Capstone Project

Play Store App Review Analysis

By
Vaithyanathan.M (Arunai Engineering College)

WHY ANALYZE THE GOOGLE PLAY STORE?

Mobile App Market is set to grow 20% by 2023

Android Apps comprise 90% of the Mobile App Market

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

- Android is the most popular operating system in the world, with over 2.5 billion active users spanning over 190 countries.
- Google Play was launched on March 6, 2012, bringing together Android Market making a shift in Google's digital distribution strategy .
- Android is the dominant mobile operating system today, more than 85% of all mobile devices running Google's OS. The Google Play Store is the largest and most popular Android app store.
- There are more than 3.5 million apps found on Google Play Store.
- The Play Store apps data has enormous potential to drive app-making businesses to success.

Actionable insights can be drawn for developers to work and capture the Android market. The main goal for our project is

1) The purpose of our project is to gather and analyze stored information on apps in the Google Play Store in order to provide insights on app features and the current state of the Android app market.
2) The objective of the project to Explore and analyze the data to discover key factors responsible for app engagement and success.

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.

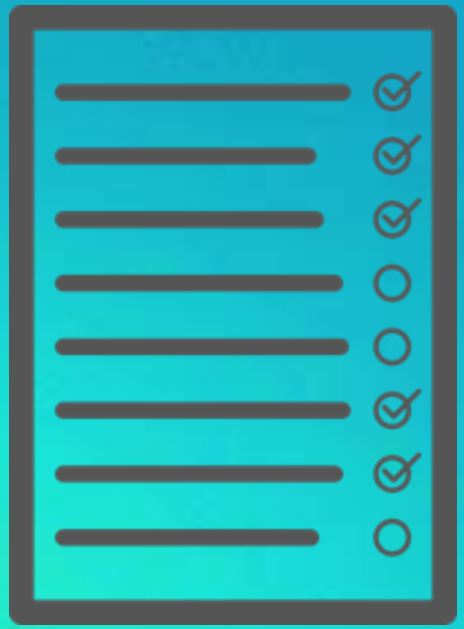
So, what factors influence an app's success?

An app is said to be successful if it has:

- A high average user rating
- A good number of positive reviews
- A good number of monthly average users
- High revenue per customer and so on.

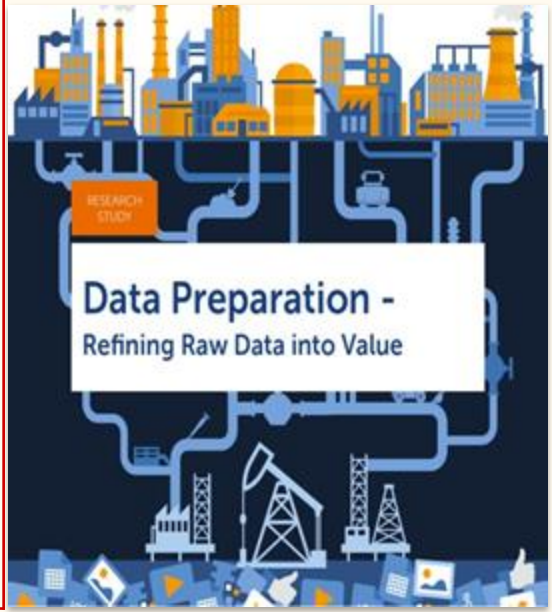
Agenda

- Introduction
- 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



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 app is supported.

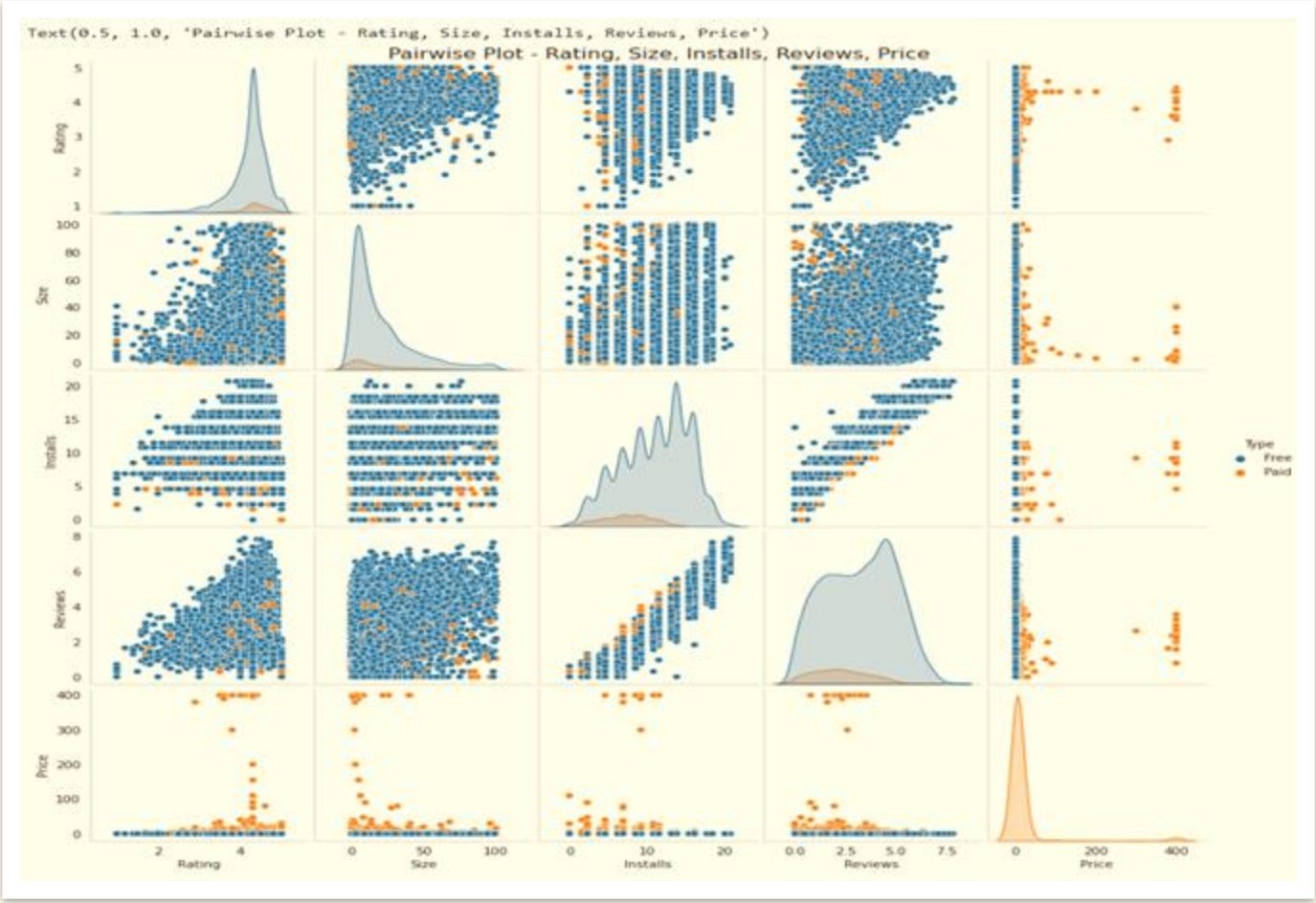
Attributes in User reviews

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

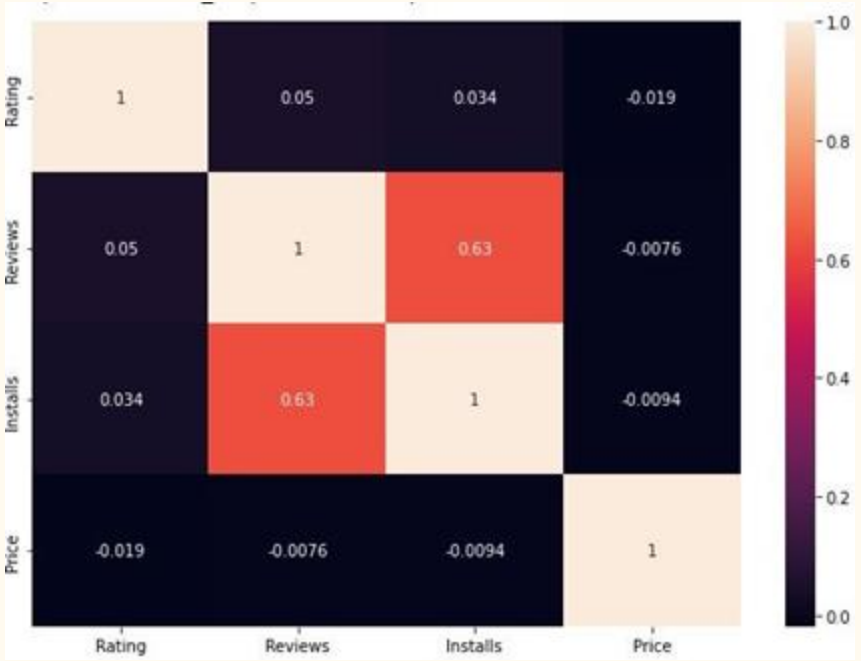
Data Cleaning	Data Exploration	Predictive Modeling
		
Understand the structure of the dataset and clean data before analysis	Uncover initial patterns, characteristics, and points of interest using visual exploration	Formulate a statistical model to forecast an outcome using relevant predictors

Pairwise Plot- Ratings, Size, Installs, Reviews, Price



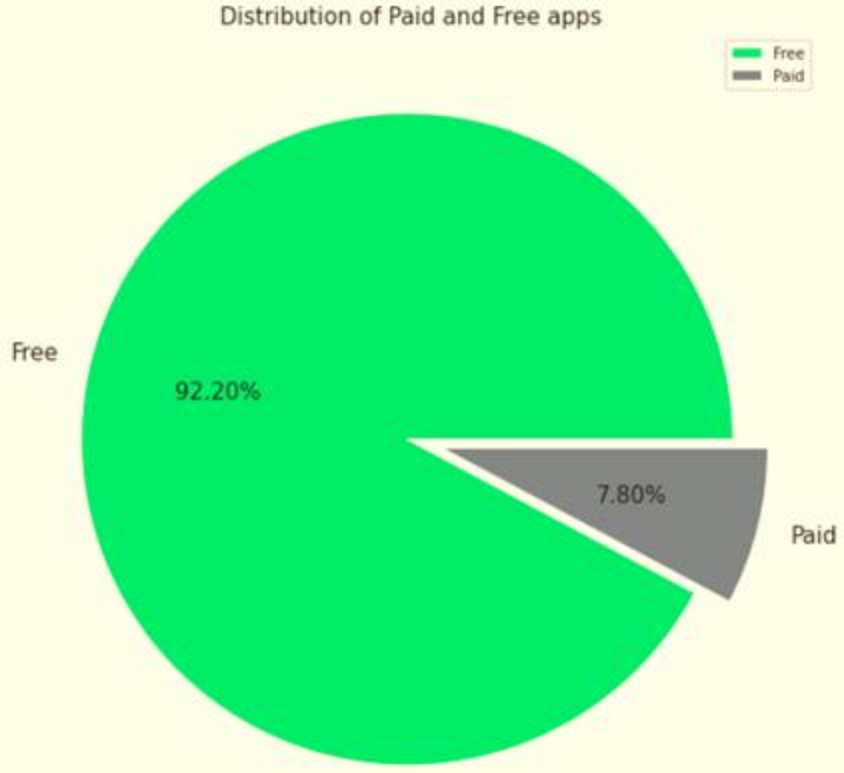
Correlation Heatmap

- There is a strong **positive** correlation between the **Reviews** and **Installs**.
- The Price is slightly **negatively** correlated with the **Rating**, **Reviews**, and **Installs**.
- The **Rating** is slightly **positively** correlated with the **Installs** and **Reviews**.



Percentage of Paid apps v/s Free apps

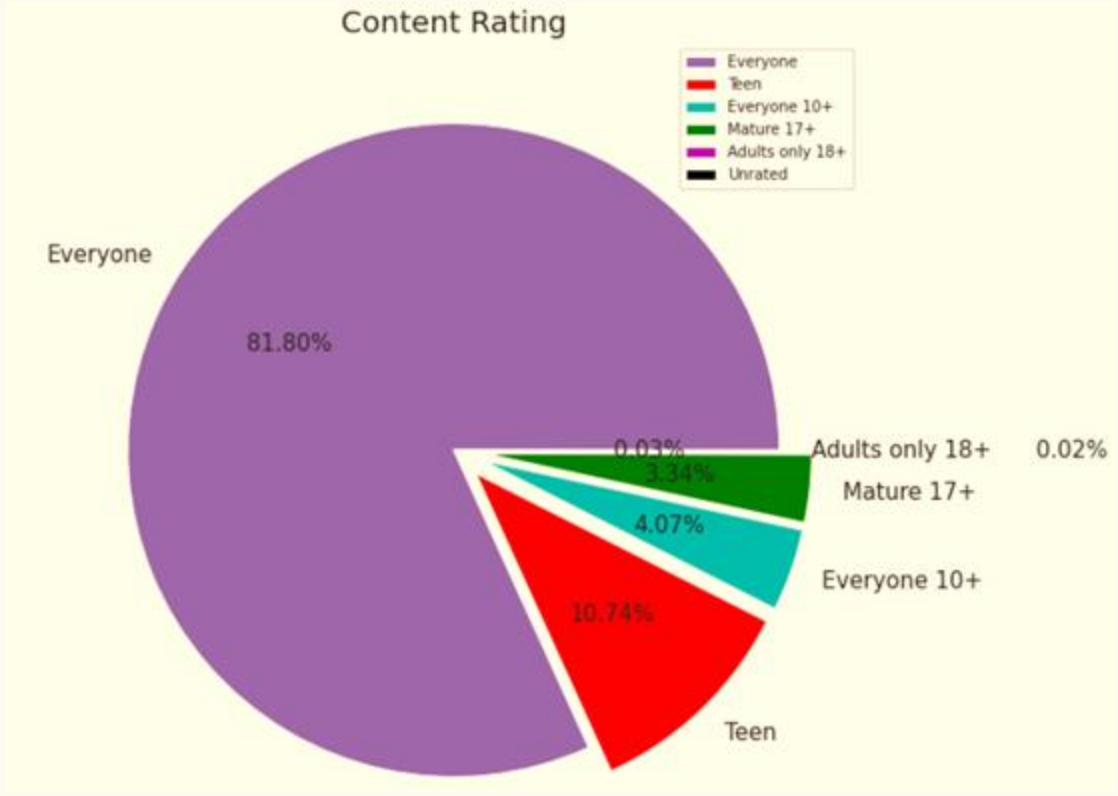
We Observed that **92.20% of Apps are free** and only **7.80% of Apps are paid** in Play store.



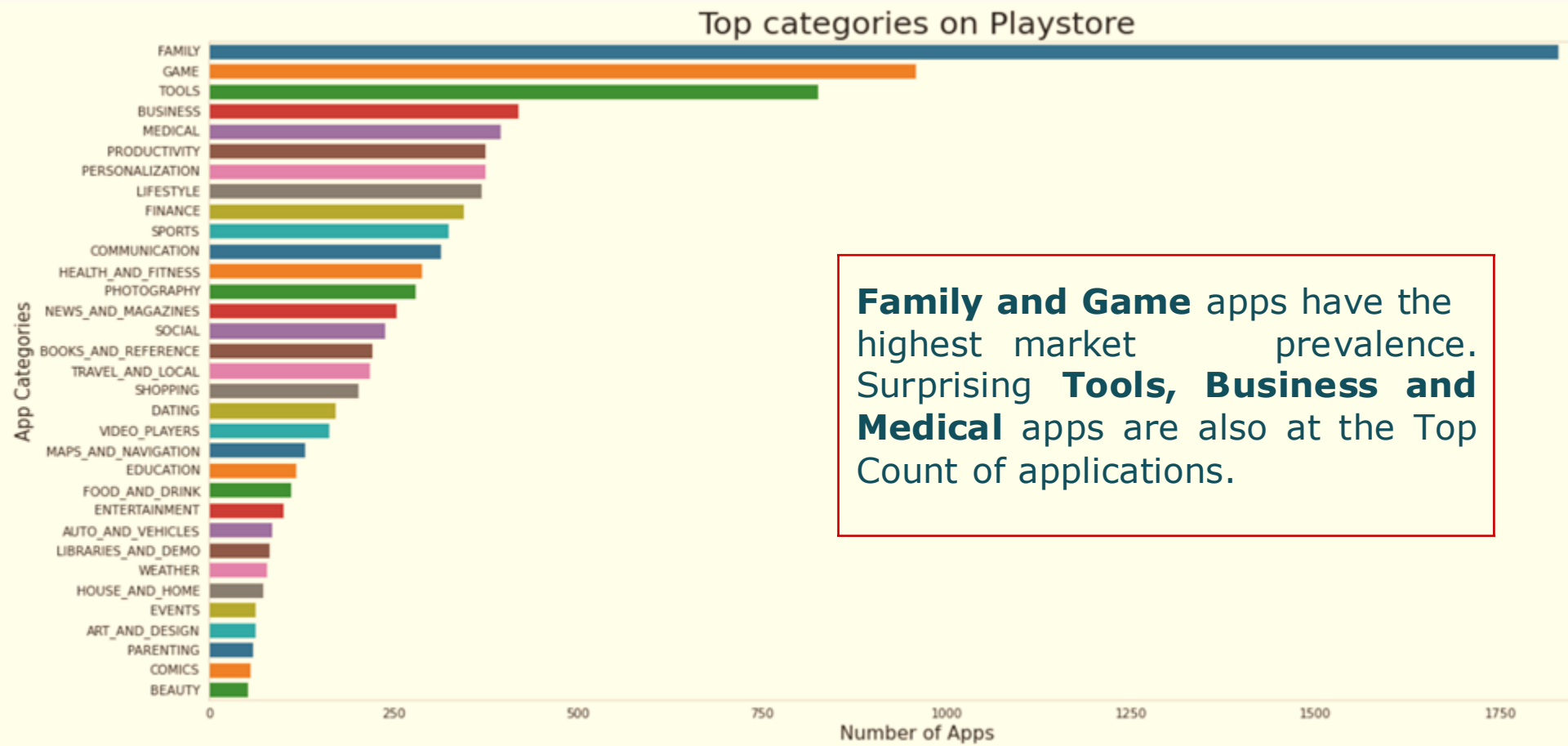
Content Rating

From the above plot we can see that Everyone category having majority of apps count.

A majority of the apps **(81.80%)** in the play store are can be used by everyone. The remaining apps have various age restrictions to use it.

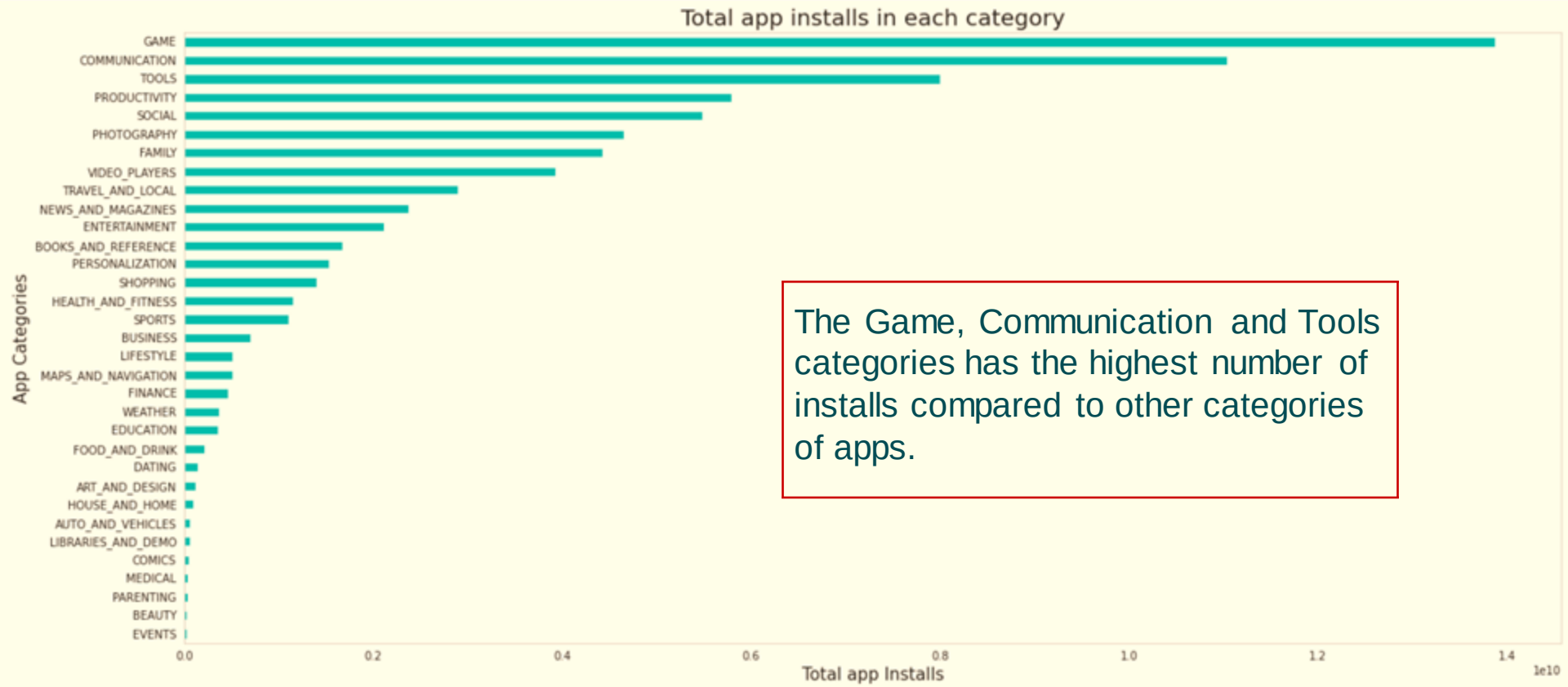


Count of Applications in each category



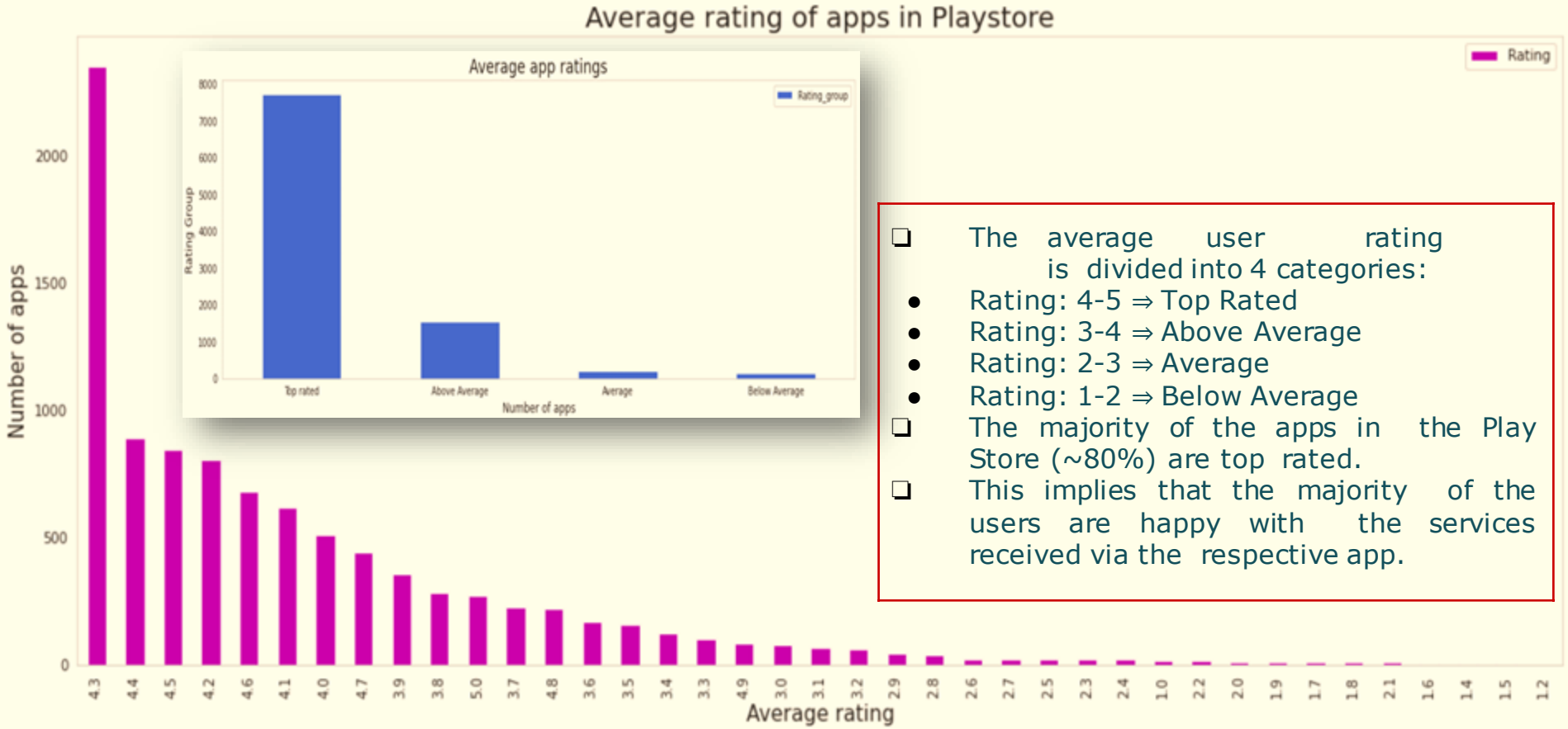
Family and Game apps have the highest market prevalence. Surprising **Tools, Business and Medical** apps are also at the Top Count of applications.

Category App's have most number of installs



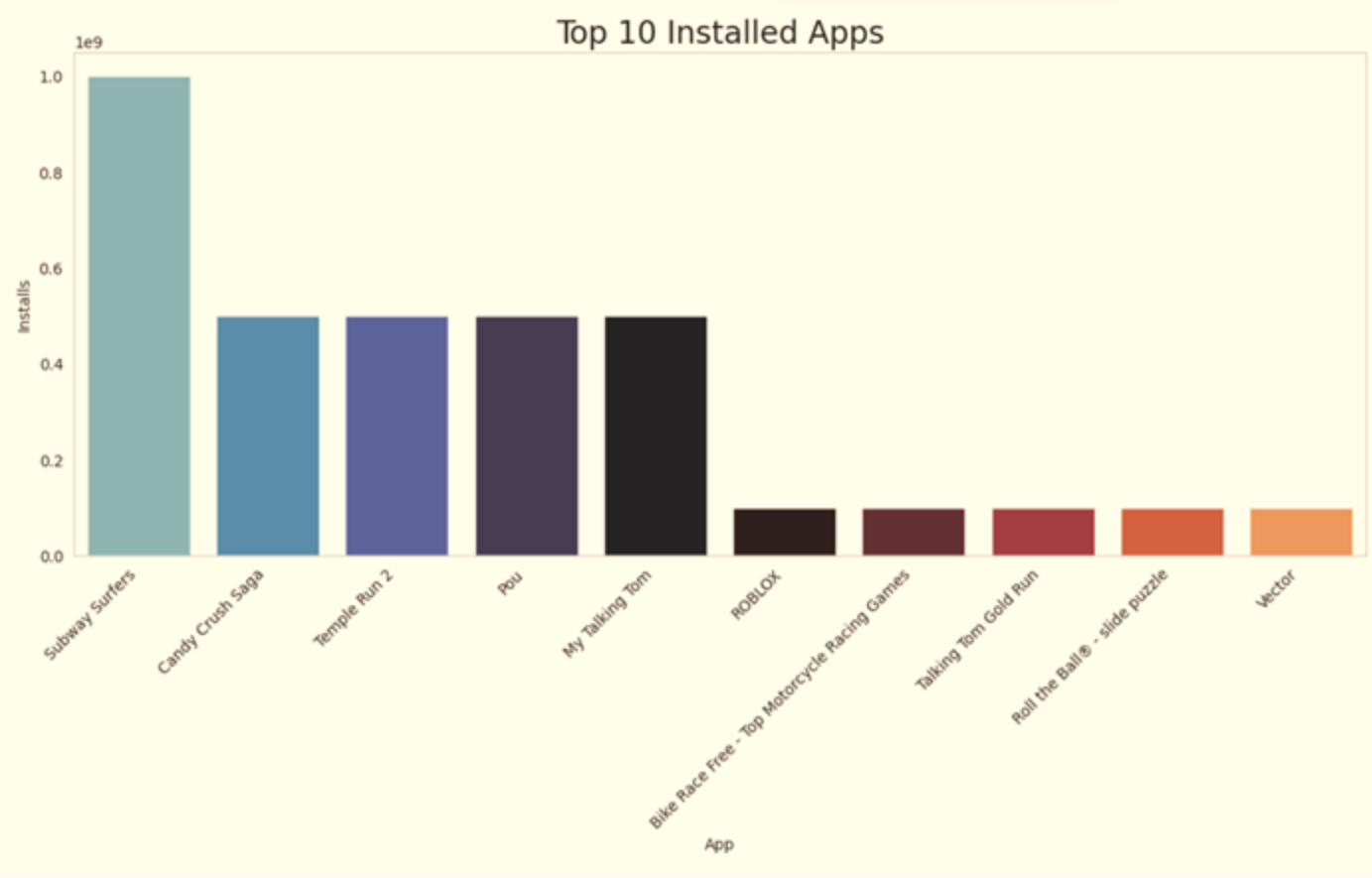
The Game, Communication and Tools categories has the highest number of installs compared to other categories of apps.

Average rating of the apps



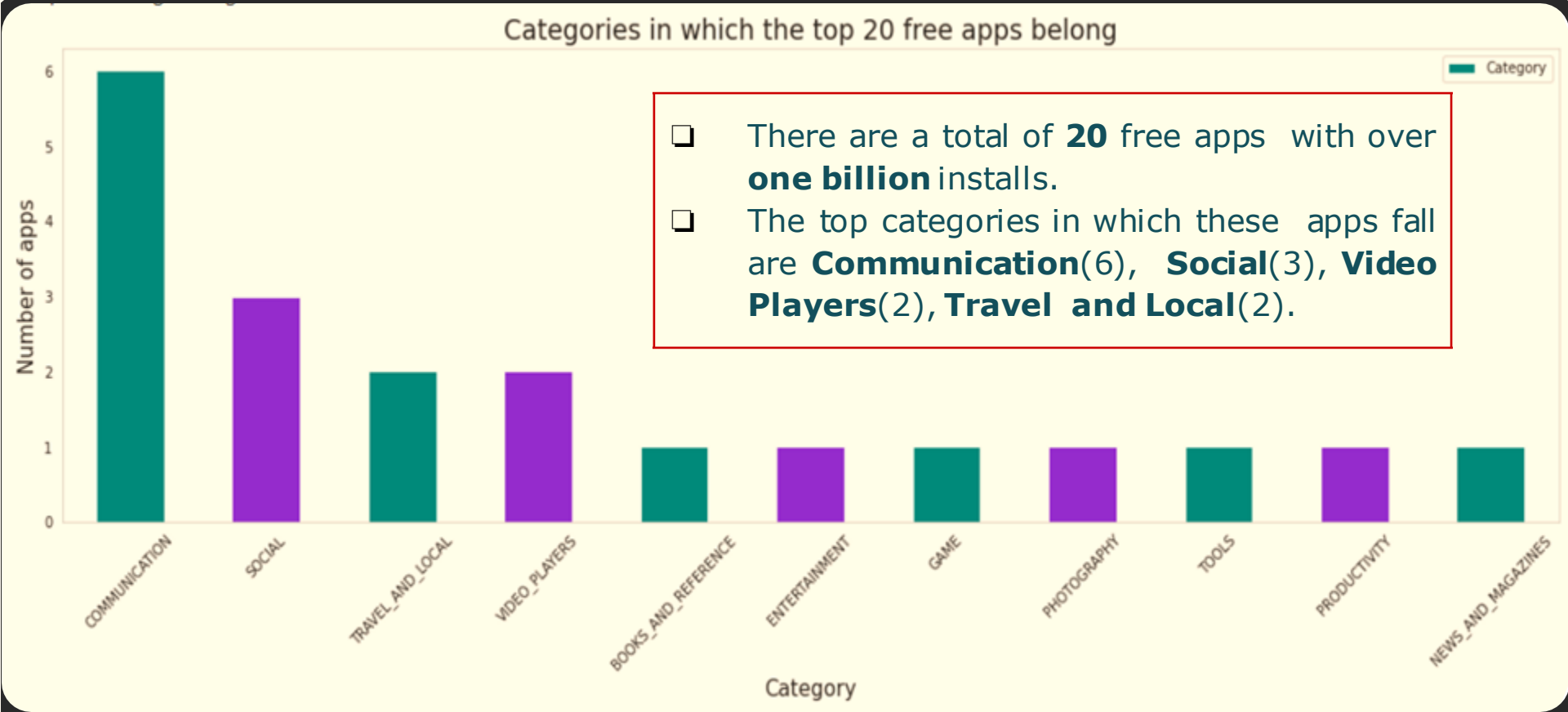
- The average user rating is divided into 4 categories:
  - Rating: 4-5 ⇒ Top Rated
  - Rating: 3-4 ⇒ Above Average
  - Rating: 2-3 ⇒ Average
  - Rating: 1-2 ⇒ Below Average
- The majority of the apps in the Play Store (~80%) are top rated.
- This implies that the majority of the users are happy with the services received via the respective app.

Top 10 installed apps in any category



This graph shows the top installed apps in the **'Games'** category. Further looking into the play store reveals that these apps are light, casual, single player games.

Top Free Apps

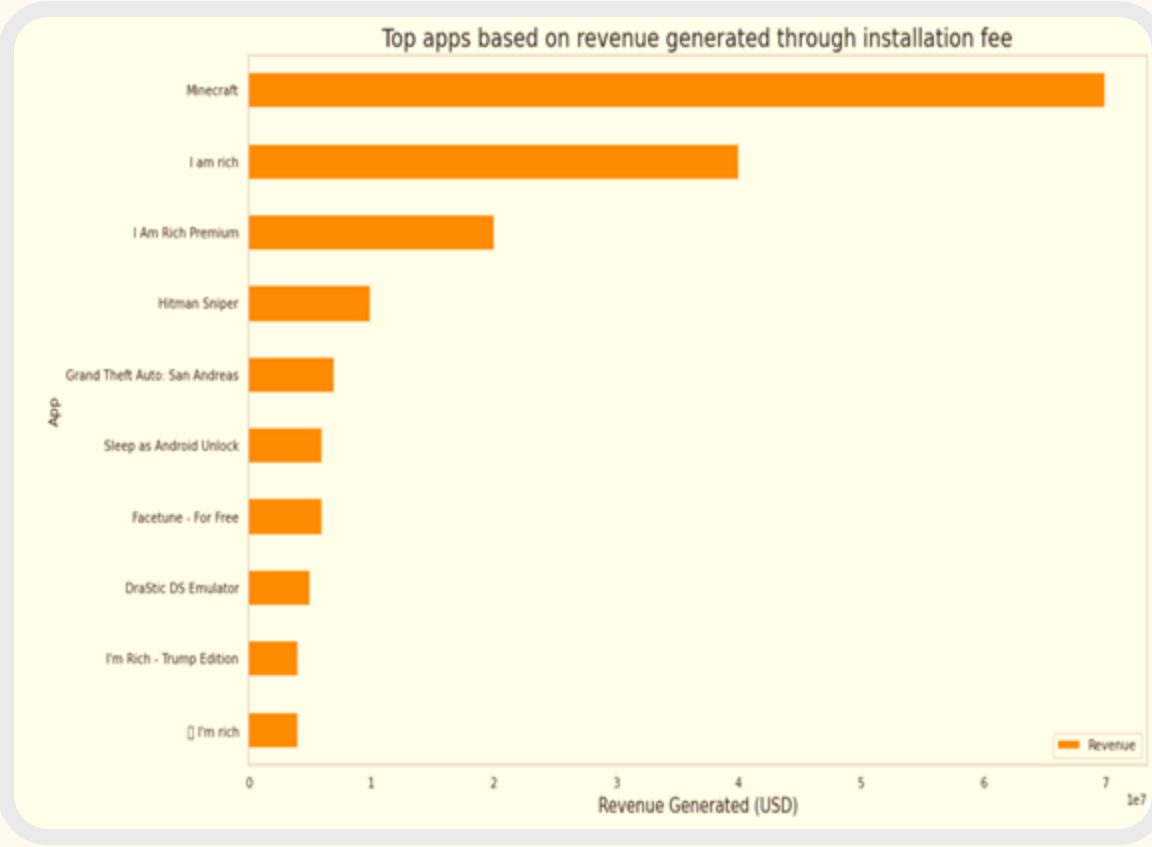


- There are a total of **20** free apps with over **one billion** installs.
- The top categories in which these apps fall are **Communication**(6), **Social**(3), **Video Players**(2), **Travel and Local**(2).

Top Paid Apps Based on Revenue Generated

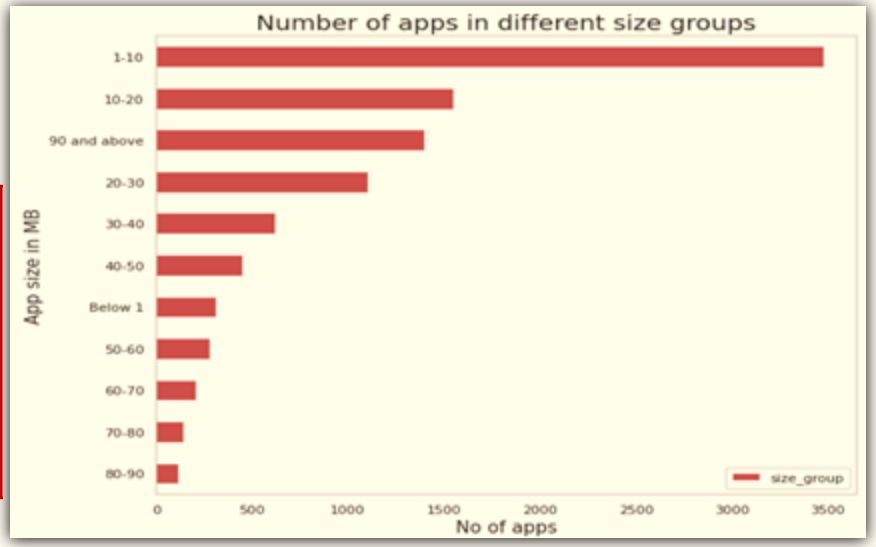
- Revenue generated is given by the formula:

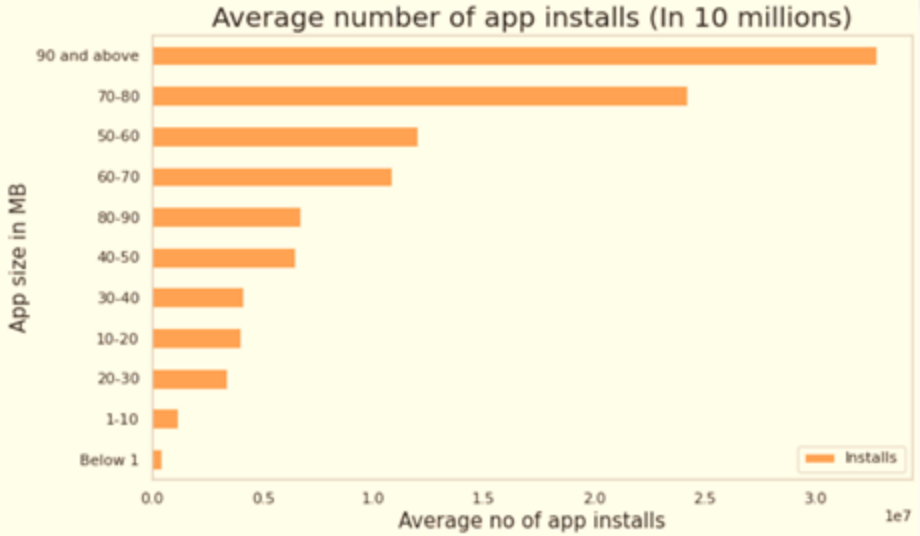
  Revenue = Installs * Price
- Note that in this case, revenue refers to the money earned only from paid app installs.
- The top categories in which these apps fall are **Lifestyle**(5), **Family**(5), **and Game**(4).
- **Minecraft**, **I am rich**, and **I am rich premium** are the top paid apps based on revenue generated.

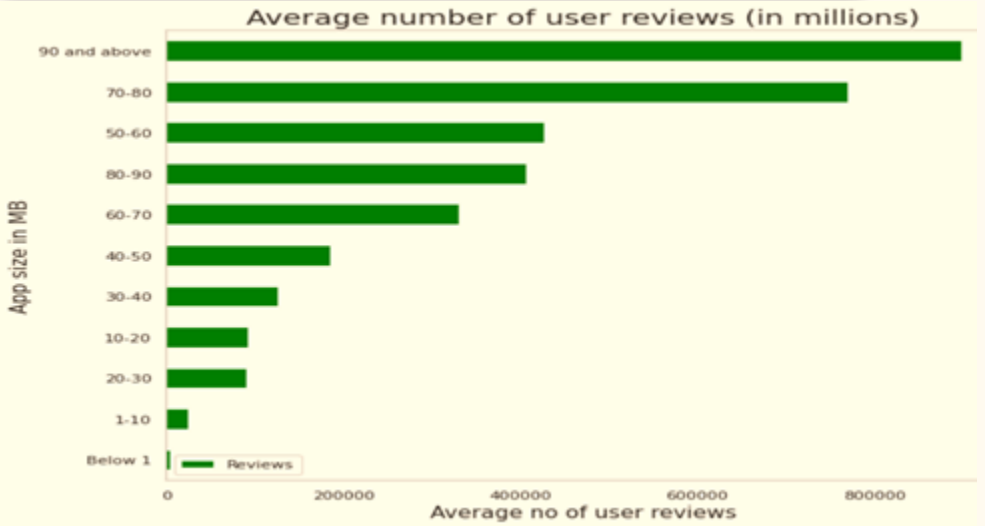


App Size Analysis

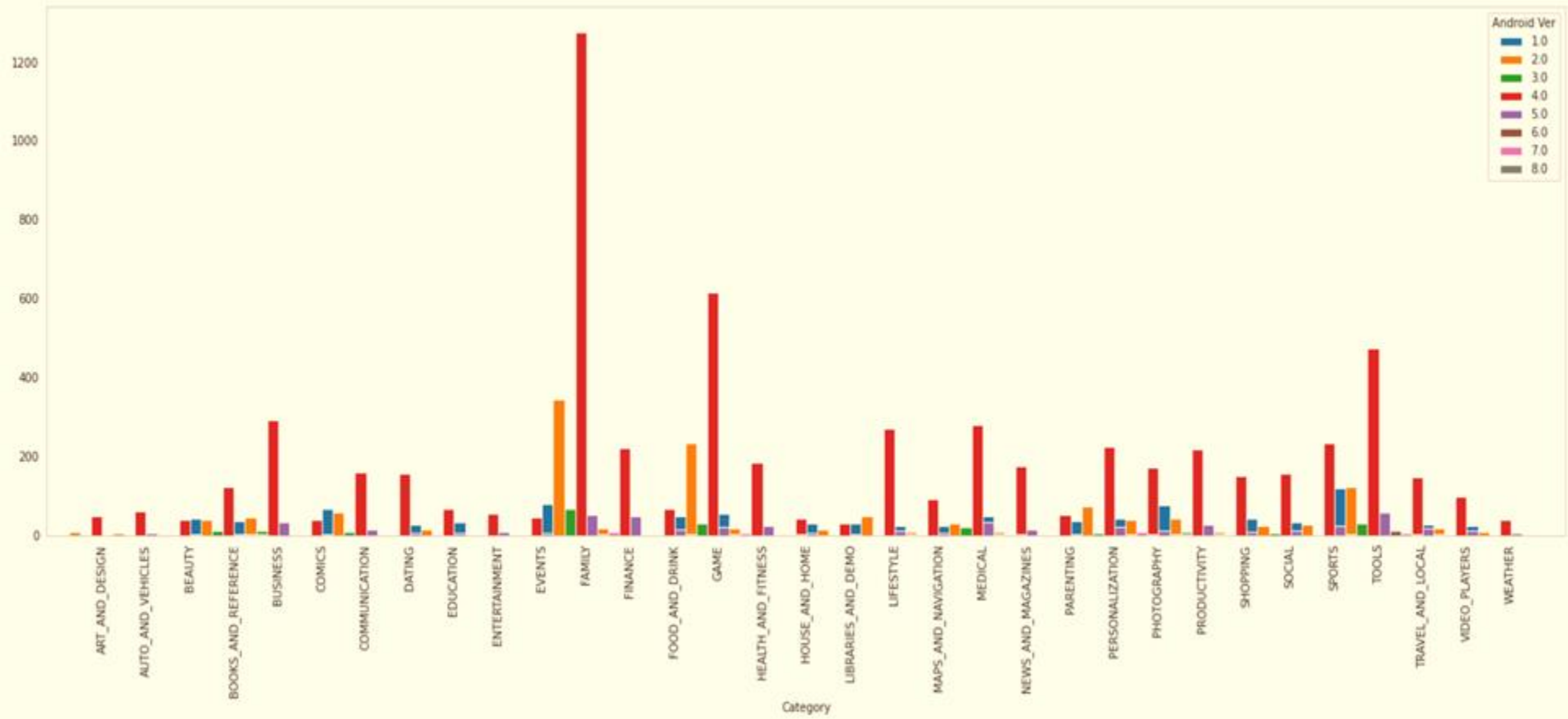
- The apps are categorized based on its size between ~0 to 100 MB in the intervals of 10 MB each.
- The total number of apps in each size category indicates the **competition**.
- Average number of **user reviews** and **average app installs** in each size category indicates the **popularity** of the respective app.



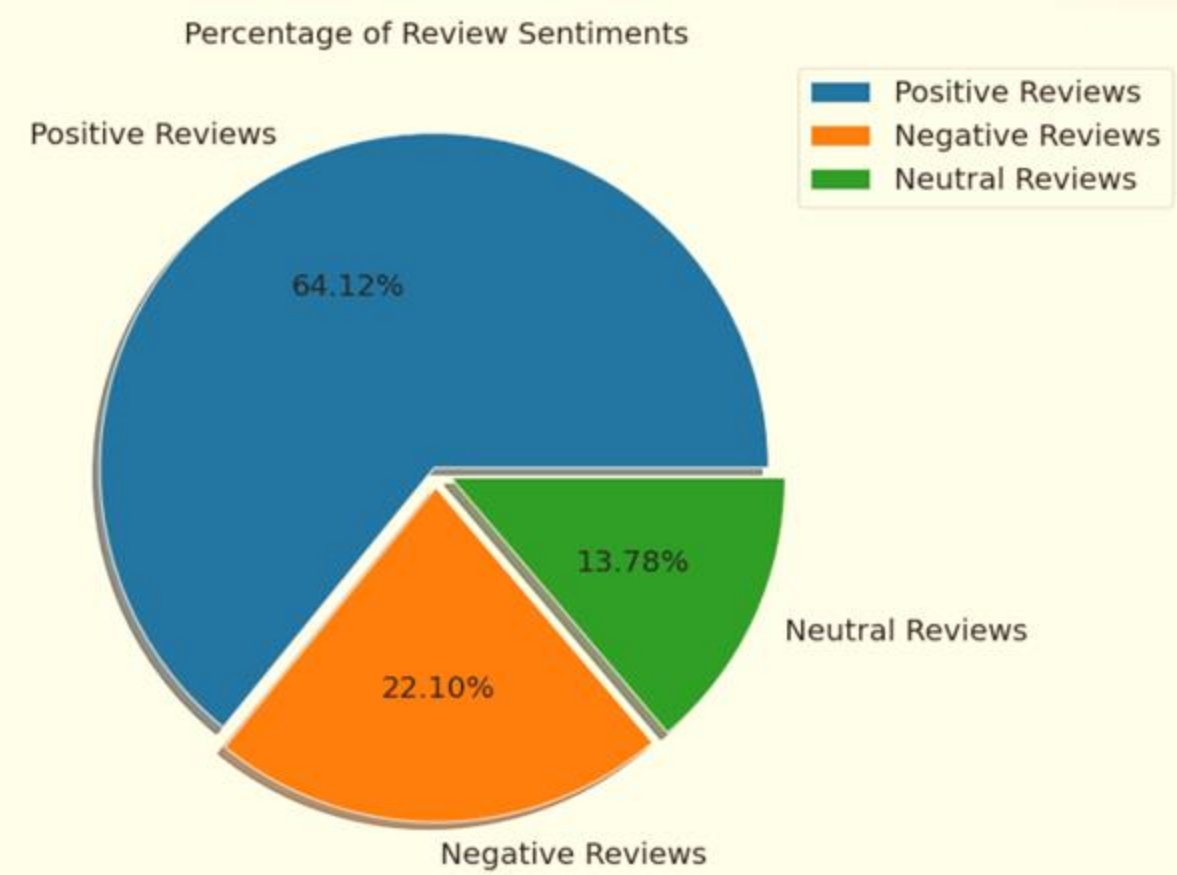




Android version based on each category



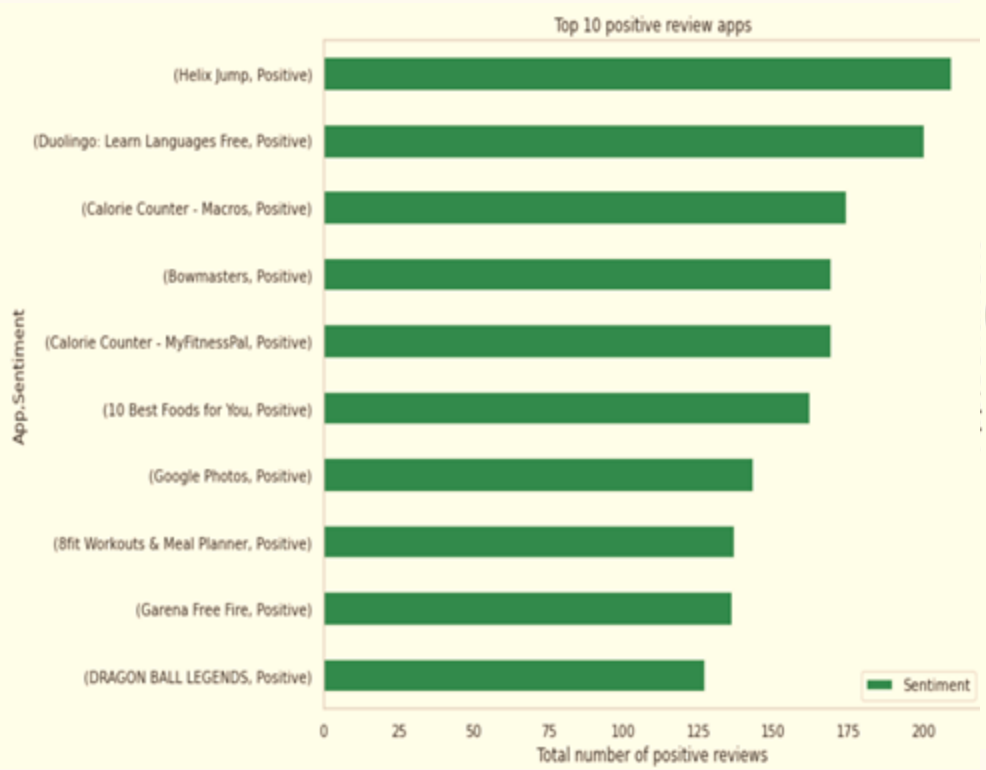
Percentage of Review Sentiments

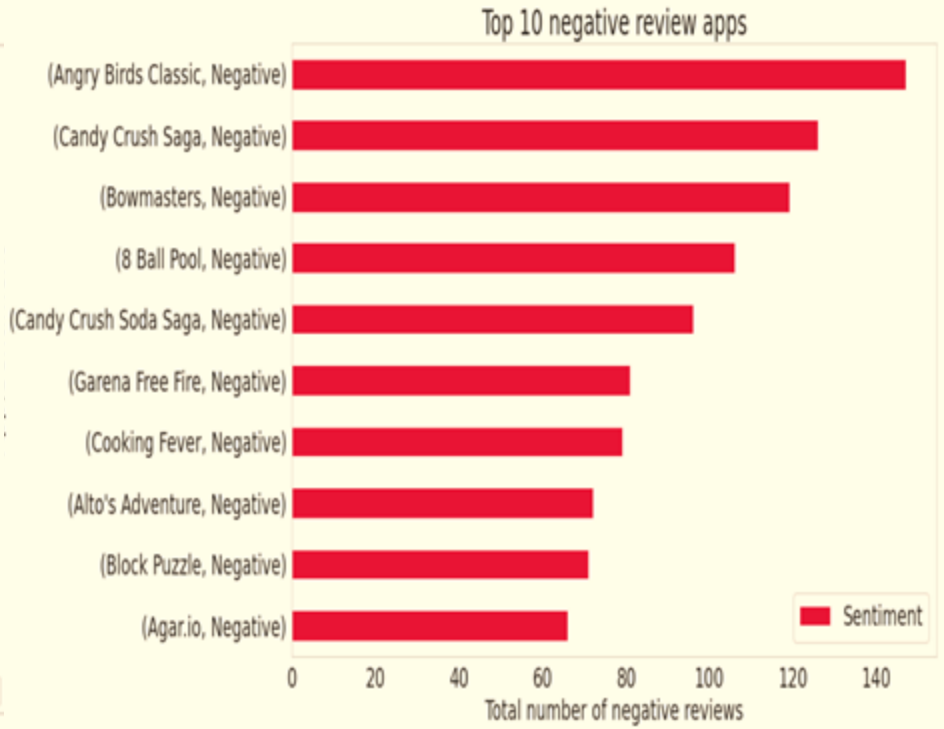


The number of **Unique** Apps from Play store and User reviews merged dataset are **816.**

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

Positive and Negative Reviews



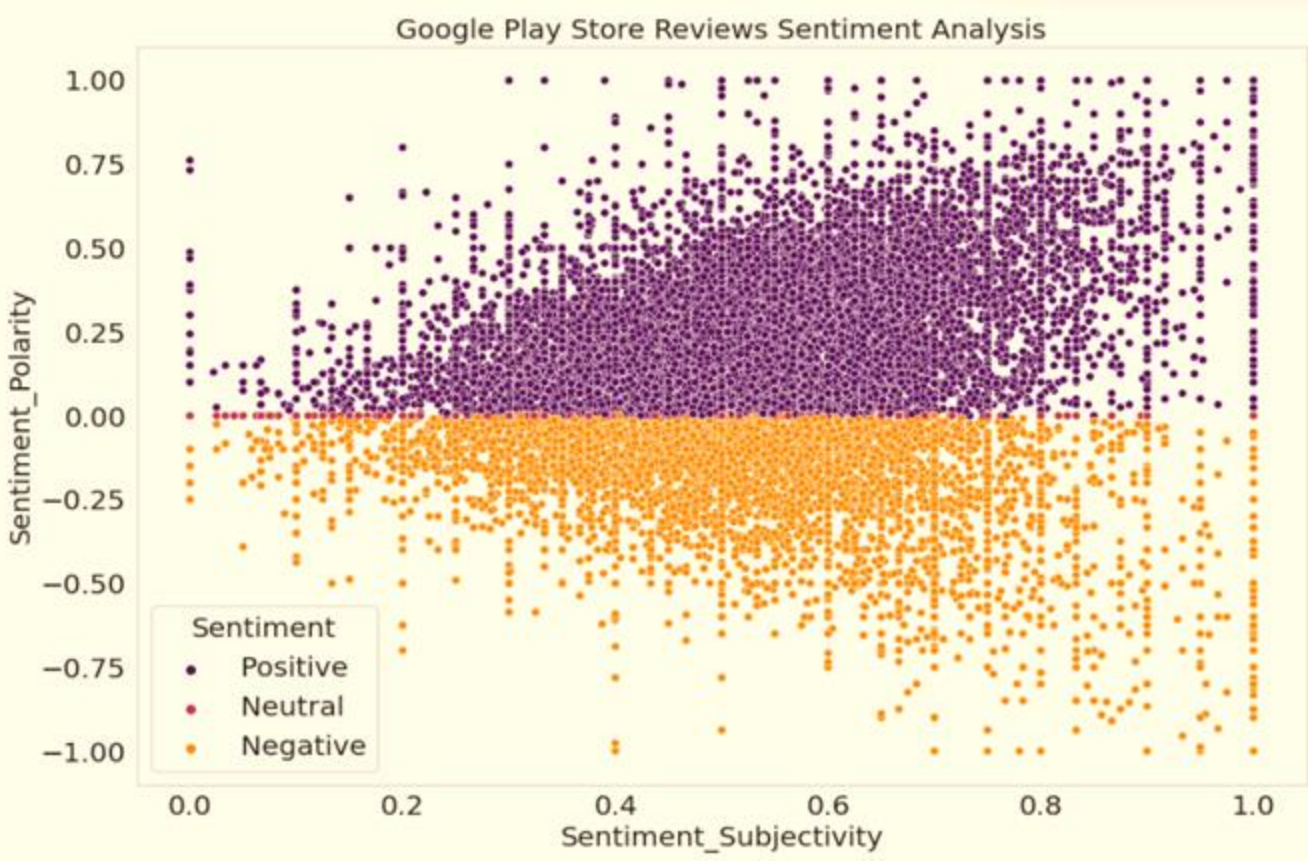


Helix Jump is a App from merged dataset has highest **209 Positive** sentiment count.

Angry Bird Classic is a app from merged dataset has highest **147 Negative** sentiment count.

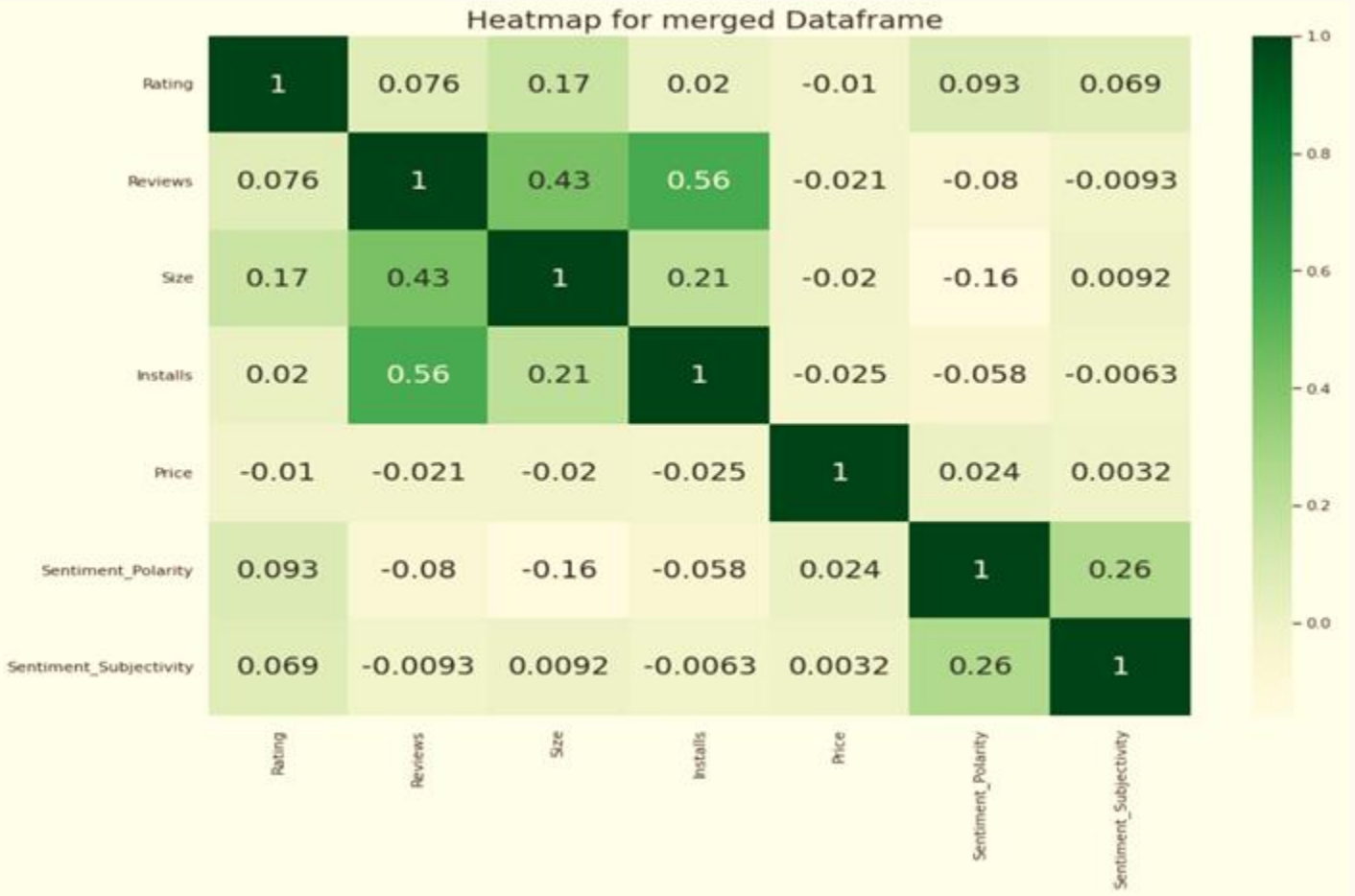
Is sentiment_subjectivity proportional to sentiment_polarity?

From the above 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

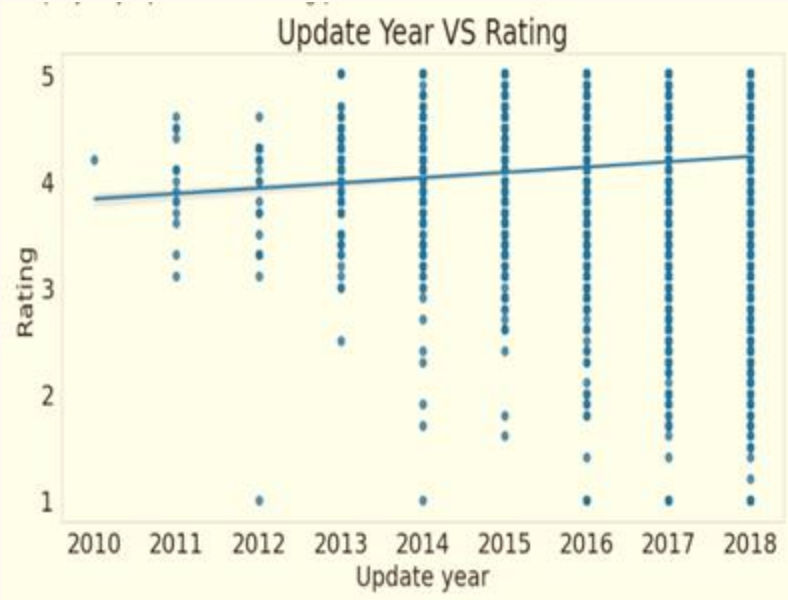


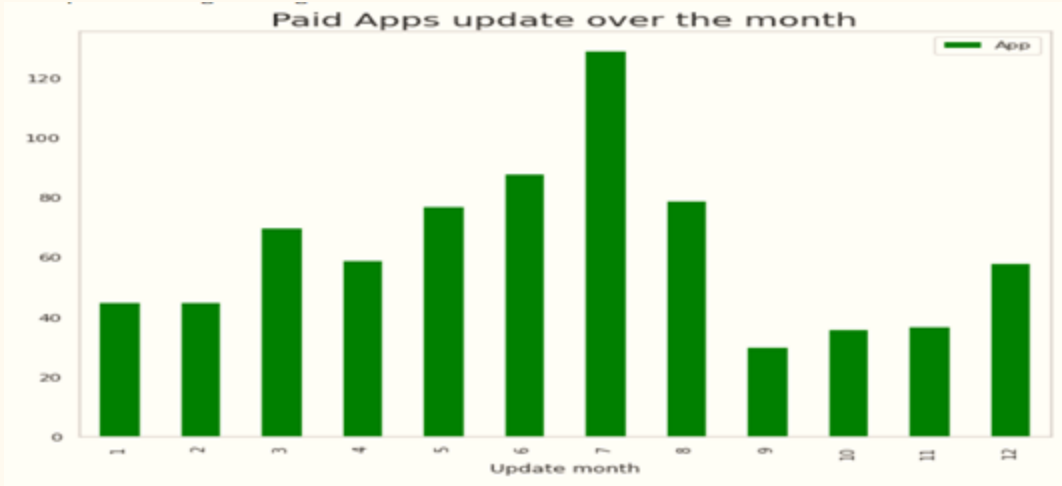
Co-Relation in merged data frame

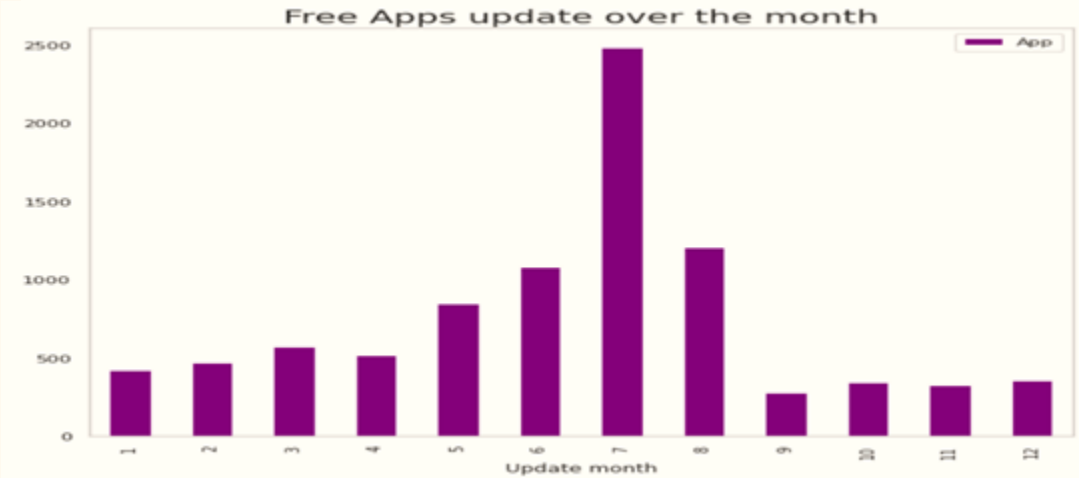
In this correlation matrix, There is not a significant relationship between Rating, Reviews, Size and Installs with respect to the Sentiment polarity and Sentiment subjectivity.



Distribution of Apps updated over the Year and Month







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