Capstone Project -1

Play Store App Reviews Analysis

Done By:

Vinit Ladse
Gaurav Bhakte
Pratiksha Kharode

WHY ANALYZE THE PLAY Store?

- Android Apps comprise 75% of the Market Share. 85% share in
- •brazil,india,turkey

Mobile App Market is set to grow 20% by 2023

•What are some interesting patterns in user behavior related to app usage & feedback

1K

What makes an App popular? Can we predict how popular it’s going to be?

Content

- Problem statement
- Introduction
- Data cleaning / null value implementation
- Data processing
- Data exploration
- Basic observation
- Insights from data
- Conclusion
- Challenges and future

Problem statement

- Google play store is mostly use app store worldwide also top global market share.
- My main objective is to find key factor responsible for app success and engagement of users.
- Thousands of new app regularly update play store of different category.
- I find distribution of every app based on their size, installs, reviews and much more.

Introduction

- Mobile industry growing rapidly, competition for apps also grown significantly so developer need to do enough research to make app success.

- The Google Play Store is found to be the largest app market in the world. It has been observed that although it generates more than double the downloads than the Apple App Store but makes only half the money compared to the App Store.

- We perform Data Cleaning over the dataset. Further we divided our project in Four main parts i.e Analysis on Play Store Data and Reviews Data,Analysis based on Cancellations, General Analysis, Data Visualization. After the data set is ready, we try to analyze the data set using different plots and remove the stuff not needed from the data set.

Data Cleaning

- Google Play store dataset has 10,841 observation of data with fields.
- Two data set 1) play store data 2) user reviews
- List of fields:

❑ App

❑ Category
❑ Rating
❑ Reviews

❑ Size

❑ Installs
❑ Type
❑ Price
❑ Content rating
❑ Genres
❑ Last updated
❑ Current version
❑ Android version

Play store data

❑ App
❑ Translated review
❑ Sentiment
❑ Sentiment polarity
❑ Sentiment subjectivity

User Reviews

Data cleaning (Contd..)

- Understand the structure of the dataset and clean data before analysis
- Finding Missing value in dataset
- Drop the null value
- Correct data type(INT,FLOAT,DATE)
- Checking outliers
- So after successfully cleaning the dataset we have 9360 columns and 13 rows

Data Processing

- The dataset collected from the Play store is semi structured or unstructured and contains significant superfluous data (defined as not contributing significant meaning).Some data type needs to change in required format as int, float, date.

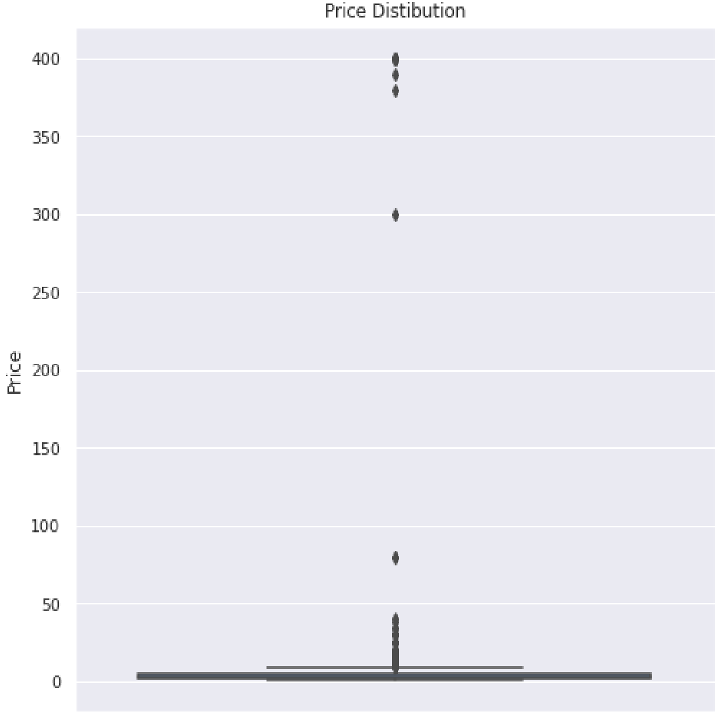
- Sizing of apps needs to convert in one measurement KB or MB. Pre-processing includes various tasks including stemming, lowercase conversion, Units, punctuation, and excluding terms.

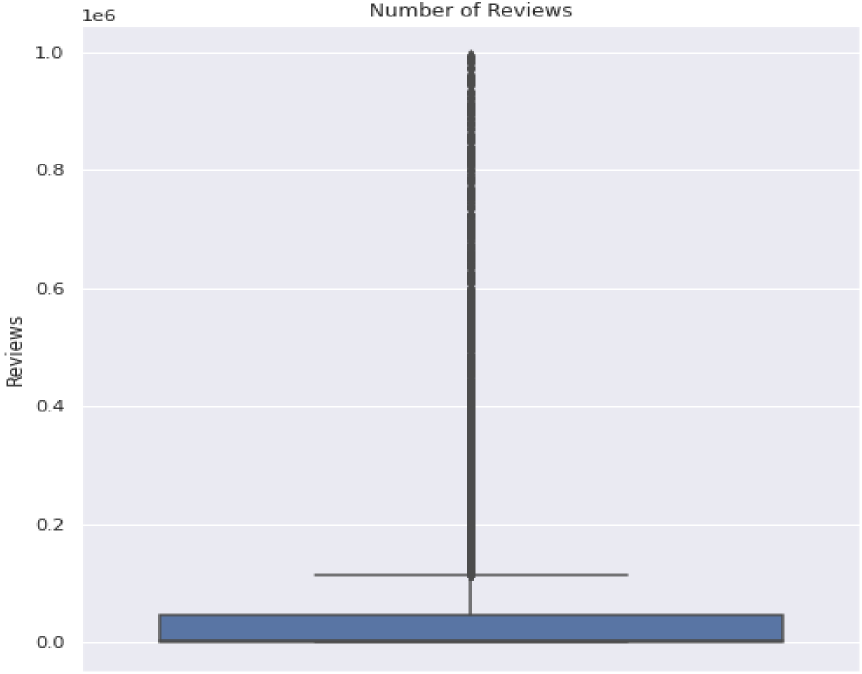
Outlier Graphs

Price Distribution

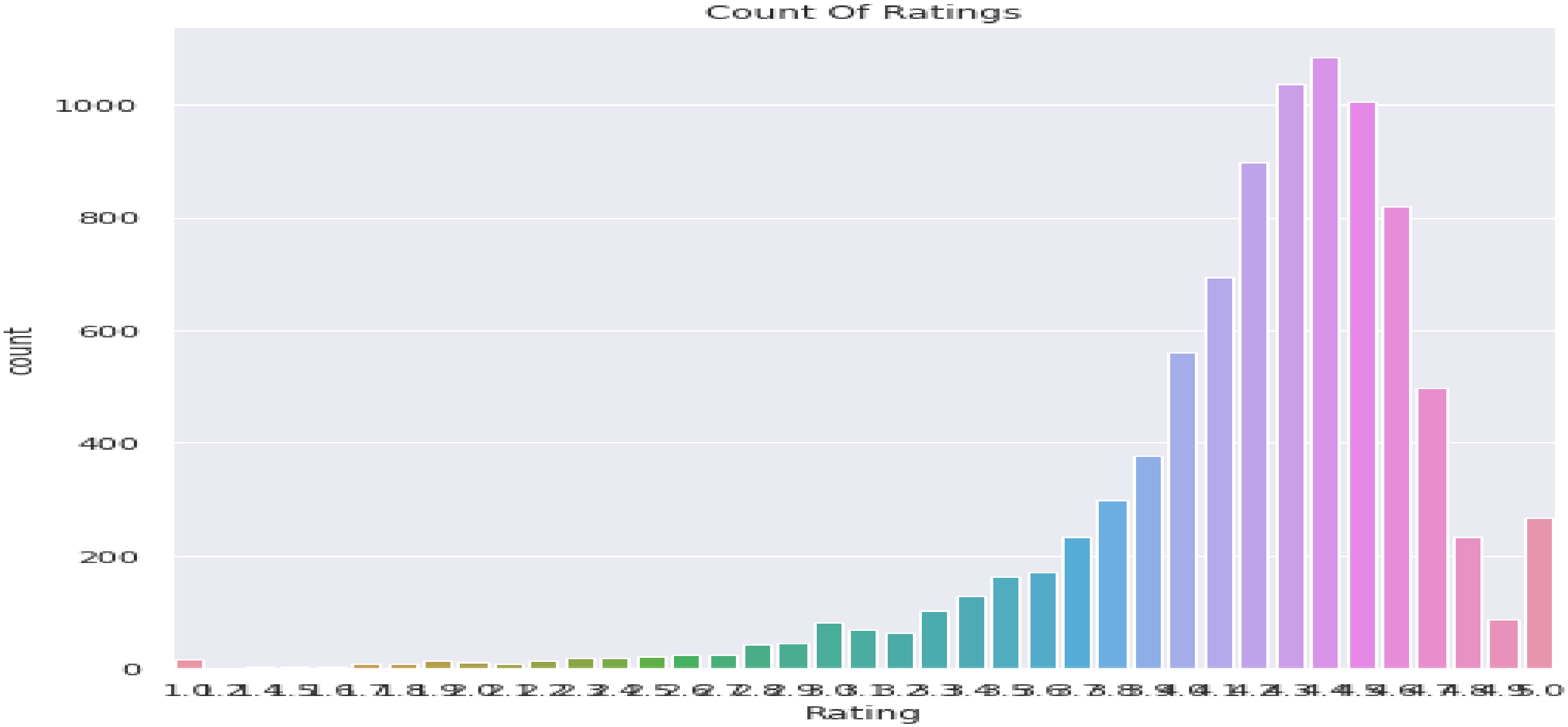
-

Number of Reviews



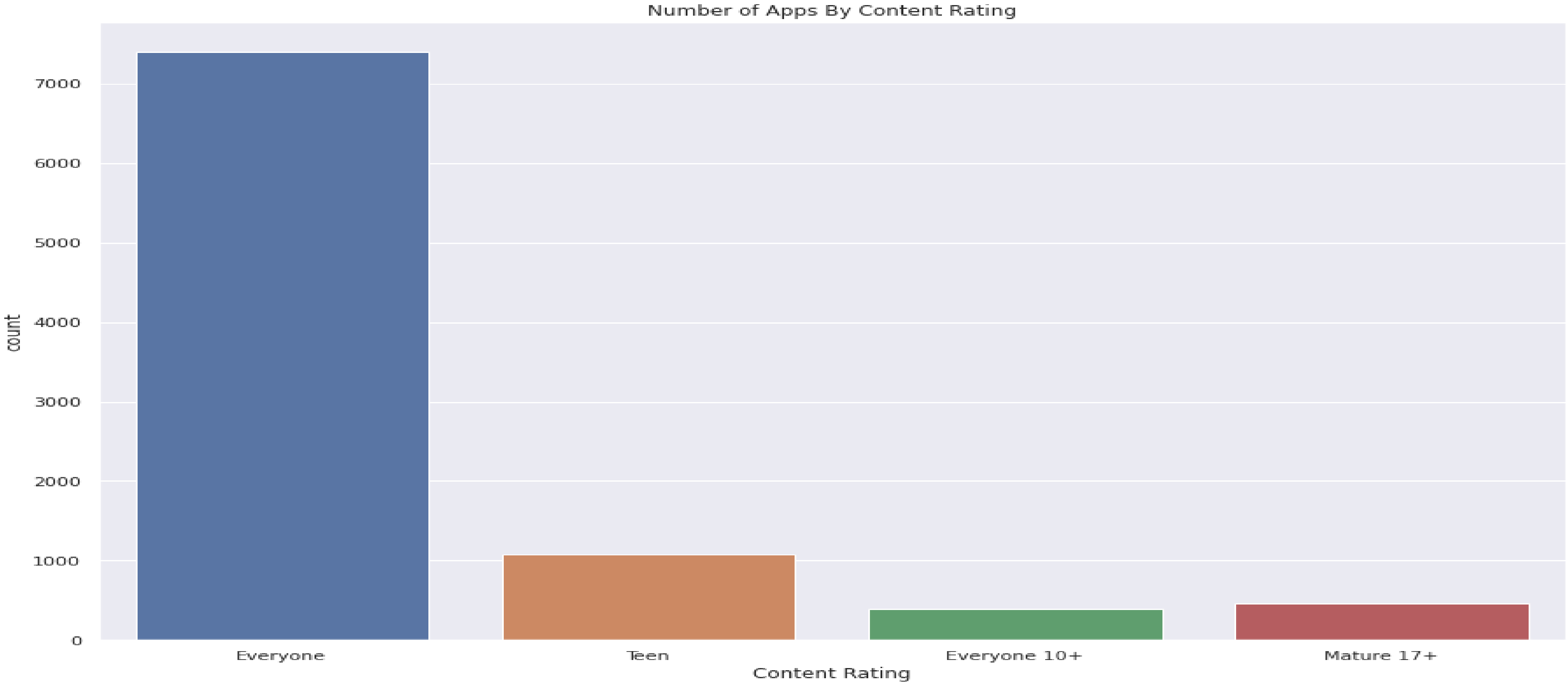


Data Exploration

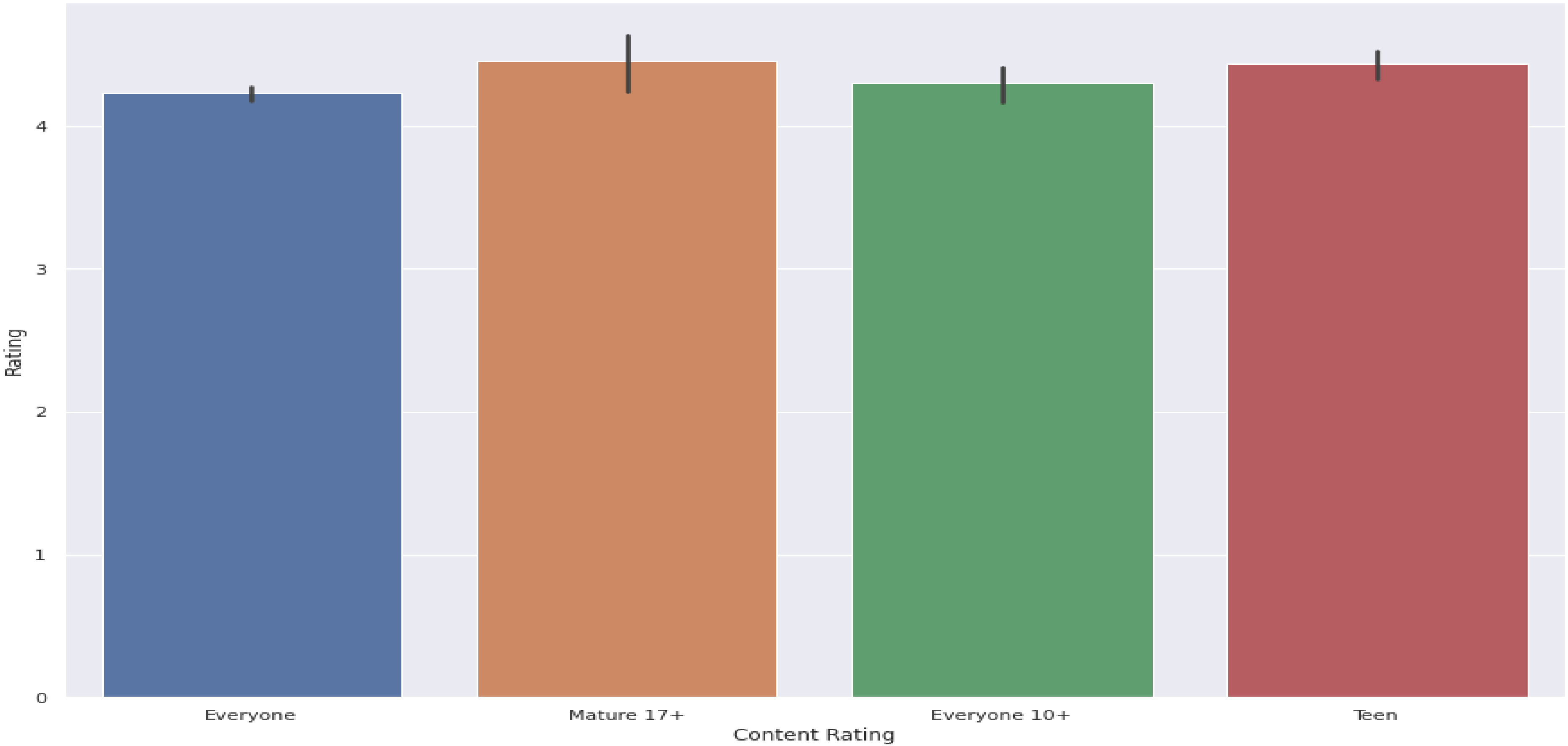


Top Content Rating Values

Content Rating vs Apps(Count)

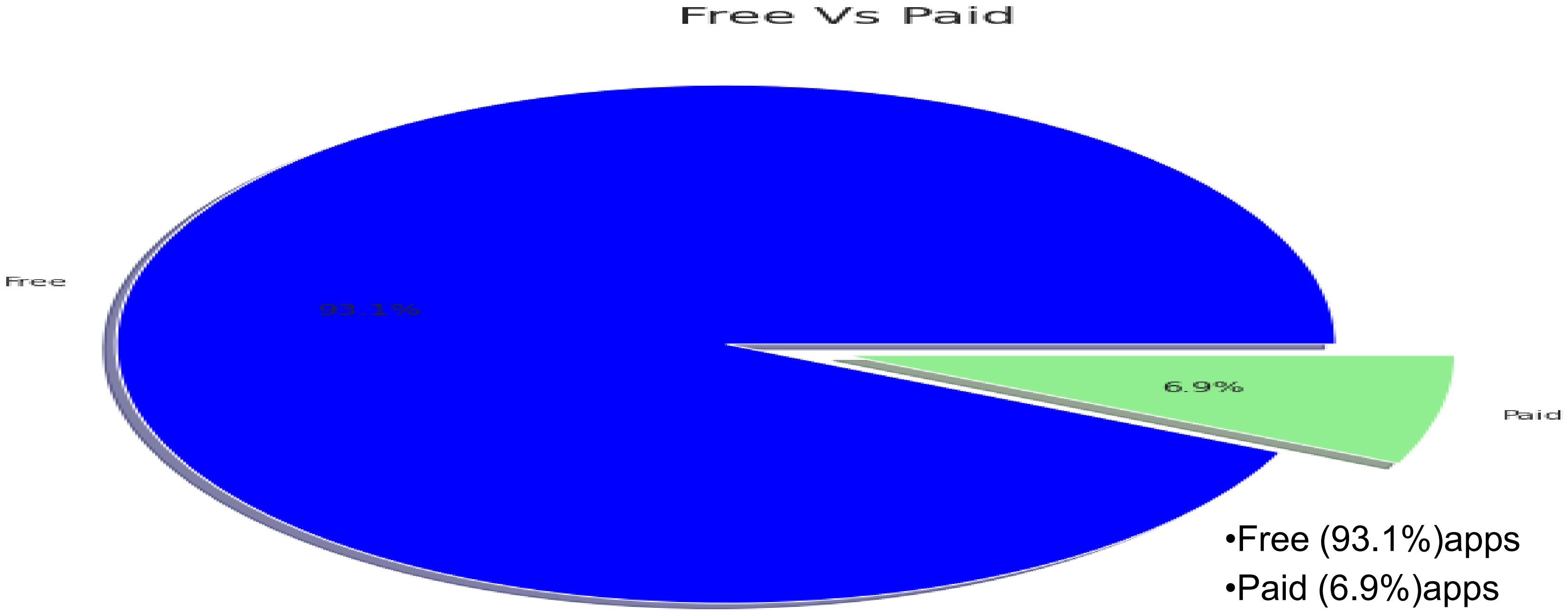


Content Rating vs Rating



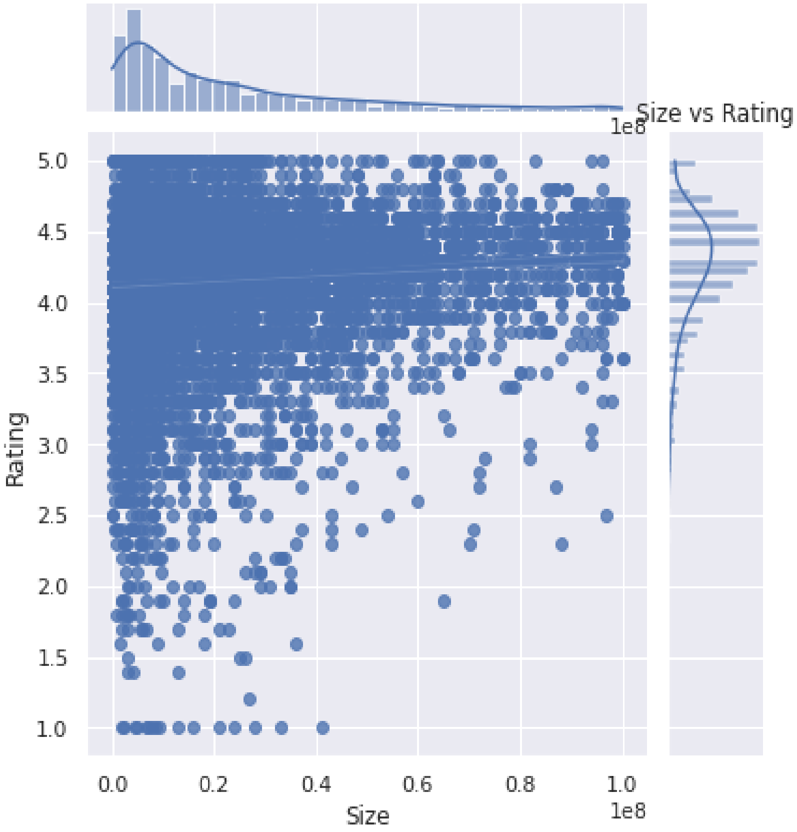
Pricing Strategies

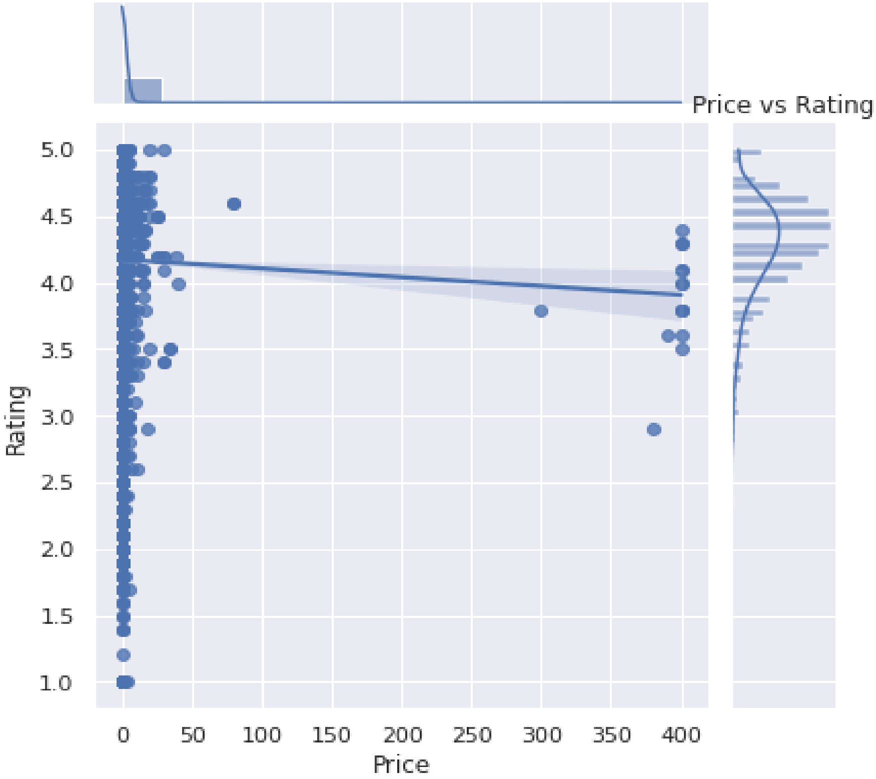
Since most Play Store apps are free, the revenue model is quite unknown and unavailable as to how the in-app purchases, in-app adverts and subscriptions leads to the success of an app. Thus, an app’s success is determined by the number of installs and the user ratings that it has received over its lifetime rather than the revenue it generated.



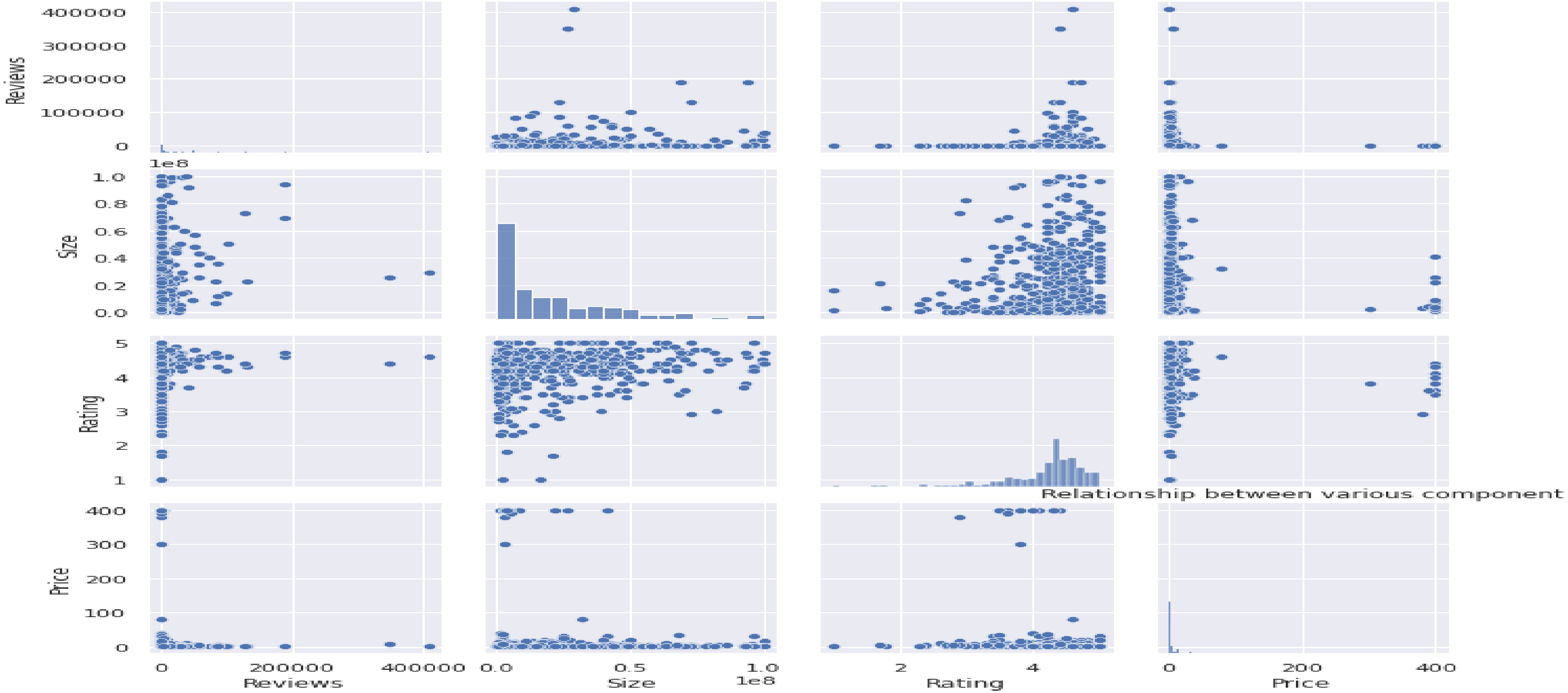
- Free (93.1%)apps
- Paid (6.9%)apps

Effect of Price and Size Vs Rating





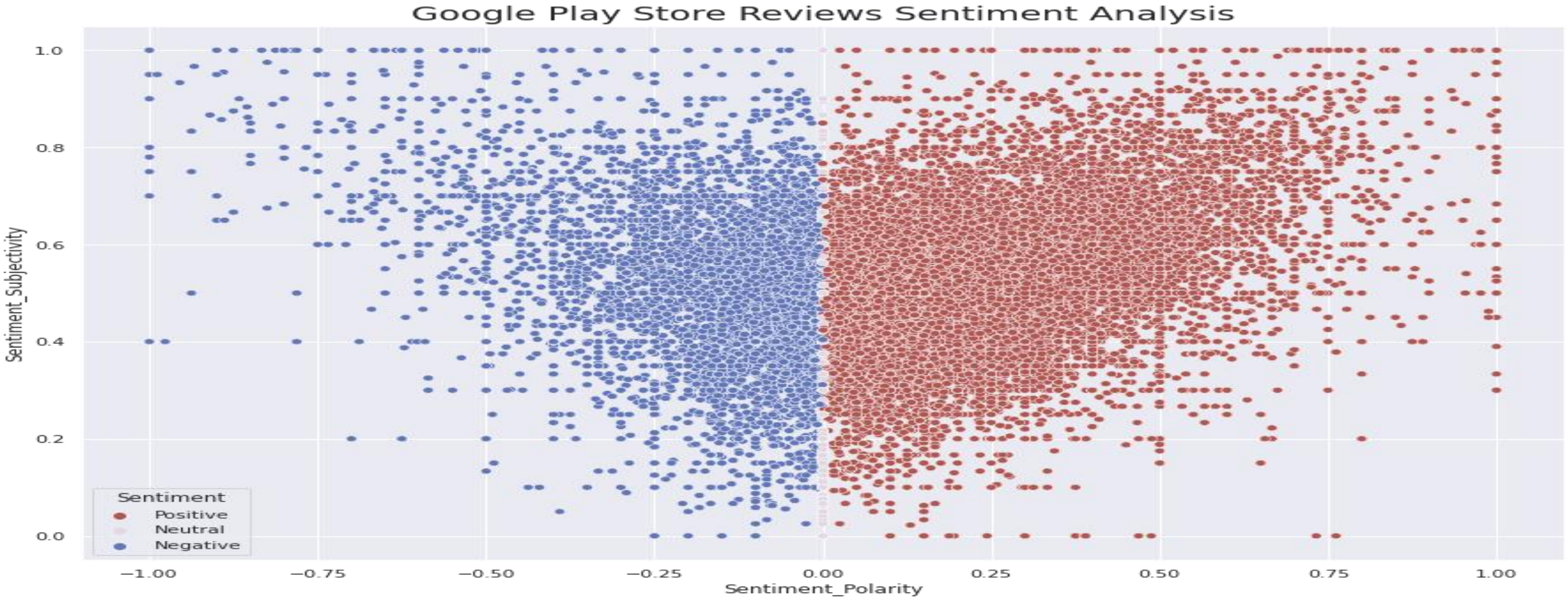
Pairplot with the columns - Reviews, Size, Rating, Price



Basic Observation
Below are some observation by doing data wrangling.

Average app rating	4.18
Top five category highest average rating	1) Events 2) Education 3) Arts and design 4) parenting 5) personalization
App with maximum reviews	Clash of clans
Top 5 app having highest reviews	1) Clash of clans 2) subway surfers 3) clash Royal 4) Candy crush 5) UC-browser
Most expensive app	I’m rich

Google Play Store Reviews Sentiment Analysis



From the above 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.

Insights from data

WORDCLOUD

- Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance.

Sentiment Polarity

- The polarity of a sentiment measures how negative or positive the context is.
- In the data that we have, the polarity ranges from -1 (most negative) to +1 (most positive).

WORD CLOUD For FREE App

WORD CLOUD For PAID App

CONCLUSION

(Data)

- That's it! We reached the end of our exercise.

- The dataset contains possibilities to deliver insights to understand customer demands better and thus help developers to popularize the product.After analysing the dataset we have got answers to some of the serious and interestings questions which any of the android users would love to know.

- We dealt with missing data and outliers, we tested some of the fundamental statistical assumptions and we even transformed category variables into dummy variables.

- That's a lot of work that Python helped us make easier. Dataset can also be used to look whether the original rating 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.

(Reviews)

- Paid apps have a slightly higher number of favourable reviews than free apps.
- Free apps get more negative and neutral feedback, suggesting a wider range of opinions.
- Clash of Clans app has most number of reviews. While Subway Surfers is most number of install app.
- 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.
- Users download a given app more if it has been reviewed by a more number of people.

Challenges

- ★ Data contain NULL/NAN values in dataset.
- ★ Main task to clean data followed by data processing.
- ★ In this project we perform EDA and discovering relationships with specific features using sentiment of users.
- ★ Some data app name etc are in gibberish form and contain duplicates.

Future

- ★ Developers can use my work for there research purpose to make app success.

THANK YOU