Capstone Project 1

Play store app review analysis

Done by:

Simran Dapke

Content

WHY ANALYZE THE PLAY STORE?

Mobile App Market is set to grow 20% by 2023

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

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

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

☆ Problem statement

- For this project I analyze Play store data of 2017-2018 ,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.

★ 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

User reviews

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

☆ 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

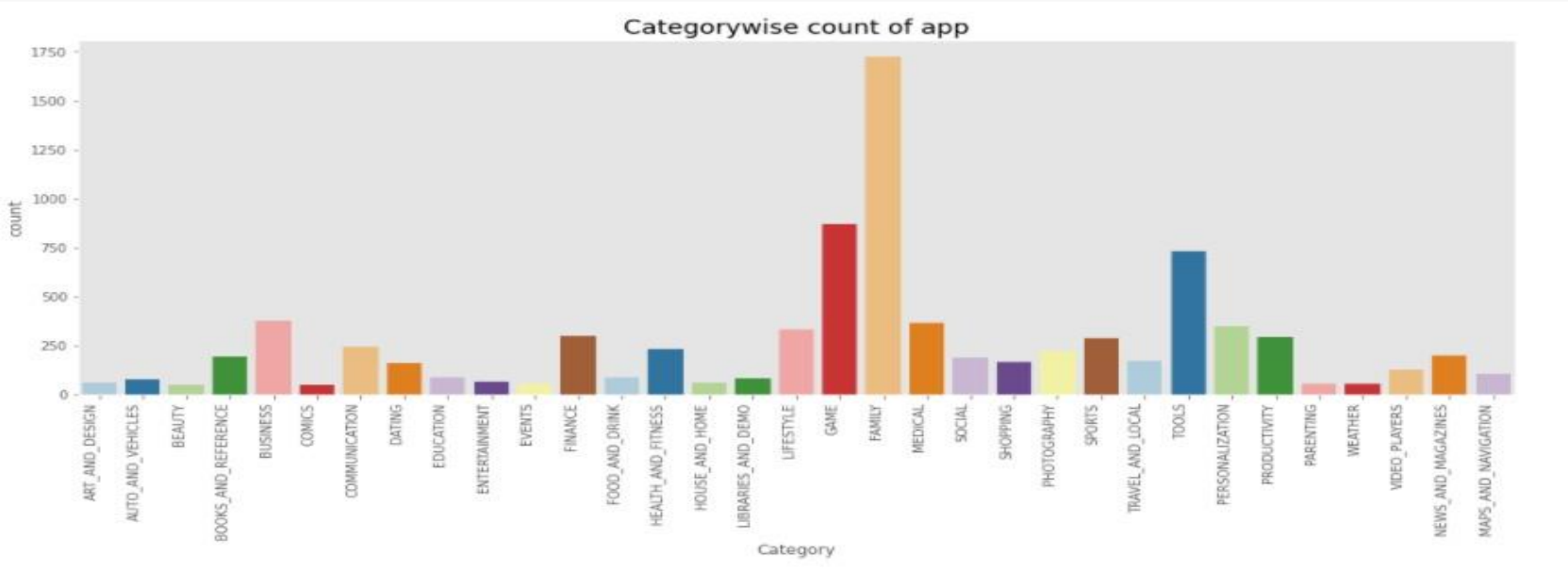
☆ Data Processing

- 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.

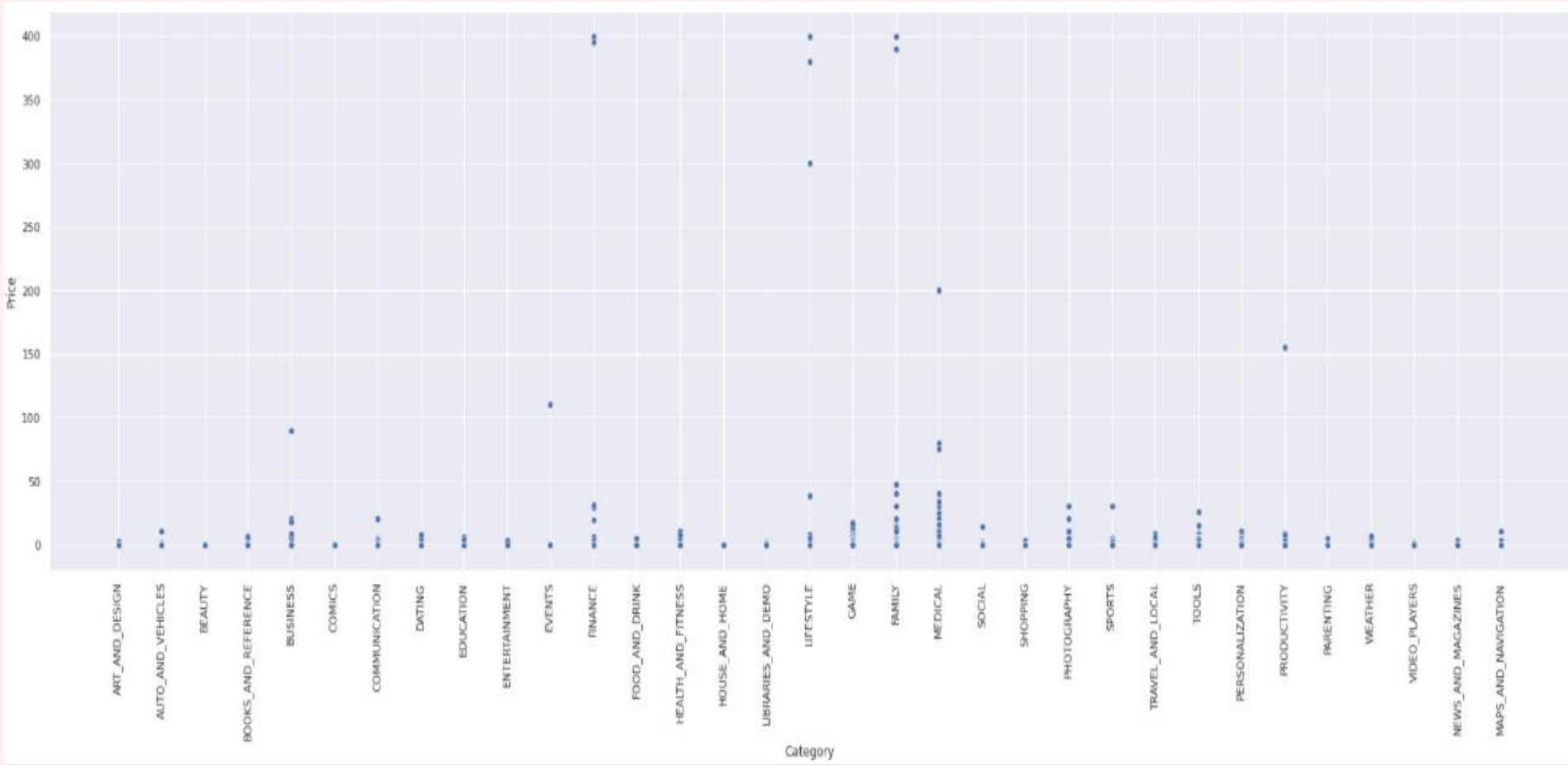
- Sizing of apps needs to convert in one measurement KB or MB. Pre-processing includes various tasks including stemming, lowercase conversion, Units, punctuation, and excluding terms.

⭐ Data Exploration

Which category contain most apps?

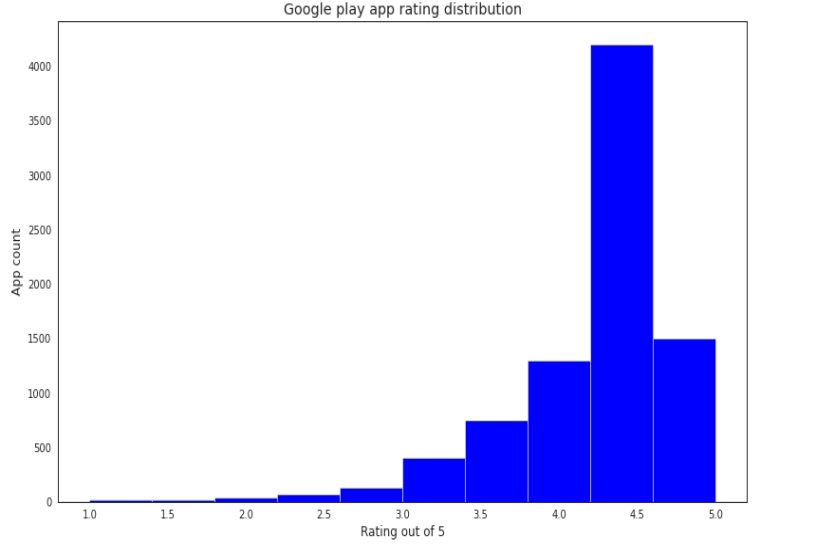


Shocking fact?



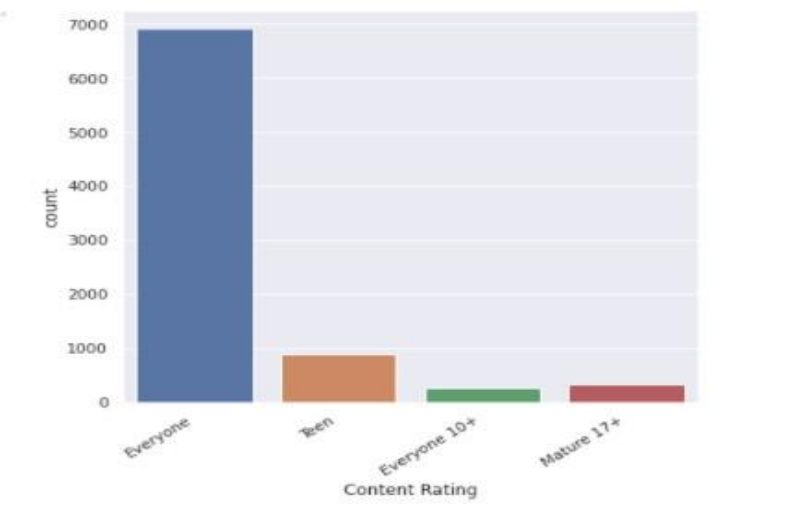
Rating of apps

Rating Distribution of app



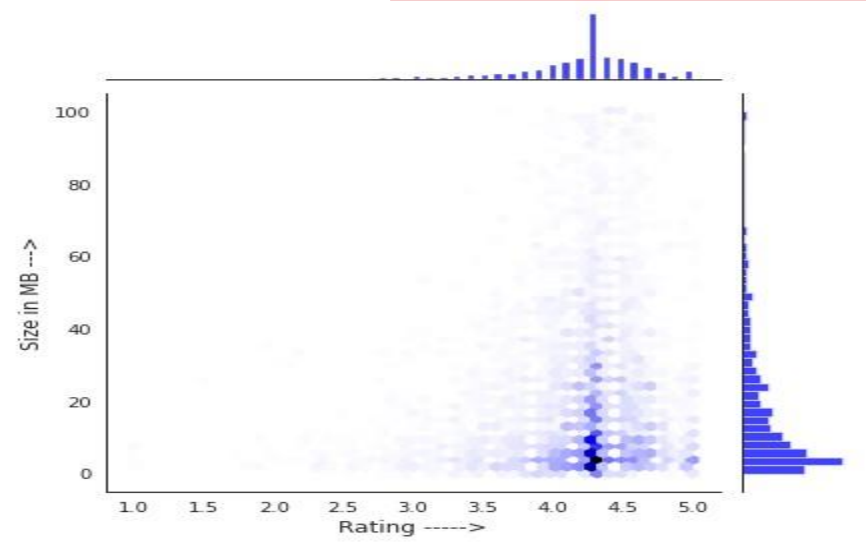
Average rating above 4.1

Content Rating of app



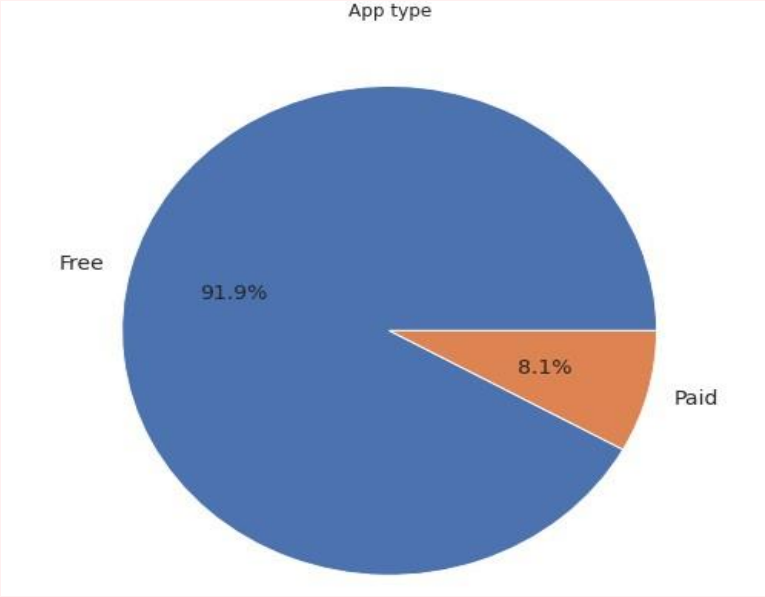
Most apps come under this everyone

Light Vs Bulky



- **Most of app size vary between 0.1MB to 40MB.**
- **Family category contain high size bulky apps.**

Pricing strategy?



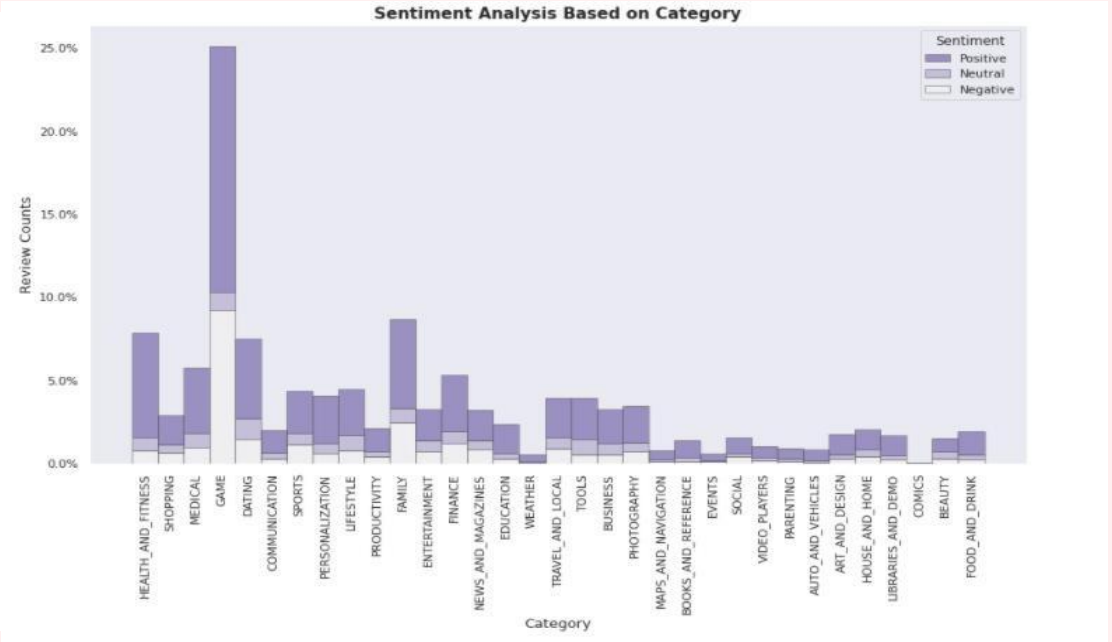
- Free (7742)apps
- Paid (680)apps

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.

Correlation

- moderate positive correlation of 0.6 exists between the number of reviews and number of downloads.
- customers tend to download a given app more if it has been reviewed by a larger number of people.

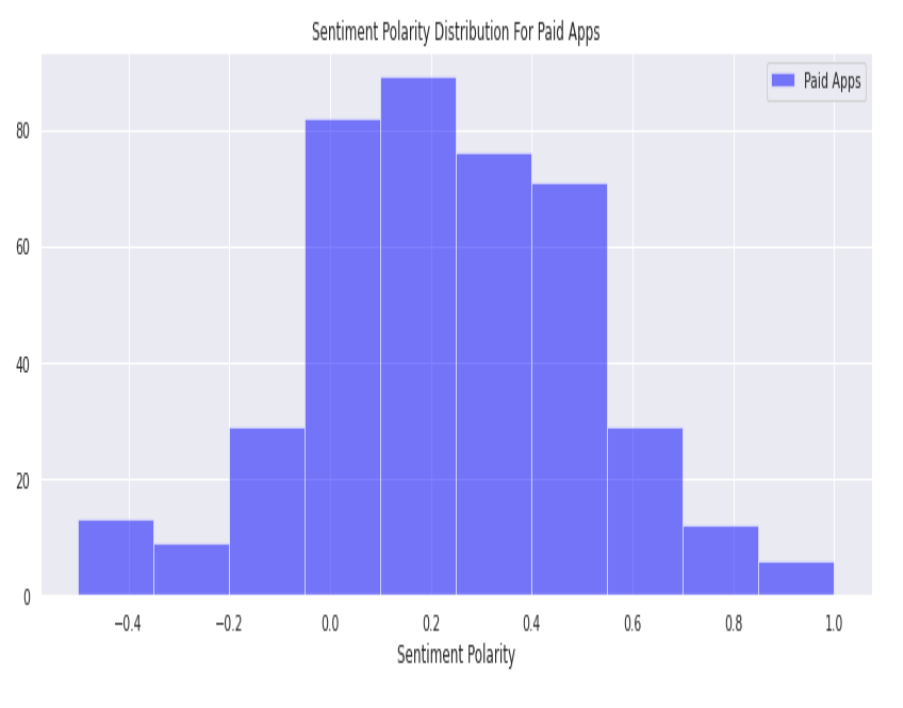
Sentiment analysis

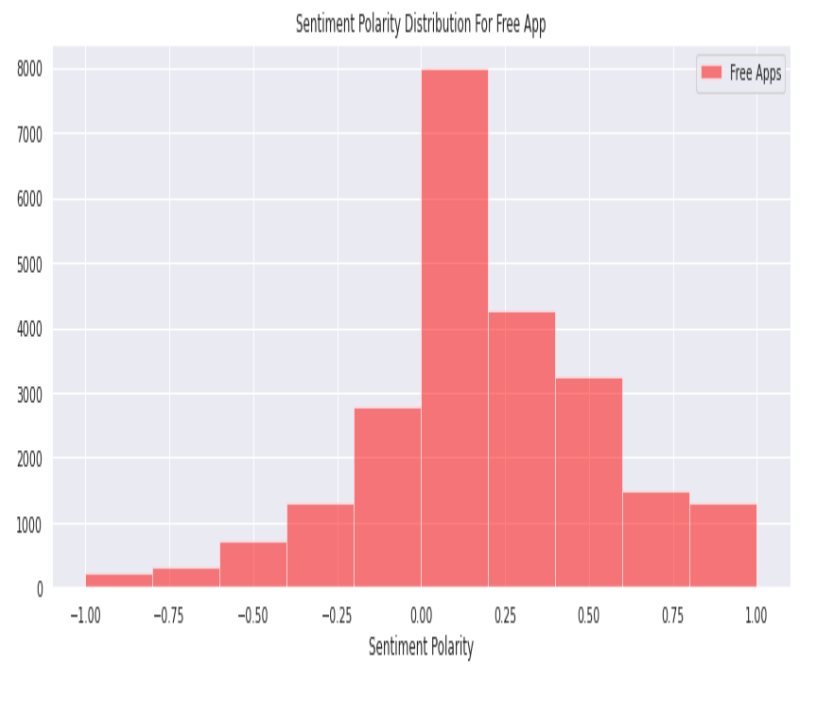


Free and paid has most positive reviews with high variance

Family, Sports and Health & Fitness apps perform the best, Having more positive reviews

Sentiment polarity distribution





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 surface 3)clash royal 4)Candy crush 5)UC-browser
❑ Most expensive app	I’m rich

☆ 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.

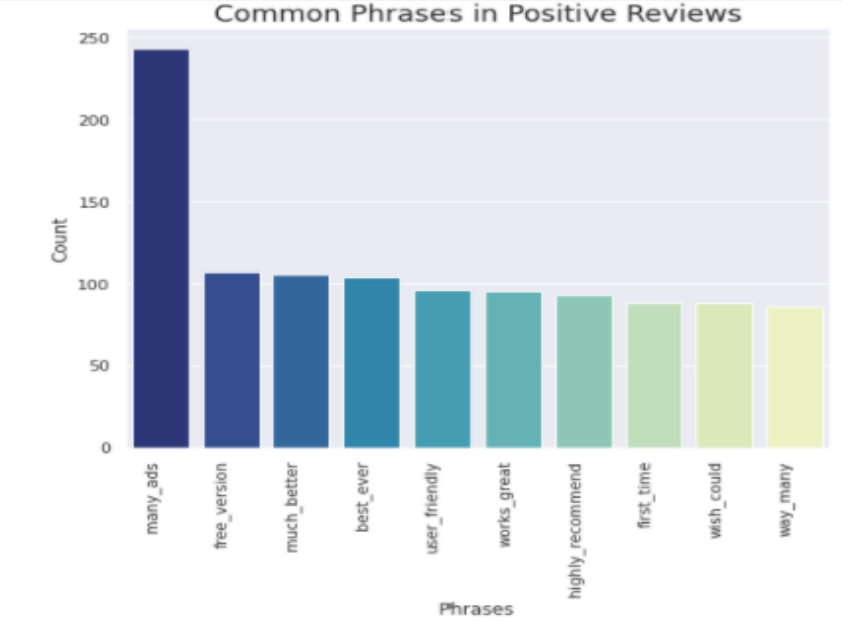
Sentiment Polarity

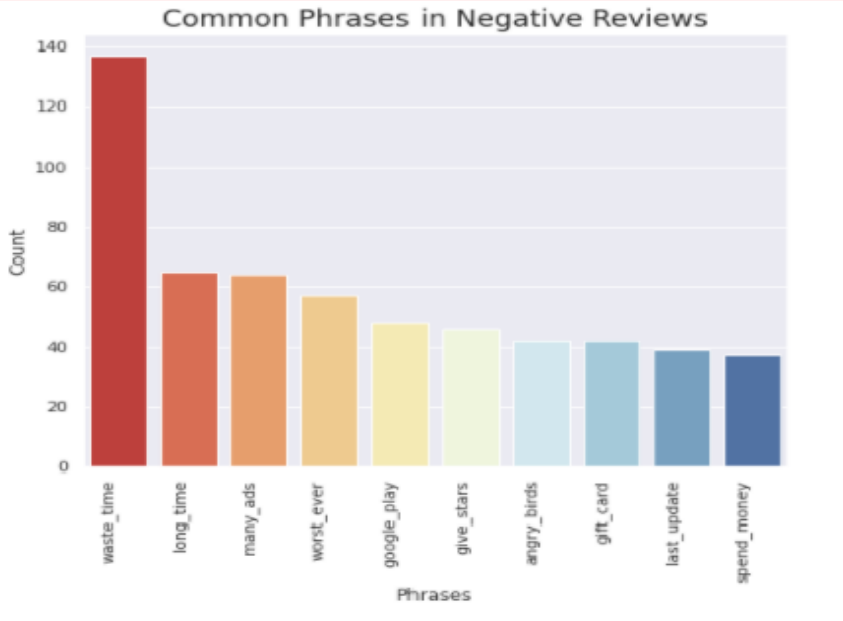
- The polarity of a sentiment measures how negative or positive the context is.
- In the data that we have, the polarity ranges from -1 (most negative) to +1 (most positive).

WORDCLOUD for FREE App

WORDCLOUD for PAID App

Highlighted reviews





CONCLUSION

- **Average rating** we see of apps on Google Play Store is **4.18**
- Most of user like light app and they pay for it also. Thus, paid app that is bulky may not do well in the market that’s reason for getting good installs.
- Most of the **top rated apps are optimally sized between ~0.1MB to ~40MB - neither too light nor too heavy.**
- **Medical and Family** apps are the most expensive and even extend upto 80$.

CONCLUSION

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

Challenges

- Data contain NULL/NAN values in dataset.
- Main task to clean data followed by data processing.
- Some data app name etc are in gibberish form and contain duplicates.

Future

- In this project I perform EDA and discovering relationships with specific features using sentiment of users.
- Developers can use my work for there research purpose to make app success.

Reference

1. Stackoverflow
2. GeeksforGeeks
3. Jovian
4. Research paper based on play store analysis.

THANK YOU