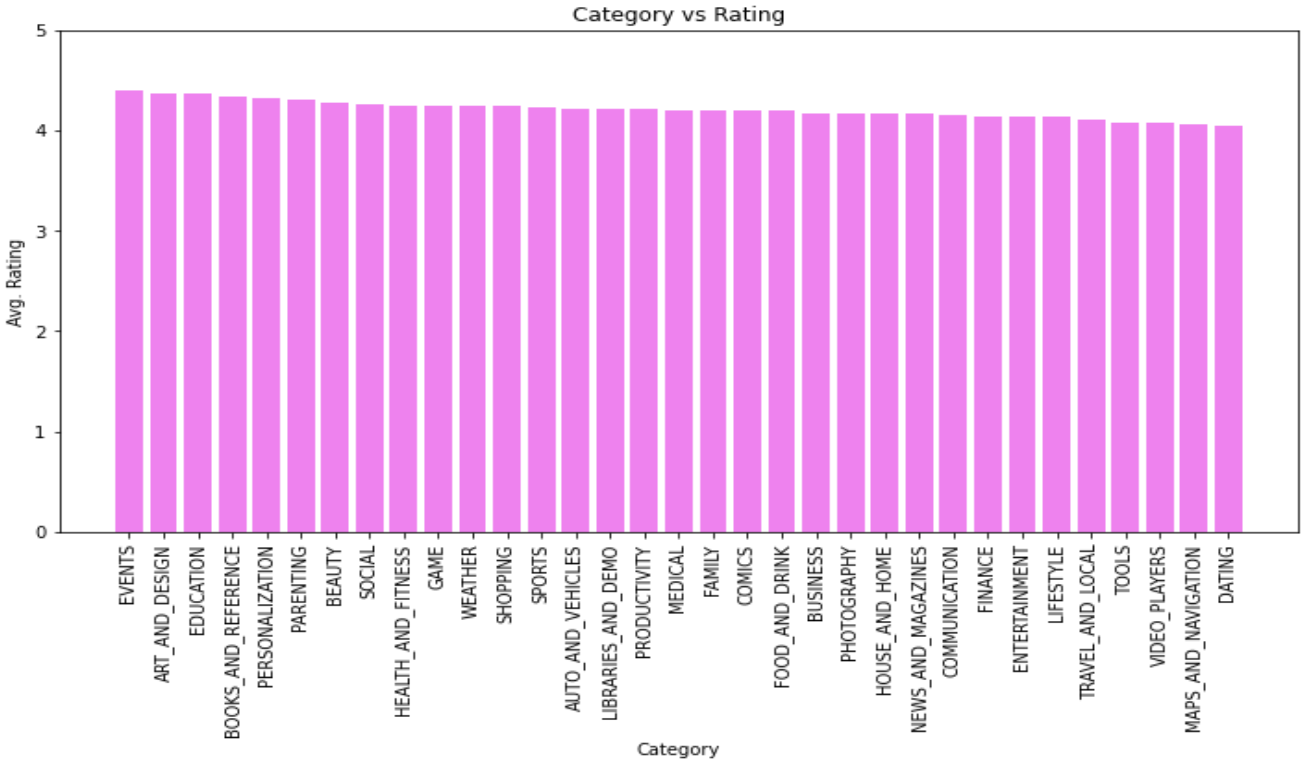
Most Install App in particular category -:

- In the graph, we can see that messenger, Skype, and WhatsApp have the most users in the communication category.

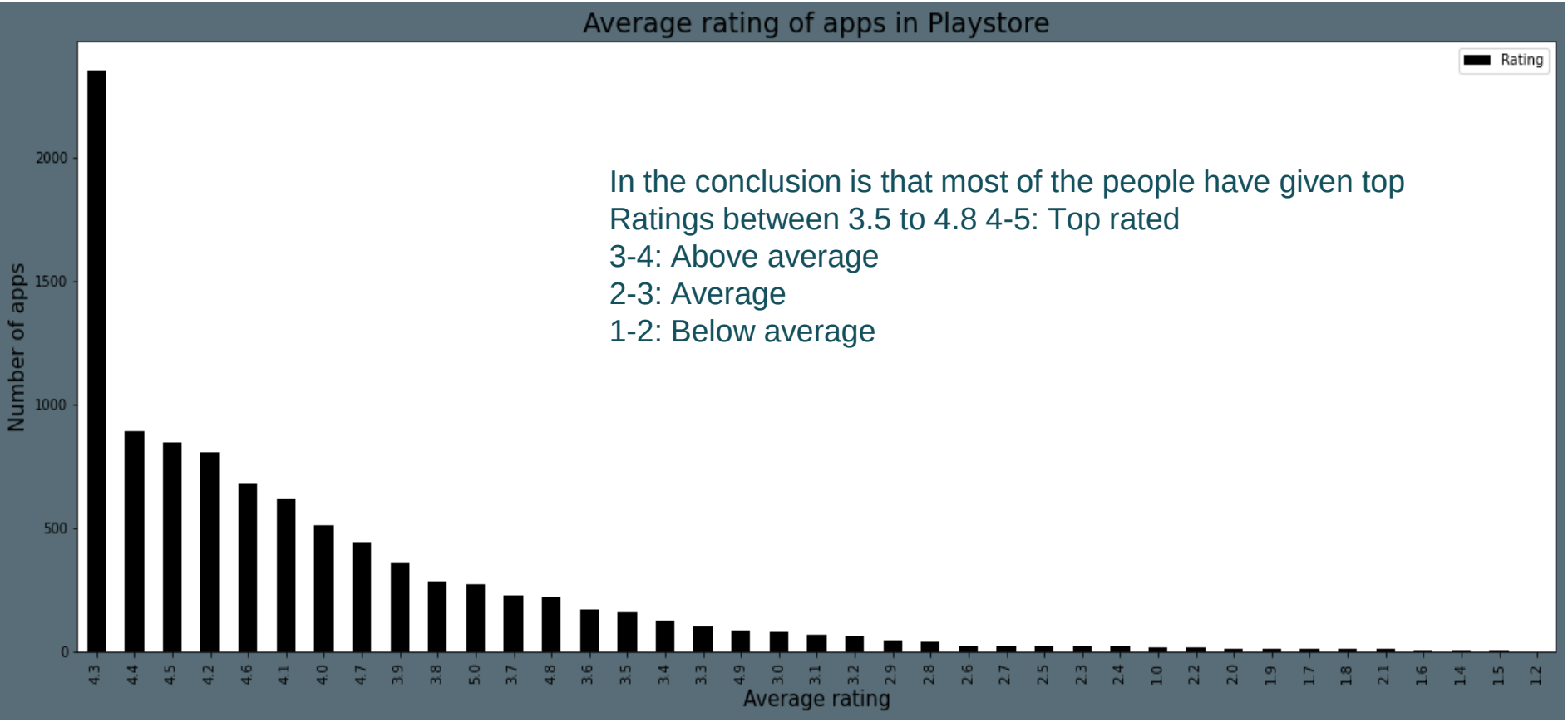


Highest Rating Category -:

- Above we saw every category has a slightly similar rating. There is not much difference.



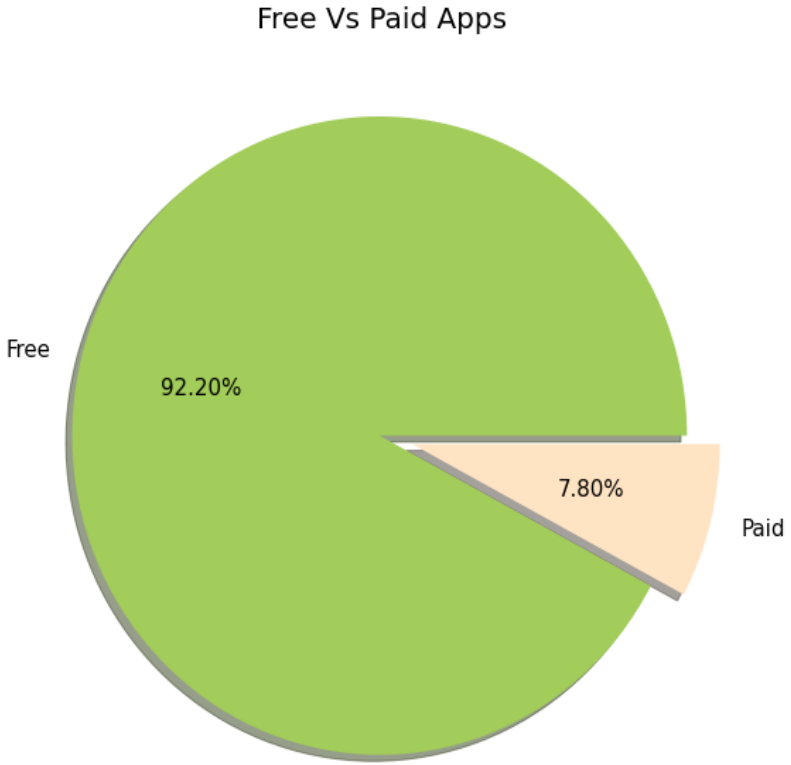
Average Ratings of Apps -:



In the conclusion is that most of the people have given top Ratings between 3.5 to 4.8 4-5: Top rated
3-4: Above average
2-3: Average
1-2: Below average

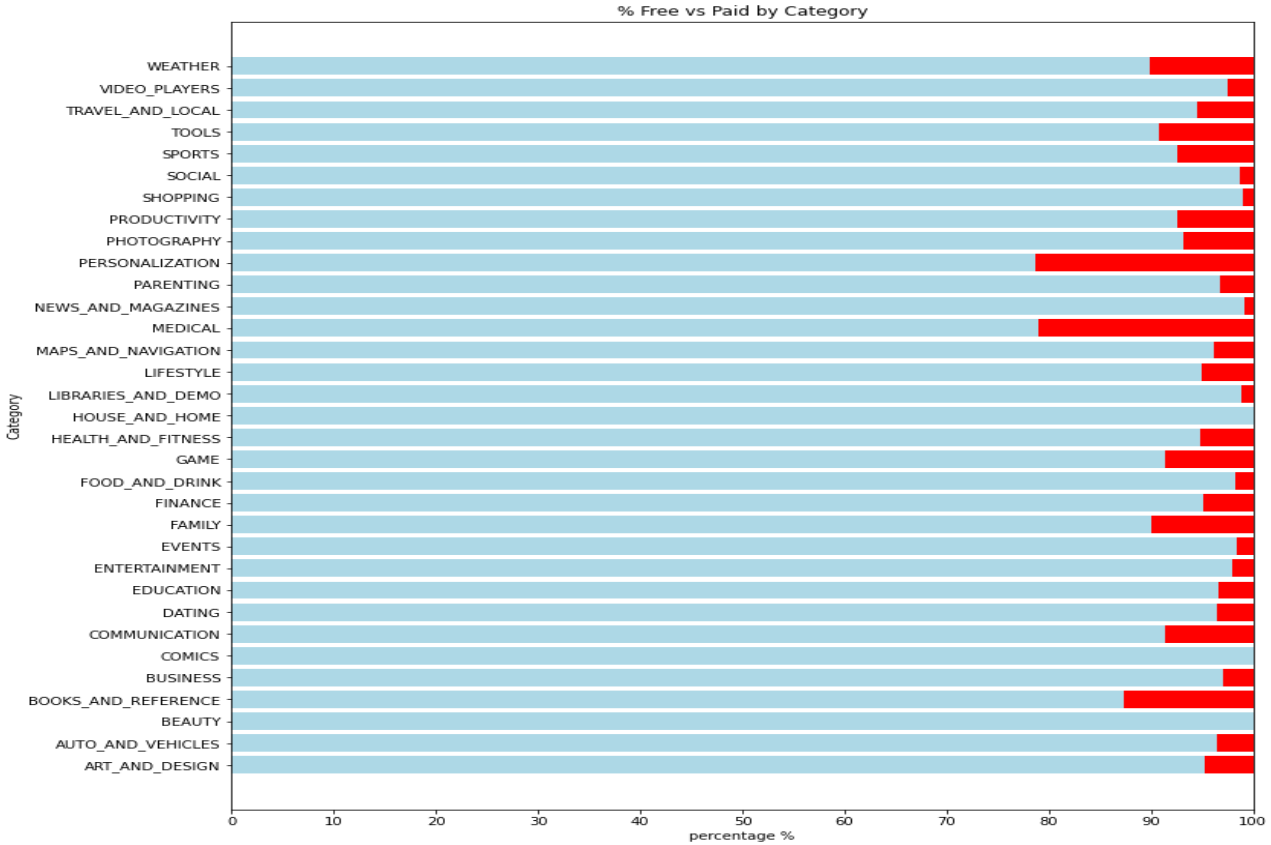
Percentage of Paid apps v/s Free apps

From the above graph we can see that 92% of apps in google play store are free and 8%are paid.
From the above data, we also see value counts of free and paid apps.

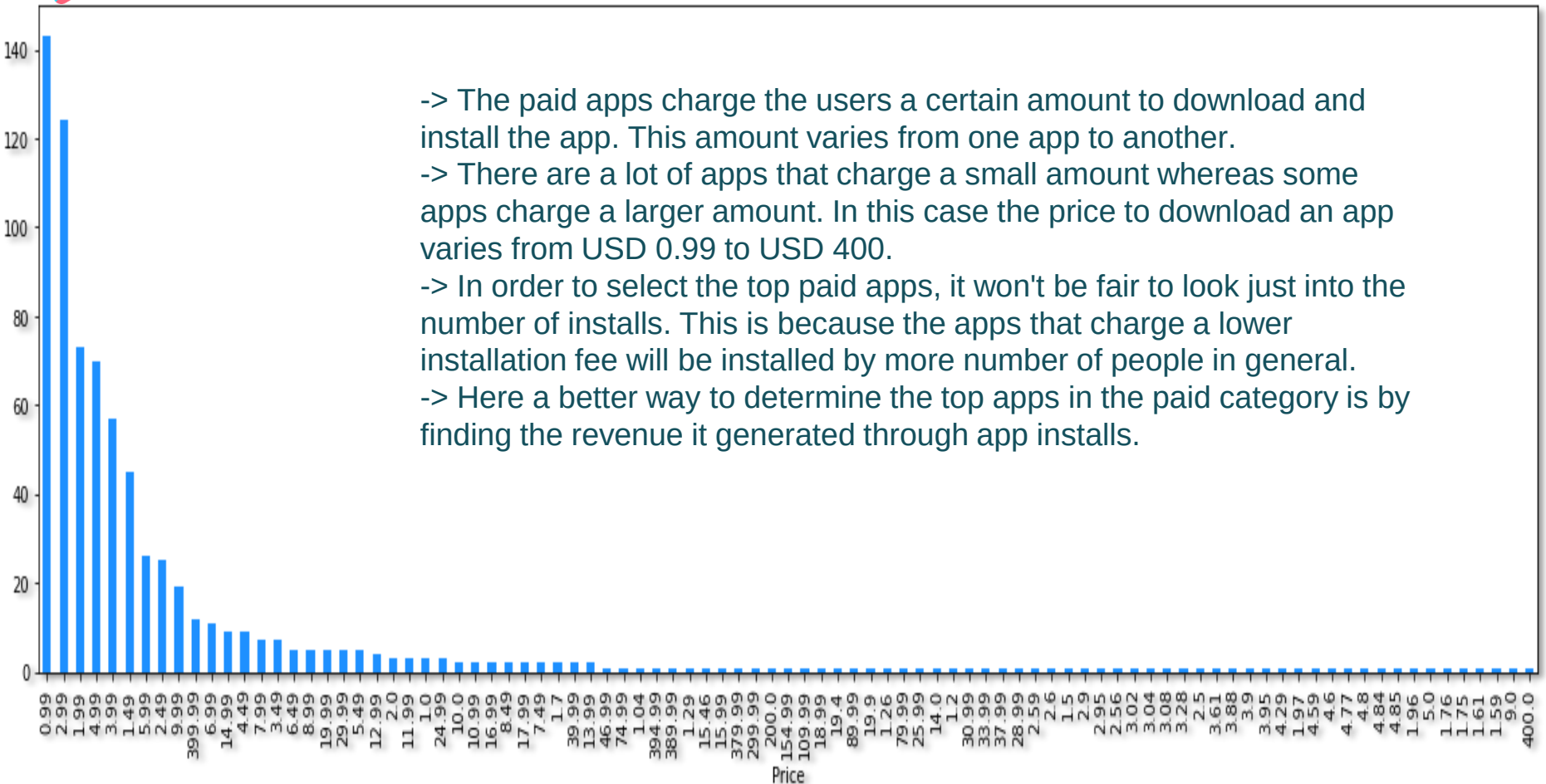


Free vs Paid category wise-:

Every category has more than 75 % of apps that are free for user

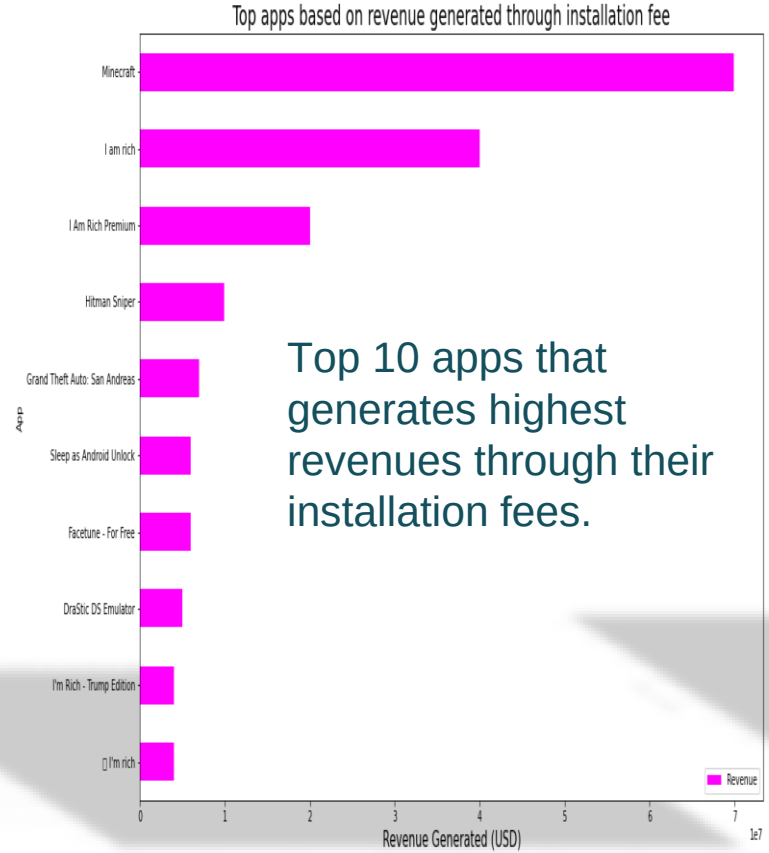


Paid Type Apps-:

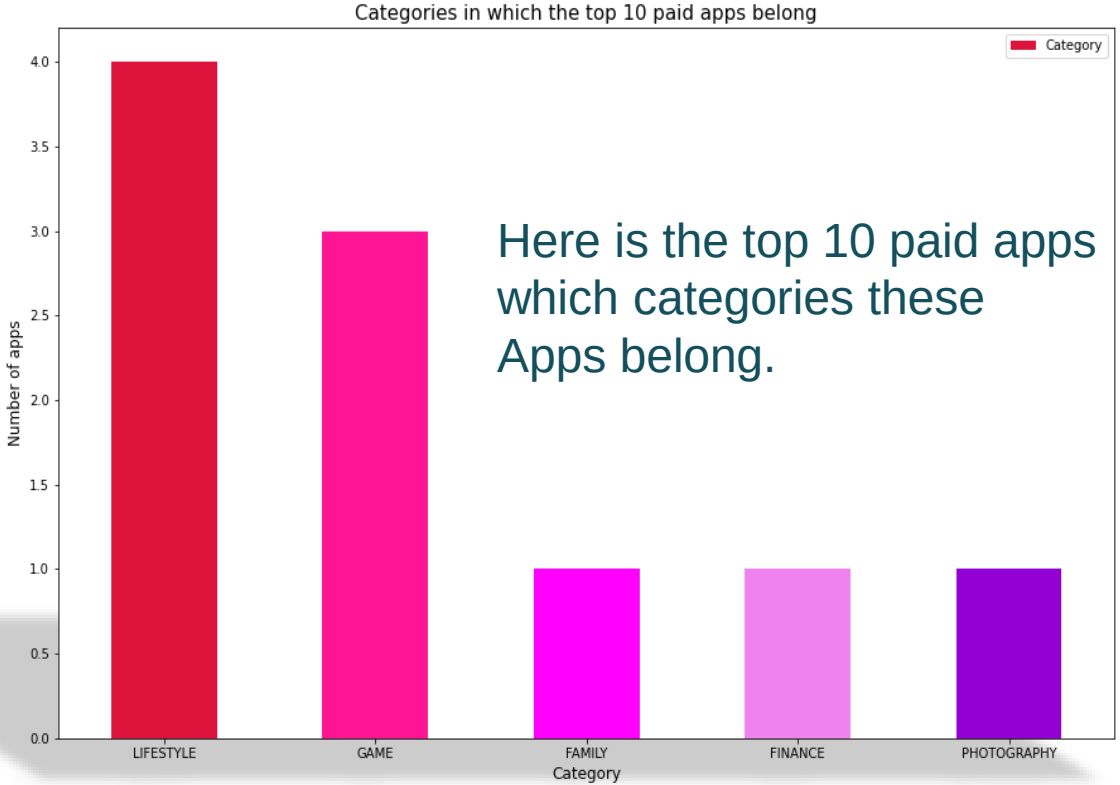


-> The paid apps charge the users a certain amount to download and install the app. This amount varies from one app to another.
-> There are a lot of apps that charge a small amount whereas some apps charge a larger amount. In this case the price to download an app varies from USD 0.99 to USD 400.
-> In order to select the top paid apps, it won't be fair to look just into the number of installs. This is because the apps that charge a lower installation fee will be installed by more number of people in general.
-> Here a better way to determine the top apps in the paid category is by finding the revenue it generated through app installs.

Revenue generated by App through the installs-:



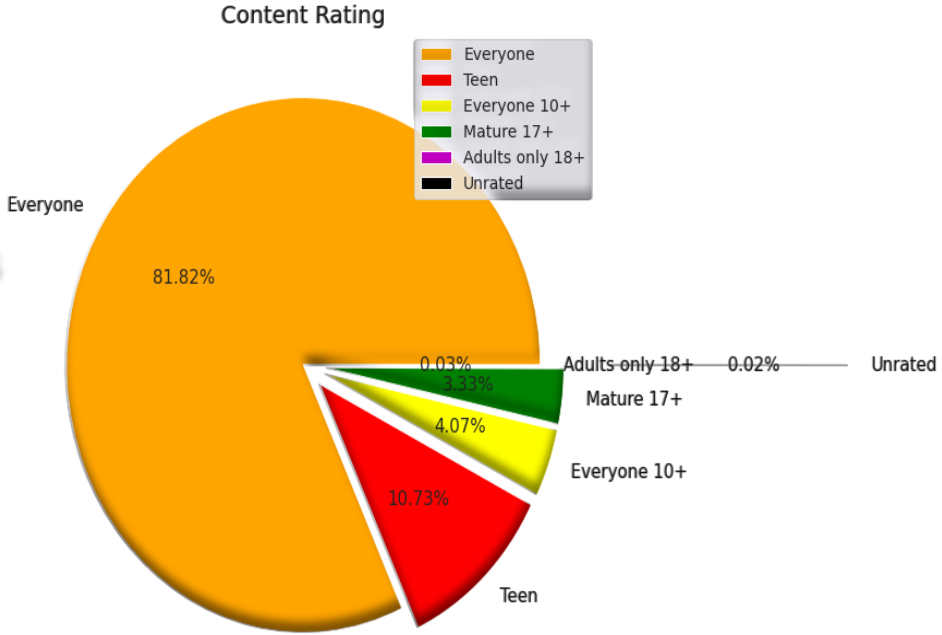
Top 10 apps that generates highest revenues through their installation fees.



Here is the top 10 paid apps which categories these Apps belong.

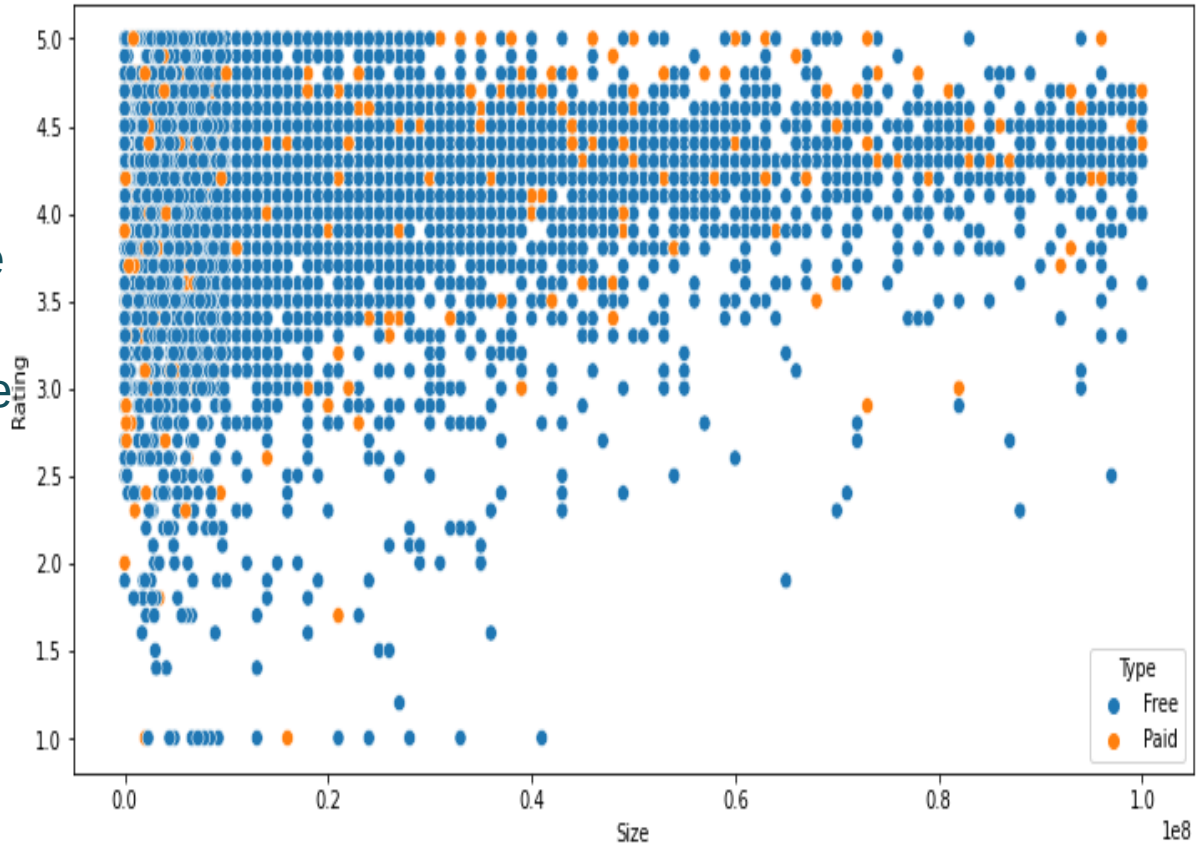
Content Rating

we can see above that 82% Ratings are given by Everyone. We can assume that least are having age restrictions to use some apps

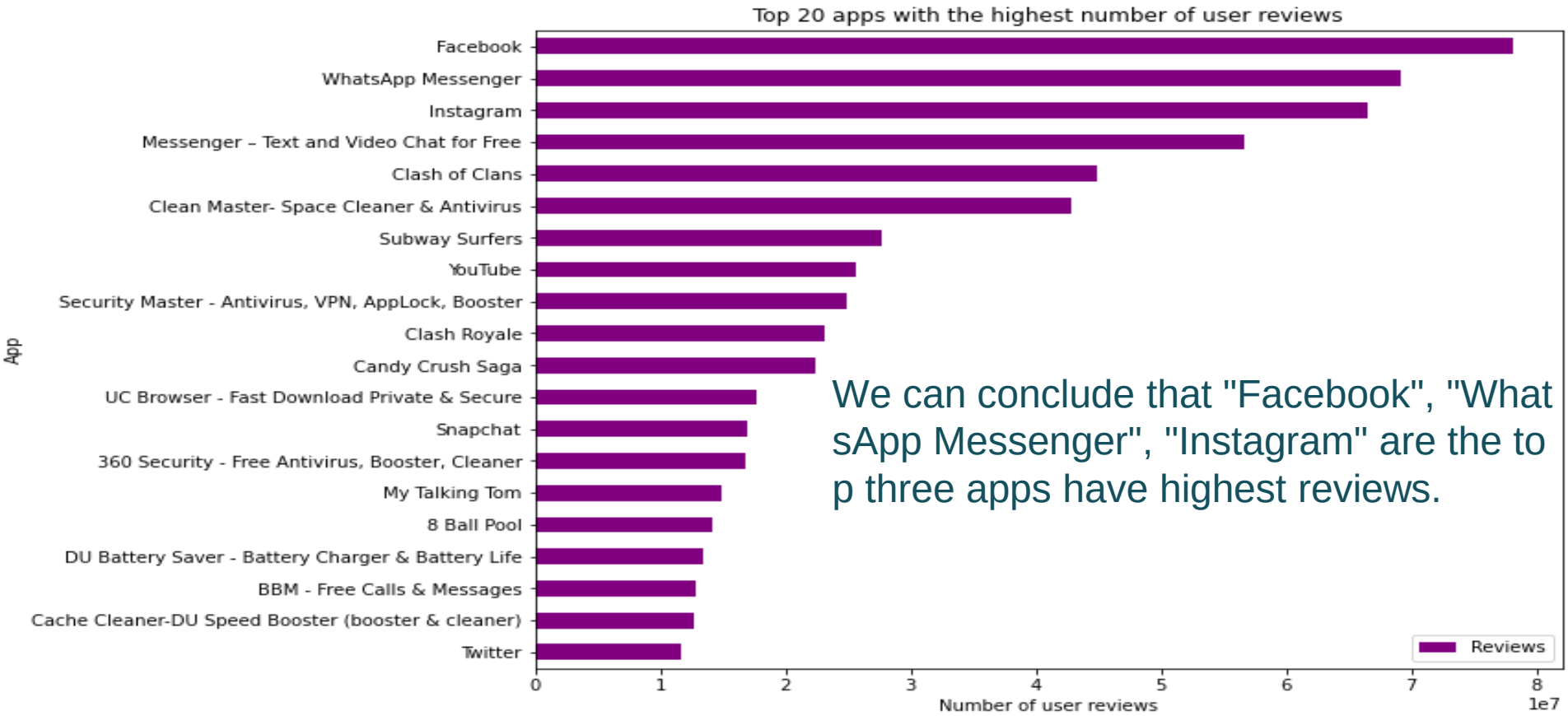


Distribution of apps in term of their rating, size and type.

From this scatter plot, we can imply that majority of the free apps are small in size and having high rating. While for paid apps, we have quite equal distribution in term on size and rating.



Highest Number of Reviews for App-:

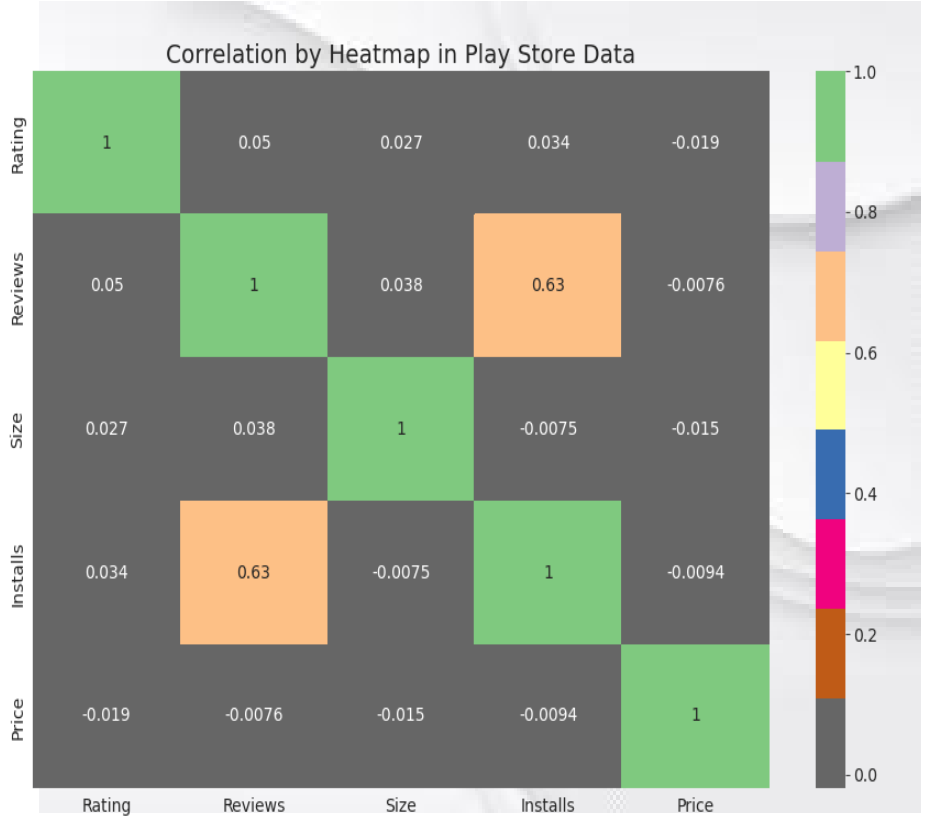


We can conclude that "Facebook", "What sApp Messenger", "Instagram" are the to p three apps have highest reviews.

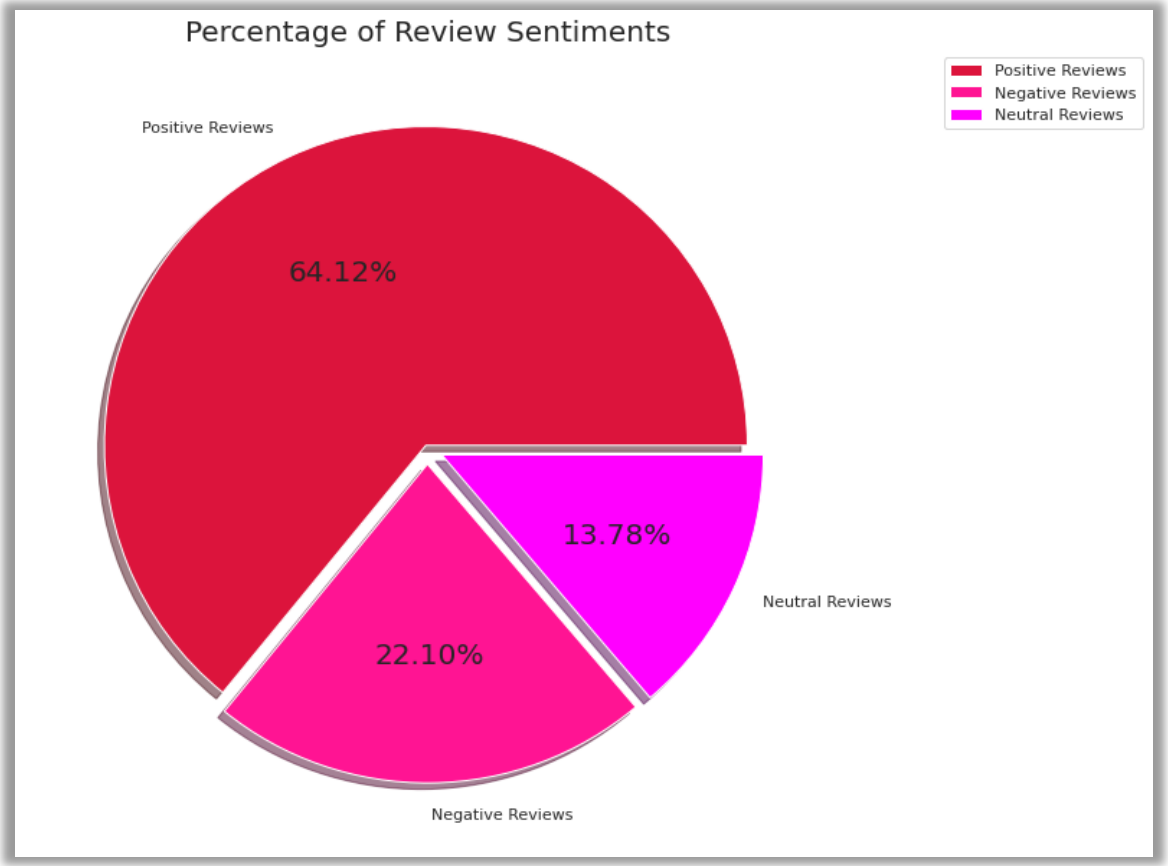
Correlation Heatmap

Clearly, we saw that reviews and installs are more correlated and the value is 0.64. 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.



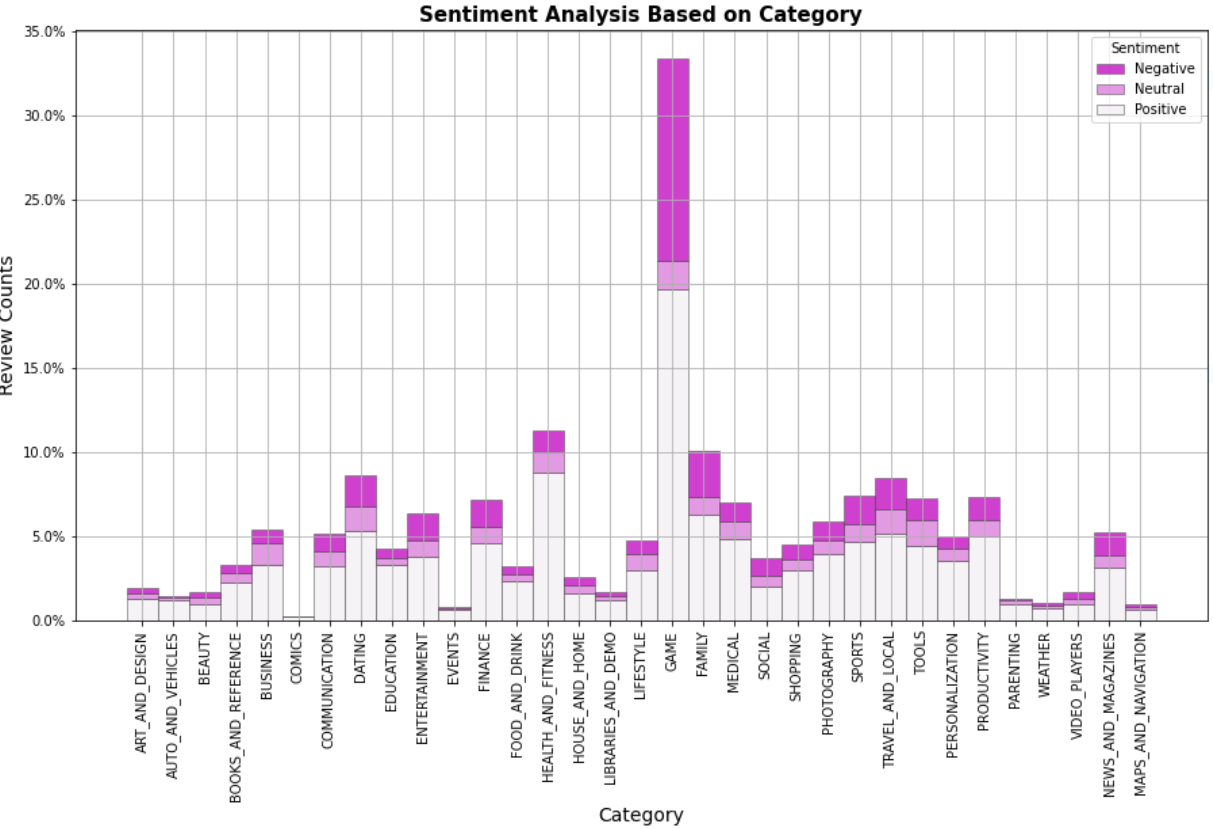
Percentage of Review Sentiments



The number of **Unique** Apps from Play store and User reviews merged dataset are **816.**

From Sentiment column**, 64%** are **Positive, 22%** are **Negative** and **14%** are **Neutral** values.

Sentiment Analysis Based on Category

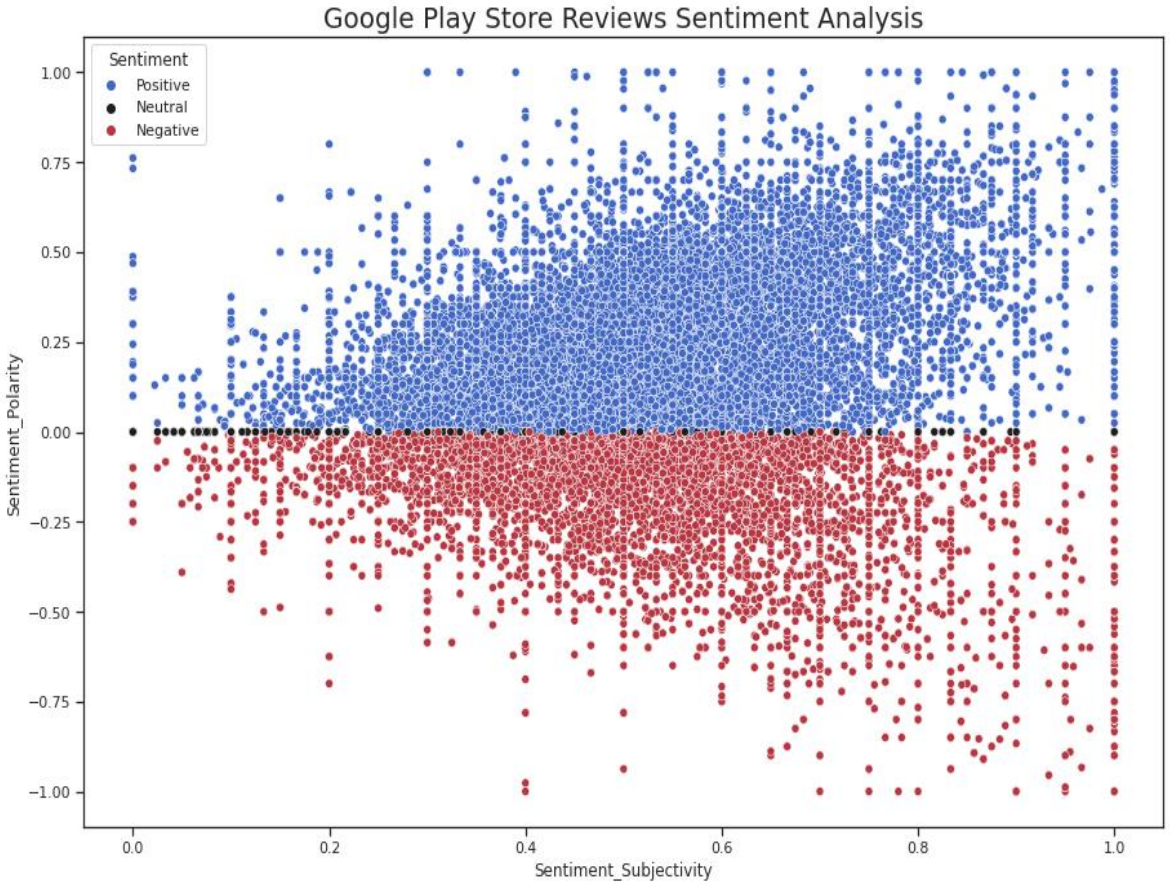


Here we notice that most of the reaction are from 'Game' category and very less from 'Comics', 'Events', 'Maps And Navigation' and 'Weather'.

-This shows that people take much interest in Game category app as compare to other apps.

Proportionality- Sentiment_subjectivity vs sentiment_polarity

From the scatter plot it can be concluded that sentiment subjectivity is not always proportional to sentiment polarity but in maximum number of case, shows a proportional behavior, when variance is too high or low



Challenges Faced -:

- Reading the dataset and comprehending the problem statement.
- Examining the business KPIs for app development and devising a solution to the problem.
- Handling the error, duplicate and NaN values in the dataset.
- Designing multiple visualizations to summarize the information in the dataset and successfully communicate the results and trends to the reader.

Conclusion’s

~ 92.19% apps are Free and 7.81% apps are paid in type.

~ 81.80% apps have Everyone content rating.

~ Events category has a highest mean rating of 4.39 and Dating category has lowest 4.05 rating.

~ Family, Game and Tools are top three categories having 1906, 926 and 829 app count.

~ Most competitive category: Family

~ Category with the highest number of installs: Game

~ Tools, Entertainment, Education, Business and Medical are top Genres.

~ 8783 Apps are having size less than or equal to 50 MB.

~ 7749 Apps has rating more than 4.0 including both type of app.

~ Overall sentiment count of merged dataset in which Positive sentiment count is 64%, Negative 22% and Neutral 14%.

Conclusion’s

~ It's good to develop a **Free type** app and having a content rating for **Everyone.**

~ Percentage of apps that are top rated = **81.80%**

~ There are **20** free apps that have been installed over a **billion** times

~ **Minecraft** is the only app in the paid category with over **10M** installs, and also has produced the most revenue only from installation fee.

~ Price, Rating, Size **has no or very less correlation** with **Sentiment Polarity**.

~ The median size of the apps in the play store is 12 MB

~ The apps whose size **varies with device** has the highest number average app installs.

~ The apps whose size is **greater than 90 MB** has the highest number of average user reviews, ie, 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.


Thank You