Capstone Project -1

Play Store Apps 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 and Turkey and many more.

Mobile App Market is set to grow 20% by 2023

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

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

Content

- Problem Statement
- Introduction
- Data Cleaning/Null values implementation
- Data Processing
- Data Exploration
- Basic Observations
- 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 applications are one of the fastest-growing segments of downloadable software application markets. Out of all of the markets we choose Google Play store due to its increasing popularity and recent fast growth.

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

Our Analysis is divided into four phases: Analysing and Organising Data, Data Cleaning, Business Analysis and Visualization, and Reviews Analysis. First, we collect the data from the Kaggle website. In the next step, we try to do data cleaning on the data set to reduce the error percentage. 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

- It is process of detecting the corrupt data, removing the irrelevant parts of the data and replacing the correct data. The actual process of data cleaning is to remove the error and validating the data.
- Data can be cross checked to remove the error. Issue can be resolved by validating the Data Preparation is the process of cleaning and transforming raw data prior to processing and analysis. It is an important step prior to processing and often involves reforming data, making correction to data, and the combining of the data sets to enrich data.
- Data cleansing is the process of detecting and correcting (or removing) corrupt or inaccurate records from the recordset, table or database and refers to identifying incomplete,incurrent, inaccurate or relevant parts of the data and then replacing, modifying or deleting the dirty or coarse data.
- We saw that the dataset contains many null or missing values

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:

Play store data

- App
- Category
- Rating
- Reviews
- Size
- Installs
- Type
- Price
- Content rating
- Genres
- Last updated
- Current version
- Android version

- App
- Translated review
- Sentiment
- Sentiment polarity
- Sentiment subjectivity

User Reviews

Data cleaning (Contd..)

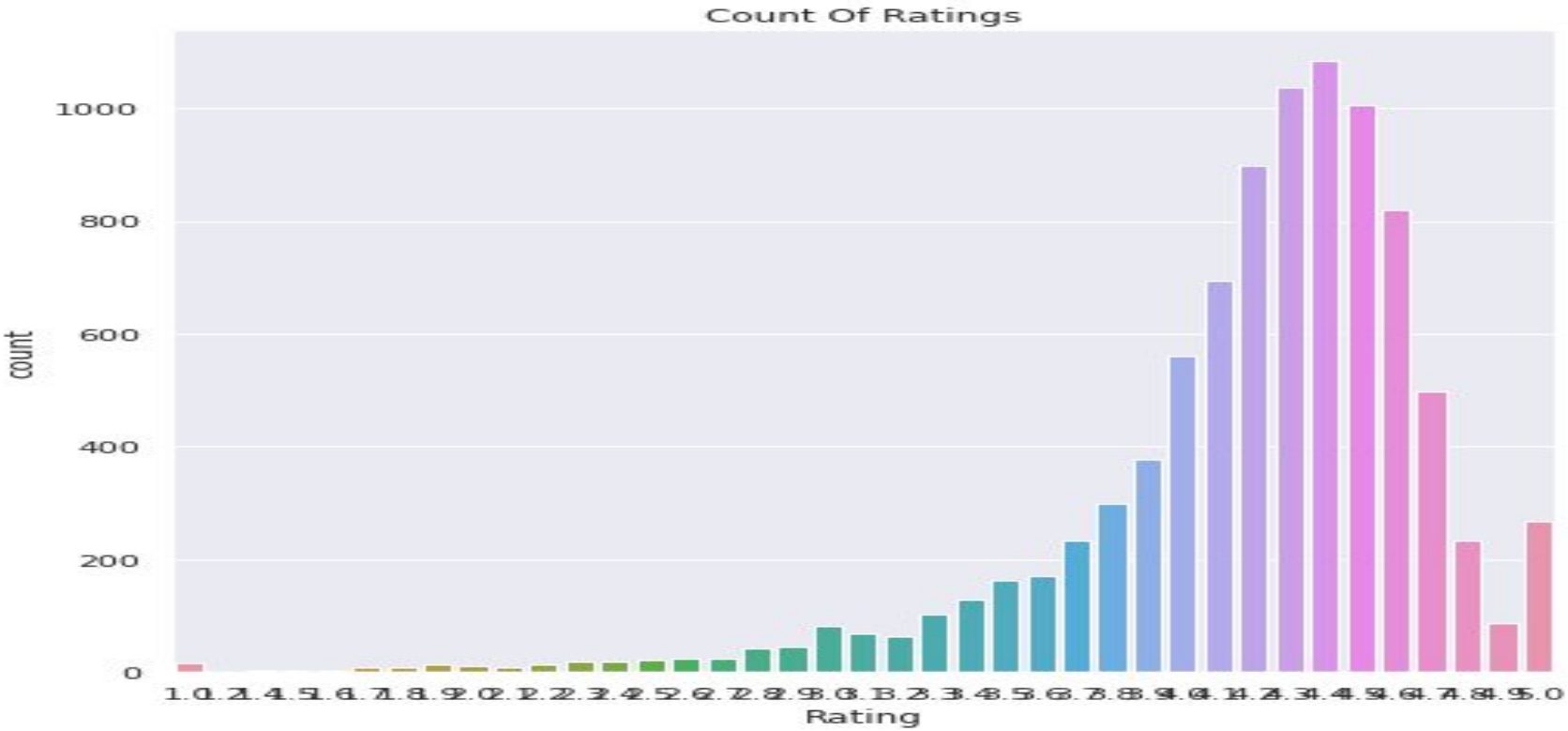
- Understand the structure of the dataset and clean data before analysis
- Finding Missing value in dataset
- Correct data type(INT,FLOAT,DATE)
- Replace null value with aggregate function (mean, mode,median)
- Checking outliers

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.

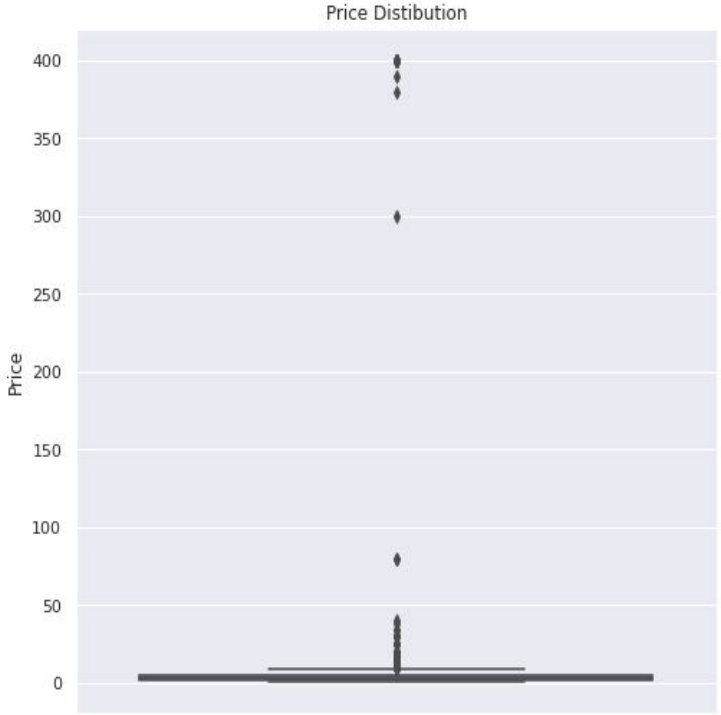
•Sizing of apps needs to convert in one measurement KB or MB. Pre-processing includes various tasks including stemming, lowercase conversion, Units, punctuation, and excluding terms.

Data Exploration

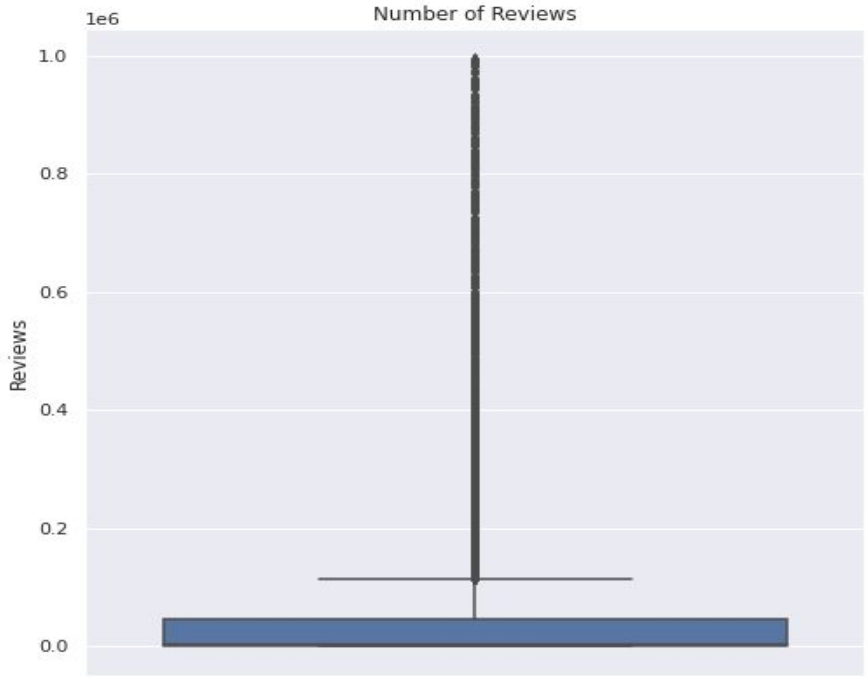


Outliers graphs

Price Distribution

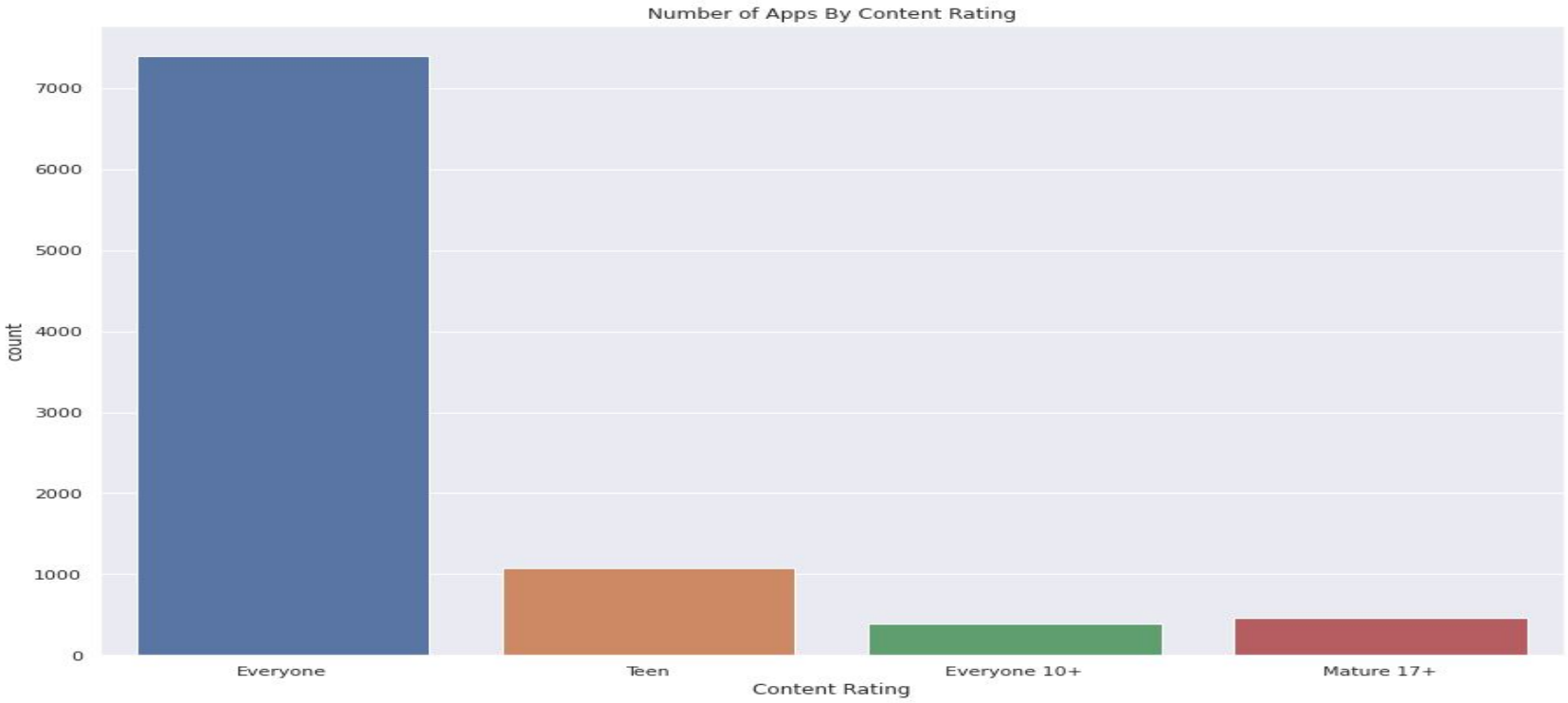


Number of Reviews

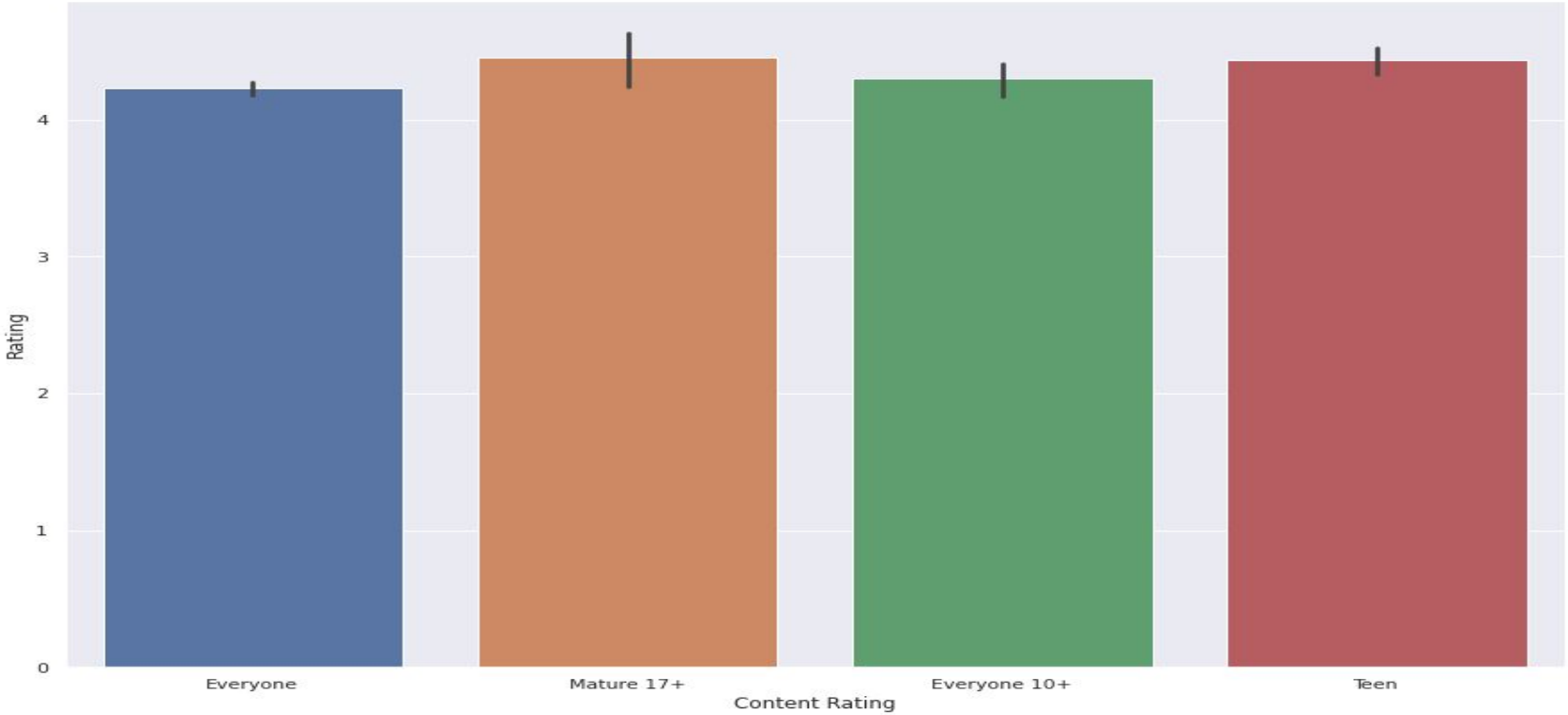


Top content Ratings values

Content Rating vs Apps(Count)

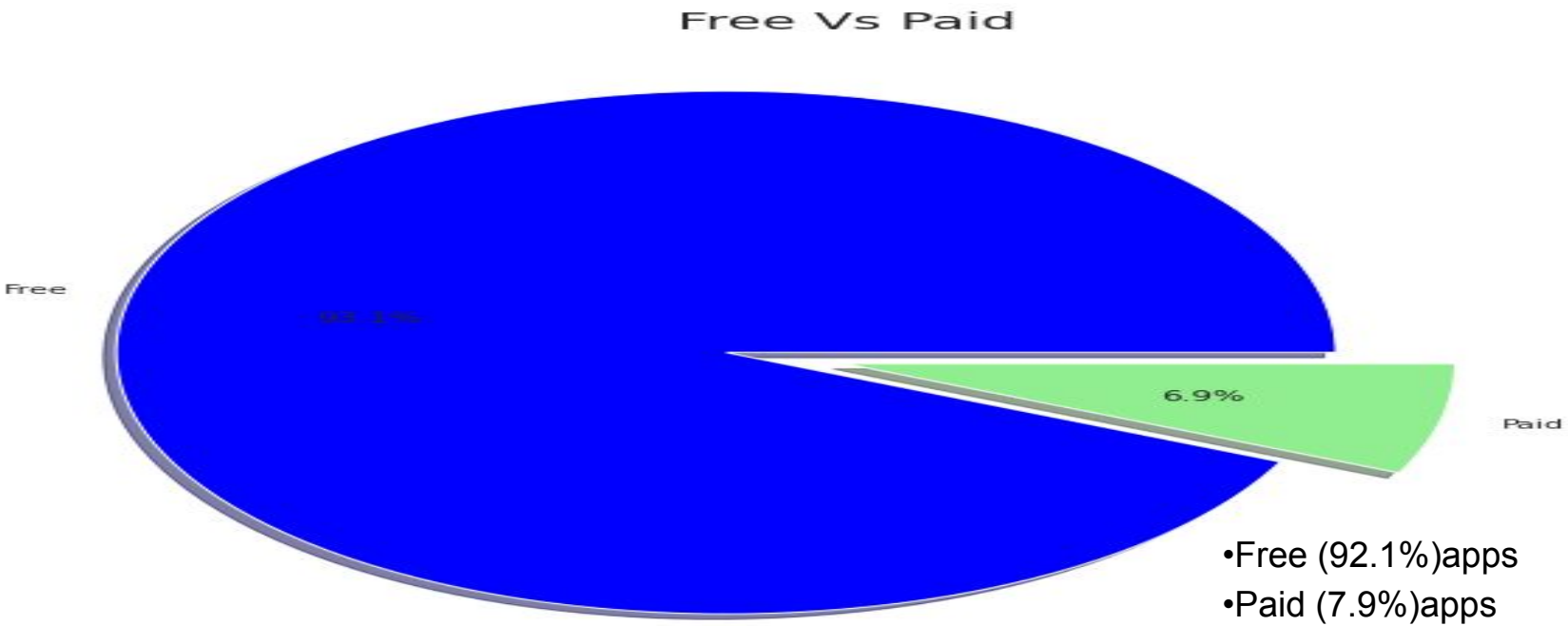


Content Rating vs Rating



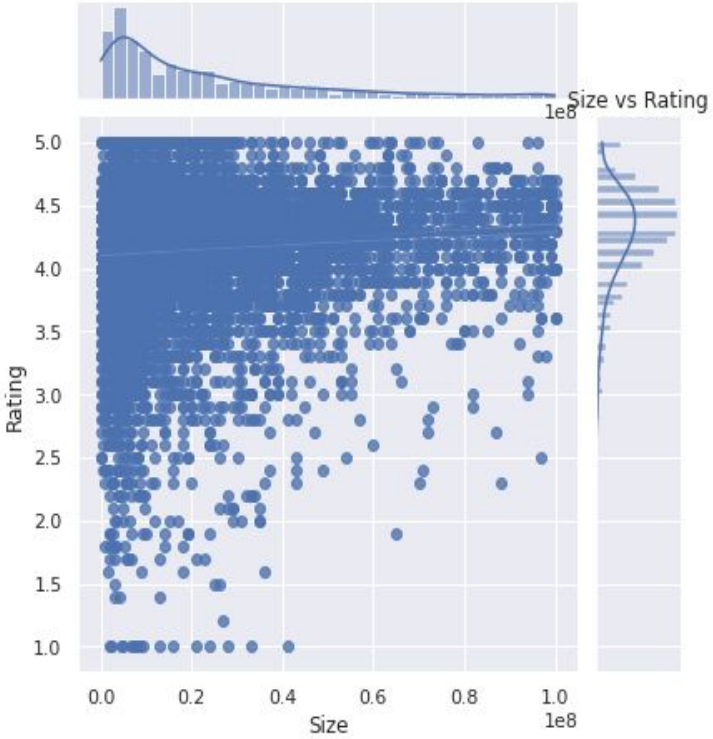
Pricing Strategies

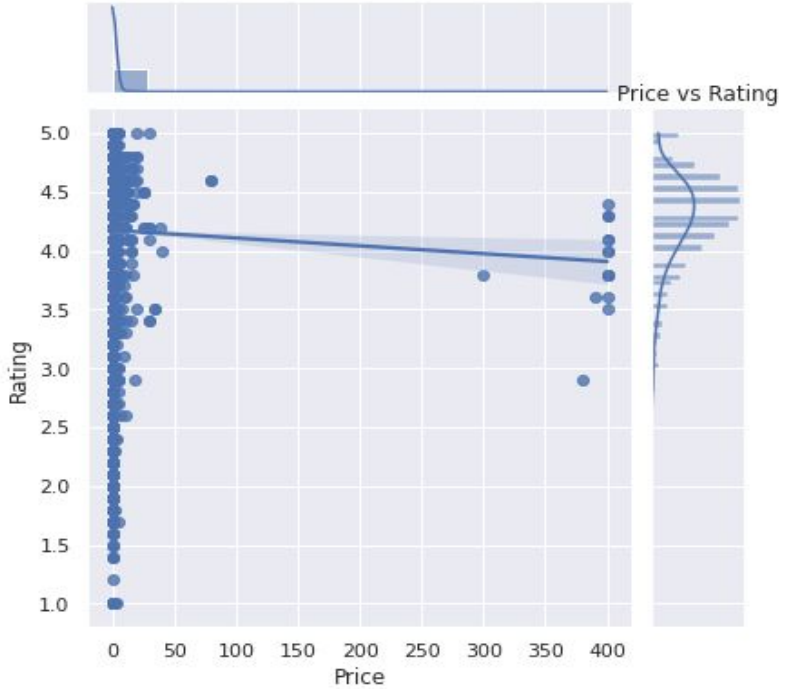
Since most Play Store apps are free, the revenue model is quite unknown and unavailable as to how the in-app purchases, in-app adverts and subscriptions leads to the success of an app. Thus, an app's success is determined by the number of installs and the user ratings that it has received over its lifetime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



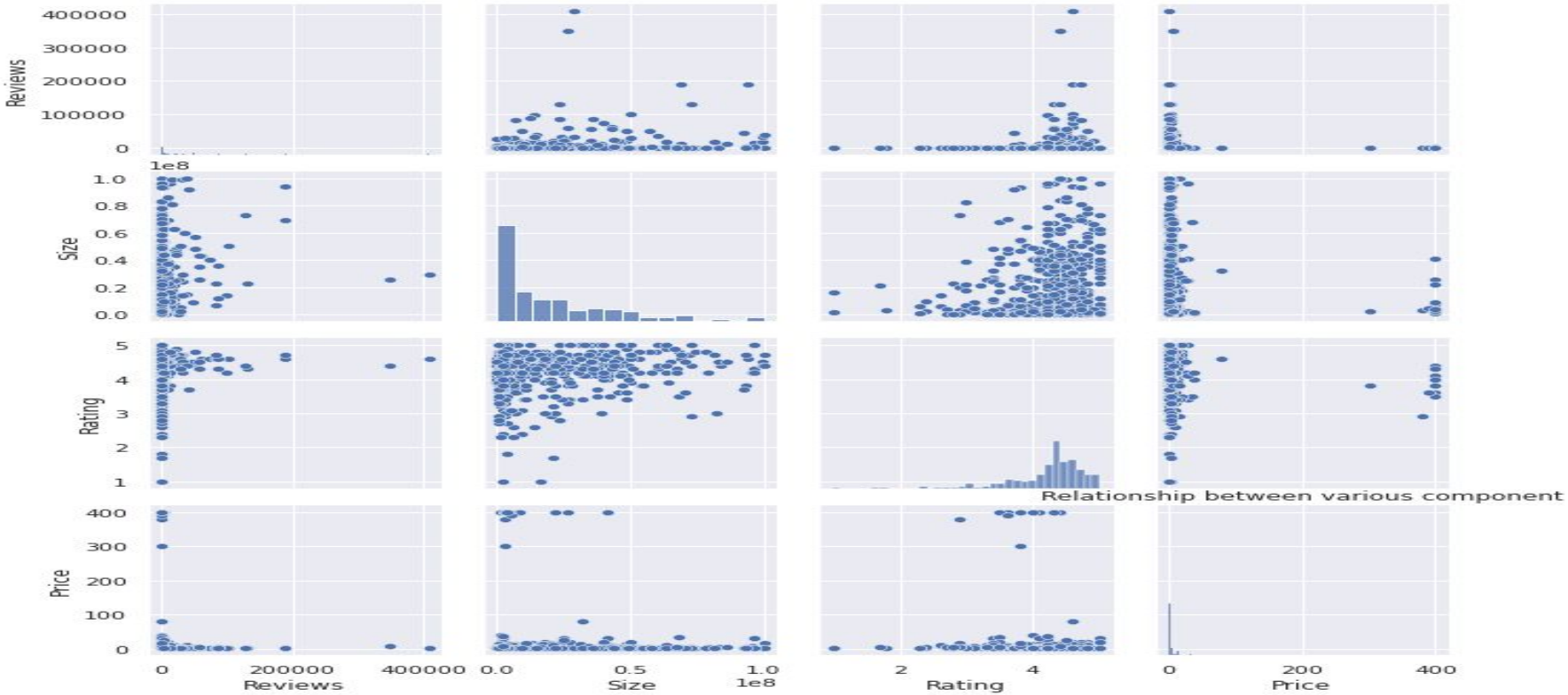
- Free (92.1%)apps
- Paid (7.9%)apps

Effect of Price and Size VS Rating





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



Basic 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

Sentiment Polarity vs sentiment subjectivity

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.

CONCLUSION

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