Capstone Project - 1

EDA on Play Store App Reviews

By
Shaloy Elshan Lewis
Data Science Trainee, AlmaBetter

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.

Data Summary

Play_Store_Data

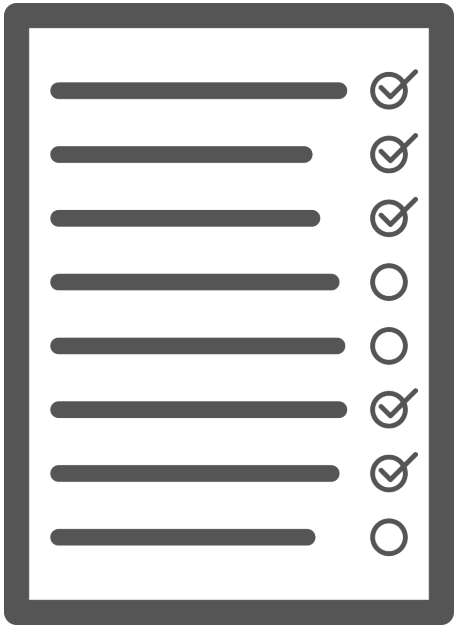
- **App**
- **Category**
- **Size**
- **Rating**
- **Reviews**
- **Installs**
- **Type**
- **Price**
- **Content Rating**
- Genres
- Last Updated
- Current Ver
- Android Ver
- **Rating Group**
- **Revenue**

User_reviews

- **App**
- **Translated Review**
- **Sentiment**
- Sentiment_Polarity
- Sentiment_Subjectivity

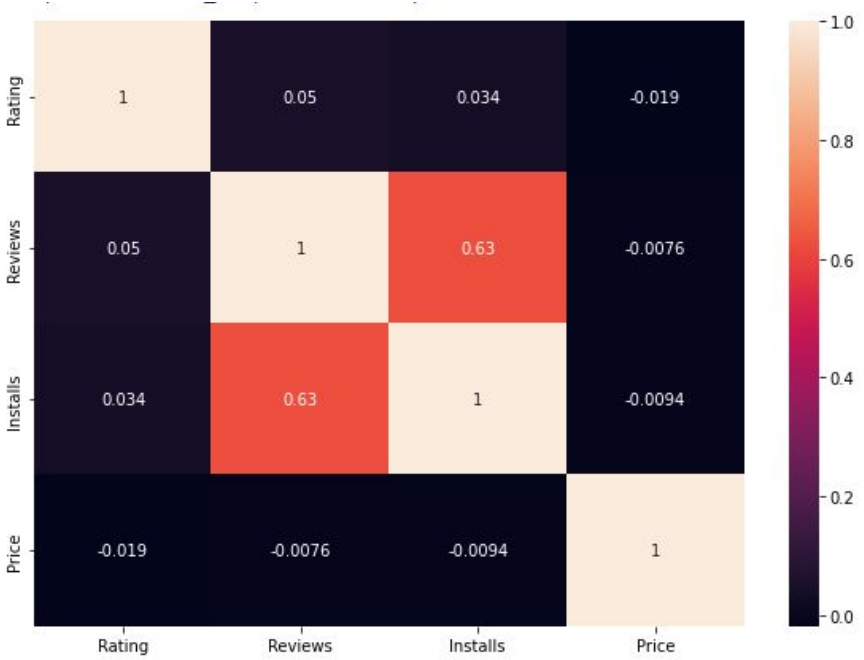
Agenda

- Correlation heatmap
- Type and Content Rating Analysis
- Categorical Analysis
- App Rating Analysis
- Top Free and Paid Apps
- Average Price of Paid Apps in Each Category
- Most Popular Apps
- App Size Analysis
- App Reviews Analysis
- Challenges Faced
- Analysis Summary



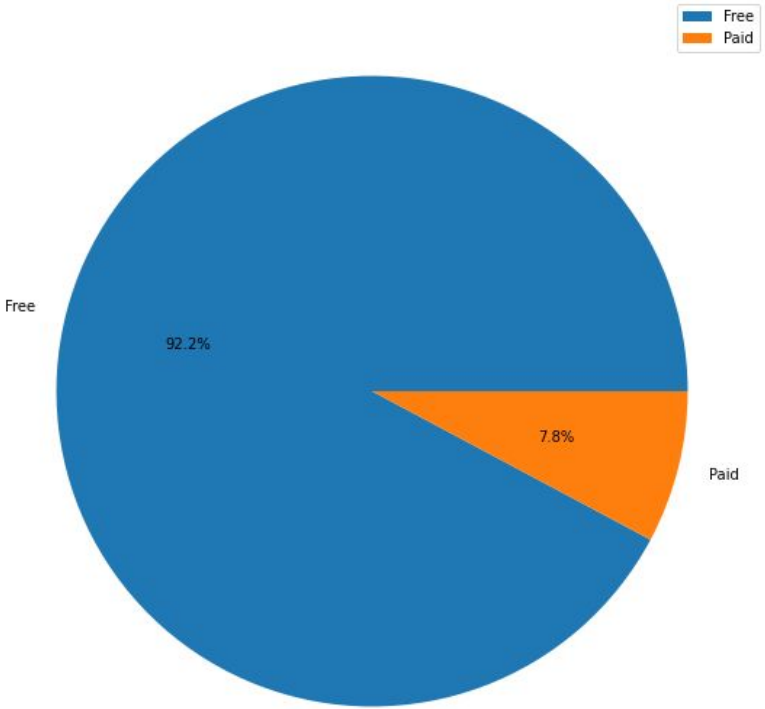
Correlation Heatmap

- 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**.

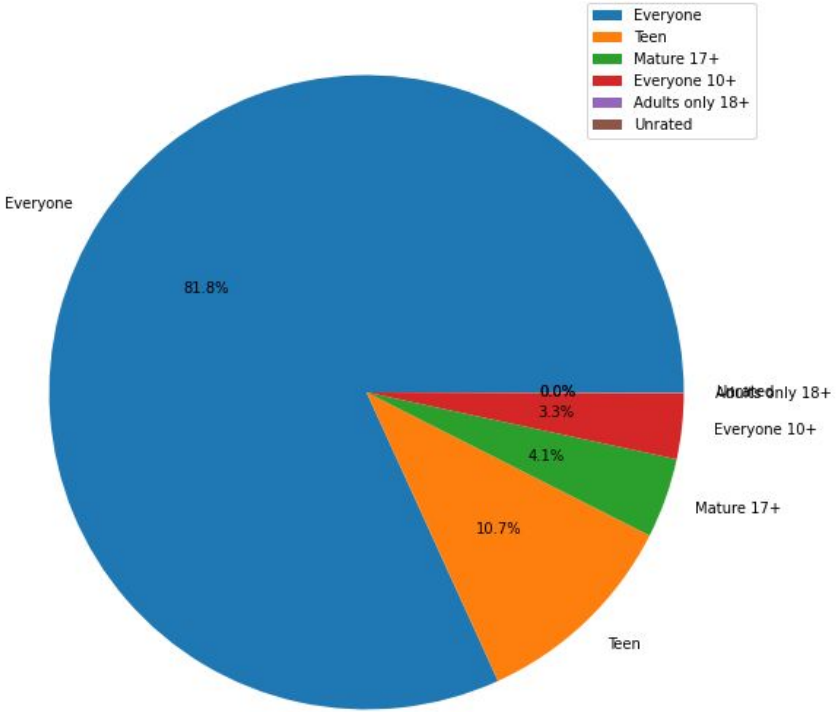


Type and Content Rating

Free and paid apps in the df

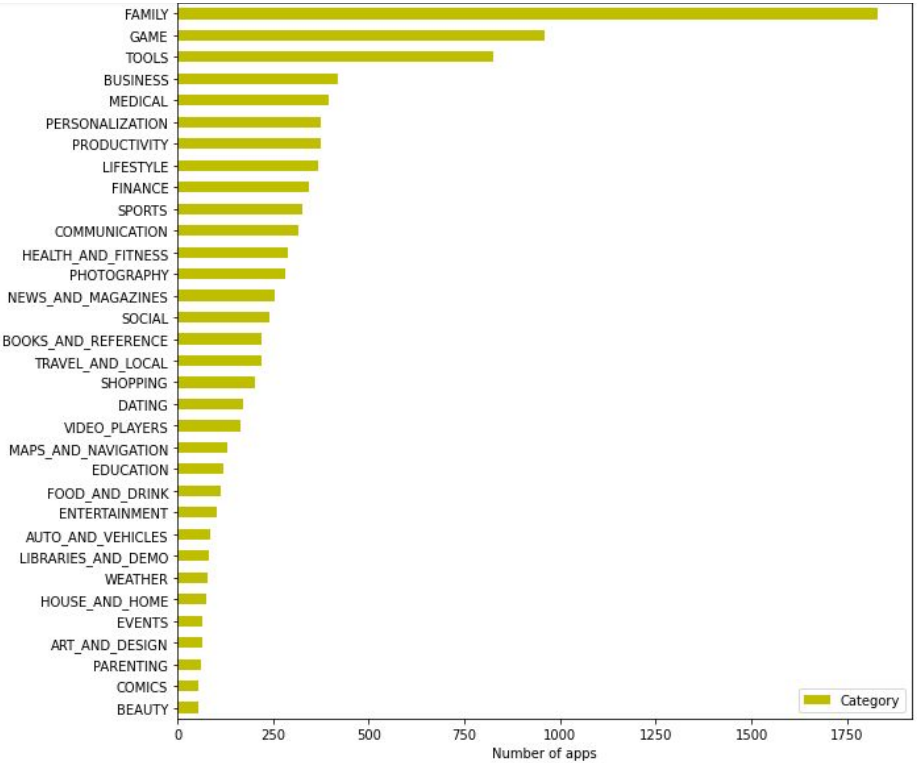


Content rating types in the df

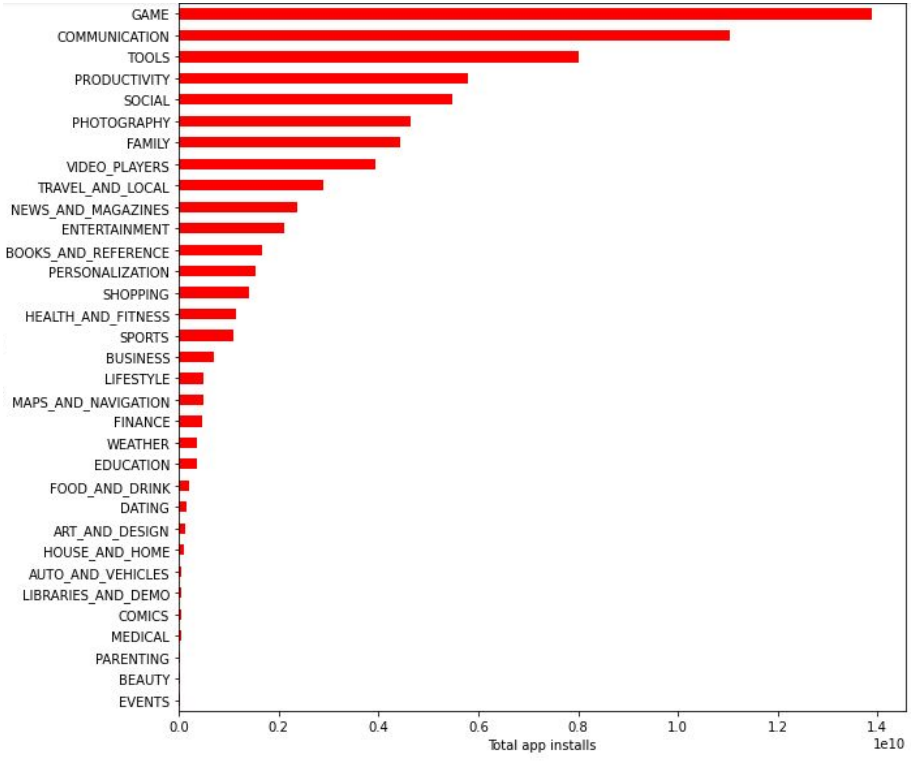


Categorical Analysis

Category vs No. of Apps

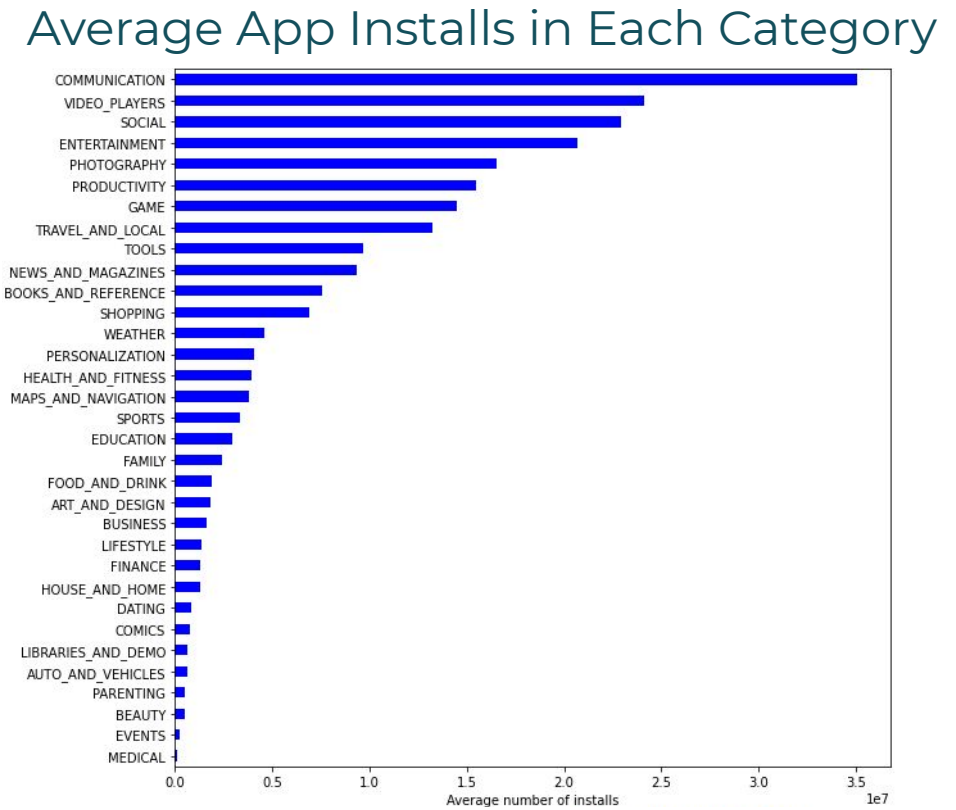


Category vs Total App Installs



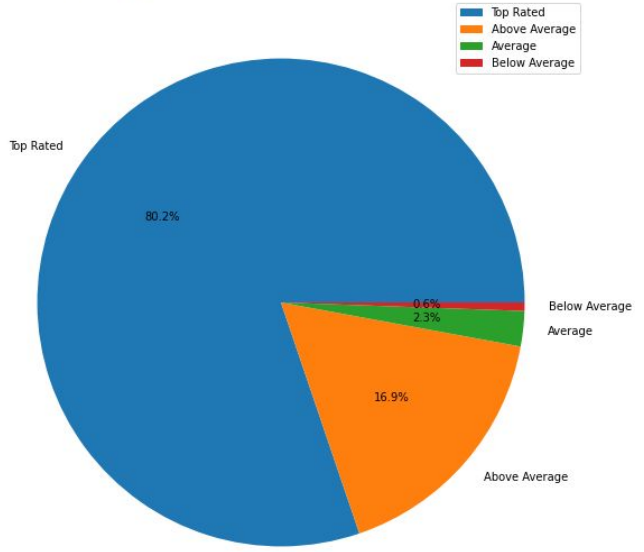
Categorical Analysis (Contd.)

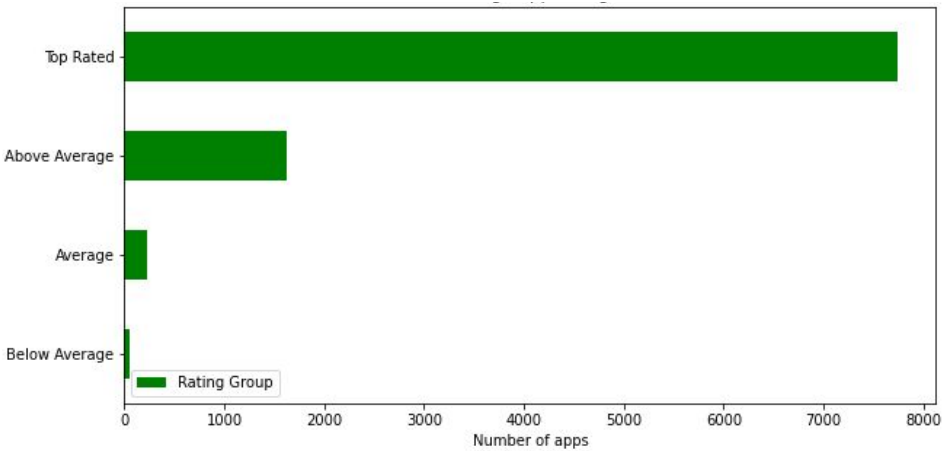
- The **Family**, **Game**, and **Tools** category has the highest number of apps.
- The **Game**, **Communication**, and **Tools** category has the highest number of app installs.
- The **Communication**, **Video Players**, and **Social** category has the highest number of average app installs



App Rating Analysis

- 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 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).

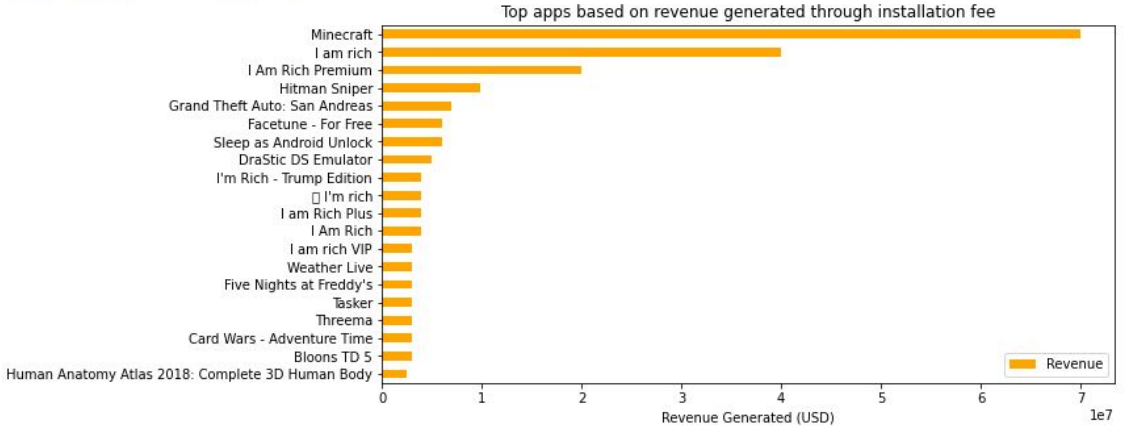
```
152                                Google Play Books
335     Messenger – Text and Video Chat for Free
336                                WhatsApp Messenger
338                      Google Chrome: Fast & Secure
340                                            Gmail
341                                          Hangouts
391                      Skype - free IM & video calls
865                                Google Play Games
1654                                  Subway Surfers
2544                                         Facebook
2545                                        Instagram
2554                                         Google+
2808                                   Google Photos
3117                      Maps - Navigate & Explore
3127                             Google Street View
3234                                          Google
3454                                   Google Drive
3665                                         YouTube
3687                          Google Play Movies & TV
3736                                     Google News
Name: App, dtype: object
```

Top Paid Apps Based on Revenue Generated

- Revenue generated is given by the formula:

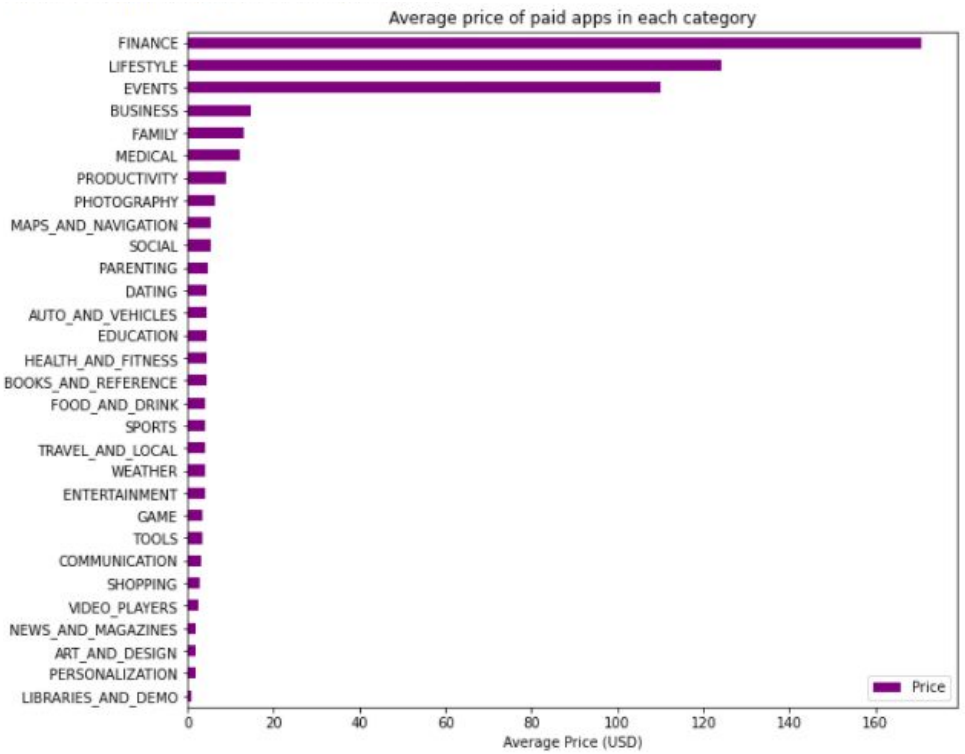
Revenue = Installs * Price

- 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.
- **Minecraft** is the only app that has over **10M** installs.



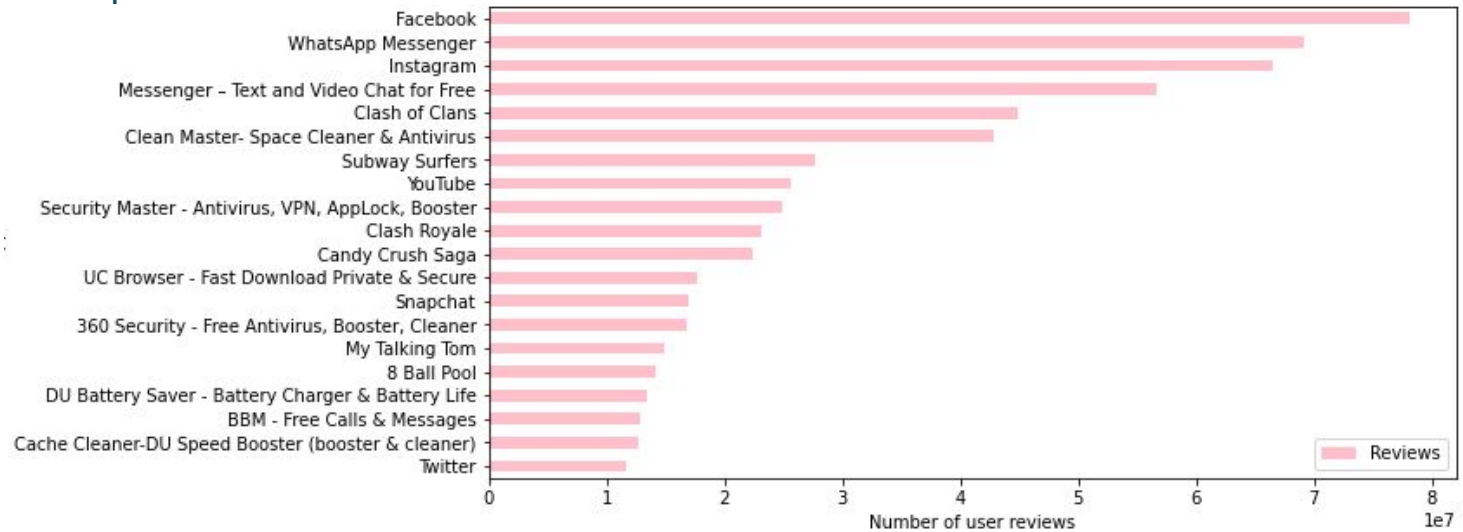
Average Price of Paid Apps in Each Category

- The paid apps in the **Finance**, **Lifestyle**, and **Events** category are on average significantly more expensive than the paid apps in other categories.



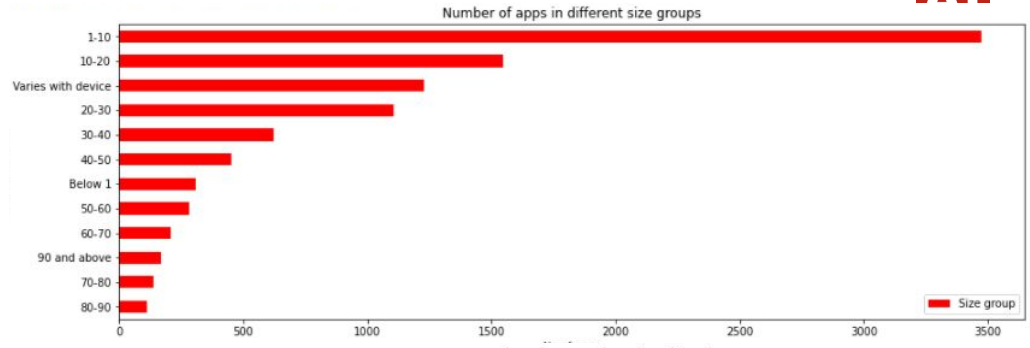
Most Popular Apps

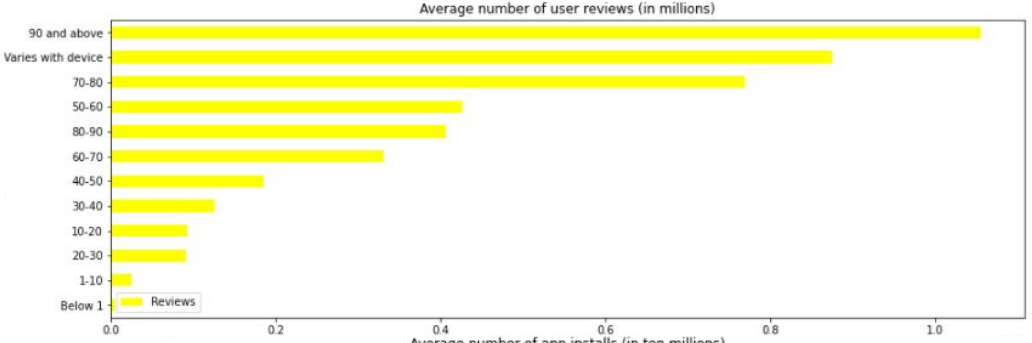
- We can state that the apps with more reviews, whether positive, negative, or neutral, are more popular than the others.
- This is because the number of user reviews indicates that these individuals have engaged with the app's content and have written their opinions on it.

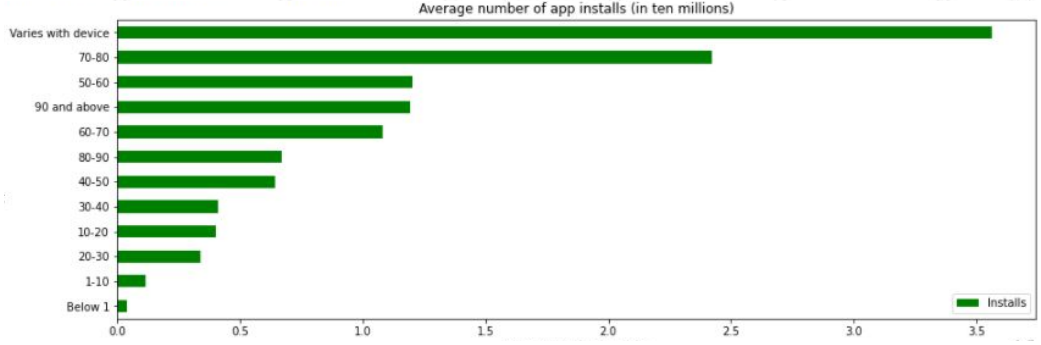


App Size Analysis

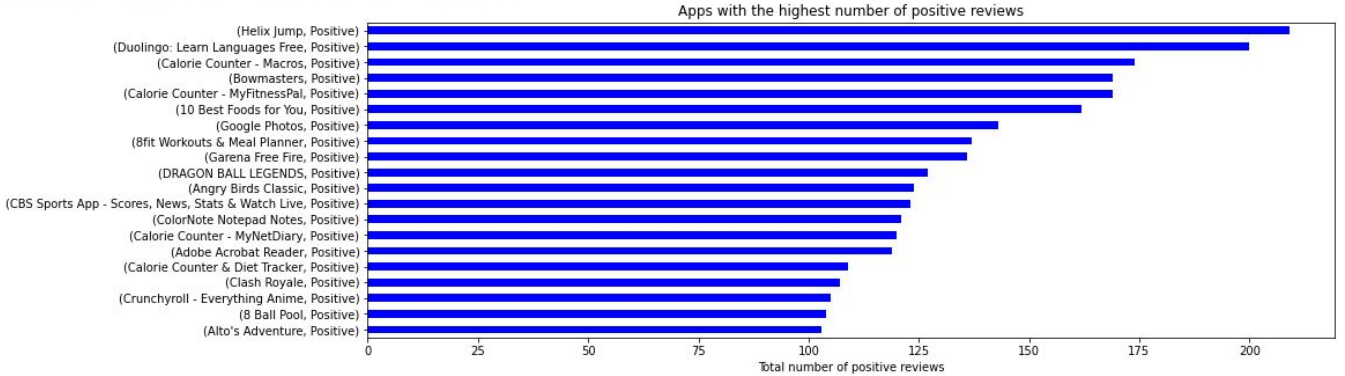
- The apps are categorized based on its size between ~0 to 100 MB in the intervals of 10 MB each.
- The total number of apps in each size category indicates the **competition**.
- Average number of **user reviews** and **average app installs** in each size category indicates the **popularity** of the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

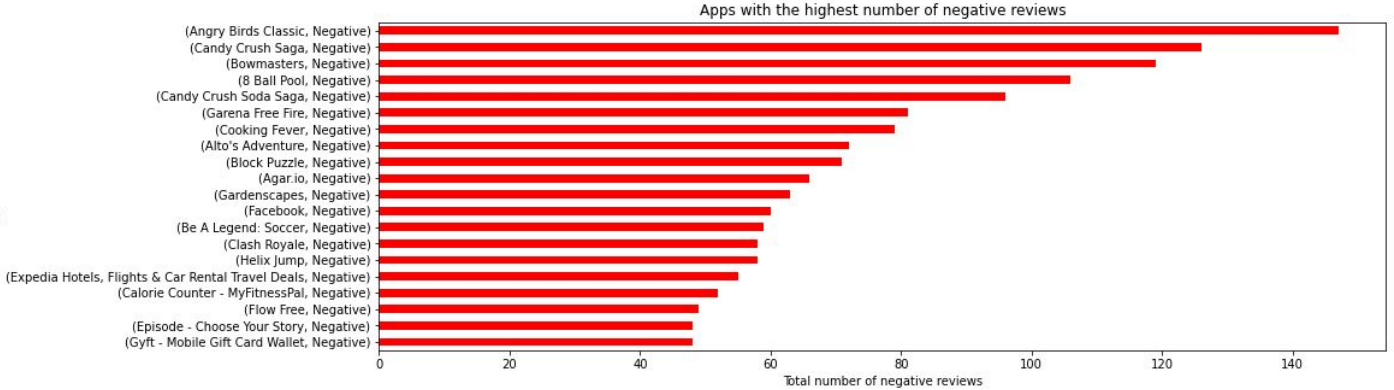






Positive and Negative Reviews





Word Cloud on translated reviews

- The word clouds is used as a visual representation of any textual data, in this case the user reviews.
- The higher the number of times a word is repeated, the bigger and bolder it gets.
- Hence the word clouds can be used to get a birds eye view of all the textual data in the dataset.

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.

Analysis Summary

- Percentage of free apps = **~92%**
- Percentage of apps with no age restrictions = **~82%**
- Most competitive category: **Family**
- Category with the highest number of installs: **Game**
- Category with the highest average app installs: **Communicaction**
- Percentage of apps that are top rated = **~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.
- There is a **positive** correlation between the **reviews** and **installs**. And also between **rating** with **installs** and **reviews**.
- **Price** is **negatively** correlated with the **rating**, **reviews**, and **installs**.

Analysis Summary (Contd.)

- Category in which the paid apps have the highest average installation fee: **Finance**
- Most popular app in the Play Store based on the number of reviews: **Facebook**
- 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