Capstone Project

Play Store App Review Analysis(EDA)

Arkopravo Pradhan

PROBLEM STATEMENT

The Play Store apps data has enormous potential to drive app-making businesses to success. Actionable insights can be drawn for developers to work on and capture the Android market. Explore and analyze the data to discover key factors responsible for app engagement and success.

Lets Analyze PlayStore Apps

1. Data summary

2. Data preprocessing and cleaning

3. Data Analysis

Description of datasets

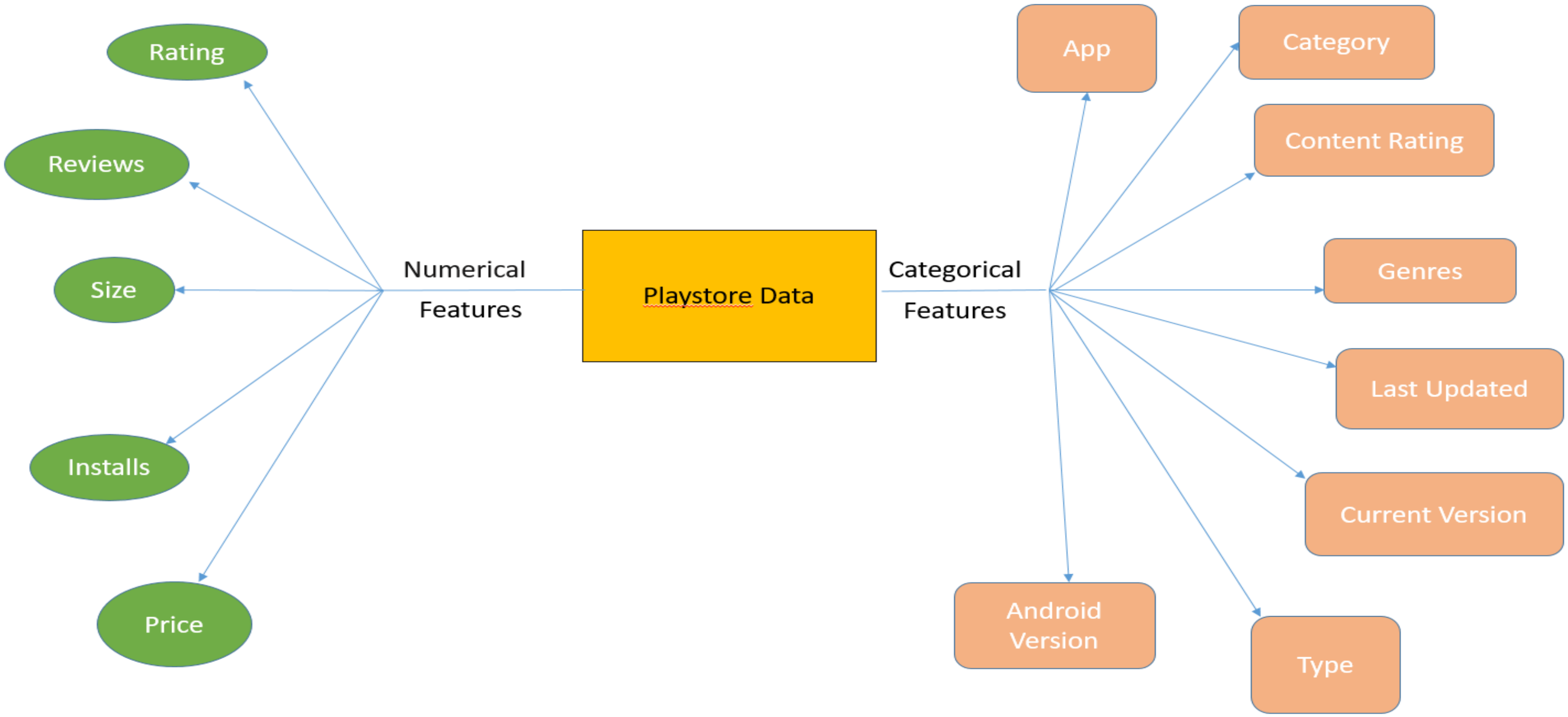
In this project we are given two datasets :-

- Playstore Data: This dataset contains app name their domain, category,geners and type having different rating and reviews . It also comporises of a price column , app' price and app version.

- User Reviews Data: This dataset contains the app name with different textual reviews and sentiment points. It is very usefull for customer sentiment analysis.

We tried our best to do the exploratory data analysis(EDA) on this datasets.

Data Summary of Playstore Dataset



Data Summary of Playstore Dataset

Numerical Features

- Rating: The column constists of numerical values, which rates the app out of 5. Rating is a user review feature. It can be used during review analysis.

- Installs: This column constists of how many times an app is installed. By the number of installs we can understand which app is most used and prefferd app among the given apps.

Data Summary of Playstore Dataset

- Reviews: The column constists of object values that how many reviews a app received . Later we changed it to numeric data type.

- Size: It constists of size of each app(space it will occupy in your device).

- Price: In playstore to use some apps we have to pay a price and some apps are free. The column consists of the price of each app.

Data Summary of Playstore Dataset

Categorical Features

- App: The name of each app. This column has multiple duplicate values. So we have kept row of an app with maximum number of reviews, assuming it to be the latest one

- Category: It gives the information that which app is under which category. It is a vital column for EDA.

- Type: The column tells us wheather the app is free of cost or not.

- Content Rating: It gives the information ,that app can used by which age group or everyone .

Data Summary of Playstore Dataset

- Genres: It gives the information that which app is under which domain. It is a vital column for EDA.
- Last Updated: It gives the information that when the app is last updated.
- Current Version: Provides the current version of each app.
- Android Version: Provides on which android version the app can be installed and used.

Data preprocessing and cleaning for Playstore Dataset

1. The first step towards data filtering is to remove 10472 index due to data mismatch in the column.

```
[ ]  df1.drop(df1.index[10472], inplace=True)
```

2. Next, in our dataset there is a column having number of installs in object format. So, we change the datatype to integer , also removed “+” and “,” from the string.

```
[ ]  df1['Installs'] = df1['Installs'].map(lambda x: x.rstrip('+'))
     df1['Installs'] = pd.to_numeric(df1['Installs'].str.replace(',',''))
```

Data preprocessing and cleaning for Playstore Dataset

3. Next, removing '$' from the values of price column which is in object format and converting it to numeric.

```
[ ]  df1['Price'] = pd.to_numeric(df1['Price'].str.replace('$',''))
```

4. Due to high variance in install column , we used log transformation on it and created “log_installs”. The log transformation reduces or removes the skewness of our original data. Log transformation also de-emphasizes outliers and allows us to potentially obtain a bell-shaped distribution. The idea is that taking the log of the data can restore symmetry to the data.

```
df1['log_installs'] = np.log2(df1['Installs'])
```

Data preprocessing and cleaning for Playstore Dataset

5. This dataset have multiple duplicate values . Each app is having identical rows with difference in number of reviews. It may have happened that for the same app, the data has been scraped in different points of time. So we have kept row of an app with maximum number of reviews, assuming it to be the latest one.

6. After that we removed “$” from reviews column and changed its datatype from object to numeric.

```
df1['Reviews'] = pd.to_numeric(df1['Reviews'].str.replace('$',''))
df1 = df1.loc[df1.groupby(['App'])['Reviews'].idxmax()]
```

Data preprocessing and cleaning for Playstore Dataset

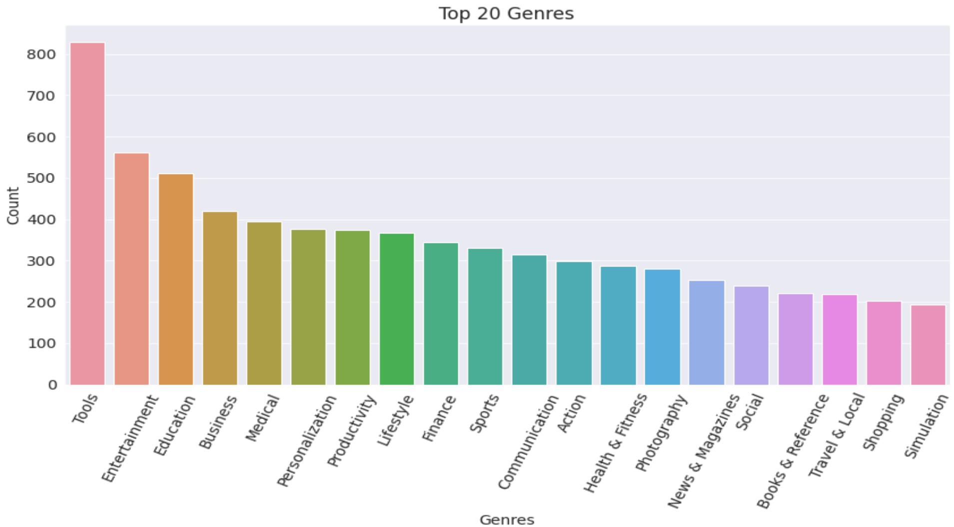
7. In the size column , unit is either MB or KB so we changed the whole column to MB and also removed the null values from the column.

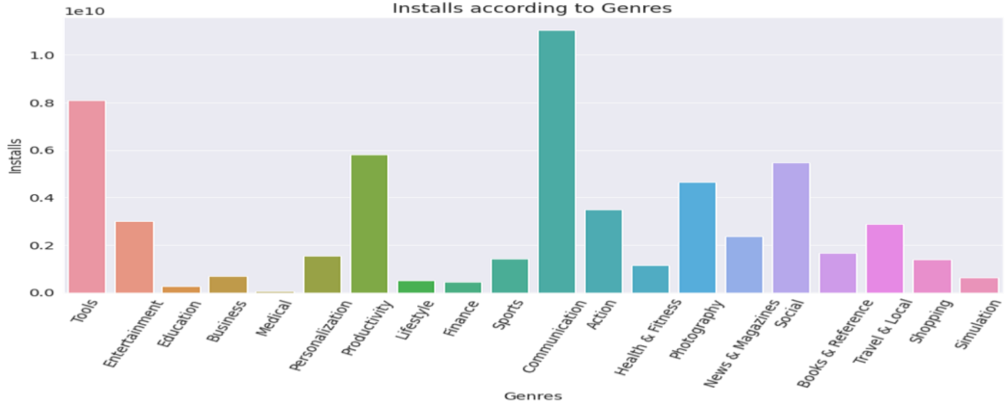
```
[45] df1['Size'] = df1['Size'].apply(lambda x: str(x).replace('Varies with device', 'NaN') if 'Varies with device' in str(x) else x)
     df1['Size'] = df1['Size'].apply(lambda x: str(x).replace('M', '') if 'M' in str(x) else x)
     df1['Size'] = df1['Size'].apply(lambda x: str(x).replace(',', '') if 'M' in str(x) else x)
     df1['Size'] = df1['Size'].apply(lambda x: float(str(x).replace('k', '')) / 1000 if 'k' in str(x) else x)
     df1['Size'] = df1['Size'].apply(lambda x: float(x))
```

```
df1.loc[df1['Size'].isnull(),'Size']=0
```

EDA of Play Store Dataset

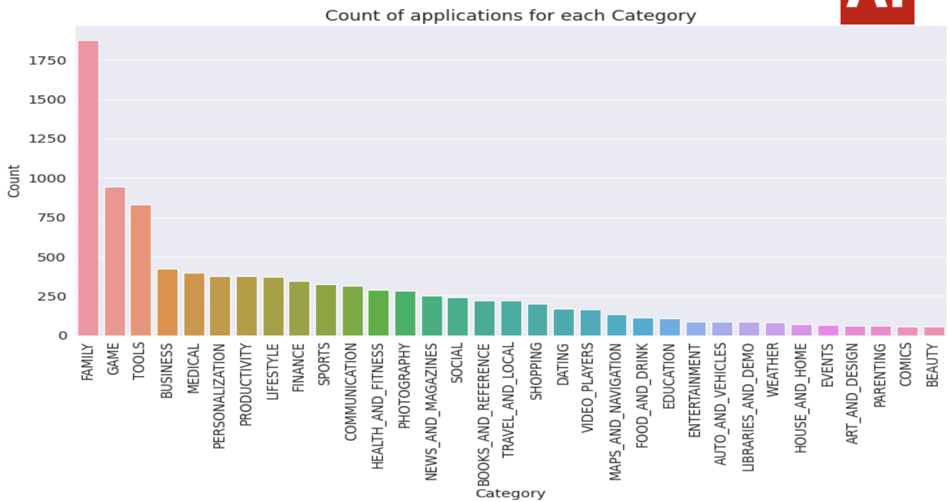
As we can see from these two plots: Maxinum number of apps present in google play store comes under Tools, Entertainment and Education Genres but as per the installation and requirement in the market plot, scenario is not the same. Maximum installed apps comes under Communication, Tools and Productivity Genres.

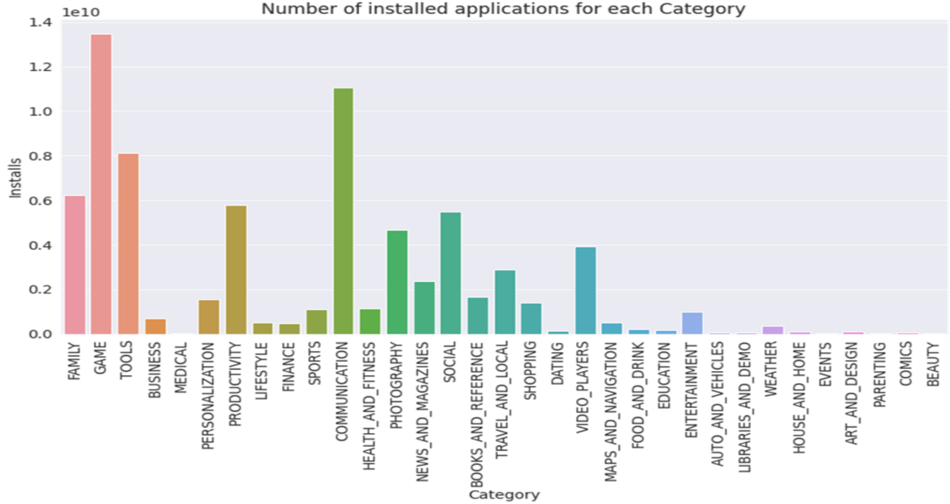




EDA of Play Store Dataset

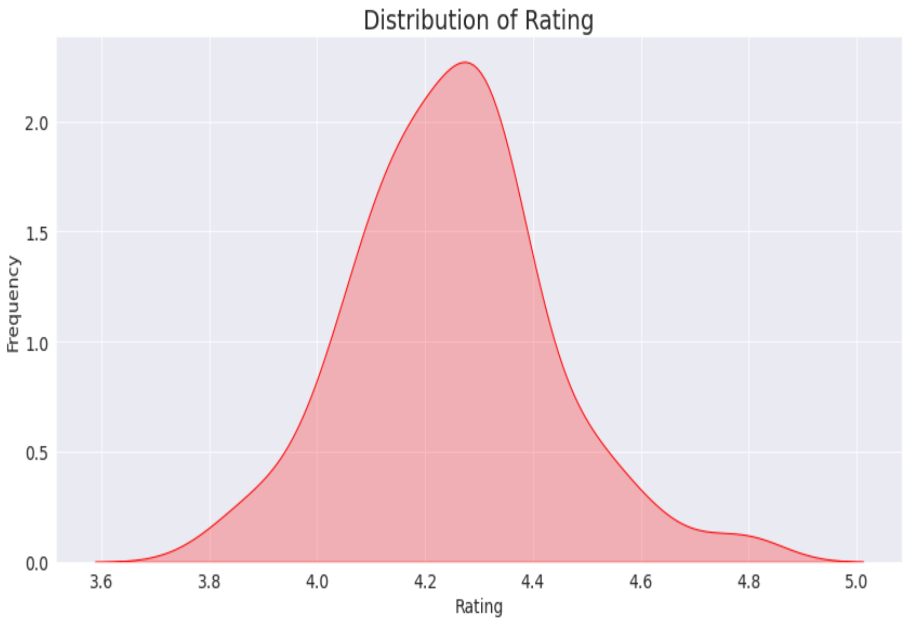
From these two plots we can conclude that, maximum number of apps present in google play store comes under Family, Games and Tools Category but as per the installations and requirements in the market place, this is not the case. Maximum installed apps comes under Games, Communication and Tools.





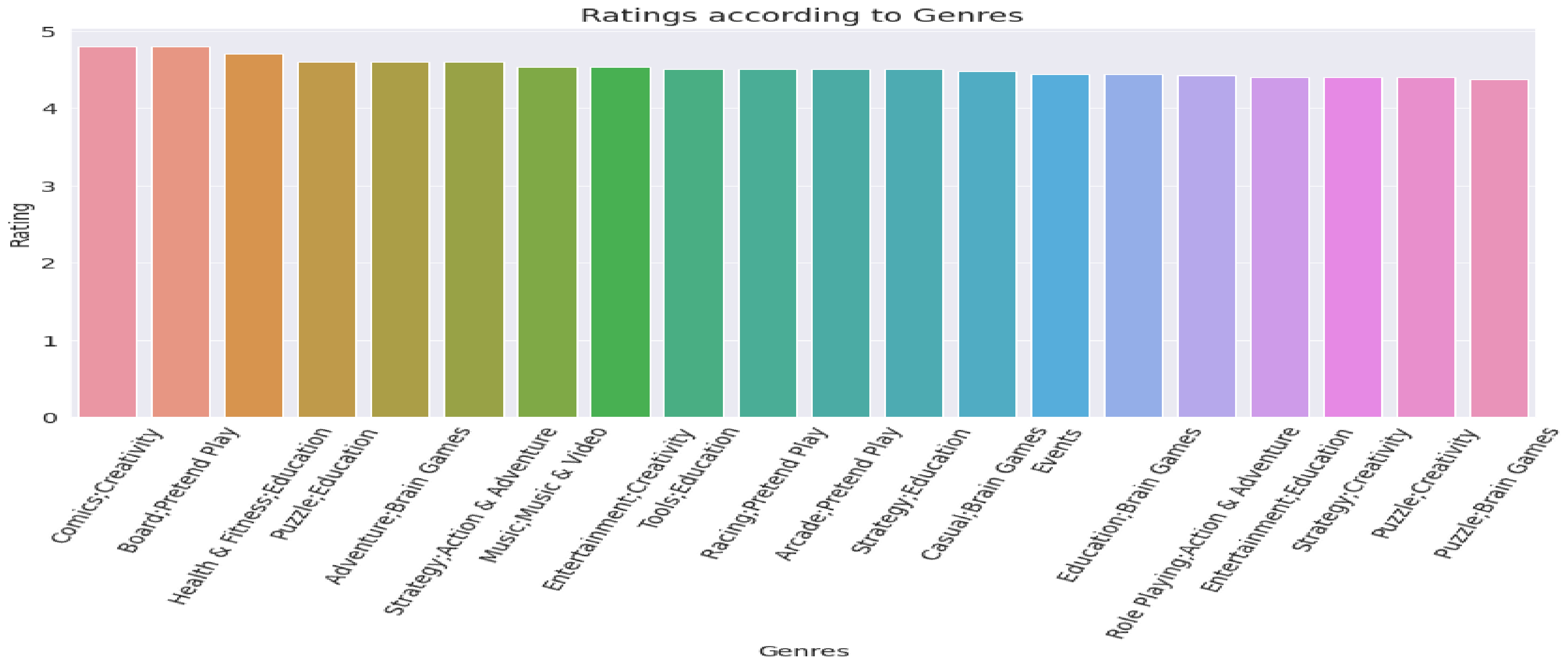
EDA of Play Store Dataset

- Average rating of application in store is around 4.3, which is very high. This plot can 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.



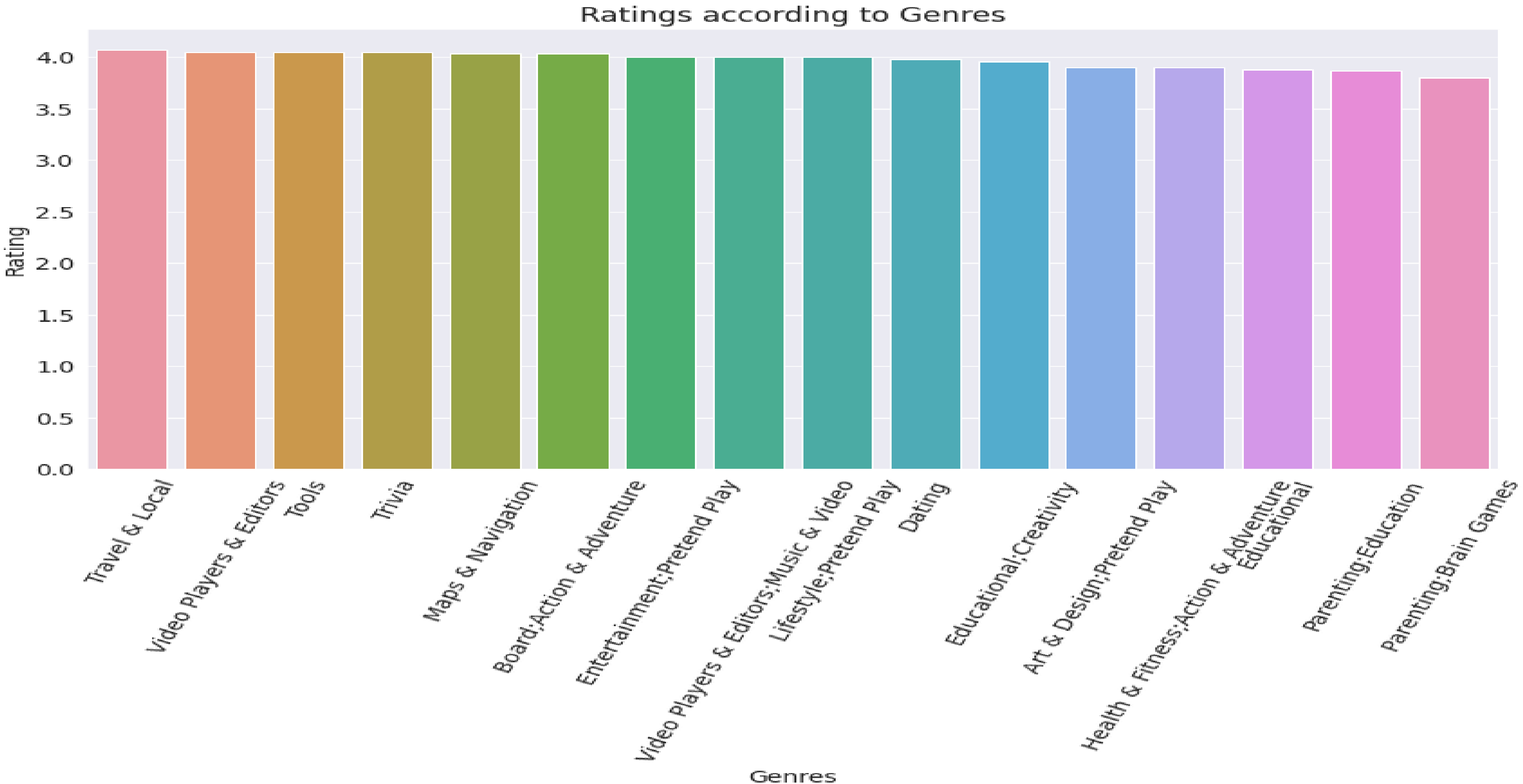
EDA of Play Store Dataset

High Rated Genres :-



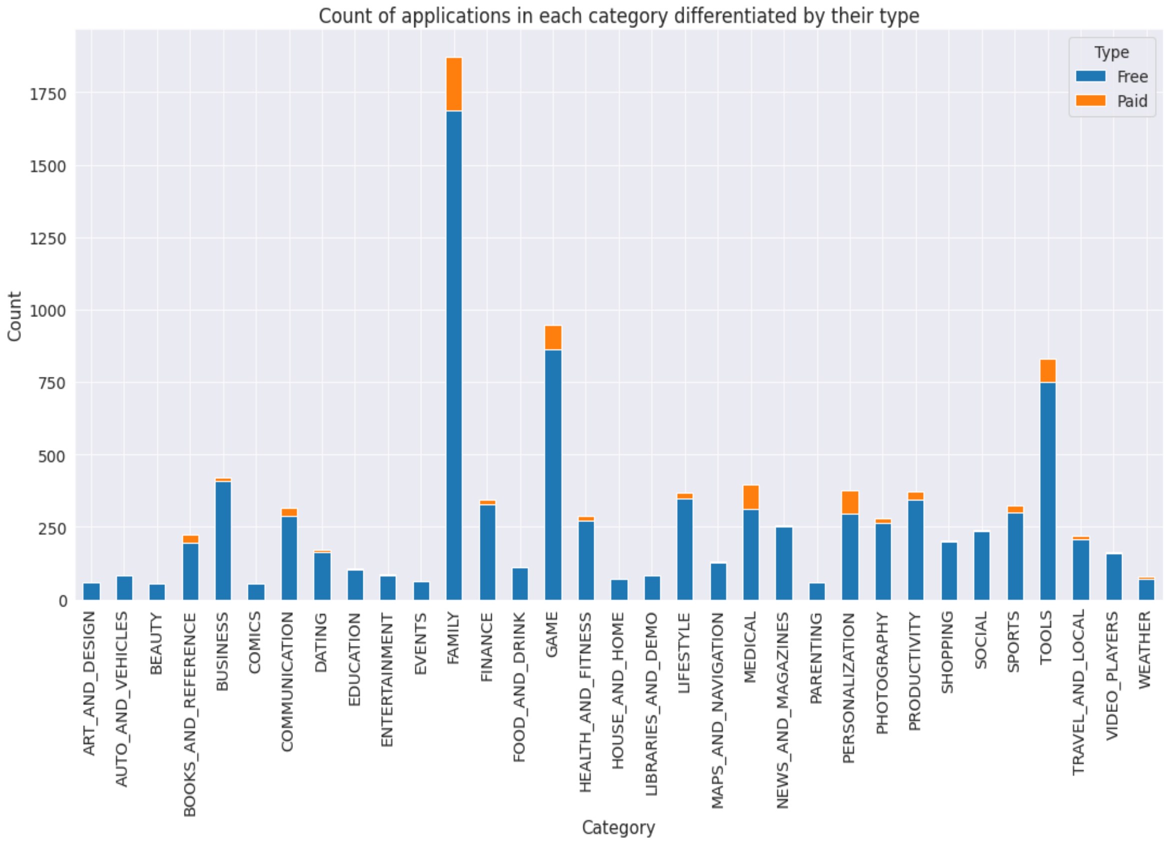
EDA of Play Store Dataset

- ***Low Rated Genres*** :-



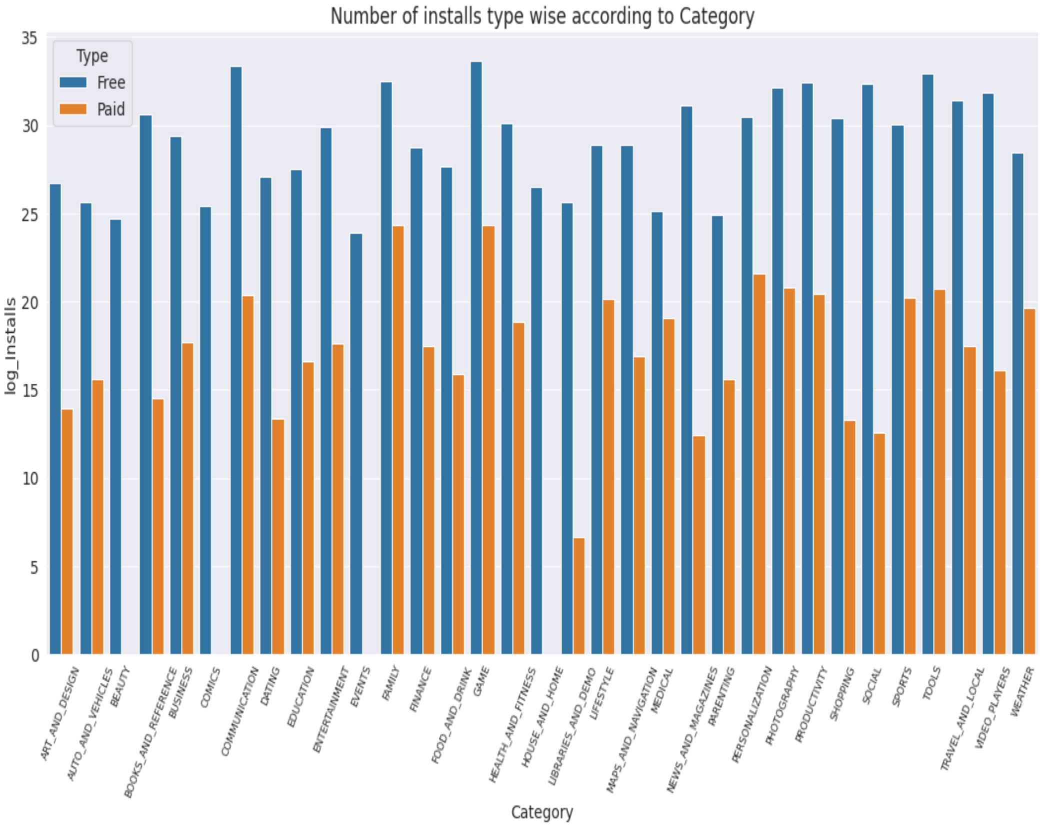
EDA of Play Store Dataset

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



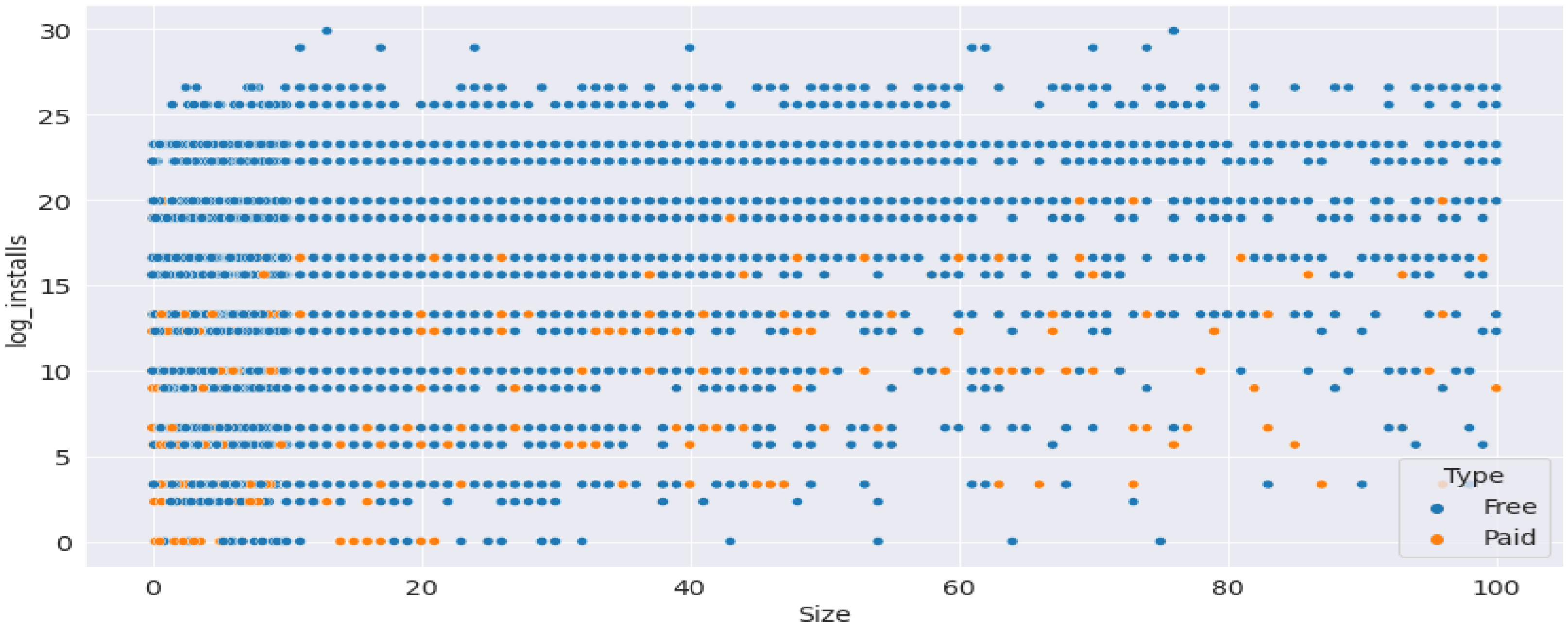
EDA of Play Store Dataset

- It can be concluded that the number of free applications installed by the user are very high when compared with the paid ones. As we have converted number of installs to it’s log, that is why the difference in the plot between free and paid apps seems to be low.

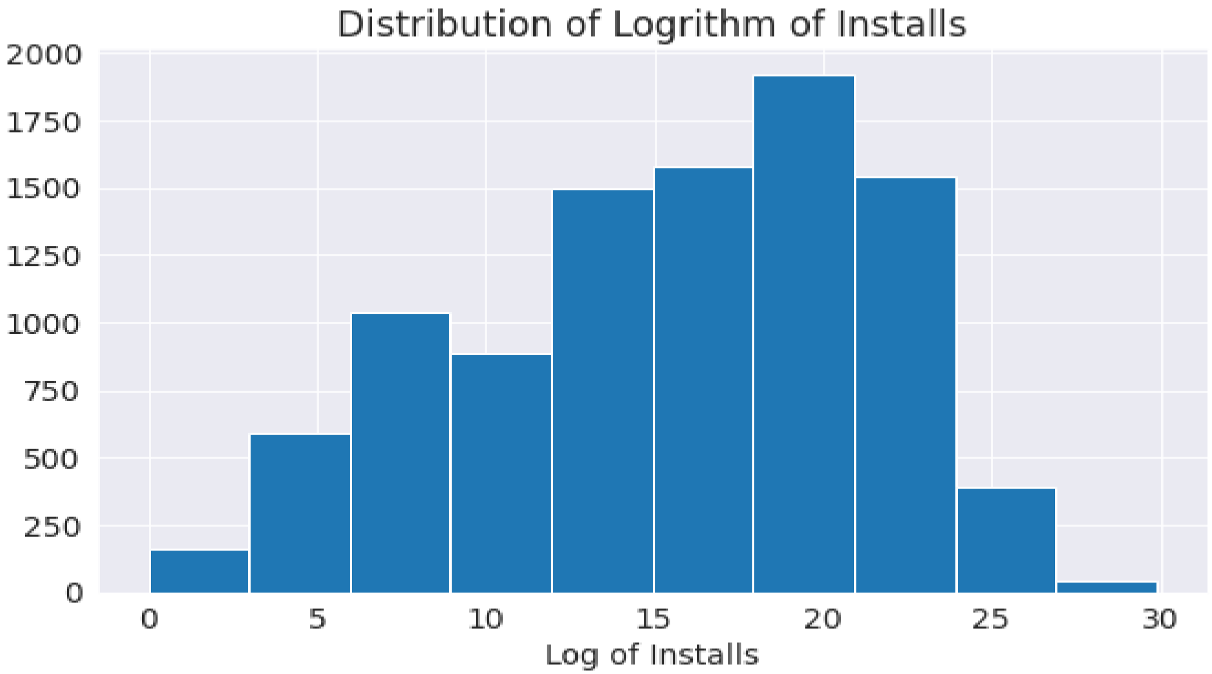


EDA of Play Store Dataset

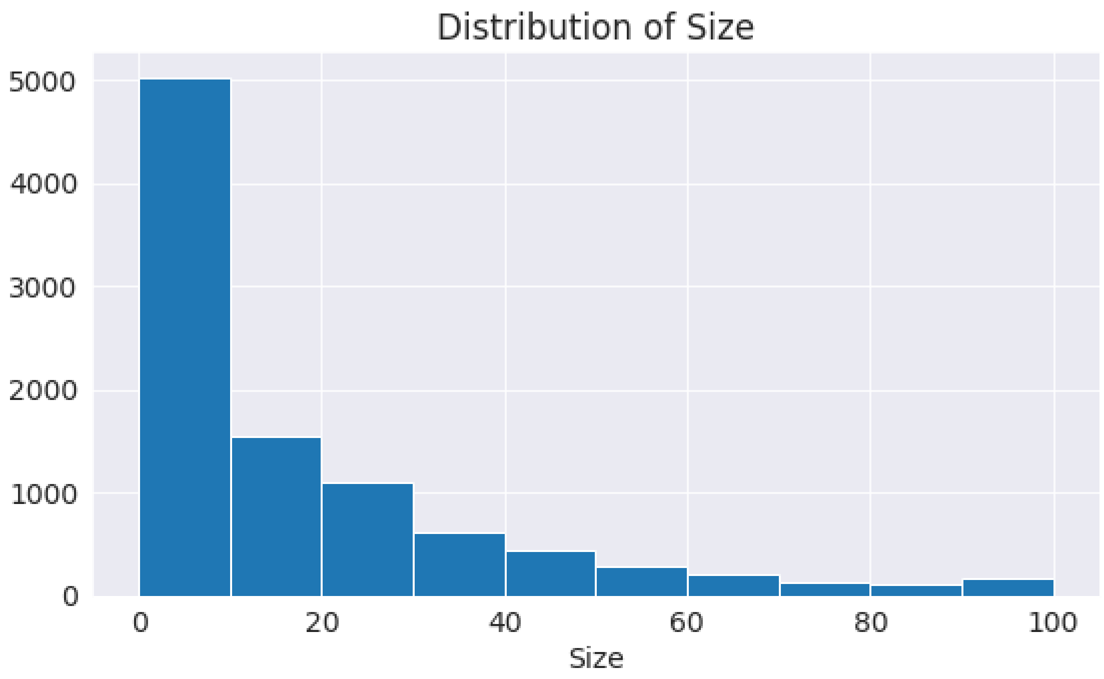
- It is clear from the plot that size may impact the number of installations. Bulky applications are less installed by the user.



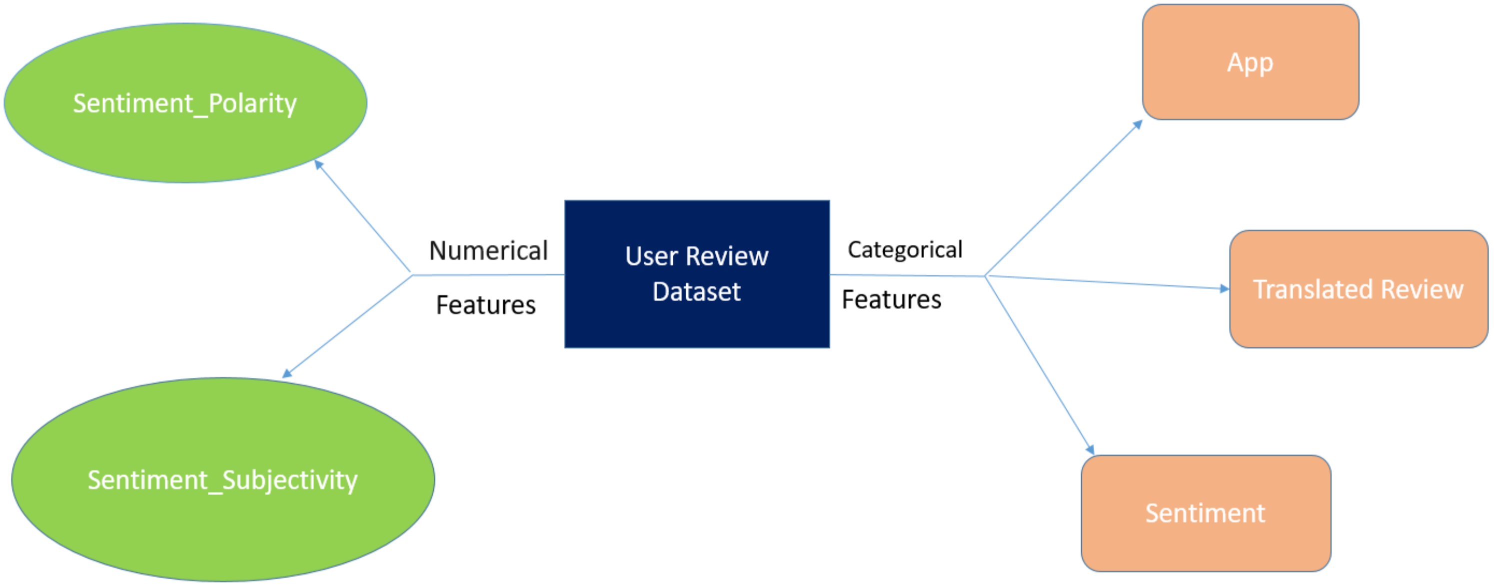
EDA of Play Store Dataset



EDA of Play Store Dataset



Data Summary of User Review Dataset



Data Summary of User Review Dataset

Numerical Features

- Sentiment_Polarity: Sentiment polarity for an element defines the orientation of the expressed sentiment, i.e., it determines if the text expresses the positive, negative or neutral sentiment of the user about the entity in consideration. It consists of numerical values and it is a vital feature for EDA.

- Sentiment_Subjectivity: Sentiment subjectivity is basically the process of determining the attitude or the emotion of the user, i.e., whether it is positive or negative or neutral. It consists of numerical values and it is a vital feature for EDA.

Data Summary of User Review Dataset

Categorical Features

- App: This column has name of the each app.
- Translated reviews: This column consists of user reviews in astring format. It is used during review analysis.
- Sentiment: It consists of or the emotion of the user, i.e., whether it is positive or negative or neutral. It plays a vital part in EDA and review analysis.

Data preprocessing and cleaning for User Review Dataset

1. Using log transformation on sentiment_count . It reduces the skewness and change the distribution to normal distribution.

```
category_sentiment['log_sentiment_count'] = np.log2(category_sentiment['Sentiment Count'])
```

2. Changing the sentiment_polarity and sentiment_subjectivity column to their absolute form.

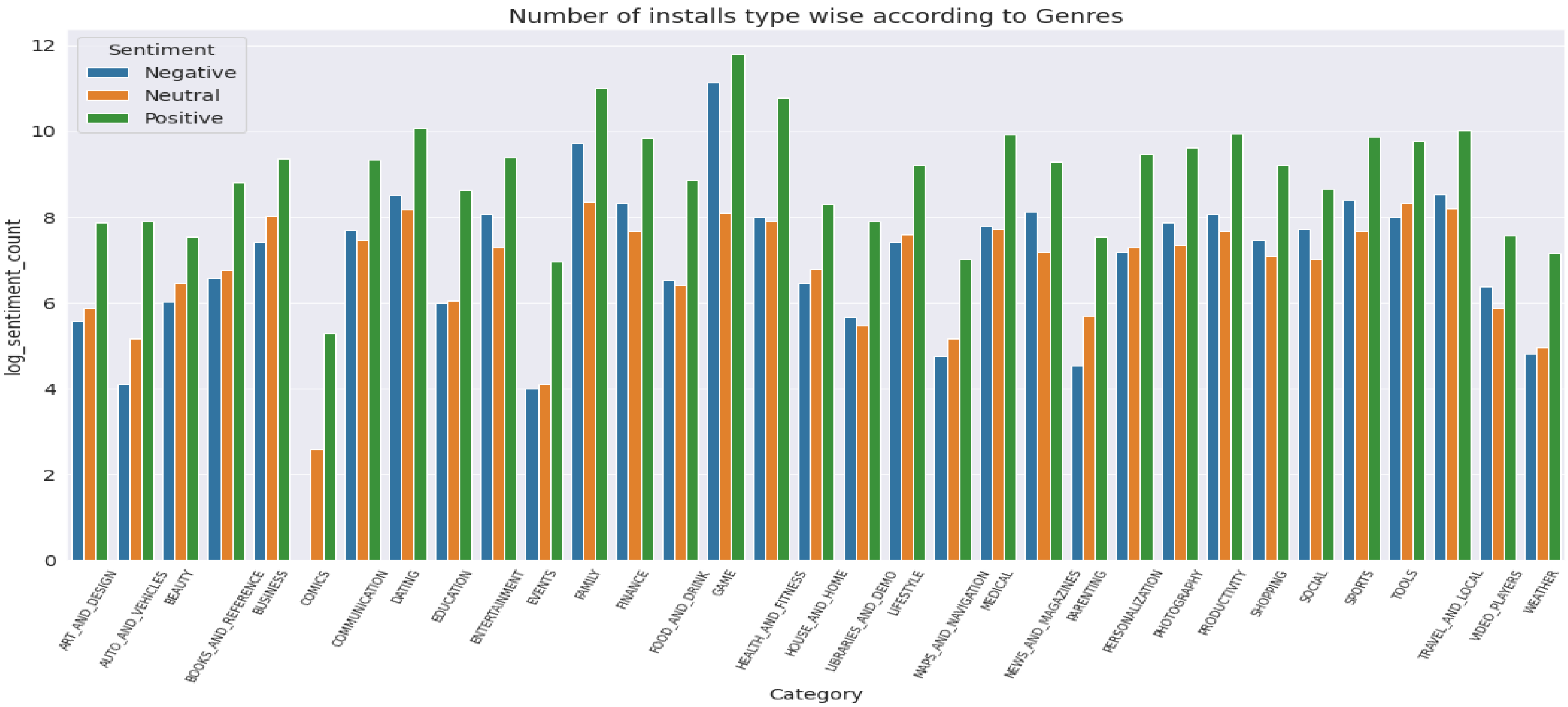
```
merged_df['Sentiment_Subjectivity'] = merged_df['Sentiment_Subjectivity'].abs()

merged_df['Sentiment_Polarity'] = merged_df['Sentiment_Polarity'].abs()
```

3. For doing EDA we used the not null values of the column.

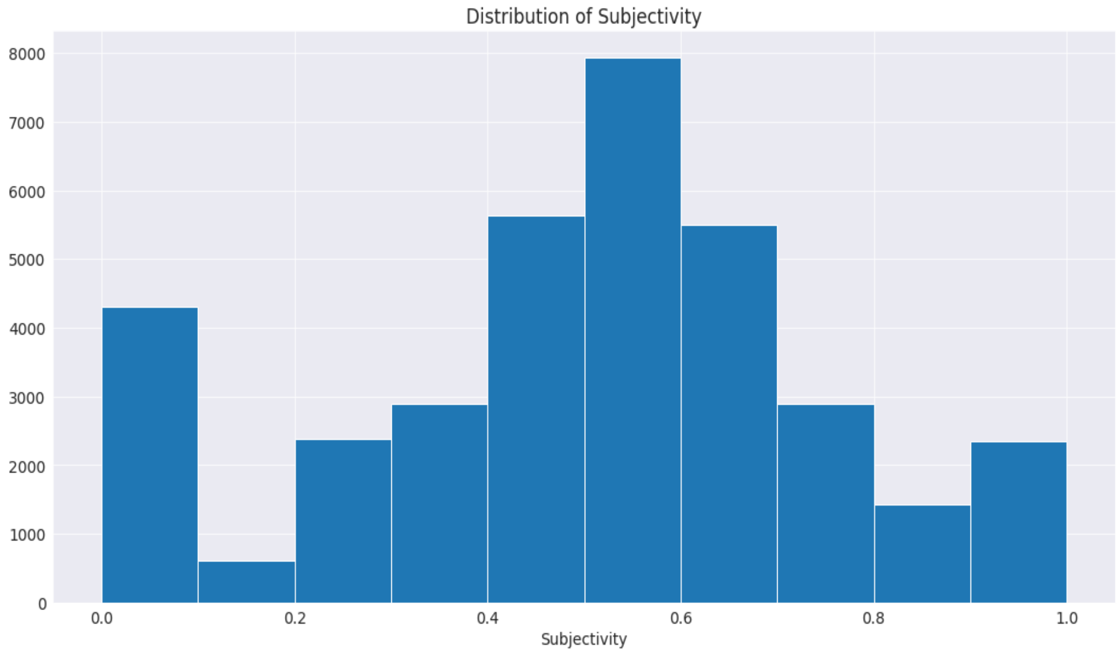
EDA of Merged Dataset

- It can be seen from the plot that the number of positive reviews are way higher than negetive and neutral ones.



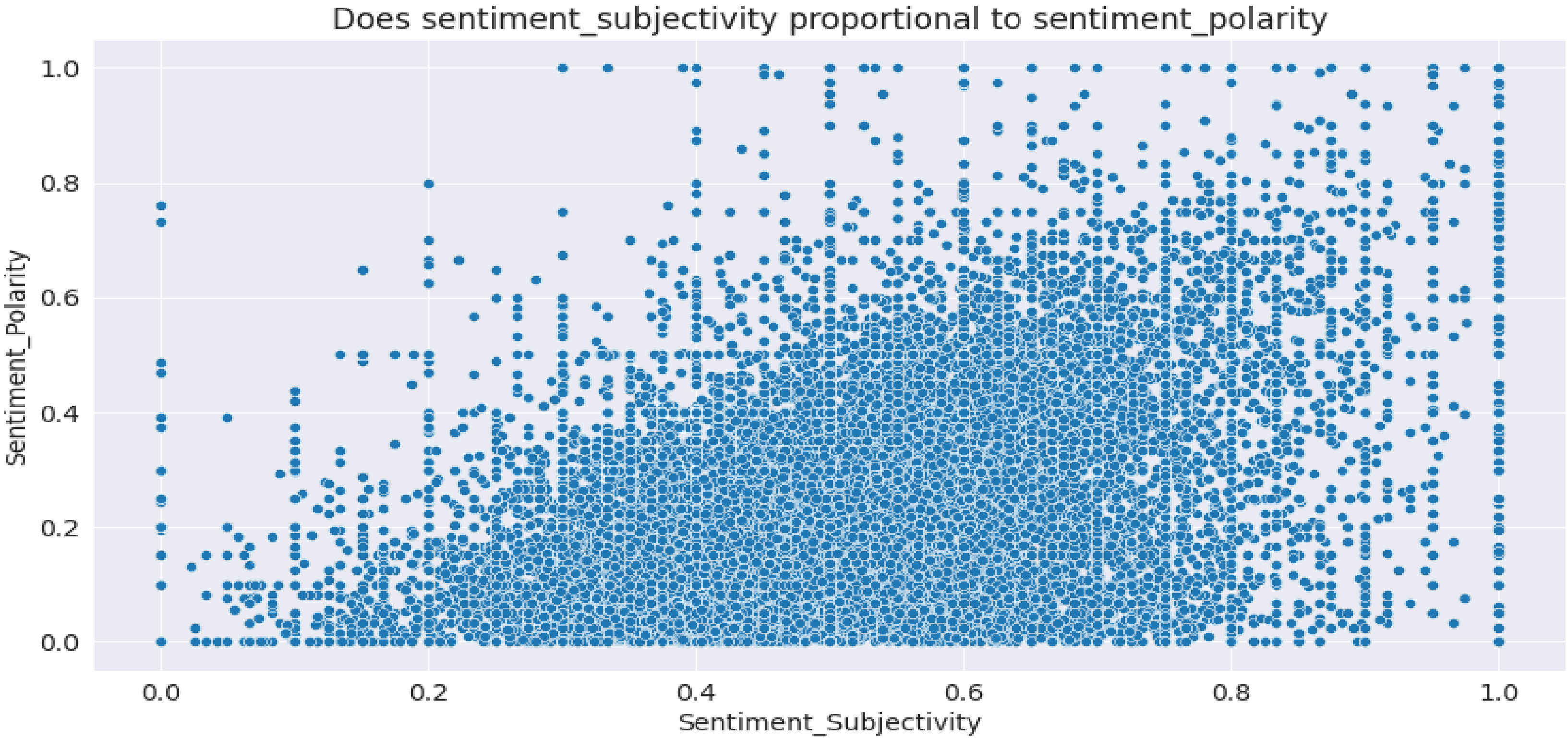
EDA of Merged Dataset

- It can be seen that maximum number of sentiment subjectivity lies between 0.4 to 0.7. From this we can conclude that maximum number of users give reviews to the applications according to their experience.



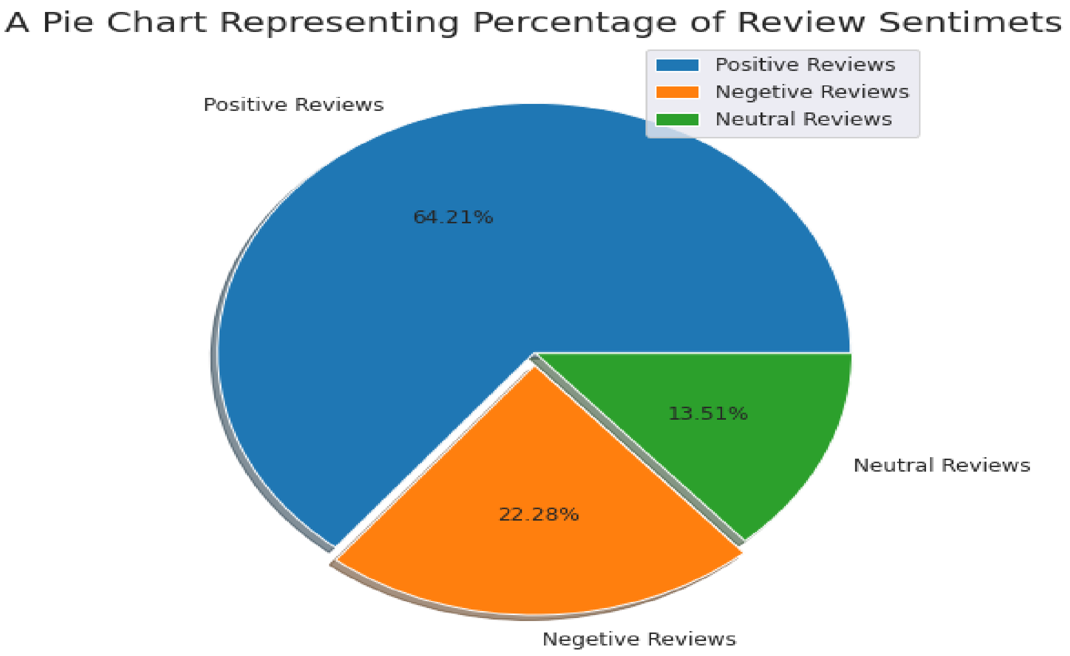
EDA of Merged Dataset

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



EDA of Merged Dataset

EDA of Merged Dataset



CONCLUSION

1. Maximum installed apps comes under communications ,tools and productivity on the basis genres.
2. Maximum installed apps comes under communications ,tools and games on the basis of category.
3. In our dataset, the majority of apps in Family, Food & Drink and Tools, as well as Social categories were free to install. At the same time Family, Sports, Tools and Medical categories had the biggest number of paid apps available for download
4. It can be concluded that maximum number of applications present in the dataset are of small size.
5. It can concluded that maximum number of sentiment subjectivity lies between 0.4 to 0.7. From this we can conclude that maximum number of users give reviews to the applications, according to their experience.
6. From the review sentiments 64.21% are positive reviews , 22.8% are negative reviews and 13.51% are neutral reviews.

Future works

- Exploring the correlation between the size of the app and the version of Andro id on the number of installs.

- Exploring reviews and sentiment of the users as per the the category of the a pplication.

- Treating the outlier of the features.


Thank You!