Capstone Project Title: Play Store App Review Analysis

TEAM MEMBERS:

Geetanjali Yadav, Shweta Sonawane

Vishwas Chole

Index

- Problem Statement
- Data Summary
- Data Pipeline
- Data Handling
- Exploratory Data Analysis(EDA)
- Conclusion

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.

Each app (row) has values for category, rating, size, and more. Another dataset contains customer reviews of the android apps.

Explore and analyze the data to discover key factors responsible for app engagement and success.

Data Summary

- Data has two different datasets viz. Play Store Data and User Review.
- The dataset Play store data has 13 different features with more than 10,000 observations while the dataset User Review has 5 different features.
- The most important features for the dataset Play store data are

  App: There are 9660 unique apps in the dataset.

  Category: Total 34 different categories are there.

Data Summary

Rating: This represents the average of the rating for each of the app.

Reviews: This represents the number of reviews for each app.

Installs: This indicates the number of installs for each of the app.

Size: It indicates the size of the app.

Type: Free indicates the app is free and Paid indicates that the app is paid.

Data Summary

Price: This feature gives the price of each of the app.

Content Rating: This features tell the rating is permissible to which age group or to every person.

Genre: This feature gives the information about the genre of the app.

Android Ver: This feature gives the android version of the app.

Last Updated: This feature gives the date of last update of the version.

Data Summary

- The User review data has four important features for the EDA:

App: There are 1074 unique apps in the dataset

Sentiment: There are three different sentiments possible for each user review.

Sentiment Polarity: This value ranges between -1 and 1.

Sentiment Subjectivity: It lies between 0 and 1.

Data Pipeline

Data Processing: In this part we have checked the data and its all features.

Data Handling: In this part, we have checked for nan values, missing values, and duplicate observations and done the refinement in the dataset.

EDA: In this part, we do some exploratory data analysis on the selected important features.

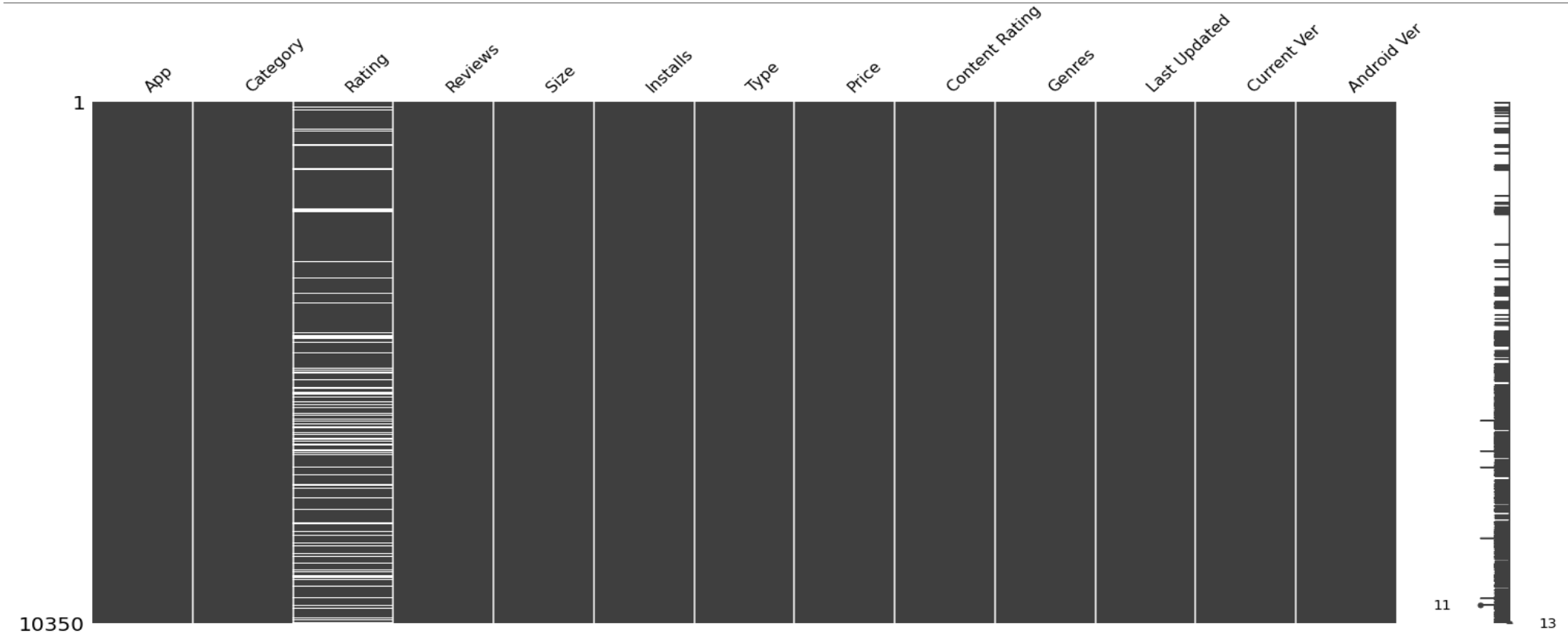
Data Visualization: Finally, we visualize the data using distinct plots for distinct features of the data, try to analyze the relationship between the features of data with the help of different plots. We also cleared all the point that why the given thing happens.

Data Handling: Duplicated Observations

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10841 entries, 0 to 10840
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   App             10841 non-null  object
 1   Category        10841 non-null  object
 2   Rating          9367 non-null   float64
 3   Reviews         10841 non-null  object
 4   Size            10841 non-null  object
 5   Installs        10841 non-null  object
 6   Type            10840 non-null  object
 7   Price           10841 non-null  object
 8   Content Rating  10840 non-null  object
 9   Genres          10841 non-null  object
 10  Last Updated    10841 non-null  object
 11  Current Ver     10833 non-null  object
 12  Android Ver     10838 non-null  object
dtypes: float64(1), object(12)
memory usage: 1.1+ MB
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 10350 entries, 0 to 10840
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   App             10350 non-null  object
 1   Category        10350 non-null  object
 2   Rating          8885 non-null   float64
 3   Reviews         10350 non-null  object
 4   Size            10350 non-null  object
 5   Installs        10350 non-null  object
 6   Type            10349 non-null  object
 7   Price           10350 non-null  object
 8   Content Rating  10349 non-null  object
 9   Genres          10350 non-null  object
 10  Last Updated    10350 non-null  object
 11  Current Ver     10342 non-null  object
 12  Android Ver     10347 non-null  object
dtypes: float64(1), object(12)
memory usage: 1.1+ MB
```

Data Handling(Missing, Nan values)



EDA(Correlation)

Observations:

1.Reviews and Installs are strongly correlated.

2.Rating and Installs shows weak correlation.

3.Installs and Price are negatively correlated.
(Installs increases price decreases)

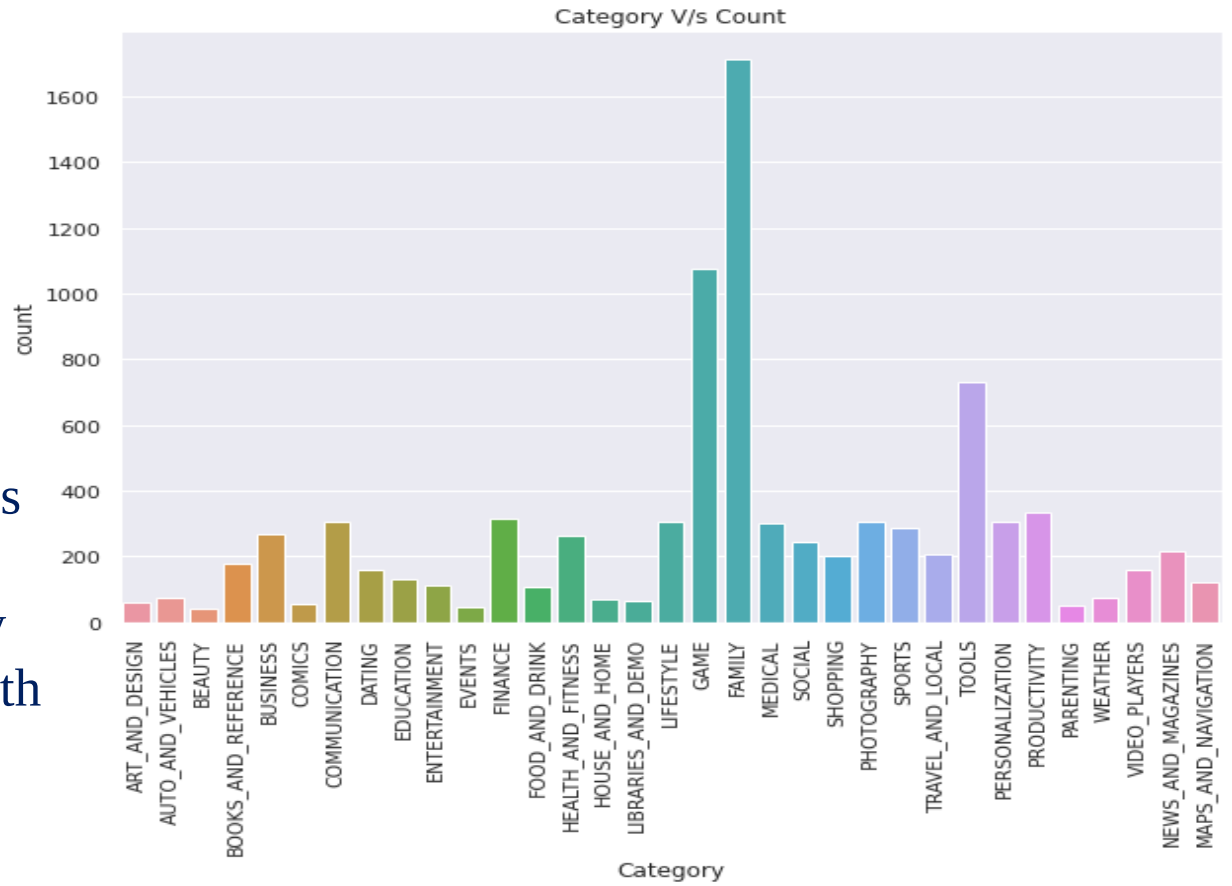
EDA(App and Category)

App Information :

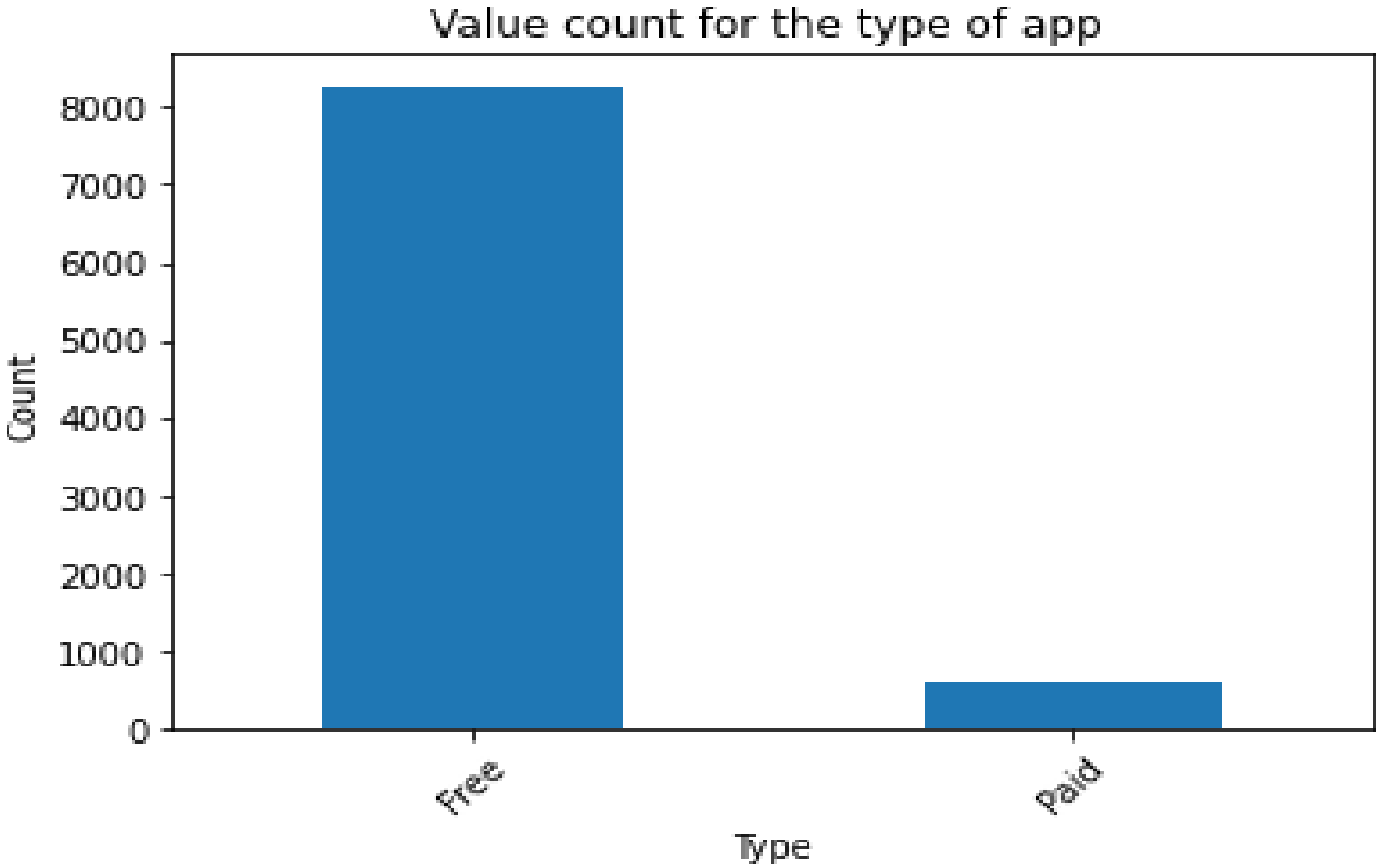
Dataset contains information of 8878 apps.

Category vs Count :

1.The Category column in our dataset has 34 different types of categories.
2.There are around 1700 app with family category, followed by game category with more than 1000 app.



EDA(Type)

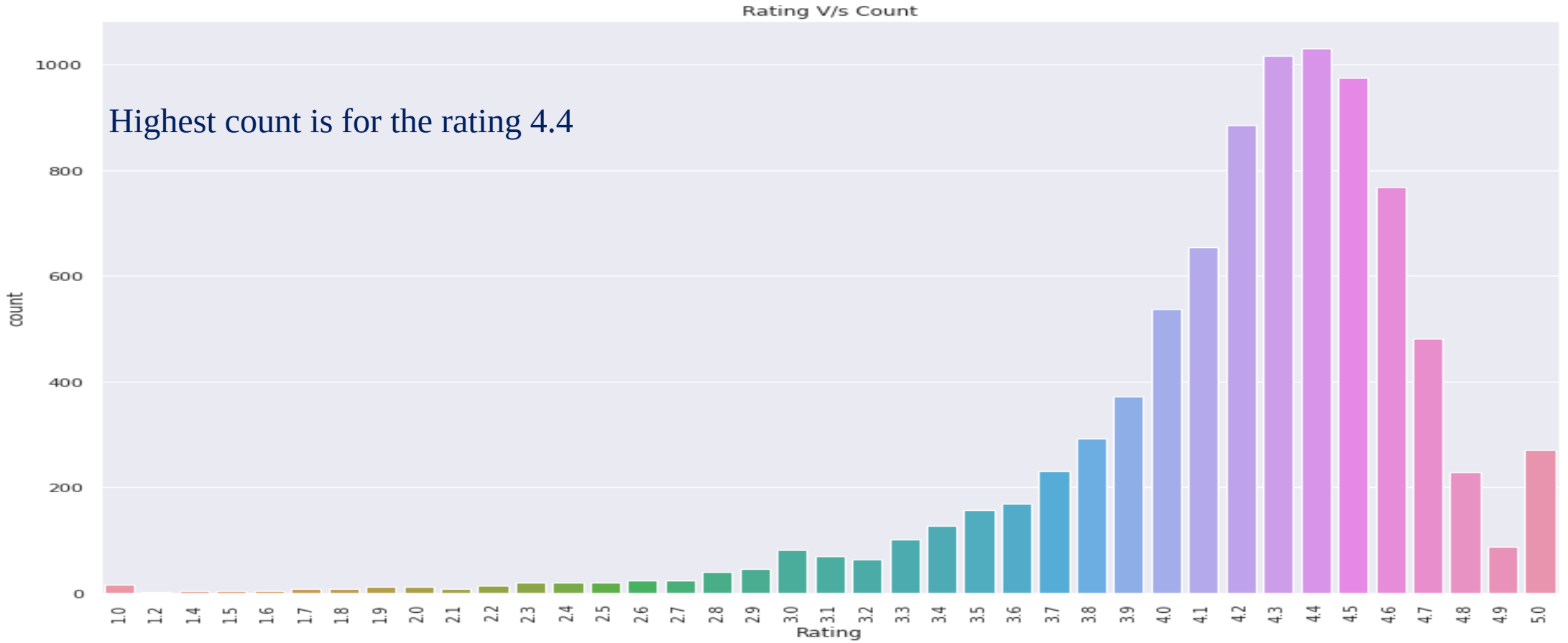


The majority of apps in the Google Play Store were free to download.

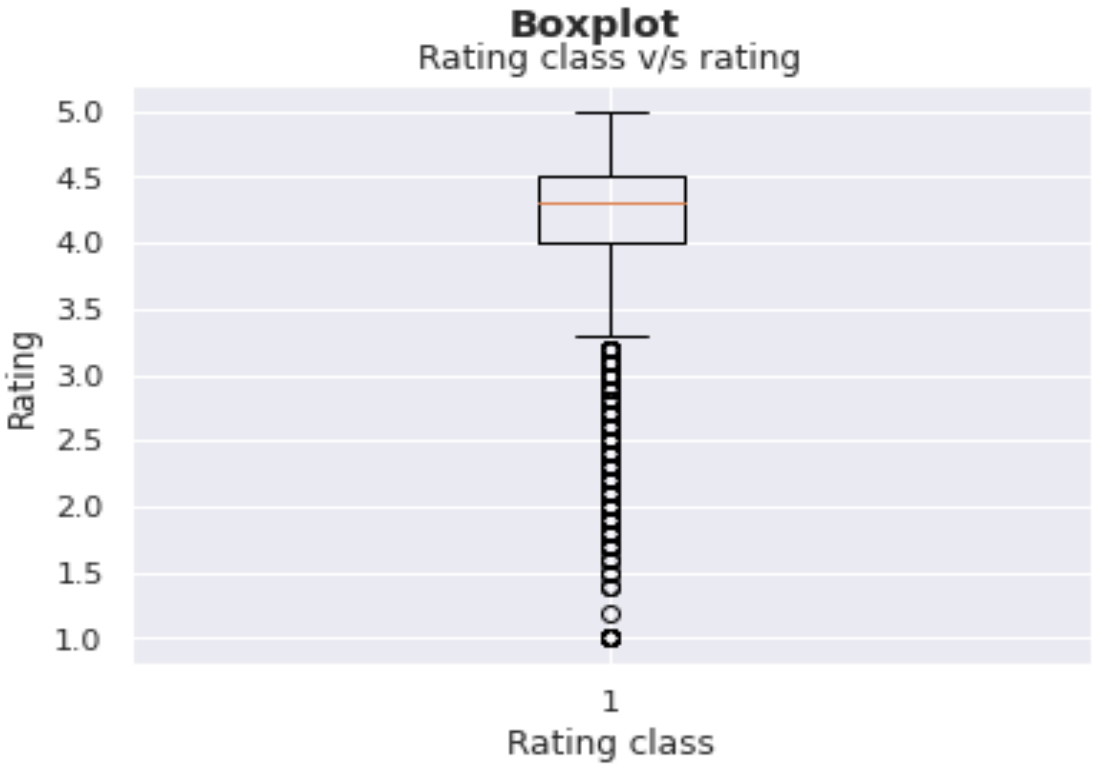
Free	**8268**
Paid	**610**

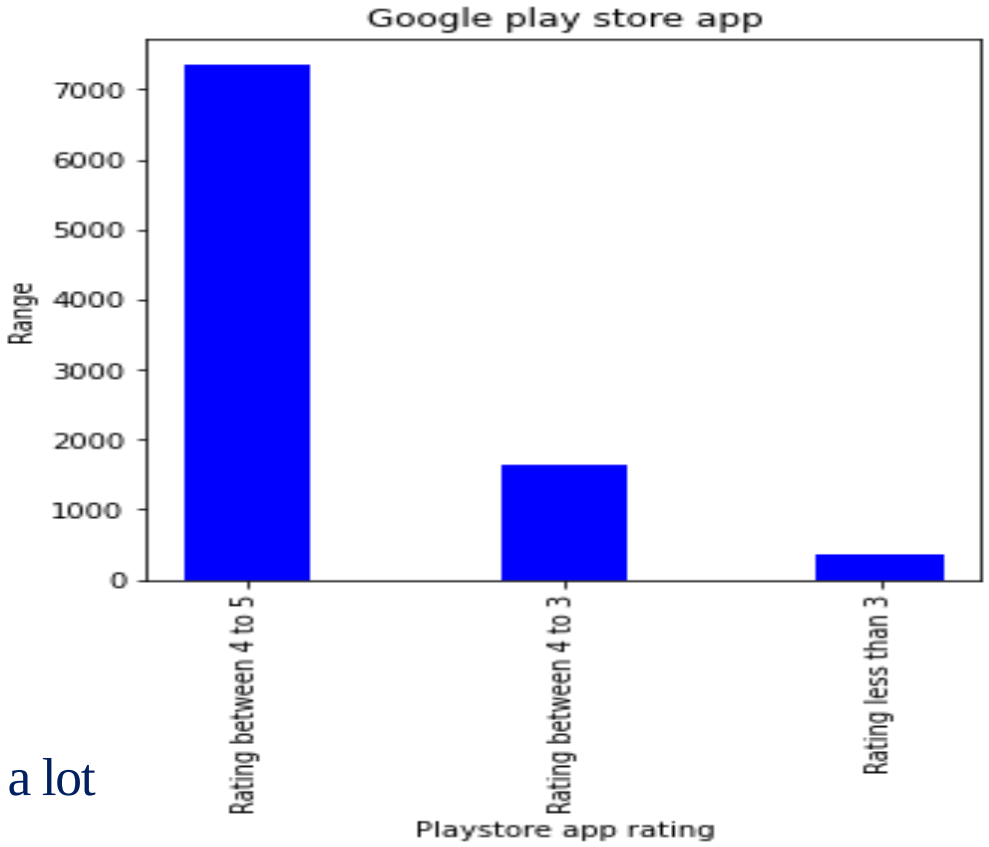
EDA

(Rating)


Rating V/s Count
Highest count is for the rating 4.4
count
Rating
0
200
400
600
800
1000
1.0
1.2
1.4
1.5
1.6
1.7
1.8
1.9
2.0
2.1
2.2
2.3
2.4
2.5
2.6
2.7
2.8
2.9
3.0
3.1
3.2
3.3
3.4
3.5
3.6
3.7
3.8
3.9
4.0
4.1
4.2
4.3
4.4
4.5
4.6
4.7
4.8
4.9
5.0

EDA (Continued)

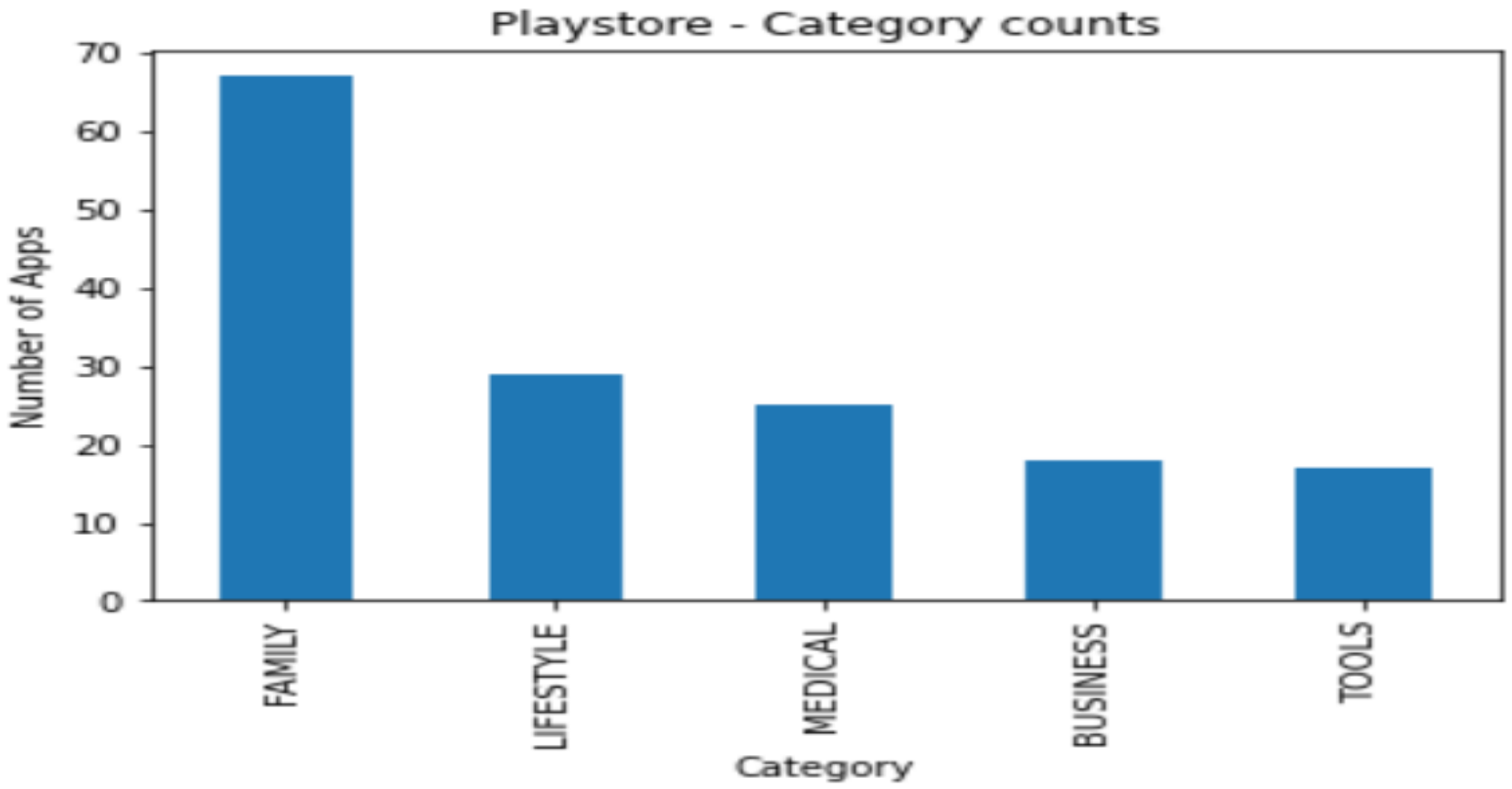




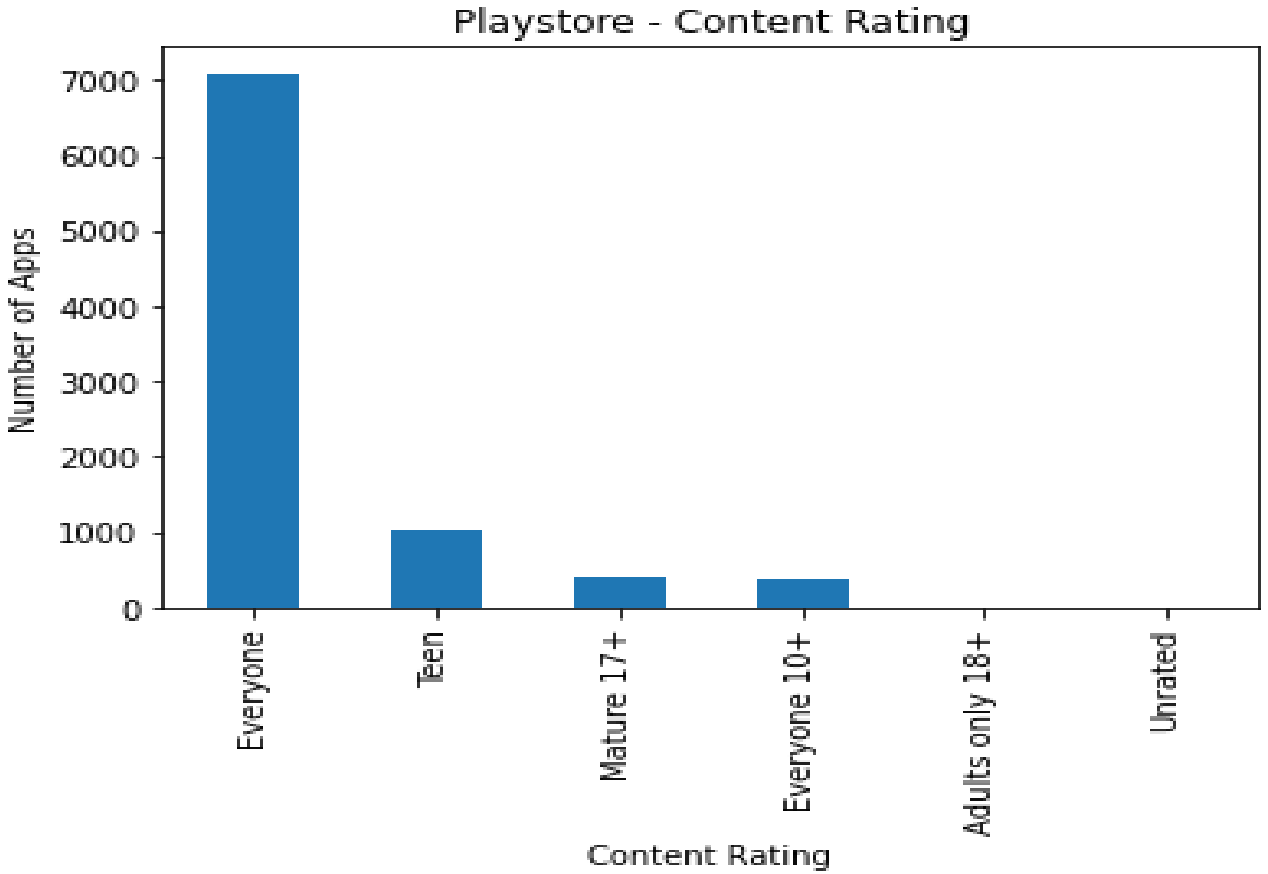
Most of the apps, clearly hold a rating above 4.0! And surprisingly a lot seem to have 5.0 rating.

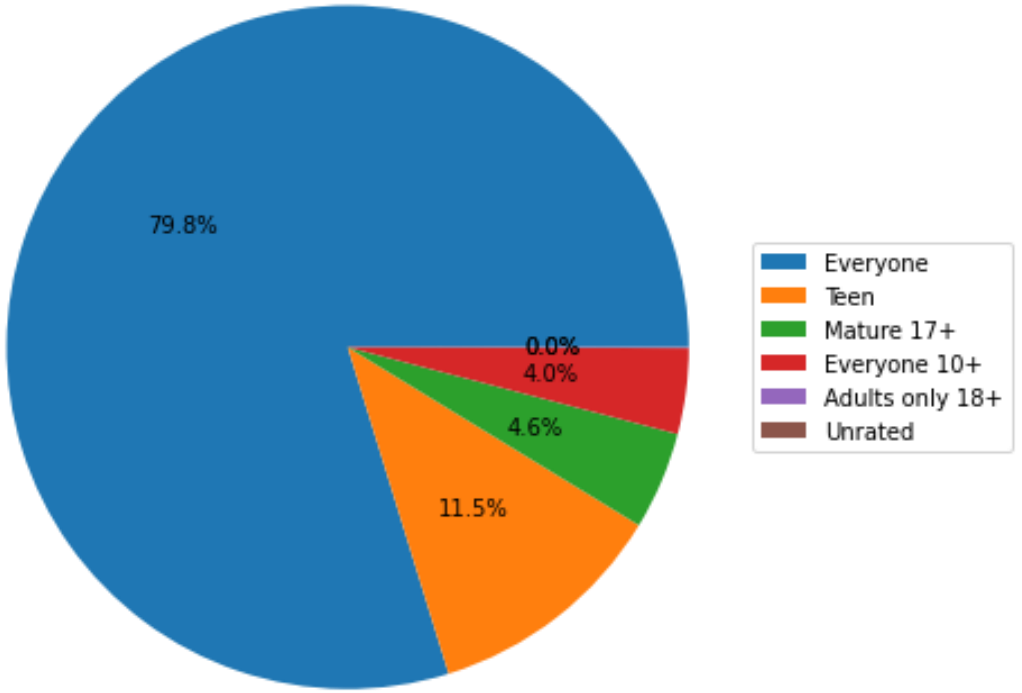
EDA

(Top 5 most dominant category w.r.t value count of app(Rating = 5))



EDA(Content Rating)

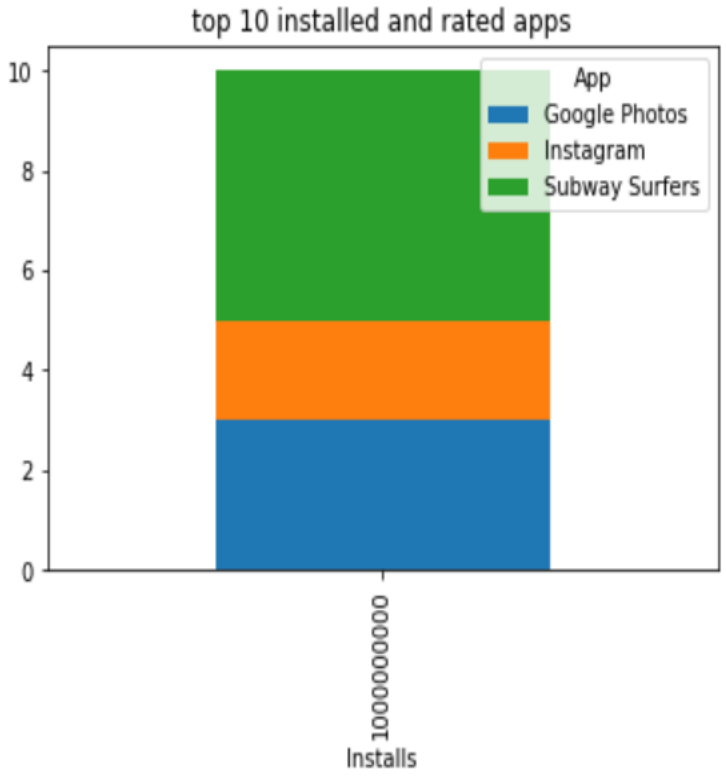




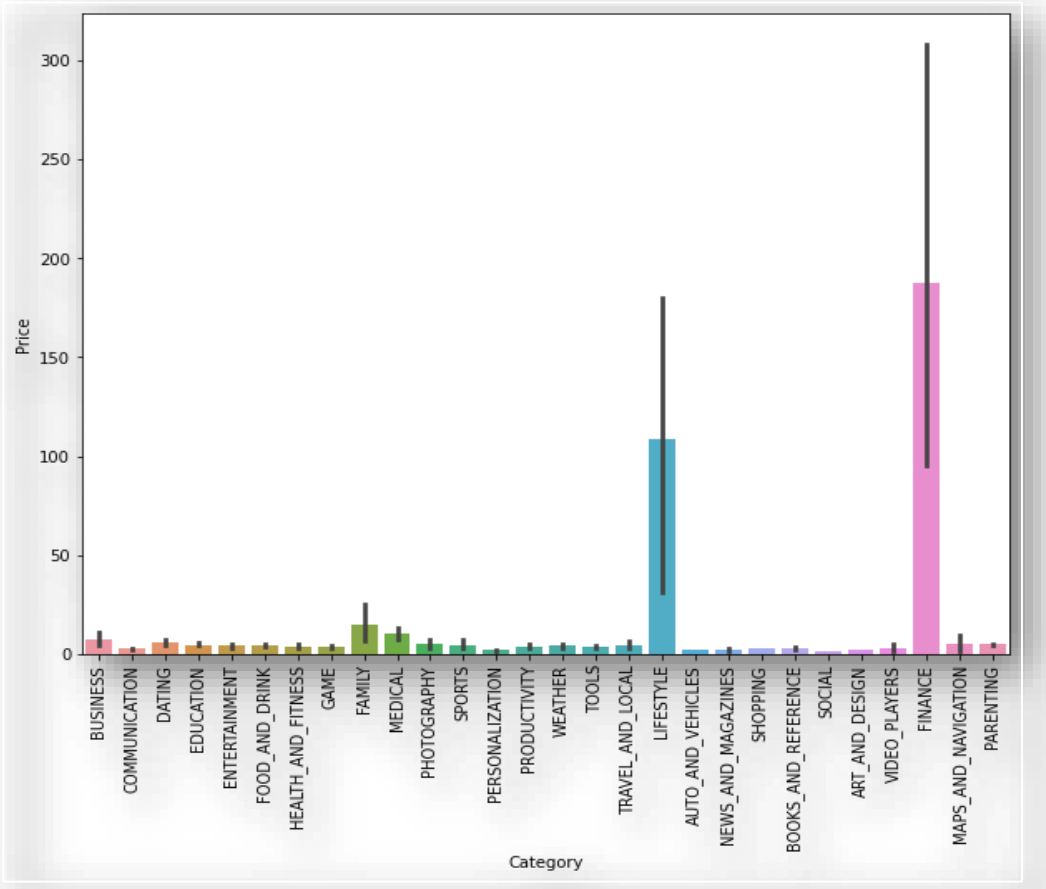
EDA
(Top 10 apps that are installed and rated)

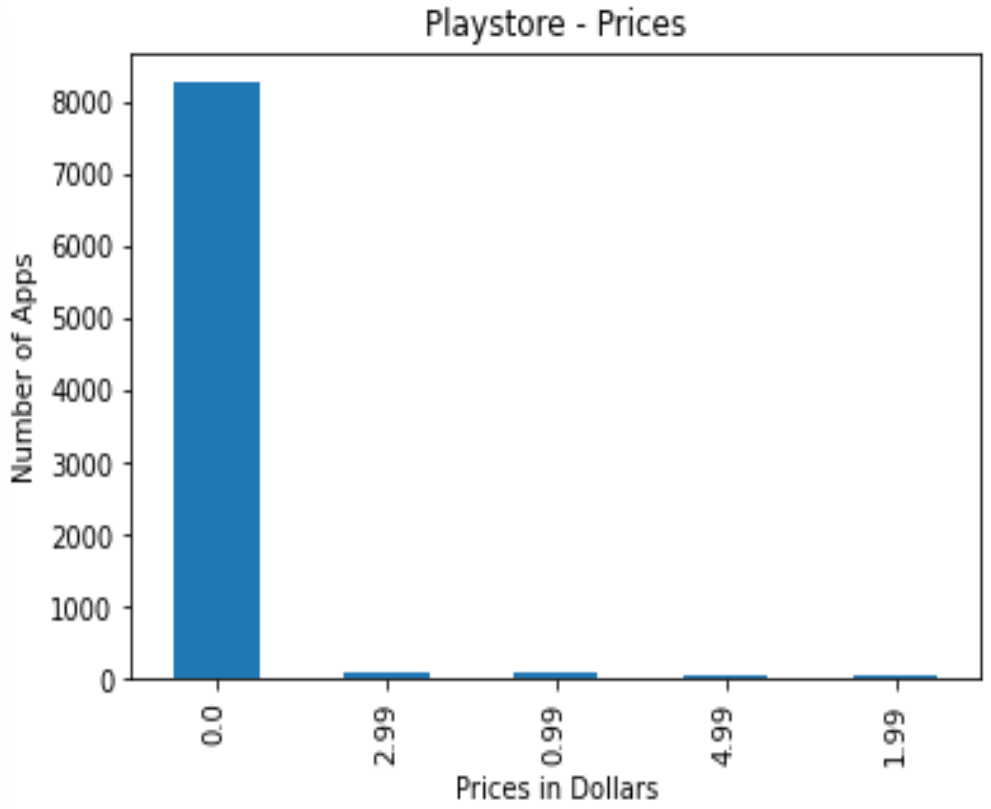
- The Google Play Store offers a wide range of applications, but most of these apps had been downloaded only by a small number of people.

- Among this app we find most rated and installed app which had been downloaded between 500 millions to 1billion+ times.



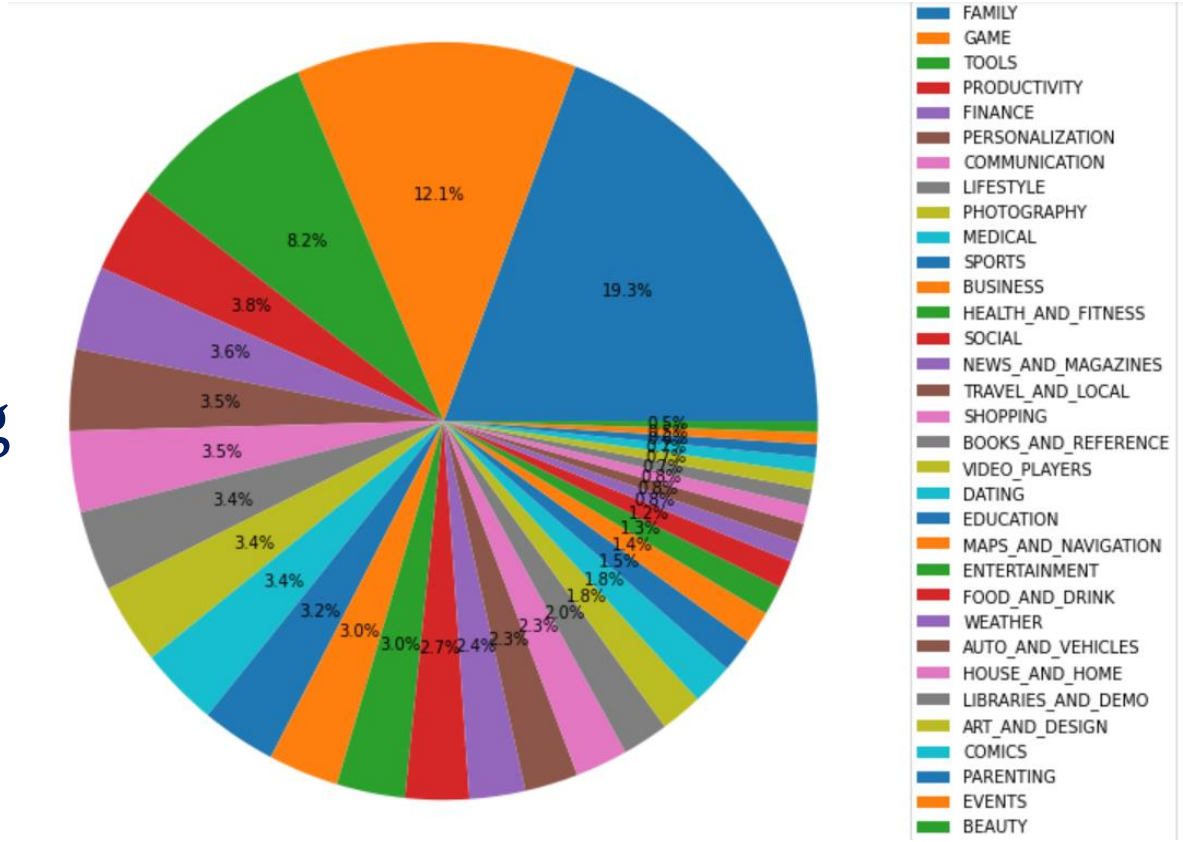
EDA(Category w.r.t Price)



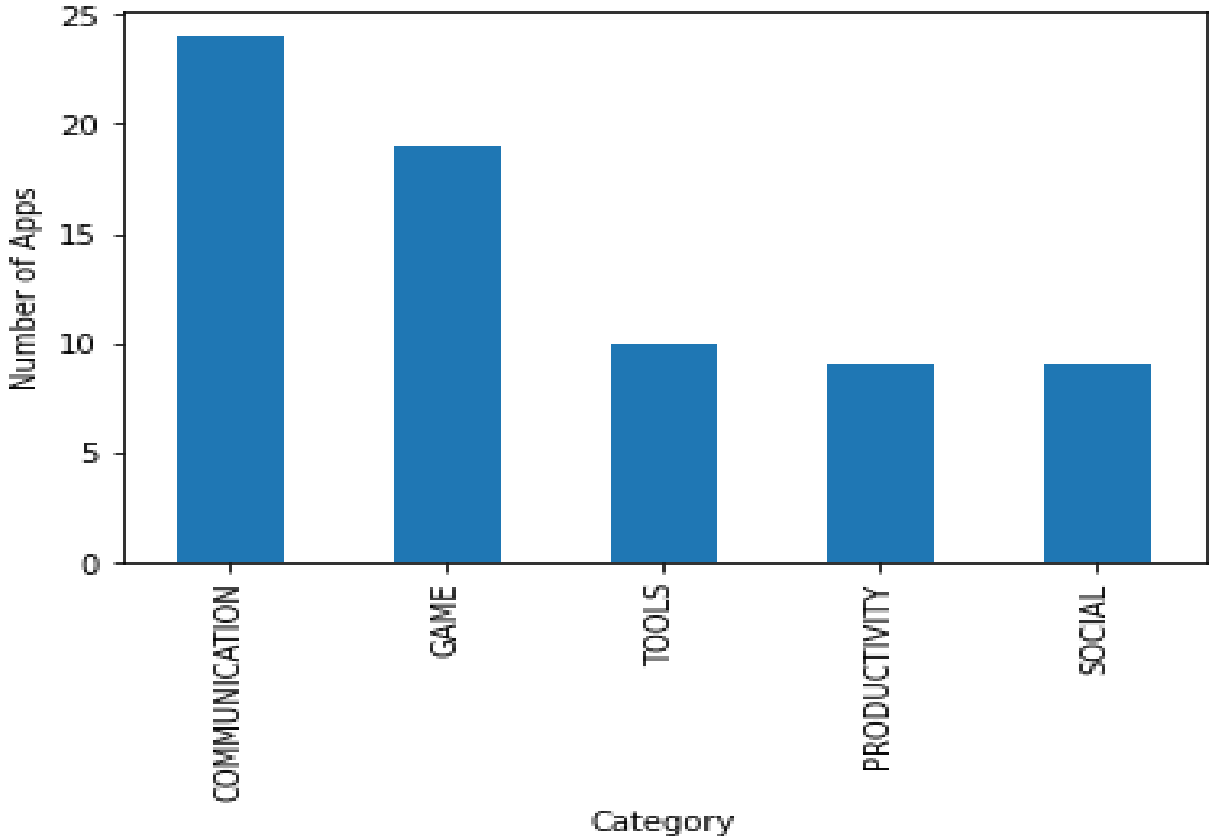


EDA(Distribution of Apps)

- Family category has 19.3% of apps which is approximately 1700 apps.
- In both paid and free apps, the category Family is having highest count for the apps that are installed.
- Second position is grabbed by category ‘Game’.

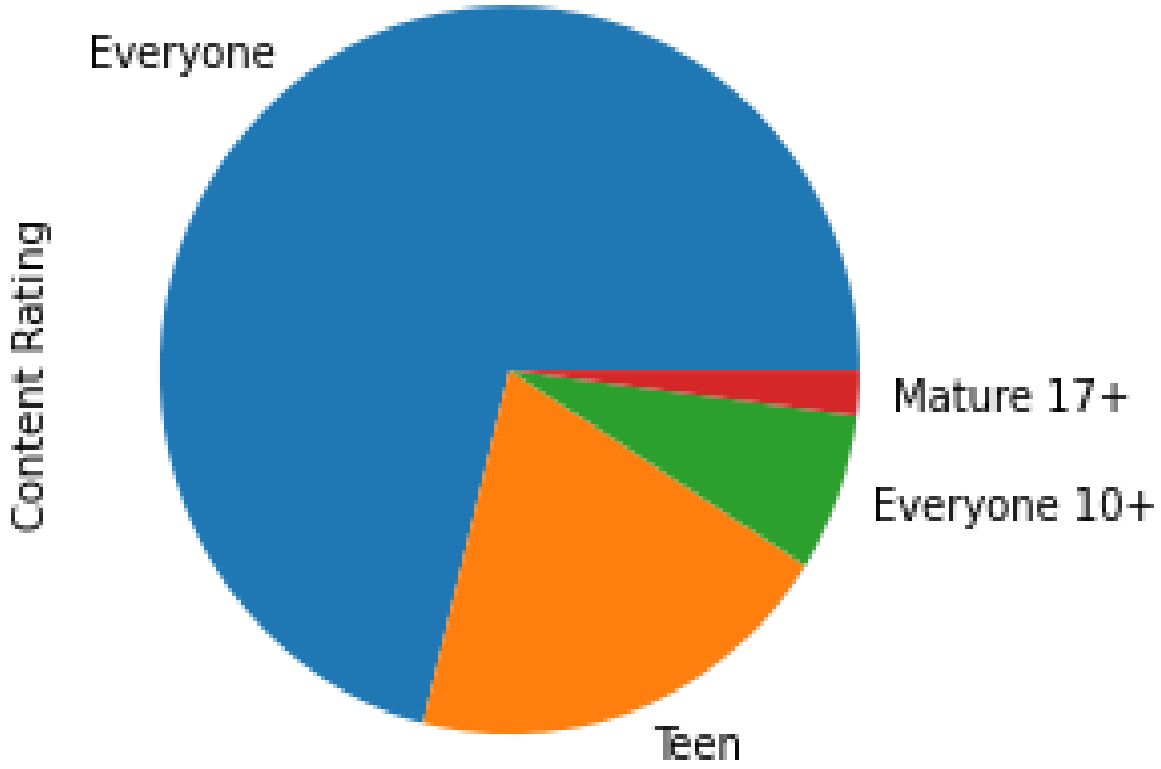


EDA
(Top apps which are rated and Installed)



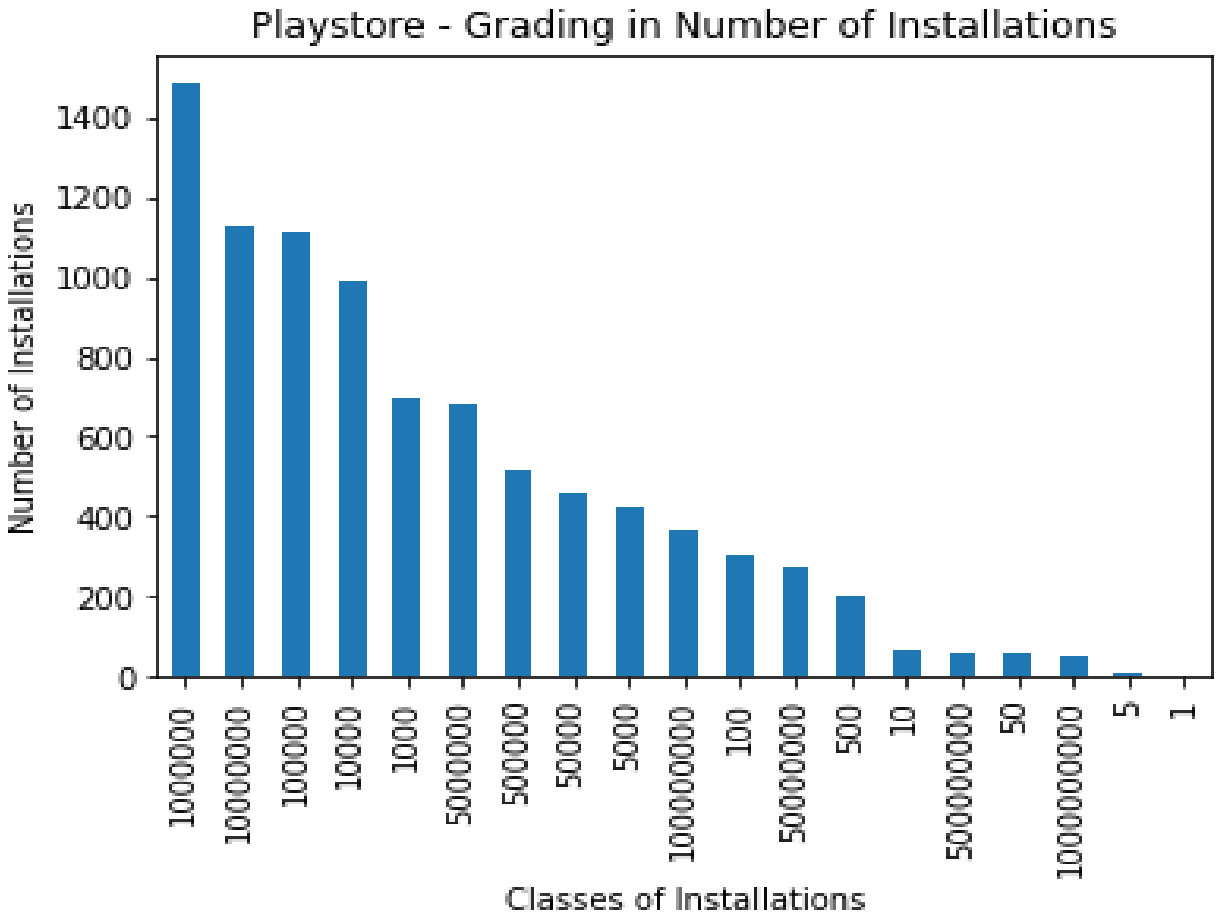
COMMUNICATION	24
GAME	19
TOOLS	10
PRODUCTIVITY	9
SOCIAL	9
NEWS_AND_MAGAZINES	7
VIDEO_PLAYERS	6
TRAVEL_AND_LOCAL	5
PHOTOGRAPHY	4
FAMILY	4
ENTERTAINMENT	1
HEALTH_AND_FITNESS	1
BOOKS_AND_REFERENCE	1

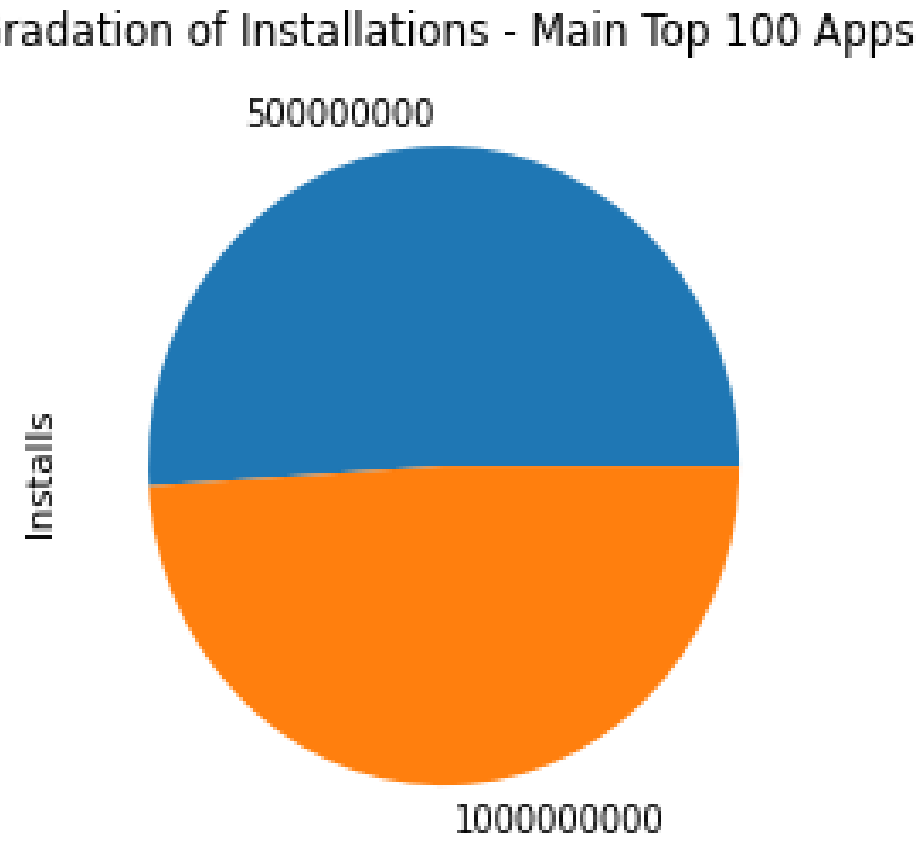
EDA(Continued)



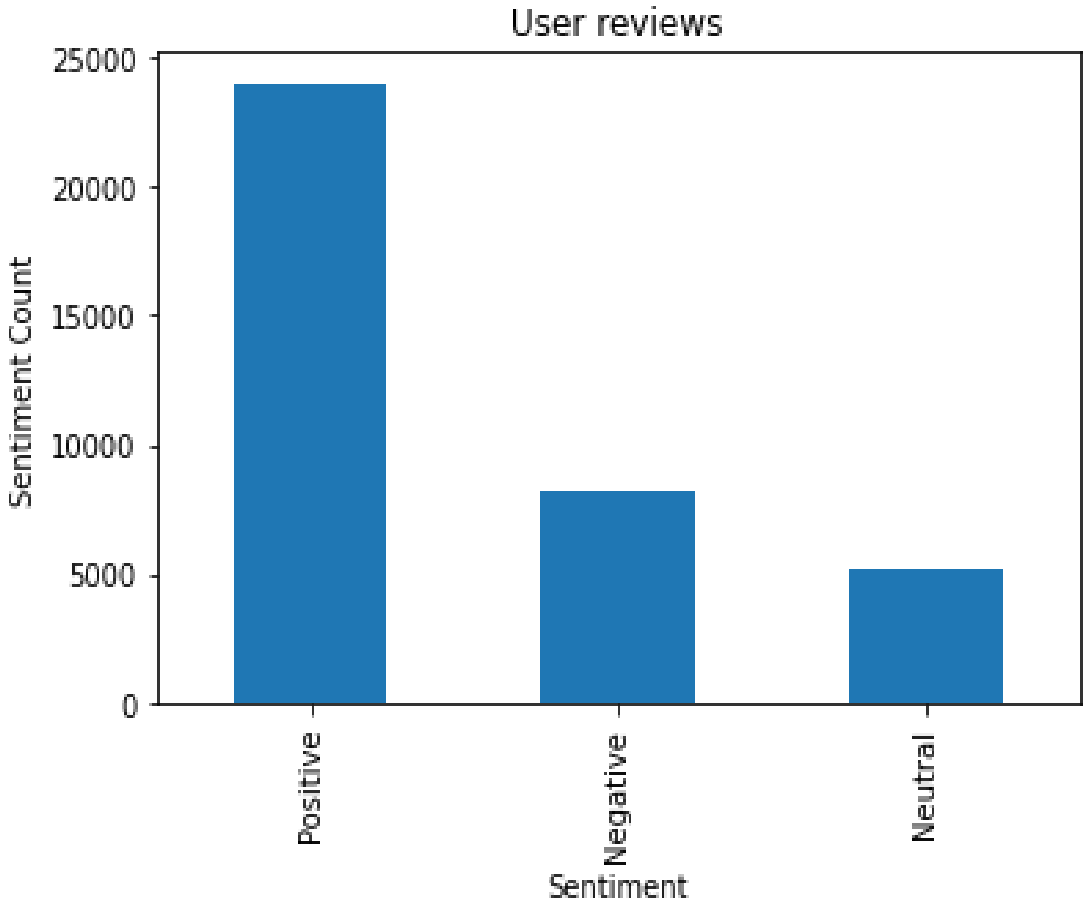

```
Everyone          71
Teen              20
Everyone 10+       7
Mature 17+         2
Name: Content Rating, dtype: int64
```

EDA(Continued)





EDA(Sentiment Analysis)



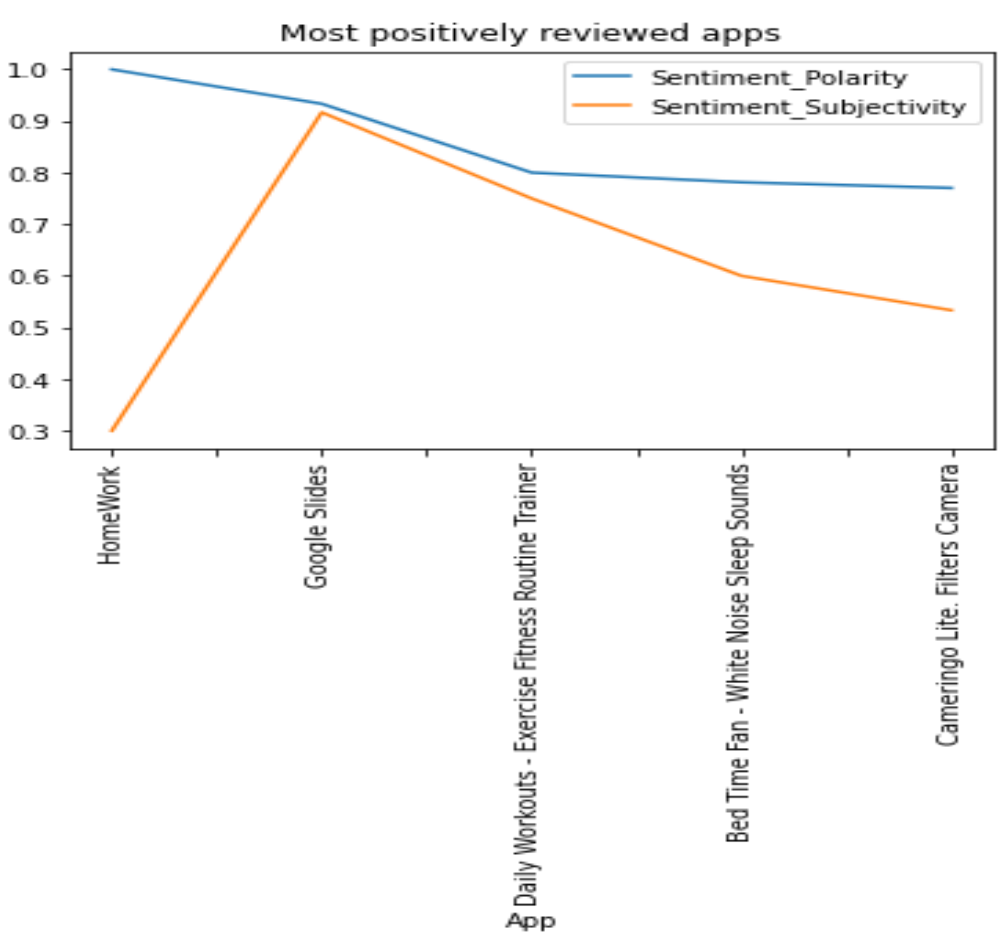
Sentiment analysis:

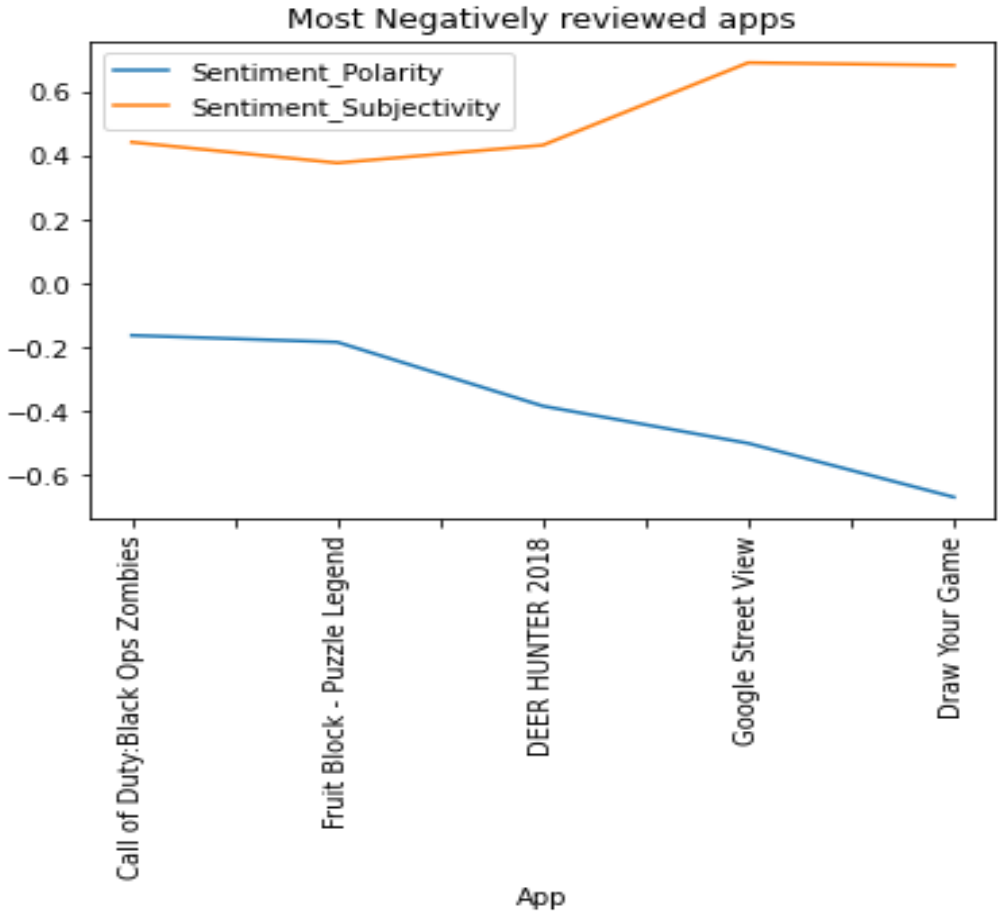
Sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document.

Classification of reviews:

1.Positive
2.Negative
3.Neutral

EDA(sentiment analysis continued)





Conclusion

Based on our observation:

1. Number of installation are directly proportional to the rating and review
2. Number of installation is inversely proportional to price.

So the installation, rating and review are the most important features for the success of app.

There is the problem of rating mismatch on a smaller scale. If this issue could be mitigated the Play Store would provide a more accurate representation of user sentiment which in turn could help developers make adjustments to their app accordingly. We also defined a success parameter for an app based on the number of installs, distribution of ratings ,Content review and price.

Suggestion through the support of EDA

- By decreasing price of app we can increase the no of installation.
- Exploring reviews and sentiment of the users as per the category of the application.

Q &A