Capstone Project-1
EDA
Play Store App Review Analysis

Team Members

Abhishek Shubham

Data Pipeline

- **Data Processing-1:** In this we have removed unnecessary features that will not be used in our analysis.

- **Data Processing-2:** In this part different data types of each columns are observed and converted from object to int, float for our further analysis.

- **Data Processing-3:** In this part we have deleted null values of some features and also for some columns like Rating we have imputed null values with the median of the column.

- **Exploratory Data Analysis:**

➢ **Univariate Analysis:** In this part, we analyzed individual features/variables and concluded inferences from them.

➢ **Multivariate Analysis:** In this part we analyzed two or more features at a time and their relation between them.

Data Summary

Play Store Dataset:

The dataset contains details of Android applications present on Google Play Store. For analysis of the mentioned data we have used Python. Our business case is to locate the best features, which we measure by Review check. There are 13 features that depict each application and an aggregate of 9649 unique observations for applications. Following variables were initially included:

1. **App** – Contains the name of the app with a short description (optional).
2. **Category** – It gives the category to the app.
3. **Rating** - It contains the average rating the respective app received from its users.
4. **Size –** : It contains the the disk space required to install the respective app.
5. **Reviews –** Number of user reviews for the app.
6. **Installs –** Number of user downloads/installs for the app.
7. **Type** – It states whether an app is free to use or paid.
8. **Price** – It gives the price payable to install the app. For free type apps, the price is zero.
9. **Content Rating** – Age group the app is targeted at.
10. **Genres -** It gives the genre(s) to which the respective app belongs.
11. **Last Updated -** Date when the app was last updated on Play Store.
12. **Current Ver. -** Current version of the app available on Play Store.
13. **Android Ver. –** Minimum required android version.

Data Summary

User Reviews Dataset:

This Dataset contains the reviews of different users for all the apps. There are 5 features that depicts the user reviews for different applications and contains total 37427 observations after data processing. Following variables were included with this dataset:

1. **Sentiment**: A view of or attitude towards a scenario, event or anything this attitude may be positive, negative and neutral. So, this column completely tells the same about different apps means that this column have the information of user's reviews(emotion, opinion) concerning different apps on Play Store.

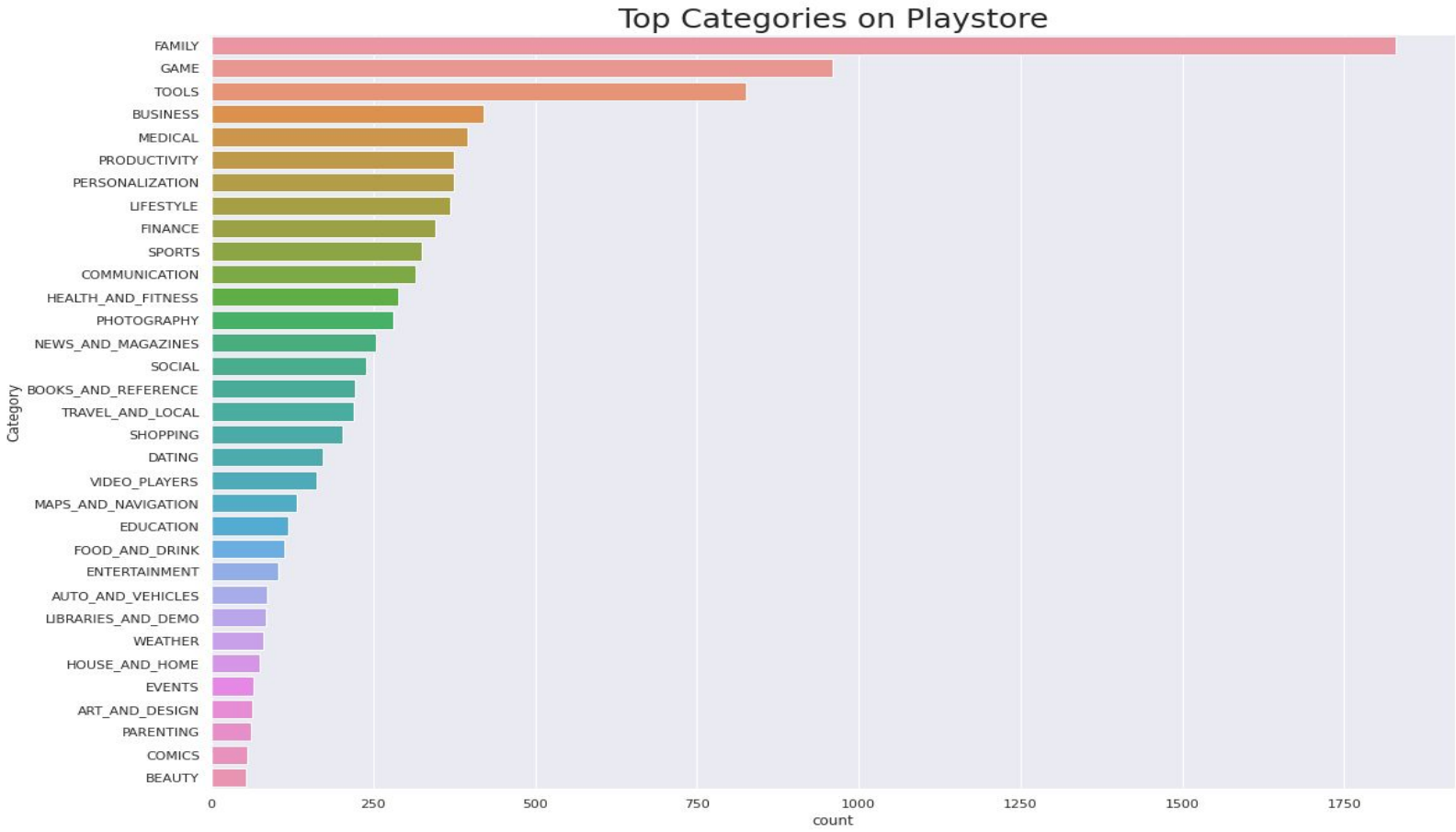
2. **Sentiment_Polarity:** It is the expression that determines the sentimental aspect of an opinion. The Sentiment polarity can be determined as float which lies in the range [-1, 1] where 1 means positive statement and -1 means negative statement.

3. **Sentiment_Subjectivity:** The subjectivity is a measure of sentiment being objective to subjective and goes from 0 to 1. Subjectivity is also a float that lies in the range [0, 1]. When it is close to 0, it is more about facts and when it increases, it comes close to be an opinion. Subjectivity refer opinion because opinion is important key to client for business concern.

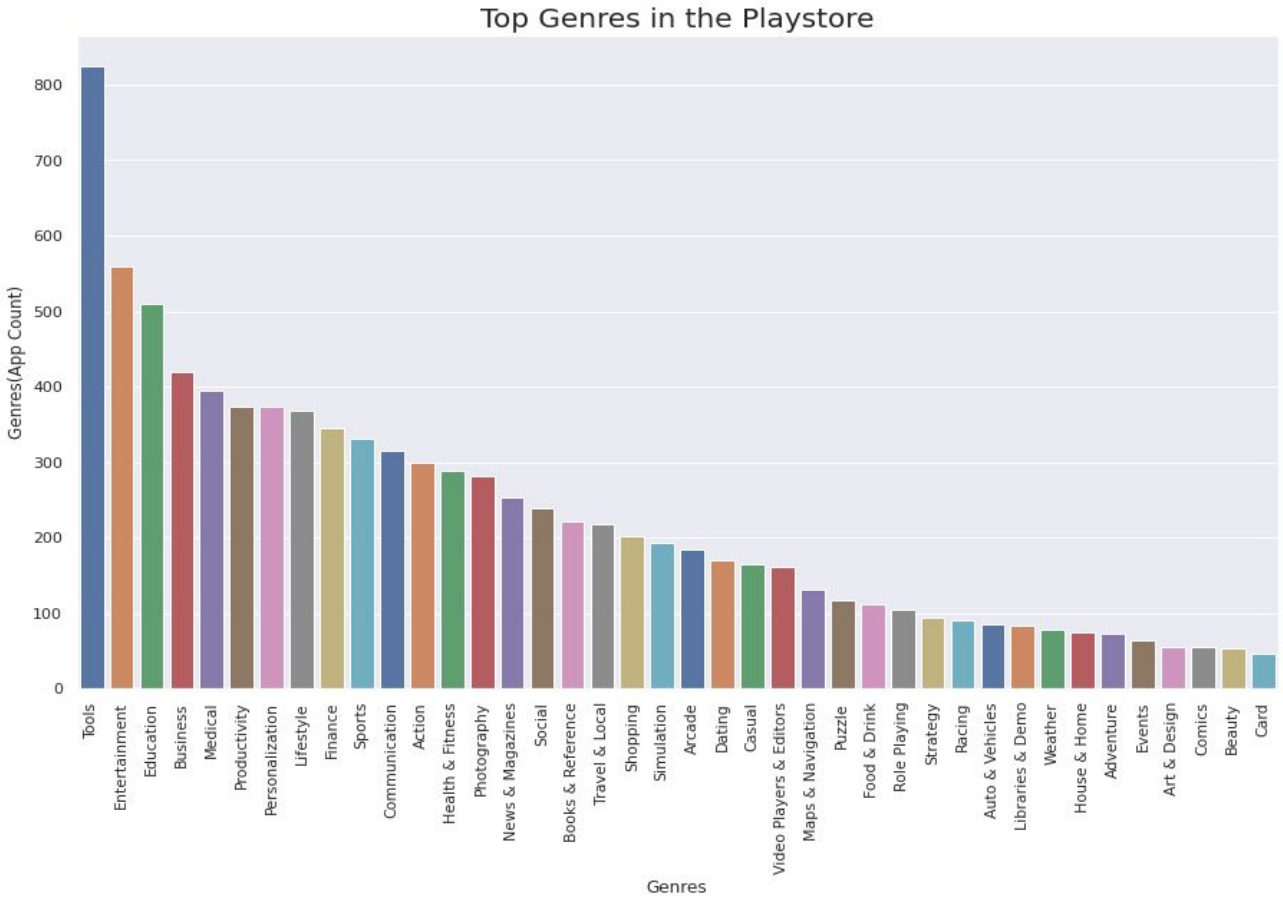
4. **Translated_Review:** It tells us about what the users feedback is about the application.

5. **App**: It tells us about the name of the application.

EDA: (Univariate Analysis)

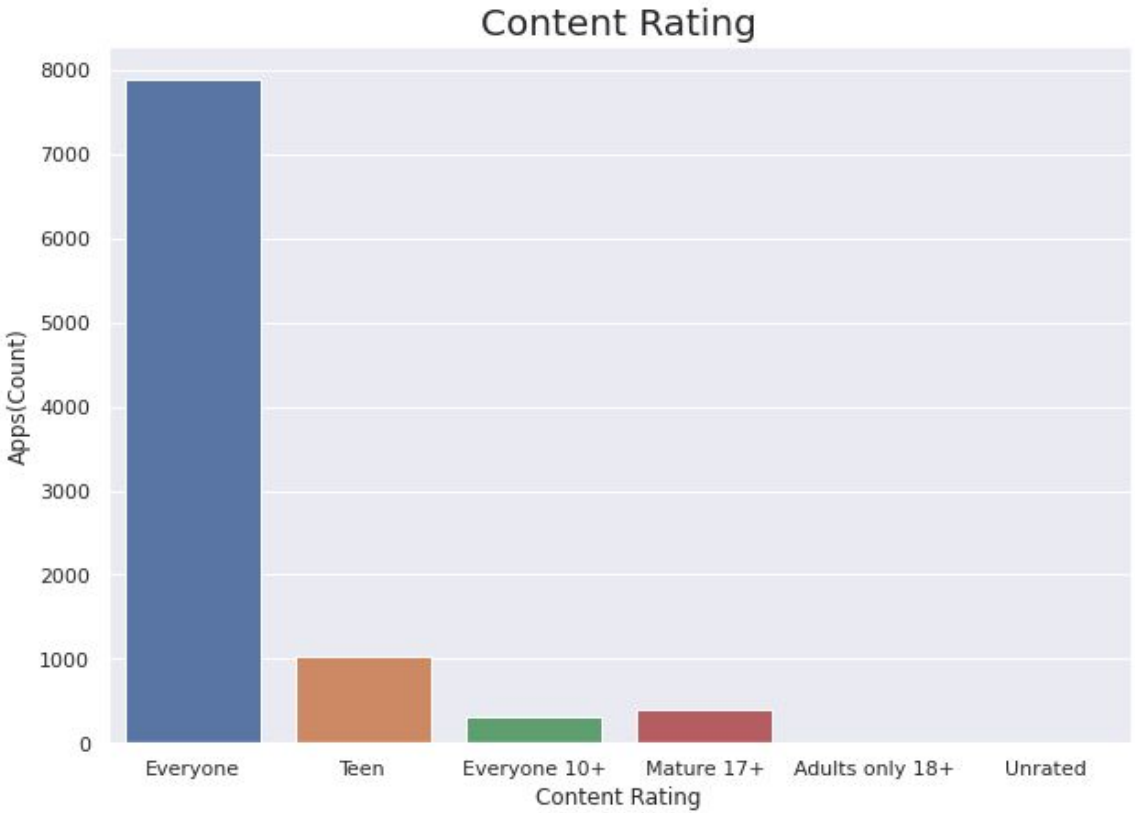


Univariate Analysis(Continued)



From this Barplot we can see that the Highest Number of Apps are found under Tools and Entertainment genres followed by Education, Productivity, Finance and many more.

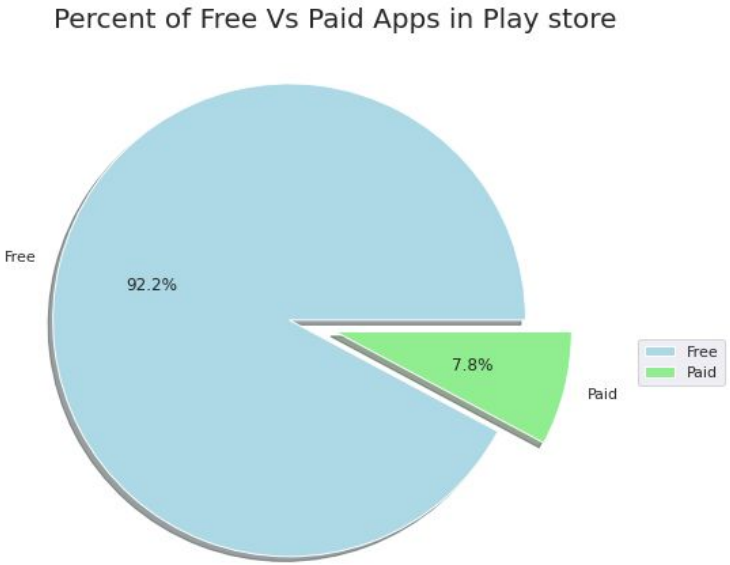
Univariate Analysis(Continued)

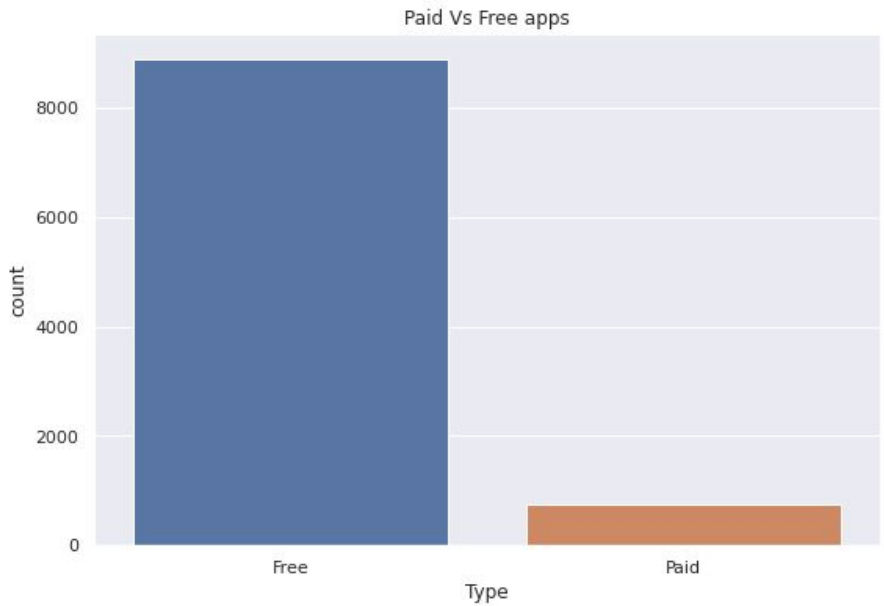


- There are 6 categories in content rating out of which majority of the apps are available for 'Everyone'. Next category of content rating apps are used by only 'Teen'.

- This shows as developers who list the apps for everyone has higher reachability and increased app engagement as compared to other content rating apps like 'Teens', 'Adults' etc.

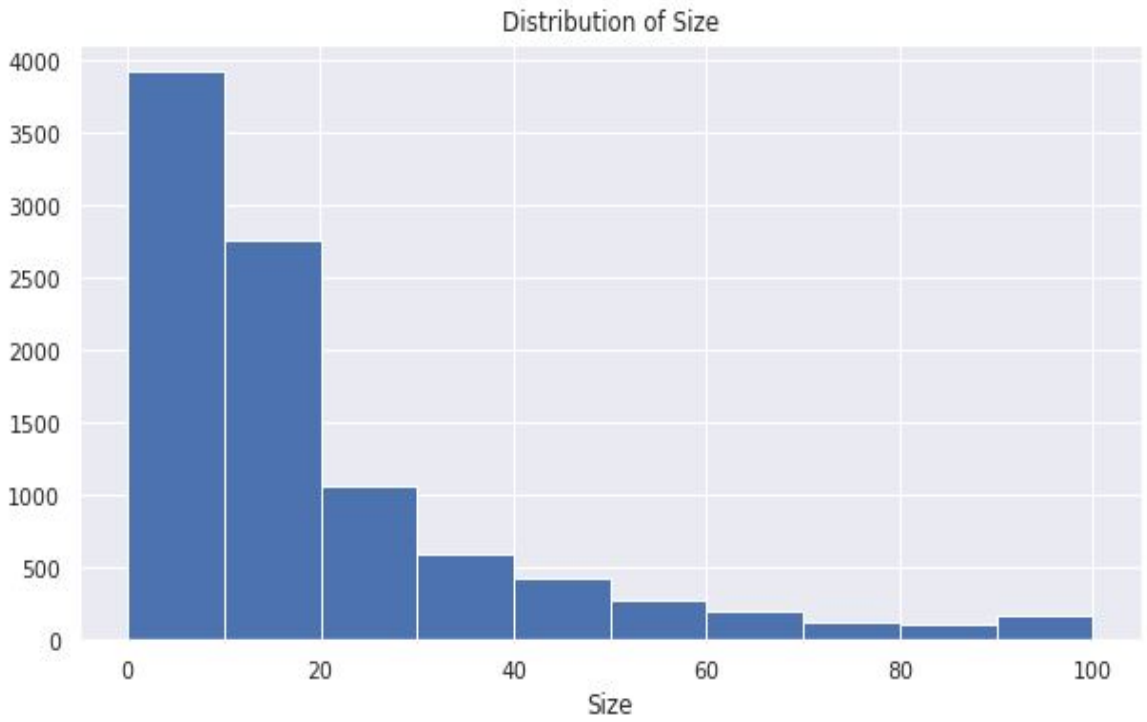
Univariate Analysis(Continued)





- There are 92.6% of free apps in the play store vs 7.8% of paid apps.
- There are more free apps than paid apps. Since most of the apps are cost free so there are more chances for the common people and many people to install the apps easily and use it for their daily needs and other uses increases the app engagement by users which positively impacts the success rate of the apps in the Play store.

Univariate Analysis(Continued)



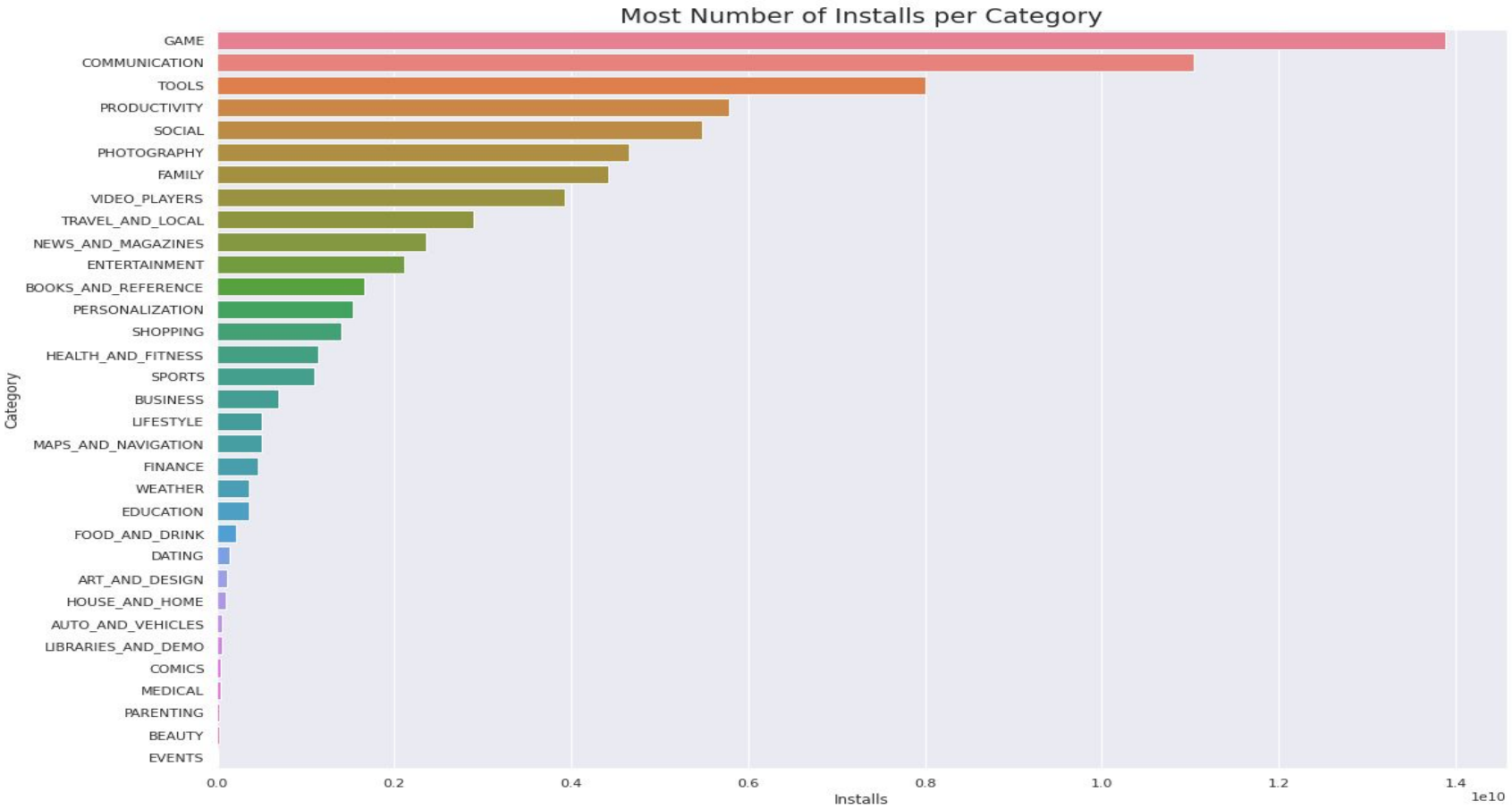
- Most of the apps size is in between 0 - 20 MB.
- There are even apps with greater size than normal. They require upto 100 MB which takes more space of a system.

Univariate Analysis(Continued)

Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud.

- From the above Word Cloud it can be seen that game is more bigger than any other, next comes good, app and so on. It can also be said that there are positive words that can be seen which indicates there are lot of apps which are given positive reviews by the Users.
- The most commonly used words from Users Translated reviews says that they have used more positive words to say about their app experience and which says it has good app engagement by its users. So this definitely tells about good user experience and good word of mouth increases the publicity for the app then it impacts the success rate of the respective apps.

Bivariate Analysis

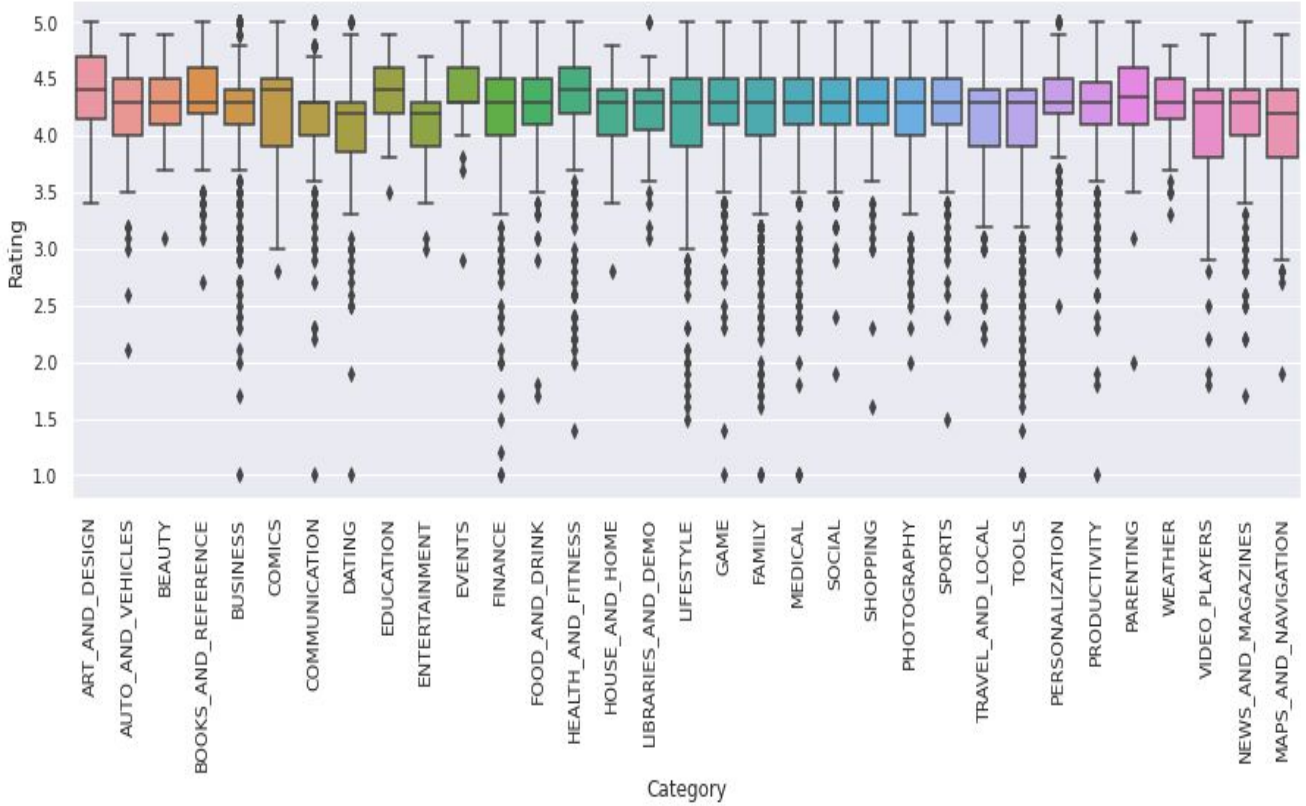


- It can be seen from above plot that the Games Category has the maximum number of Installs followed by Communication and Tools.

- The Category 'Games' has almost 14 Billion installs, next is Communication which has around 11 Billion installs then we have Tools which has around 8 Billion installs. Rest other Categories has installs less than 6 Billion.

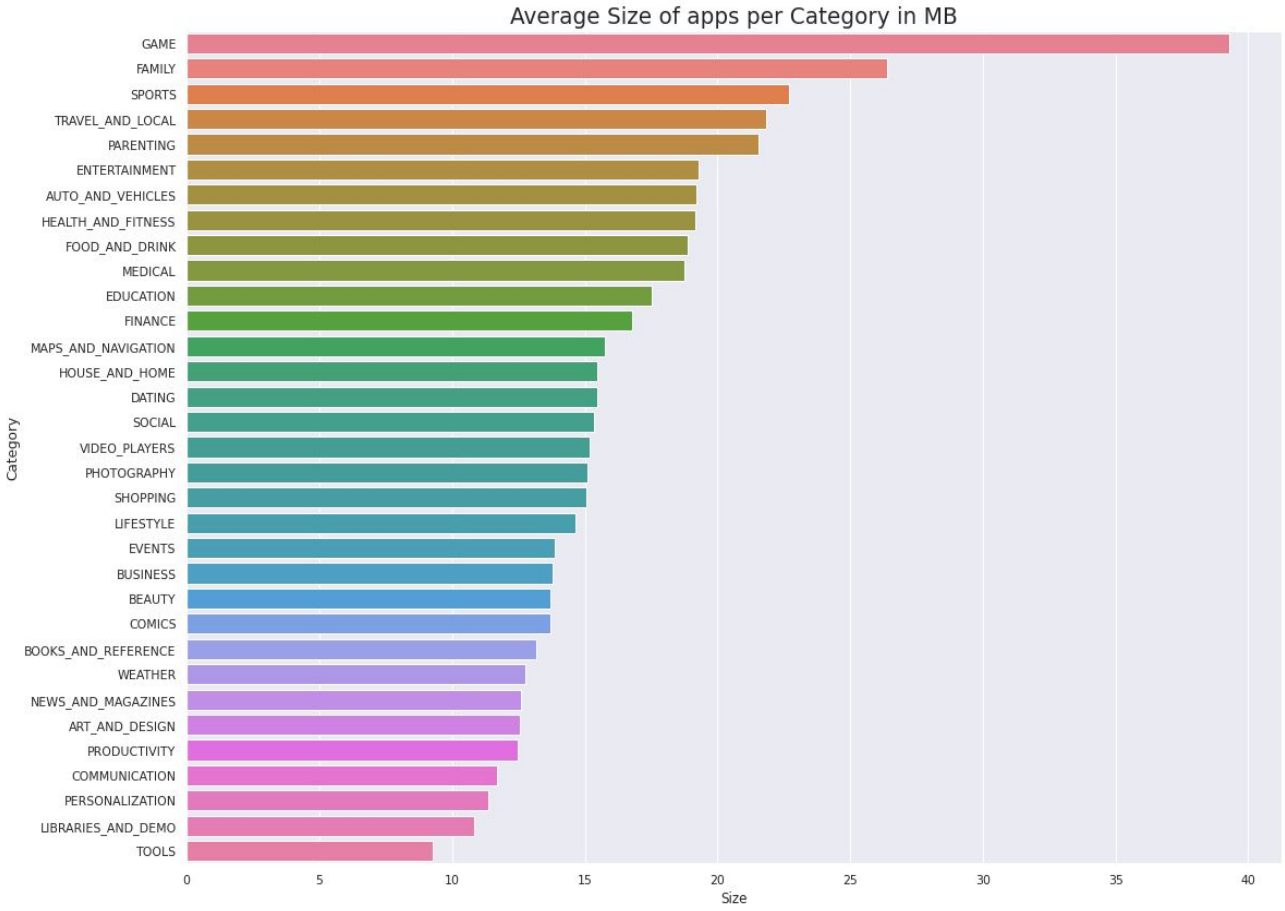
- From this barplot we can infer that the developers who make apps related to Games, Communication and tools has wide customer base as these categories has the maximum numbers of installs.

Bivariate Analysis(Continued)



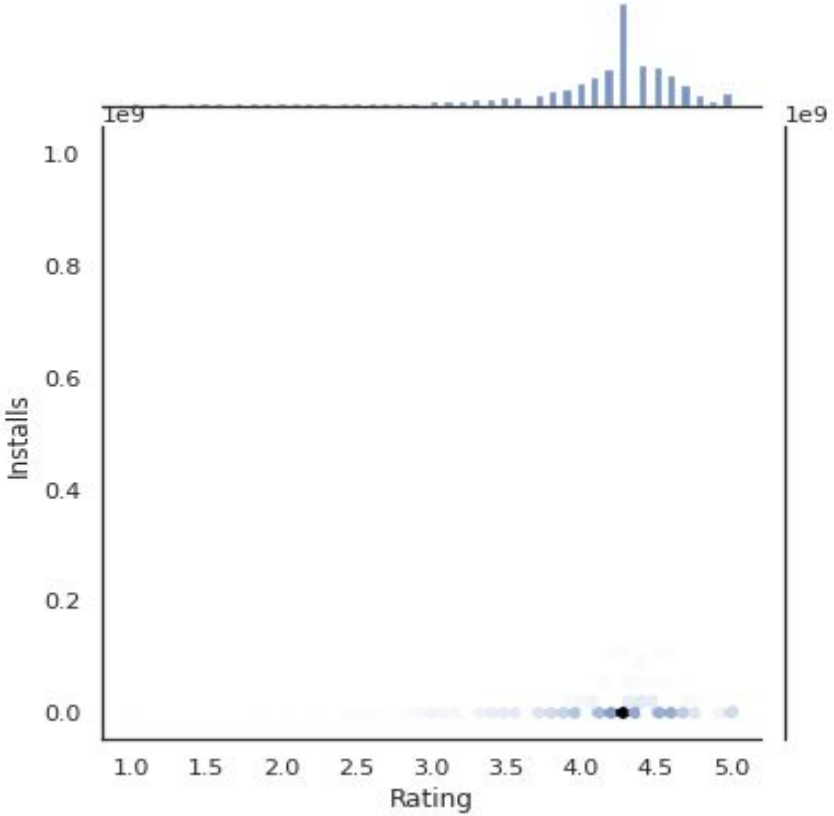
From this boxplot it can be seen that the median rating of all the categories are greater than 4.0. By knowing the reviews and improving the app features, it can help to improve the app engagement. Also by Rating it can be said that this is a good sign for high success rate of the apps.

Bivariate Analysis(Continued)



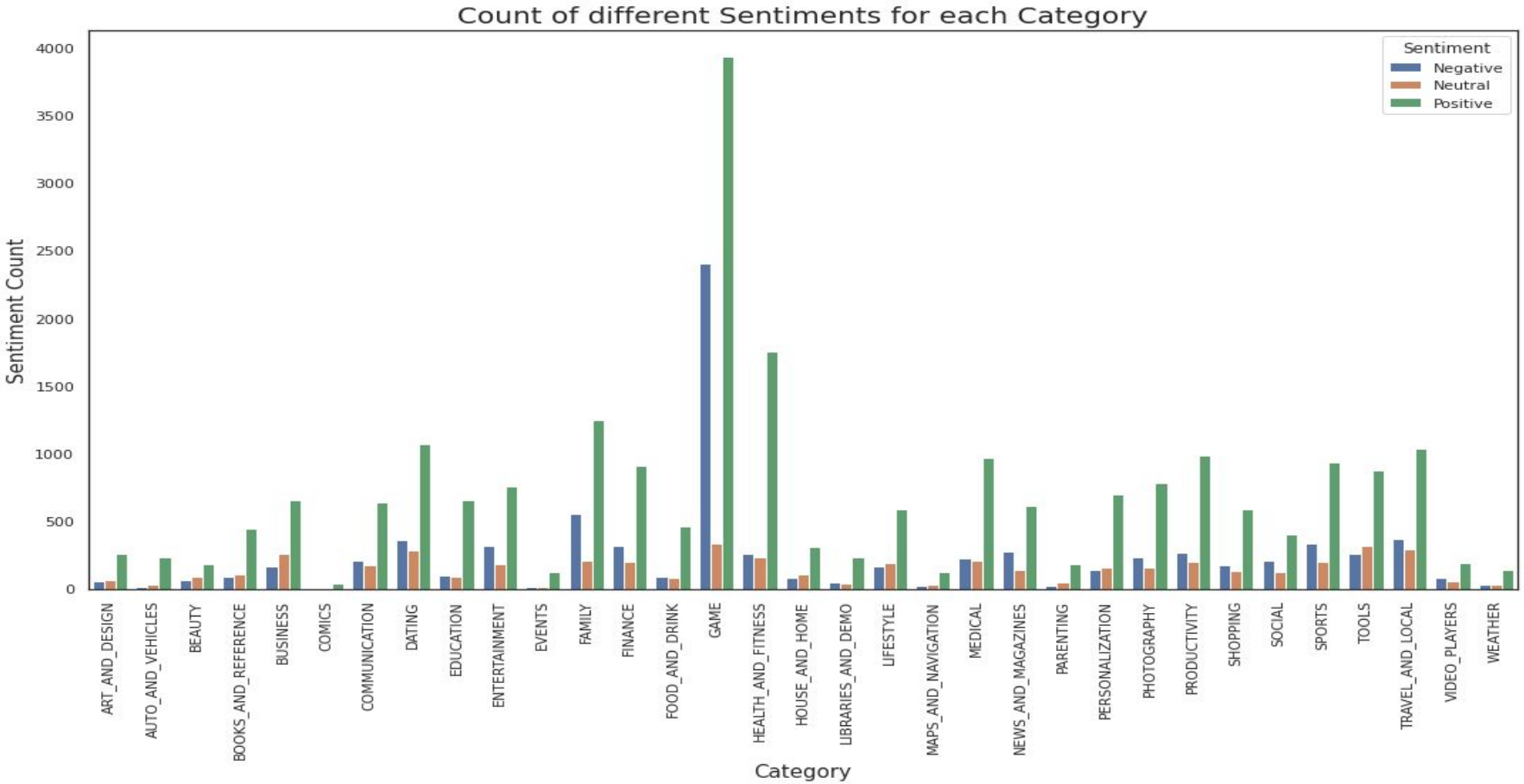
From this plot we can infer that the Game category has high average size of apps which is around 40 MB followed by Family and Sports category.

Bivariate Analysis(Continued)



From this jointplot it can be seen that high rated apps have high number of installations thus increasing the success rate of an app.

Bivariate Analysis(Continued)



Correlation Matrix

As per the correlation Matrix we can see that there exists a positive correlation between 'Installs' and 'Reviews' columns which means if the number of installation of apps increases then the reviews for the app will tend to increase too. Although the reviews can be positive, negative or neutral but the probability of positive reviews are higher than the negative or neutral reviews as seen in the above analysis.

Analysis Summary

- Percentage of free apps = 92.2%
- Percentage of Paid apps = 7.8%
- Most competitive category: Family
- Top Genres: Tools
- Category with the highest number of installs and reviews: Game
- Category with the highest average app installs: Communicaction
- Category in which the paid apps have the highest average installation fee: Finance
- The median size of all apps in the play store is 13 MB.
- User reviews contains mostly positive words and the most used words are Game, good, app, great etc according to WordCloud.
- There are more number of Positive reviews as compared to Negative and Neutral reviews.
- Median rating of all the categories are greater than 4.0
- There is Positive correlation between Installs and Reviews columns

Conclusion

- After the through analysis it can be said that the most important features that helps in predicting the success rate of an app are the Rating, reviews, Installs and type of an app.
- The features in Users review dataset that would help in the success rate and app engagement are Sentiment then Sentiment polarity and subjectivity.
- There are many categories of apps present in the play store and the apps that are high in particular category is Communication, Social and Gaming apps. It can be inferred that these apps are more successful and have high app engagement.
- From the analysis it is seen that there are good number of apps with positive reviews than negative and neutral reviews.
- There are more apps in the Play store than that are given reviews by the users. There are more ratings given to an app than the reviews. As per our analysis we can also conclude that highest rating and large number of reviews increases the success.

Takeaways for App Developer

- Developing apps related to the least categories as they are not explored much. Like events and beauty.
- Most of the apps are Free, so focusing on free app is more important.
- Focusing more on content available for Everyone will increase the chances of getting the highest installs.
- The sentiments of the user keep varying as they keep using the app, so it's important to focus more on users needs and features and make changes in the apps accordingly.

Thank you