Capstone Project - 1

PlayStore App Review Analysis

Team Members

Priya Debrani

Vikash Kumar

Let’s Start Analysis

- Defining problem statement
- Data Pre-processing
- EDA
- Data Visualization
- Conclusion
- Business Approach

Content :

- Introduction
- Problem definition
- Description of Dataset
- Data cleaning
- Data Analysis & Visualization
- Important points get after Data visualization
- Work on Sentiment Analysis (2nd Data Set)
- Conclusion

Data Pipeline

- **Data processing-1**: In this first part we've removed unnecessary features. Since there were nearly many columns with all null values.
- **Data processing-2**: In this part, we manually go through each features selected from part 1, And encoded the categorical features ,changed the columns containing date time values .
- **EDA**: In in this part, we do some exploratory data analysis (EDA) on the features selected in part-1 and 2 to see the trend.
- **Create a model**: Finally, In this last but not the last part, we create models. Creating a model is also not an easy task. It's also an iterative process. we show how to start with a with a simple model, then slowly add complexity for better performance.

Data Summary

There are two dataset: PlayStore Data & User Review data

1. **Play Store Data:**

- **App** : The name of the app
- **Category** : The category of the app
- **Rating** : The rating of the app in the Play Store
- **Reviews** : The number of reviews of the app
- **Size** : The size of the app
- **Install** : The number of installs of the app
- **Type** : The type of the app (Free/Paid)
- **Content Rating** :The appropriate target audience of the app
- **Genres** : The genre of the app
- **Last Updated** : The date when the app was last updated
- **Current Ver** : The current version of the app
- **Android Ver** : The minimum Android version required to run the app

Data Summary

2. User Review Data:

- **App** : An app name
- **Sentiment** : Sentiment given to an app by users (i.e Positive,Neutral, Negative)
- **Sentiment Polarity** :The polarity of sentiment measures how negative or positive the context is. In the data we have, the polarity ranges from +1(Positive) to -1(Negative).
- **Sentiment Subjectivity** :The subjectivity of a sentiment is how likely that sentiment is to be based on data or factual information, versus personal opinions or public notions.

Variable Summary:-

- **App** : The name of the app
- **Category** : The category of the app
- **Rating** : The rating of the app in the Play Store
- **Reviews** : The number of reviews of the app
- **Size** : The size of the app
- **Installs** : The number of installs of the app
- **Type** : The type of the app (Free/Paid)
- **Price** : The price of the app (0 if it is Free)
- **Content Rating** :The appropriate target audience of the app
- **Genres** : The genre of the app
- **Last Updated** : The date when the app was last updated
- **Current Ver** : The current version of the app
- **Android Ver** : The minimum Android version required to run the app

Dataset Info & Selection

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 122662 entries, 0 to 122661
Data columns (total 14 columns):
 #   Column                  Non-Null Count   Dtype  
---  ------                  --------------   -----  
 0   App                     122662 non-null  object 
 1   Category                122662 non-null  object 
 2   Rating                  122662 non-null  float64
 3   Reviews                 122662 non-null  int64  
 4   Size                    75432 non-null   float64
 5   Installs                122662 non-null  int64  
 6   Type                    122662 non-null  object 
 7   Price                   122662 non-null  float64
 8   Content Rating          122662 non-null  object 
 9   Genres                  122662 non-null  object 
 10  Translated_Review       72605 non-null   object 
 11  Sentiment               72615 non-null   object 
 12  Sentiment_Polarity      72615 non-null   float64
 13  Sentiment_Subjectivity  72615 non-null   float64
dtypes: float64(5), int64(2), object(7)
memory usage: 14.0+ MB
```

Data Cleaning

Data cleaning is not just a case of removing erroneous data, although that's often part of it. The majority of work goes into detecting rogue data and (wherever possible) correcting it.

Data Cleaning Step:

- **Removing unwanted observation :** Duplicate/redundant or irrelevant values deletion.
- **Missing Data handling** : Fixing issue of unknown missing values.
- **Structural error solving** : Fixing problems with mislabeled classes classes, datatype names of features, same attribute with different name etc.

Preparing dataset for modeling

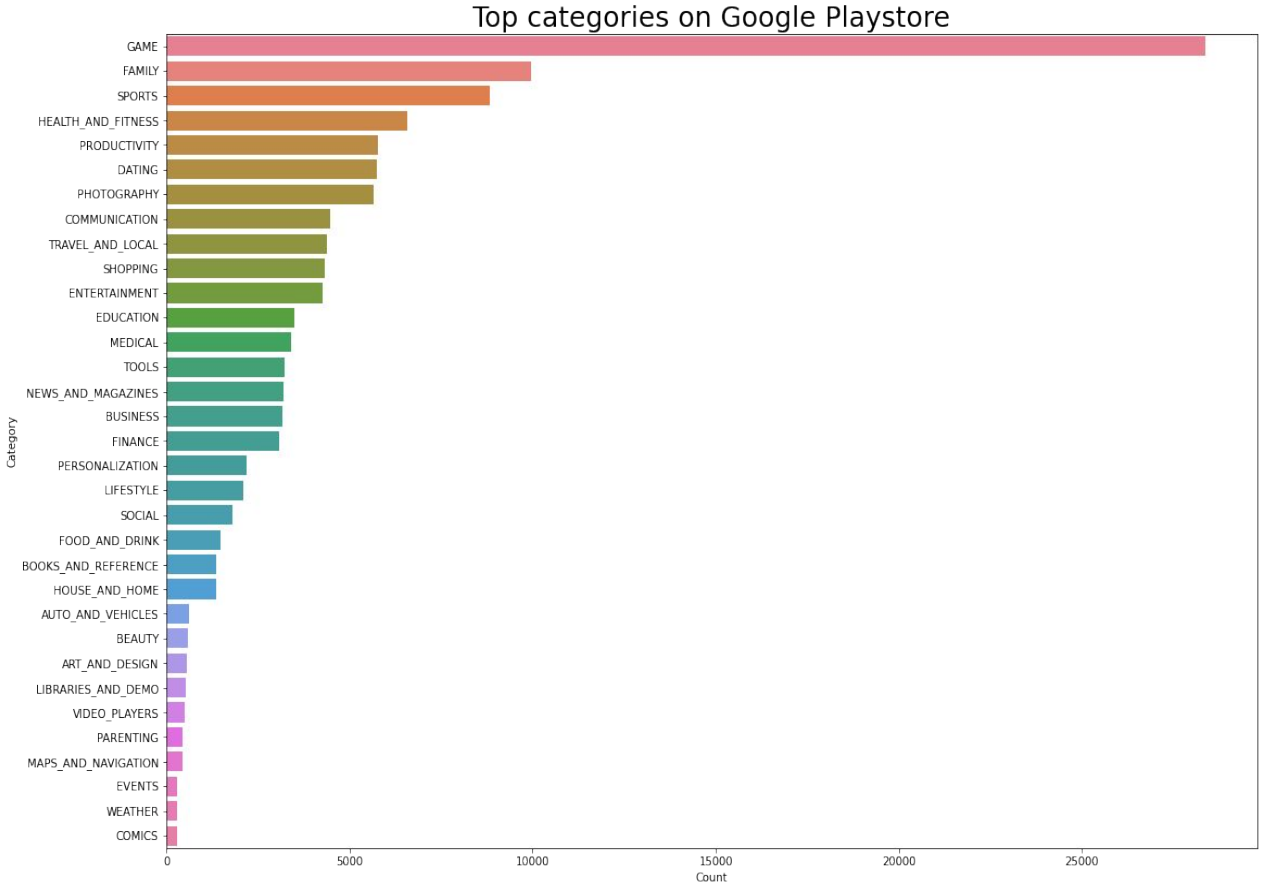
index	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Translated_Review	Sentiment	Sentiment_Polarity	Sentiment_Subjectivity
0	Coloring book moana	ART_AND_DESIGN	3.9	967	14000000.0	500000	Free	0.0	Everyone	Art & Design;Pretend Play	A kid's excessive ads. The types ads allowed app, let alone kids	Negative	-0.25	1.0
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14000000.0	500000	Free	0.0	Everyone	Art & Design;Pretend Play	It bad >:(	Negative	-0.7249999999999999	0.8333333333333333
2	Coloring book moana	ART_AND_DESIGN	3.9	967	14000000.0	500000	Free	0.0	Everyone	Art & Design;Pretend Play	like	Neutral	0.0	0.0
3	Coloring book moana	ART_AND_DESIGN	3.9	967	14000000.0	500000	Free	0.0	Everyone	Art & Design;Pretend Play	NaN	NaN	NaN	NaN
4	Coloring book moana	ART_AND_DESIGN	3.9	967	14000000.0	500000	Free	0.0	Everyone	Art & Design;Pretend Play	I love colors inspyering	Positive	0.5	0.6
5	Coloring book moana	ART_AND_DESIGN	3.9	967	14000000.0	500000	Free	0.0	Everyone	Art & Design;Pretend Play	I hate	Negative	-0.8	0.9
6	Coloring book moana	ART_AND_DESIGN	3.9	967	14000000.0	500000	Free	0.0	Everyone	Art & Design;Pretend Play	NaN	NaN	NaN	NaN
7	Coloring book moana	ART_AND_DESIGN	3.9	967	14000000.0	500000	Free	0.0	Everyone	Art & Design;Pretend Play	omgggggg	Neutral	0.0	0.0
8	Coloring book moana	ART_AND_DESIGN	3.9	967	14000000.0	500000	Free	0.0	Everyone	Art & Design;Pretend Play	It cute.	Positive	0.5	1.0
9	Coloring book moana	ART_AND_DESIGN	3.9	967	14000000.0	500000	Free	0.0	Everyone	Art & Design;Pretend Play	I love	Positive	0.5	0.6

Show 10 per page

Top categories on google playstore

Insights:

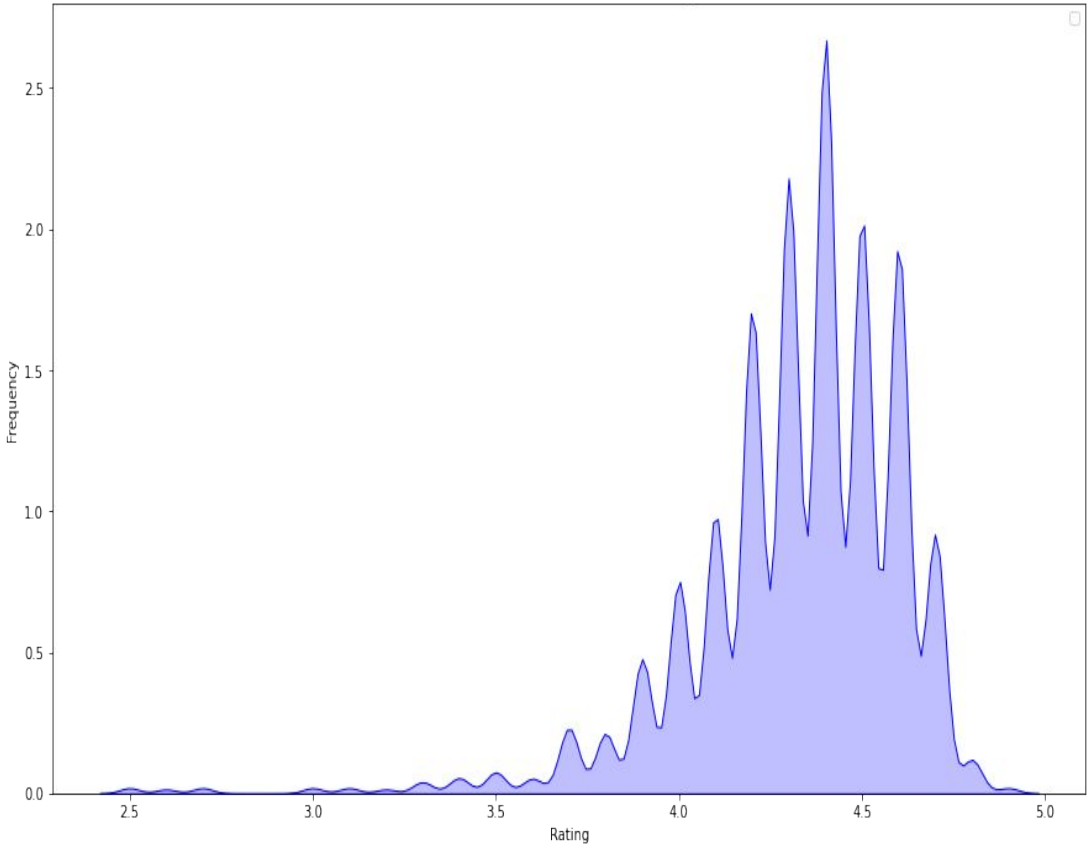
There are all total of 33 categories in the dataset from the above output we can come to the conclusion that in the play store most of the apps are under Game & Family category and least are of Events, Weather & Comics Category.



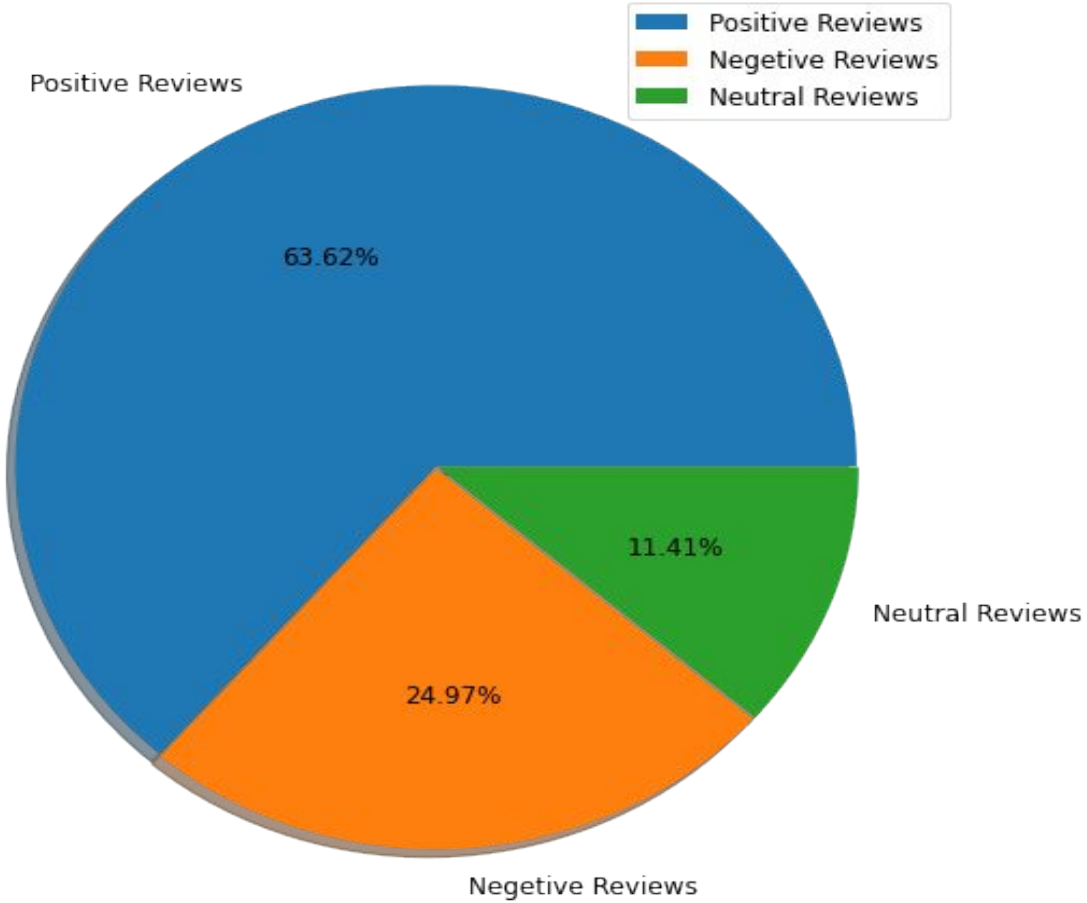
Distribution Rating:-

Insight :

we can come to the conclusion that most of the apps in the google play store are rated between 3.8 to 4.5.



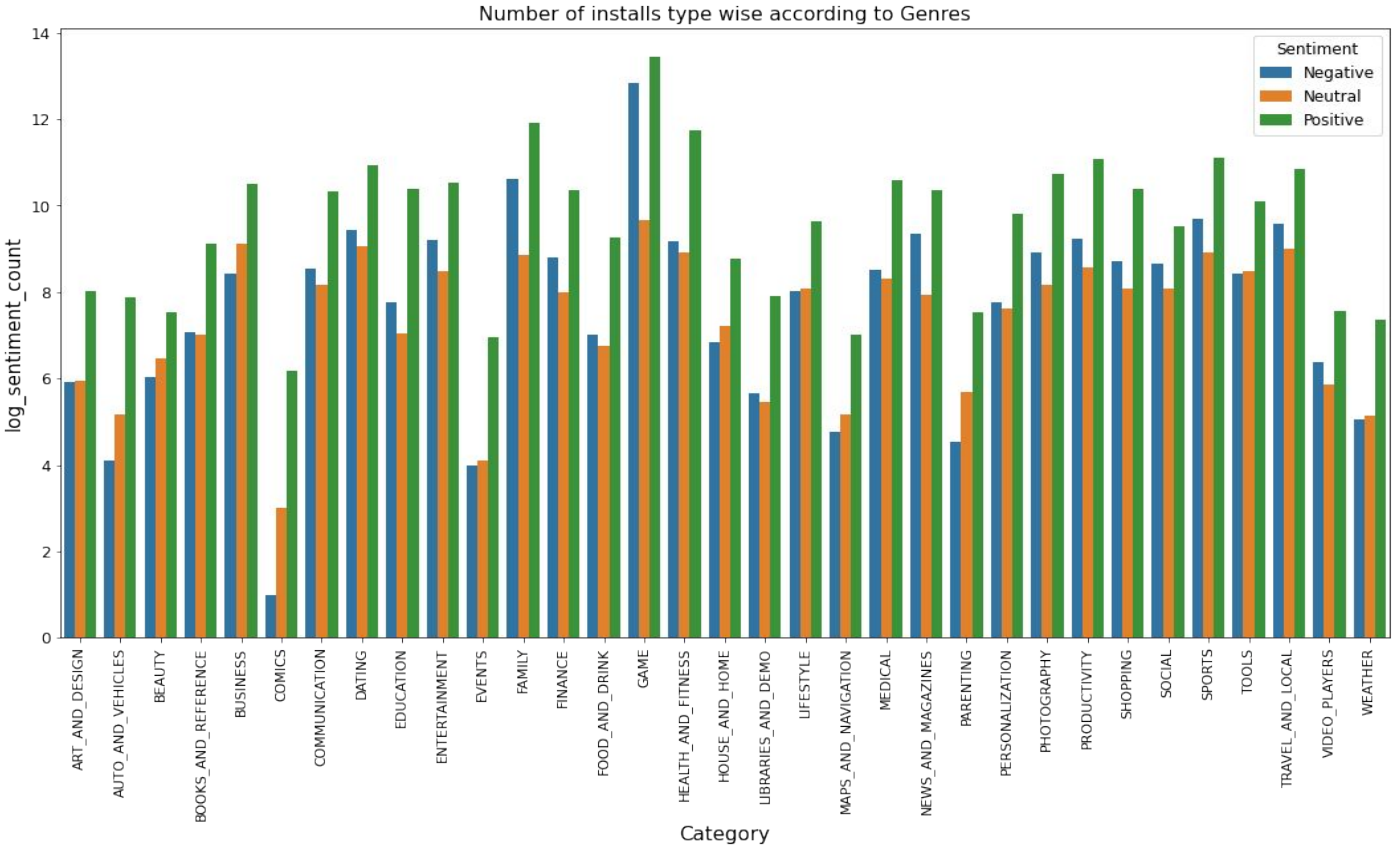
A Pie Chart Representing Percentage of Review Sentiments



Insights :

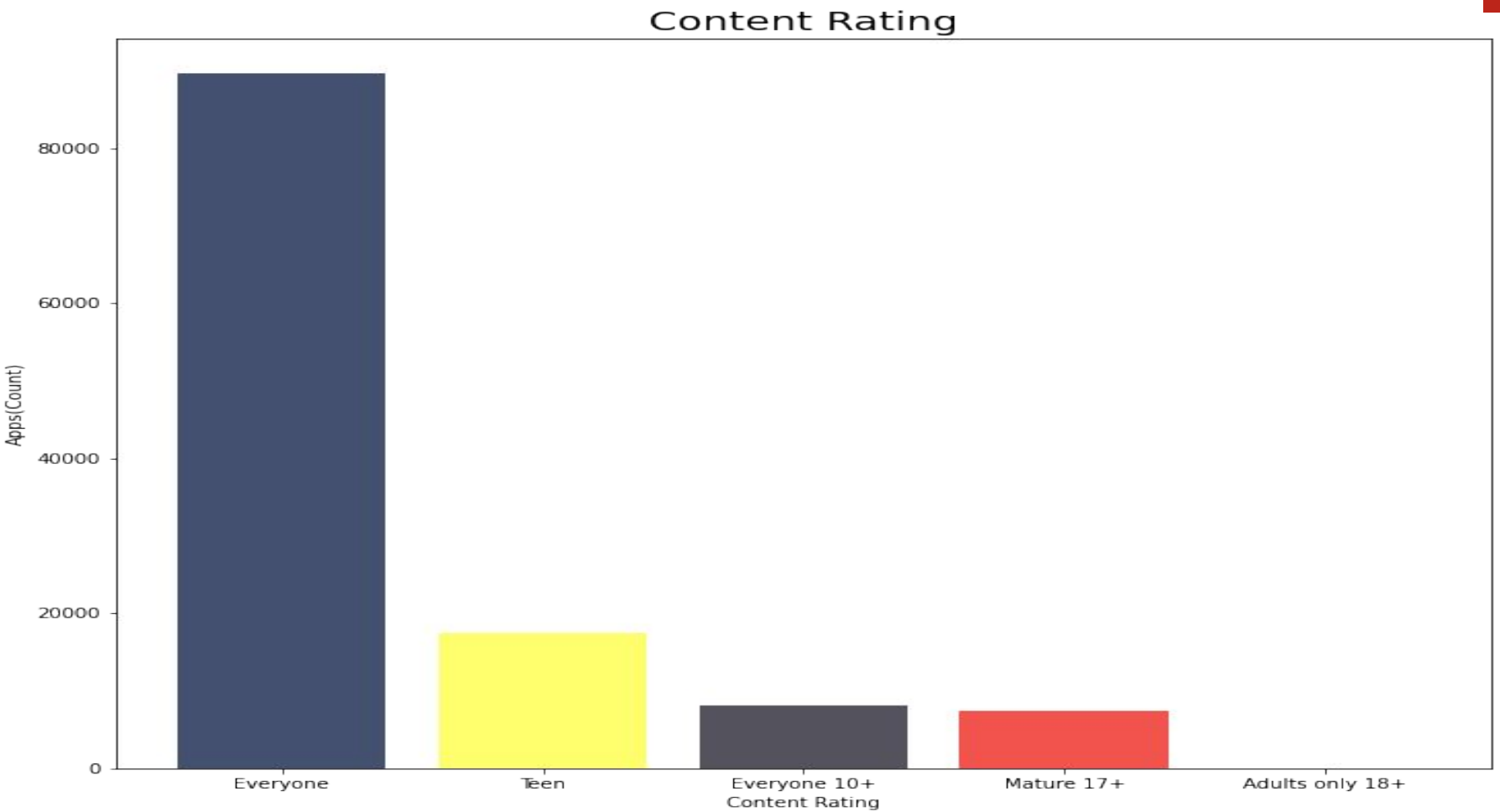
The insights found from the above chart is that 63.62% reviews are positive reviews, 24.97% are negative reviews and 11.41% are neutral reviews.

Number Of Installs Type Wise:-

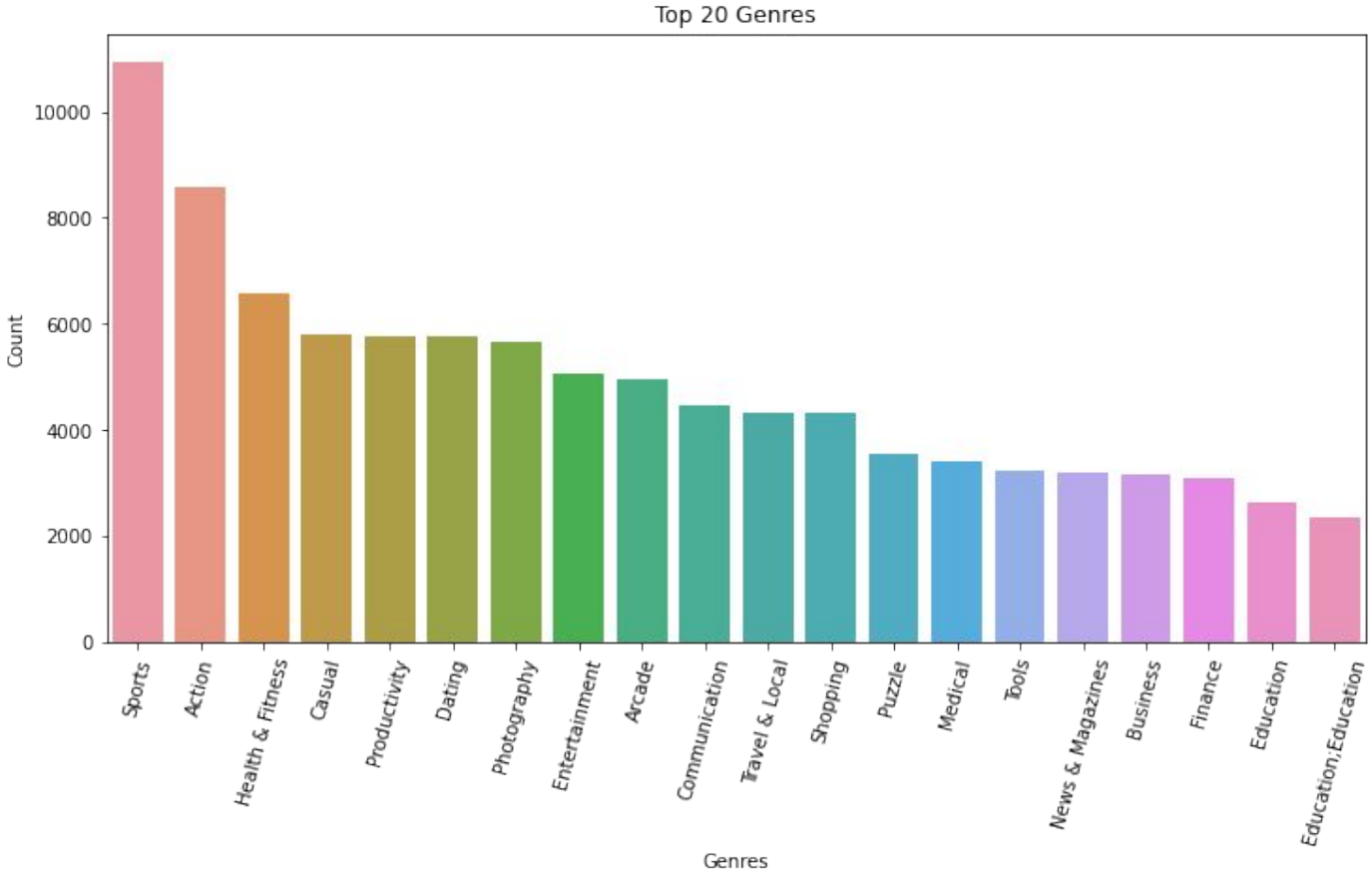


Insights :

It can be concluded that the number of free applications installed by the user are high when compared with the paid ones.


Content Rating
80000
60000
40000
20000
0
Apps(Count)
Everyone
Teen
Everyone 10+
Mature 17+
Adults only 18+
Content Rating

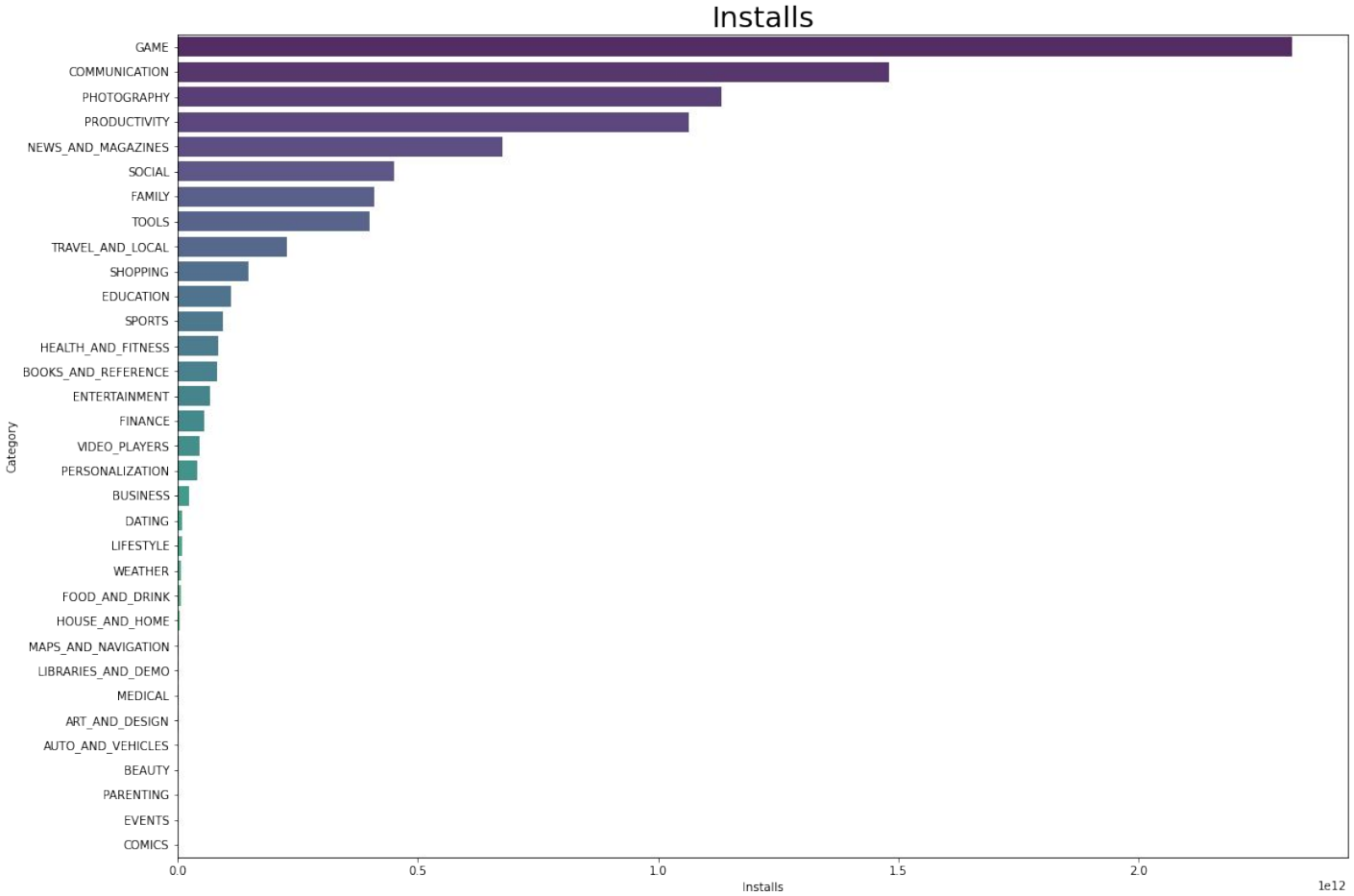
Top 20 Genres according to categories:-



Insights :

we can see that in the Sports, Action,Health & Fitness and Causal has the highest installs. In the same way by passing different category names to the function, we can get the top 20 installed apps.

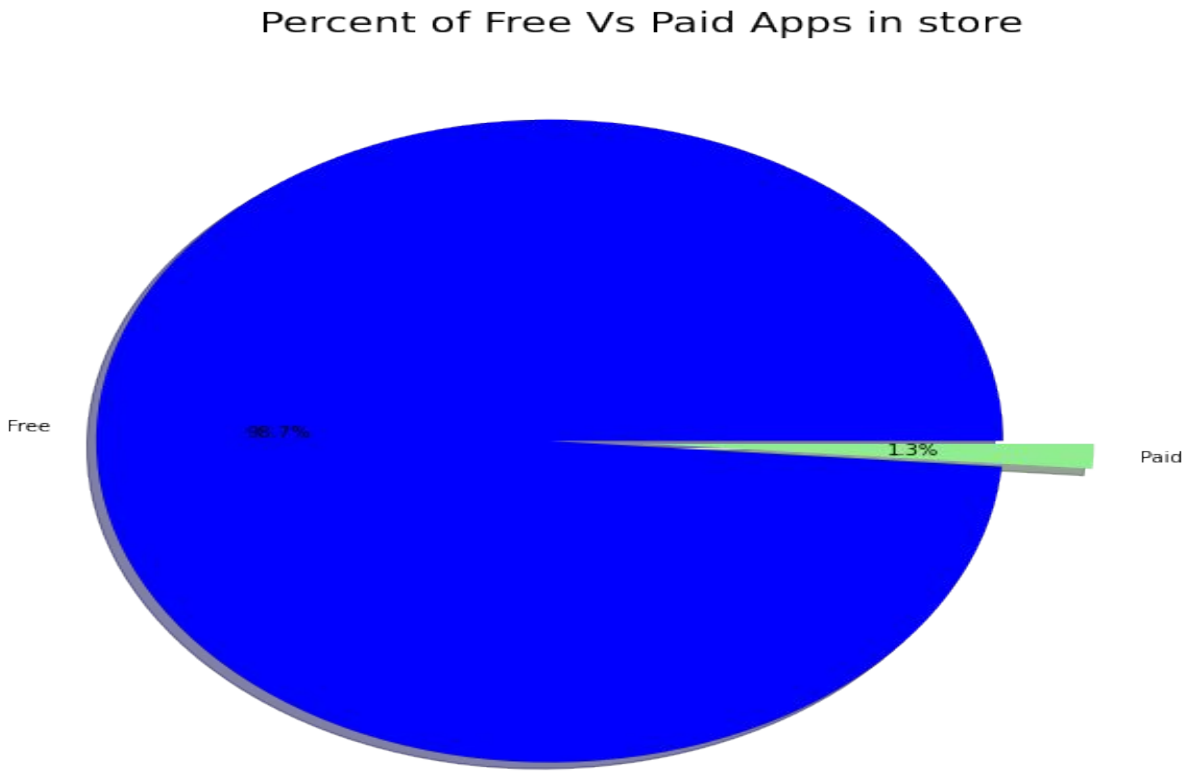
Top Installs:-



Insights :

It can be interpreted that the top categories with the highest installs are Game, Communication,Photography, Productivity News & Magazines, & Social.

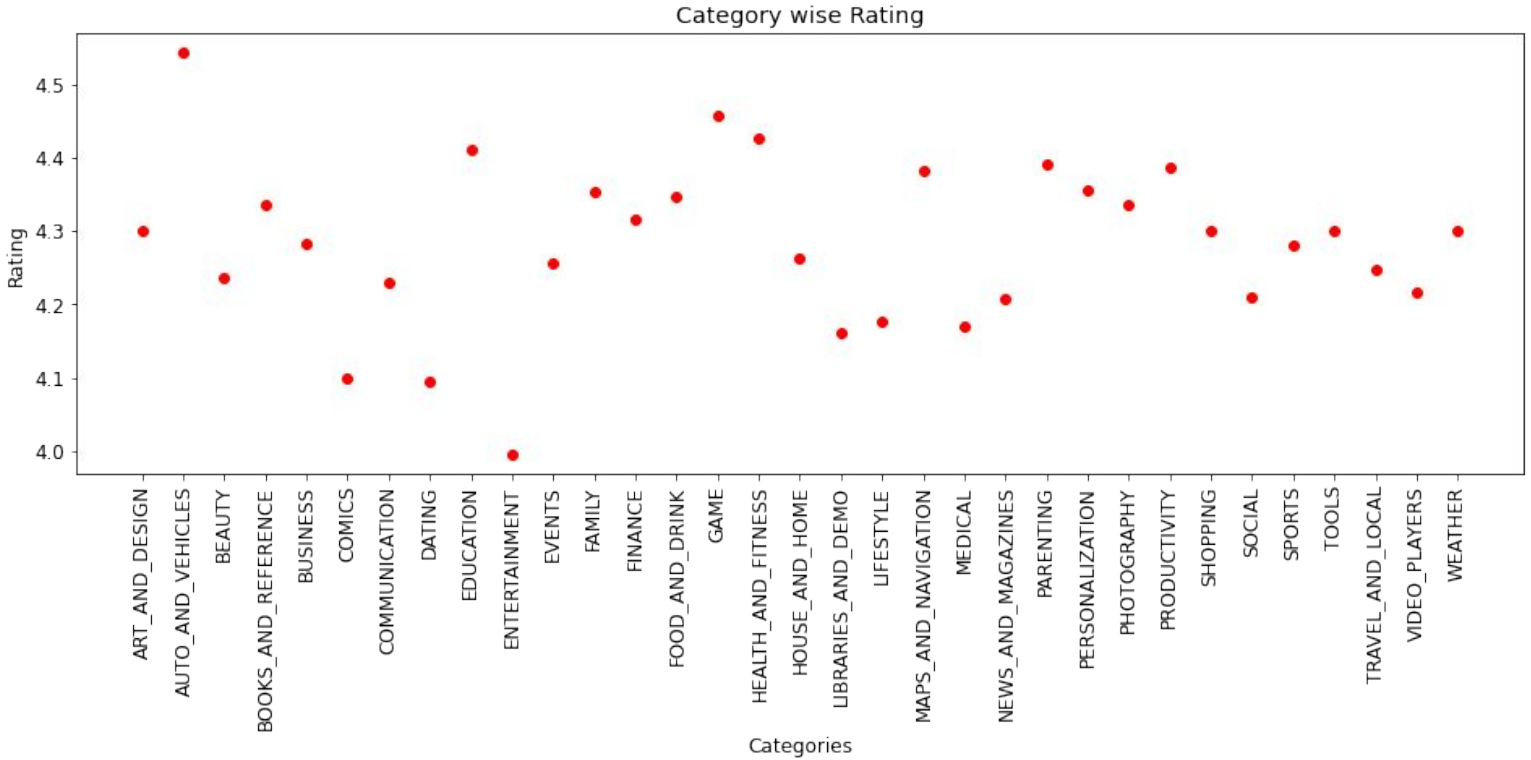
Percent of Free Vs Paid Apps:-



Insights :

we can see that 98.7%(Approx.) of apps in the google play store are free and 1.3%(Approx.) are paid.

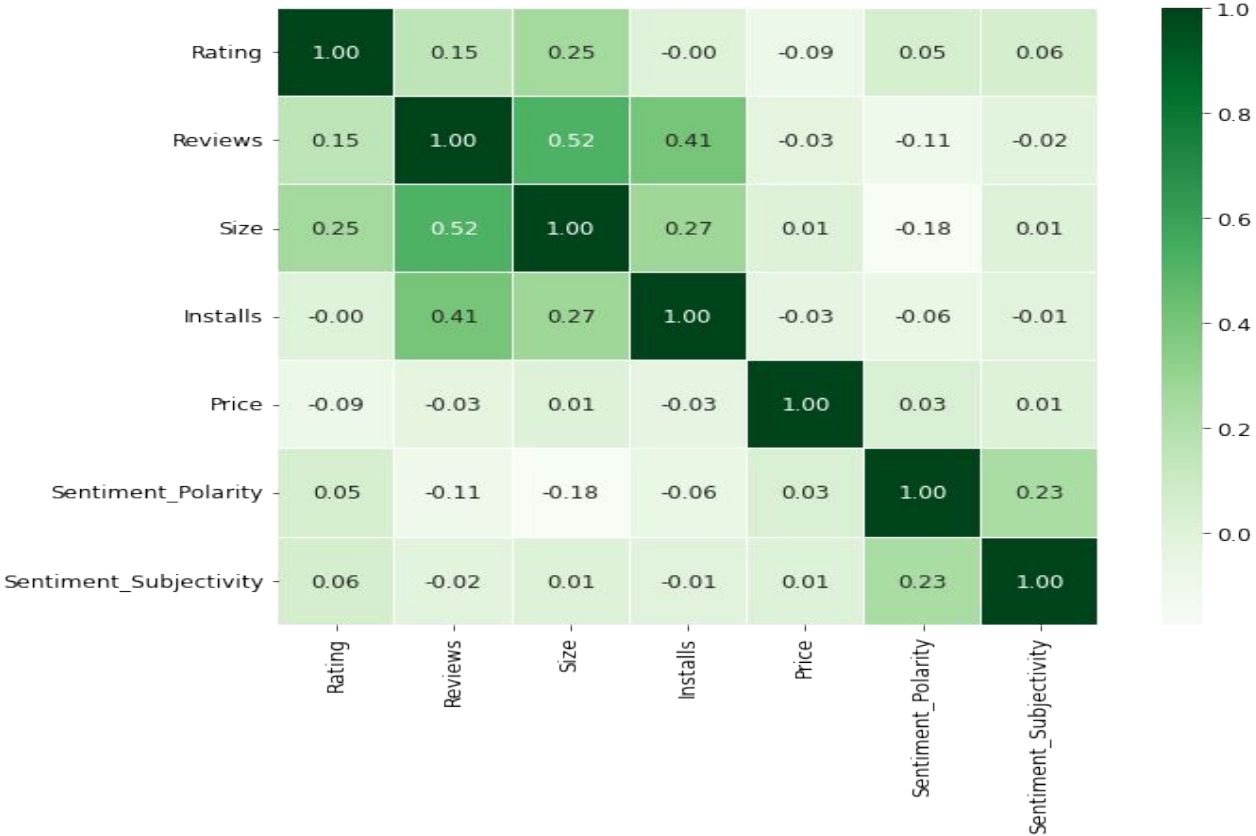
Category Wise Rating:-



Insights:

The insights found from the above data is that Auto_and_vehicles and Games category has gotten the best ratings.

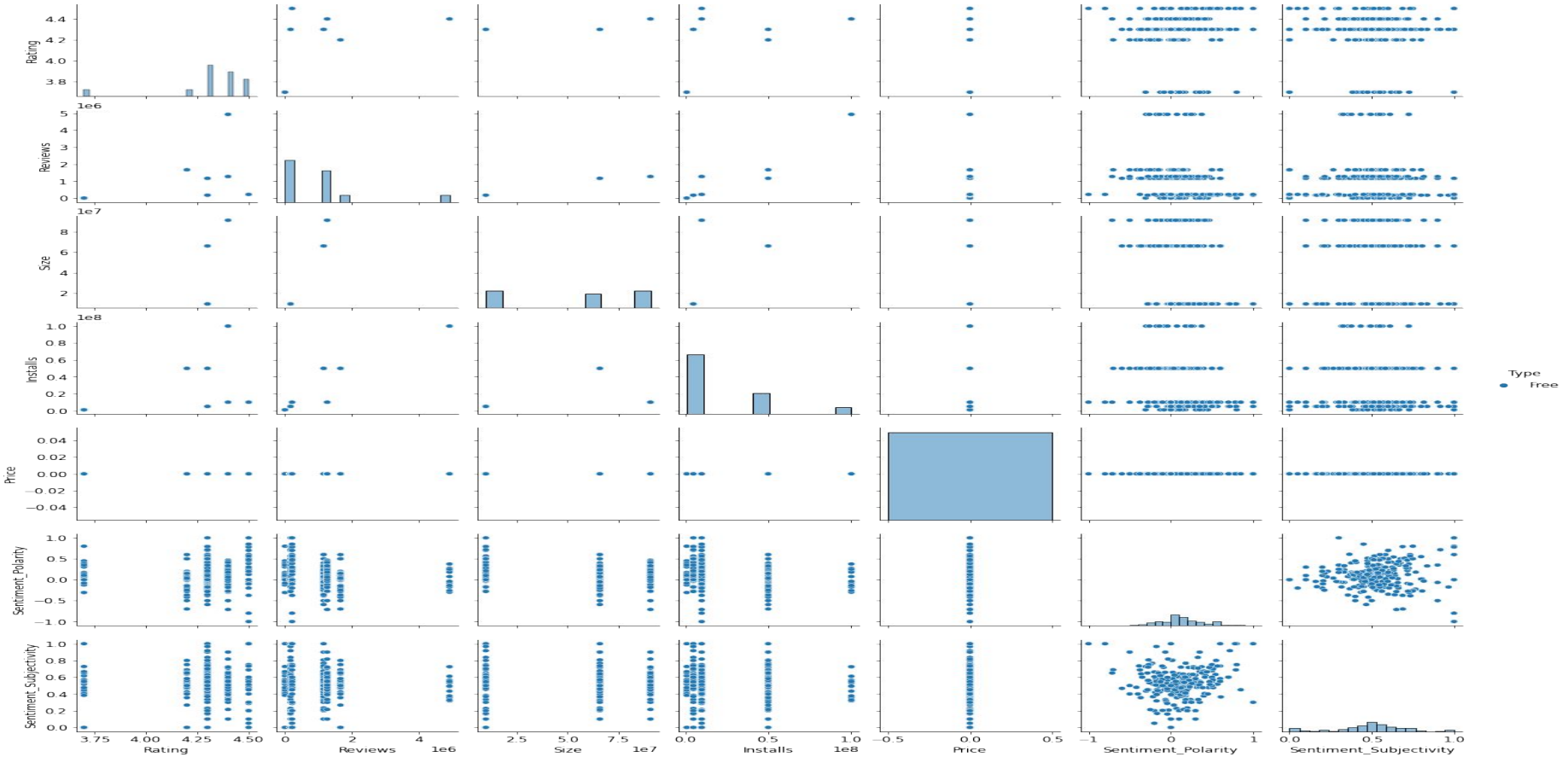
Heatmap



Insights :

A moderate positive correlation of 0.52 exists between the number of reviews and Size. This means that customers tend to download a given app more if it has been reviewed by a larger number of people. This also means that many active users who download an app usually also leave back a review or feedback. So, getting your app reviewed by more people maybe a good idea to increase your app's capture in the market!

Pairplot


Rating
Reviews
Size
Installs
Price
Sentiment_Polarity
Sentiment_Subjectivity
1e6
1e7
1e8
Type
Free

Conclusion

- The Google Play Store Apps report provides some useful details regarding he trending of the apps in the play store. As per the graphs visualizations shown above, most of the trending apps (in terms of users' installs) are from the categories like GAME, COMMUNICATION, and TOOL even though the amount of available apps from these categories are twice as much lesser than the category FAMILY but still used most.
- The trending of these apps are most probably due to their nature of being able to entertain or assist the user. Besides, it also shows a good trend where we can see that developers from these categories are focusing on the quality instead of the quantity of the apps.
- Other than that, the charts shown above actually implies that most of the apps having good ratings of above 4.0 are mostly confirmed to have high amount of reviews and user installs.

- The size and price shouldn't reflect that apps with high rating are mostly big in size and pricey as by looking at the graphs they are most probably are due to some minority. Furthermore, most of the apps that are having high amount of reviews are from the categories of SOCIAL,COMMUNICATION and GAME like Facebook, WhatsApp Messenger, Instagram, Messenger — Text and Video Chat for Free, Clash of Clans, google apps etc.

- Even though apps from the categories like GAME, SOCIAL, COMMUNICATION and TOOL of having the highest amount of installs, rating and reviews are reflecting the current trend of Android users, they are not even appearing as category in the top 5 most expensive apps in the store .
  As a conclusion, we learn that the current trend in the Android market are mostly from these categories which either assisting, communicating or entertaining apps.

Some important points we get:

Average rating of (active) apps on Google Play Store is 4.17.
If we see individually app wise the communication app like facebook and whatsapp get highly reviewed app it shown that people regularly active on that and give their feedback also on that.

Medical and Family apps are the most expensive and even extend up to 80$.

Users tend to download a given app more if it has been reviewed by a large number of people.

More than half users rate Family, Sports and Health & Fitness apps positively. Apps for games and social media get mixed reviews, with 50 percent positive and 50 percent negative responses.