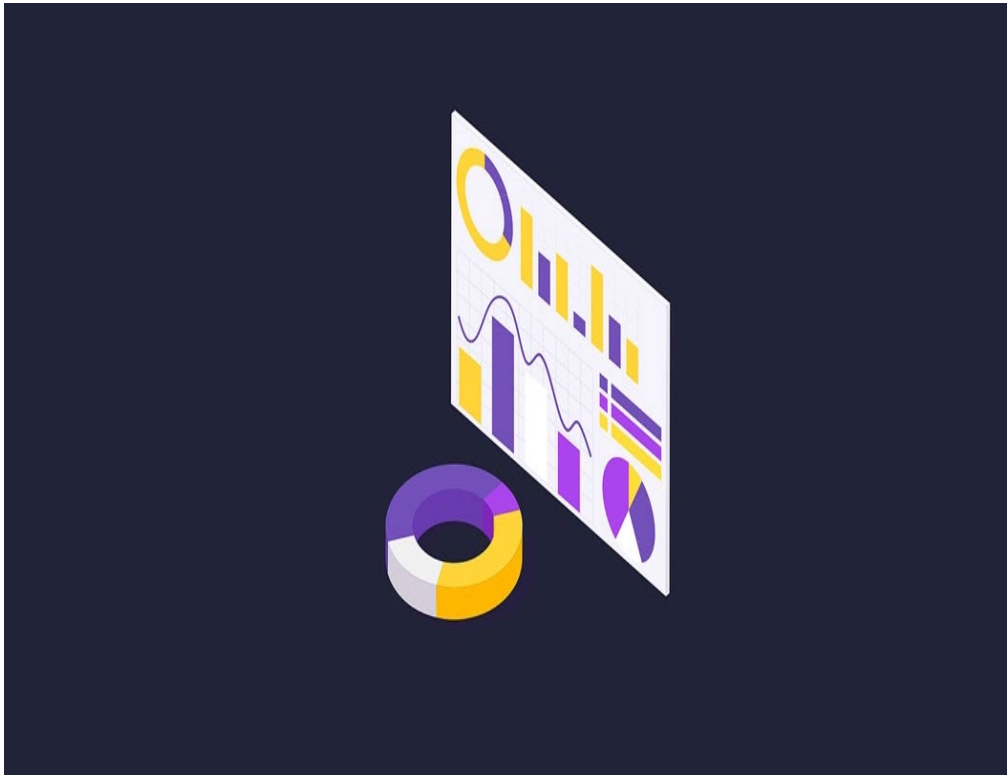
Capstone Project-I

EDA-Play Store App Review Analysis

By-Mahima Phalkey

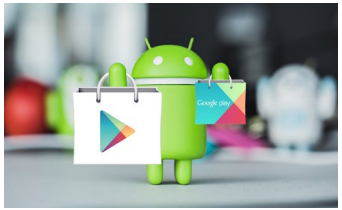
CONTENTS:

1. Agenda
2. Problem Statement
3. Introduction
4. Data Summary
5. Understanding Our Data
6. Data Cleaning
7. Data Processing
8. Observations
9. Data Visualizations
10. Conclusion
11. Challenges
12. Reference



AGENDA:

- ✓ The objective of this project is to deliver insights to understand customer demands better and thus help developers to popularize the product.
- ✓ With this project, we hope to be able to predict users' number of installs so that app developers and investors can choose the next big thing
- ✓ Dataset is of 10k Play Store apps for analyzing the Android market.

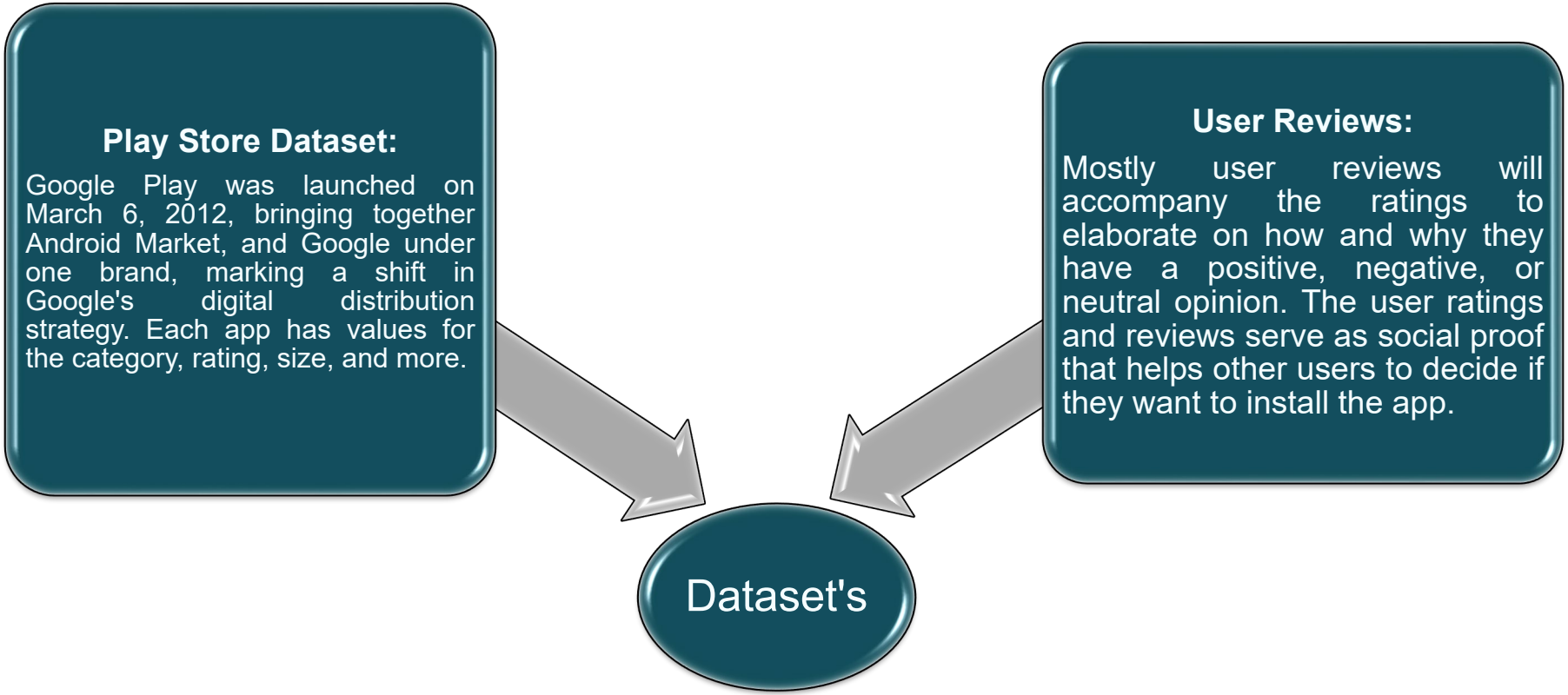


- ✓ This dataset contains details of different applications and reviews from different users.
- ✓ Beta focus groups may be run by companies, or app developers may find out what testers think of their product and receive feedback.
- ✓ The feedback and some prior knowledge about the app are used to predict how successful the app will be.

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 has values for the category, rating, size, and more.
- Another dataset contains customer reviews of the android apps
- In user has given reviews which are measured in positive, negative, or neutral sentiment.
- I have to explore both datasets to get better insights for apps success.

INTRODUCTION:



DATA SUMMARY

PLAY STORE DATASET

- App: Name of the application.
- Category: Apps are divided into different categories that categories are mentioned in this column such as Games, Travel, Books, Family, etc.
- Rating: Rating is been given out of 5 which is calculated as the avg ratings given by the user.
- Reviews: Number of reviews received from the user.
- Current_Ver: Latest version of the app.
- Size_in_MB: Memory Size of the application in MB.
- Installs: The number of installations of that particular application.
- Type: Whether that app is paid or free of cost.
- Price_in_dollars: Price of the app if not free.
- Content_Rating: This column specifies who can access this application that can be teen, everyone, mature 17+, etc.

DATA SUMMARY(CONTD…)

USER REVIEWS DATA SET

- Genres: The subcategory of the application.
- Last_Updated: The last updated date of the app.
- Android_Ver: Minimum Android version supported by the app.
- Sentiment_Subjectivity: This value ranges between 0 to 1. Lower values indicate factual information and higher values represent values indicating personal or public opinions.
- App: Name of the application.
- Translated_Review: Review in English by the user.
- Sentiment: The result of the review is either positive or Negative.
- Sentiment_Polarity: It expresses the sentiment which ranges between -1 to 1 where -1 represents negative review and 1 represents positive and others depend on the type of review.

UNDERSTANDING OUR DATA

- In the Playstore dataset there are 10,841 rows and 13 columns.
- In the User Reviews dataset there are 64,295 rows and 5 columns.
- There are 1,181 duplicate apps in play store dataset
- Except Rating column other columns are having data type as an object.
- 1,074 unique apps are having reviews in User Reviews data set
- In User Review there are 2 columns with datatype float and others are of object

DATA CLEANING

Null Values /Nan values Removal

- In the play store dataset, there were null values present in Rating, Type, Content_Rating, Android_Ver, Current_Ver these columns were having null values.
- Rating is numerical data so I have used the median to fill the null values whereas other columns are categorical values for that I have used mode to fill the null values.

Removing Duplicates

- There were duplicates present in the play store dataset.
- For that, I have used the Installs column.
- Firstly, I have changed the data type of the Installs column and then removed duplicates by keeping only those rows which are having a maximum number of Installs.

DATA CLEANING(CONTD…)

Data Type Correction

- Converting the columns to proper data type by removing unnecessary data from columns such as $ sign from Price column and making that as float type from object. Updating last updated column data type as date-time.
- Removing emojis from app name, Size column was having 2 types of size in KB and MB converting that to MB, etc.

Outlier Detection

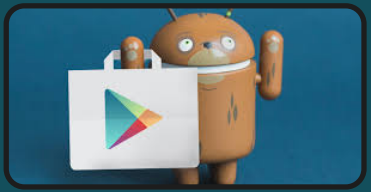
- An outlier is a point or set of points that are different from other points.
- Sometimes they can be very high or very low.
- It's often a good idea to detect and remove the outliers.
- In Play store dataset there was one outlier with rating as 7 which was removed while removing duplicates using Installs.

DATA PROCESSING

- The data provided is not structured data. We need to process that data into the proper format.
- In the play store dataset in the Size column the data given was in two formats in KB and MB so we need to process this data into one format i.e., in MB.
- In Android Ver and Current Ver column hyphen was present and some characters were present in the column so we need to process that as well by removing them.
- In Android Ver and up was present which was not useful while analyzing so I removed that as well.

OBSERVATIONS

1. There are 266 apps with the highest rating i.e. 5.
2. Travel and Local is the category which is having the highest number of ratings as 5.
3. Average rating is 4.1

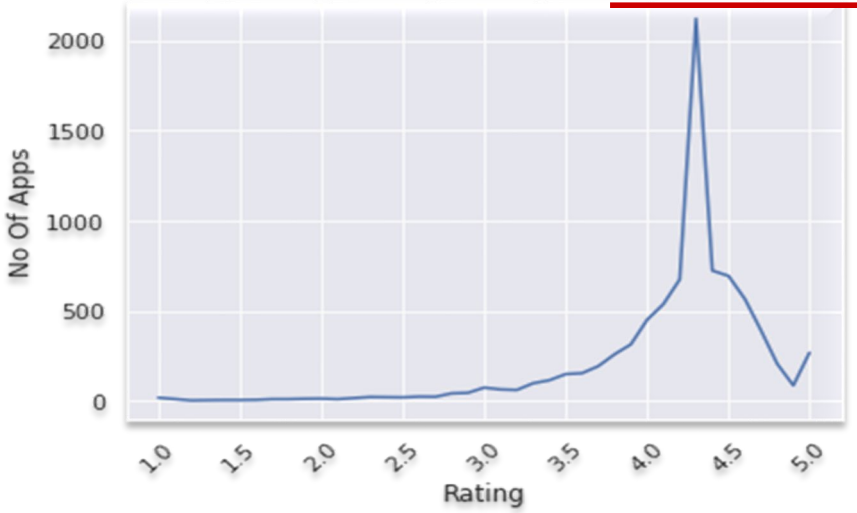


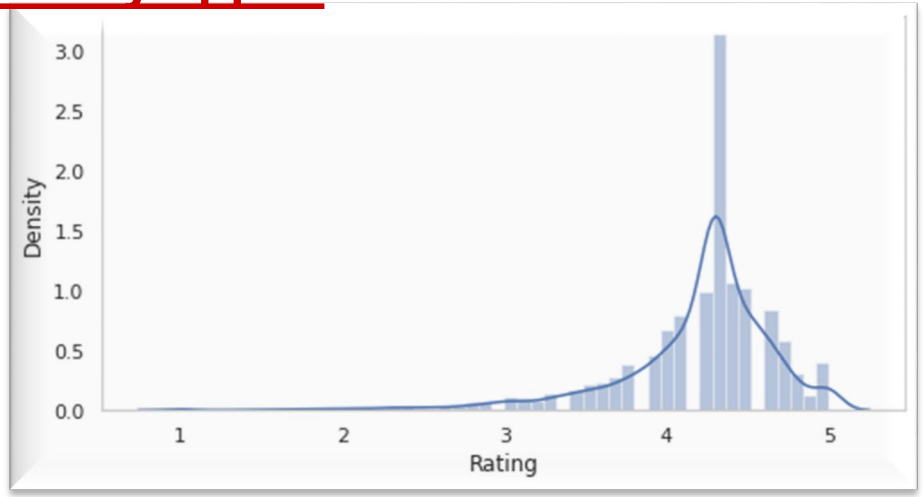
1. **There are 7747 free apps and 683 paid apps.**
2. **“I'm Rich - Trump Edition" this app is having the highest price i.e 400 $**

1. **"Clash of Clans" is the app with the highest review which comes under the game category which is free of cost.**
2. **The first 4 apps with reviews are of GAME category Clash of Clans, Subway Surfers, Clash Royale, Candy Crush Saga**

DATA VISUALIZATIONS

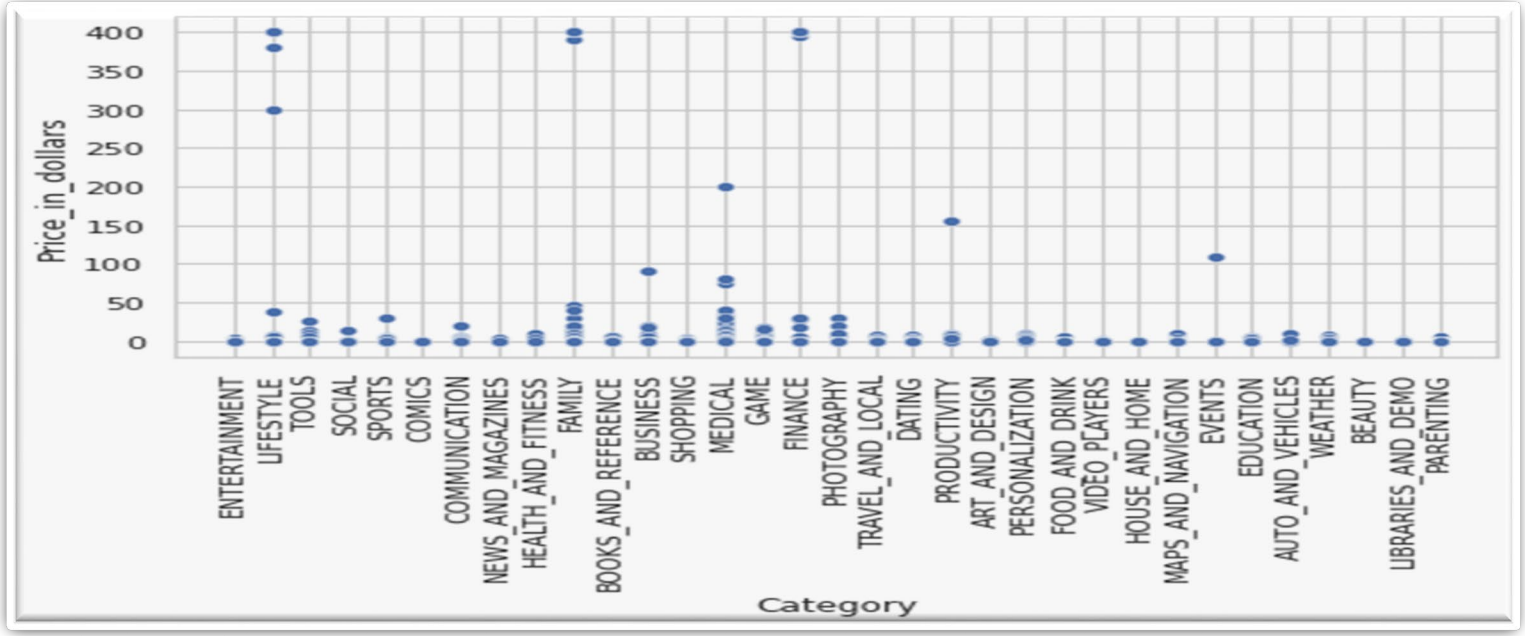
What is the maximum number of Rating and for how many apps?





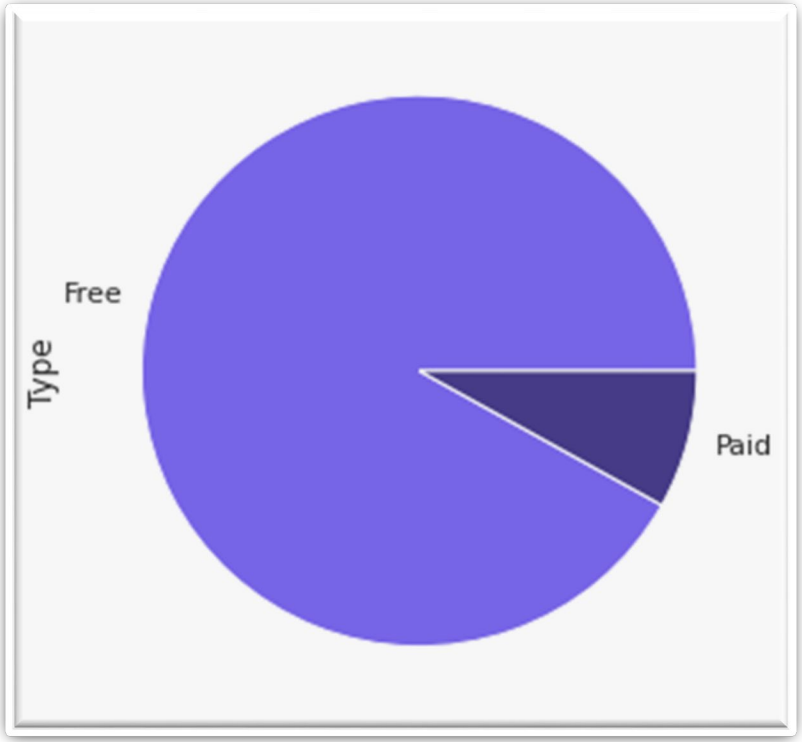
- The highest number of apps are having ratings between 4.0 and 4.5. There are 2000+ apps with ratings between 4.0 to 4.5.

Which Category has the more priced apps?



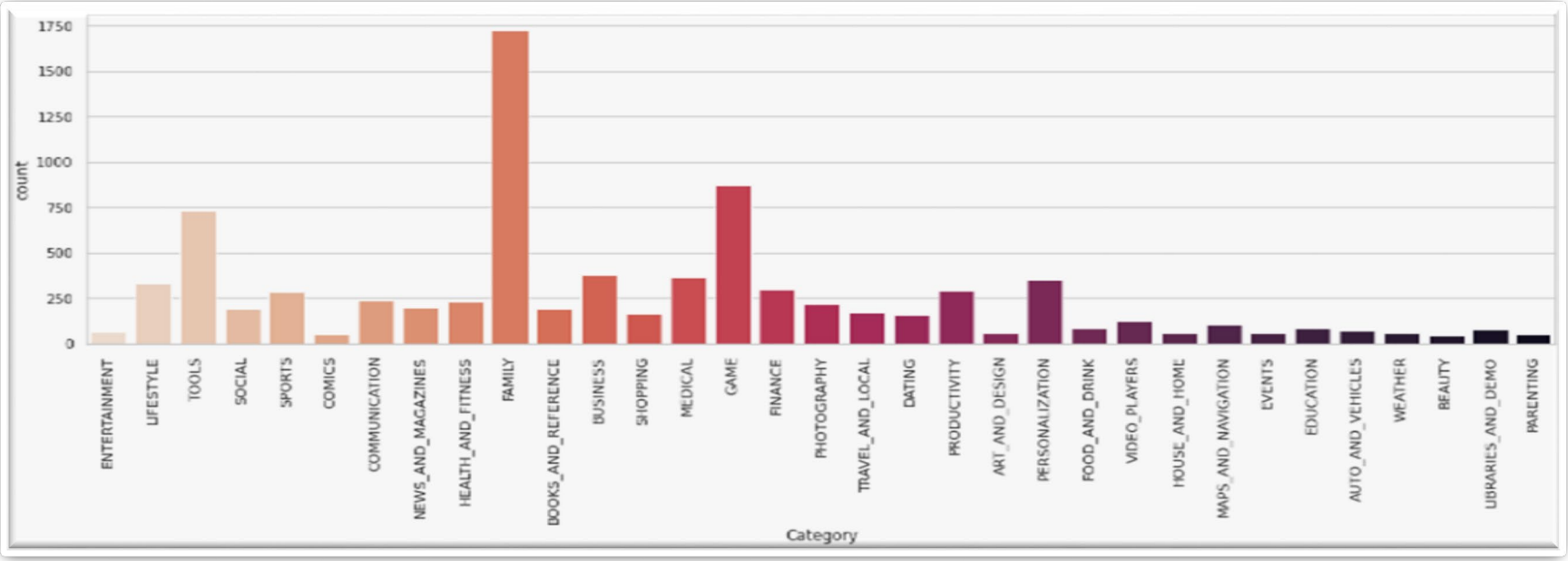
- **Family, Lifestyle and Finance categories are having highest price. Most of the categories are having prices less than or equal to 50 $.**

FREE VS PAID APPS



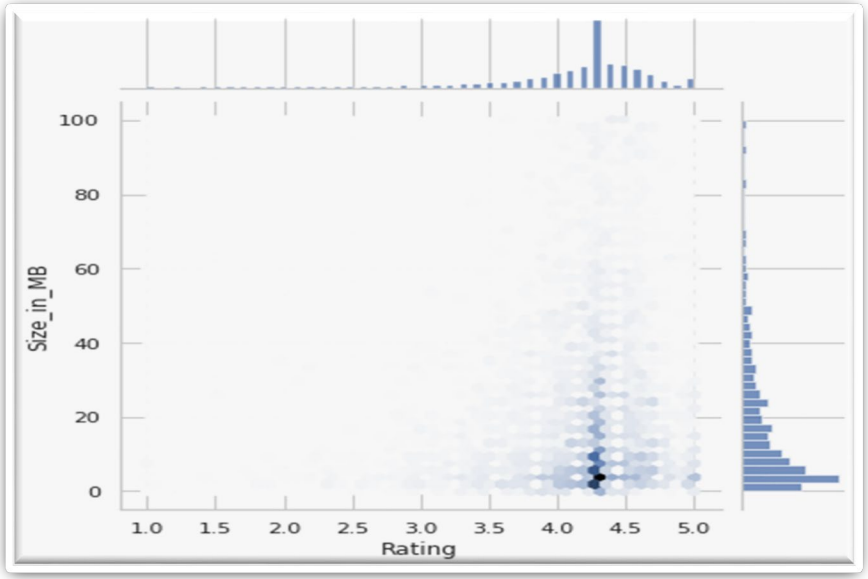
- Free apps are more as compared to Paid apps.
- Nearly 91 % of apps are free of cost 7747 apps are free of cost and 683 are paid apps.

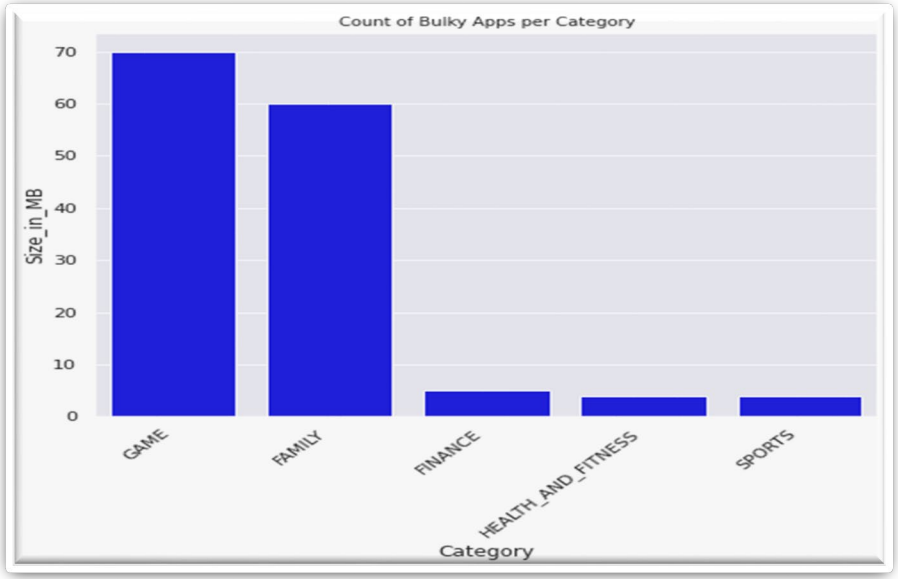
What is the count of Categories?



- From this plot, we can see that most of the apps that belong to the family category is having the highest count. Game is the second-highest category.

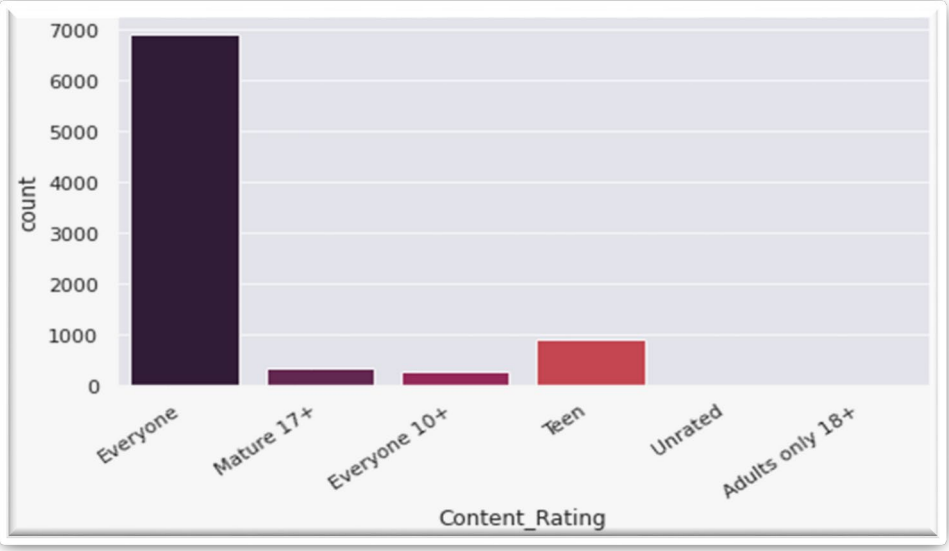
HOW MANY APPS ARE LIGHT OR BULKY?





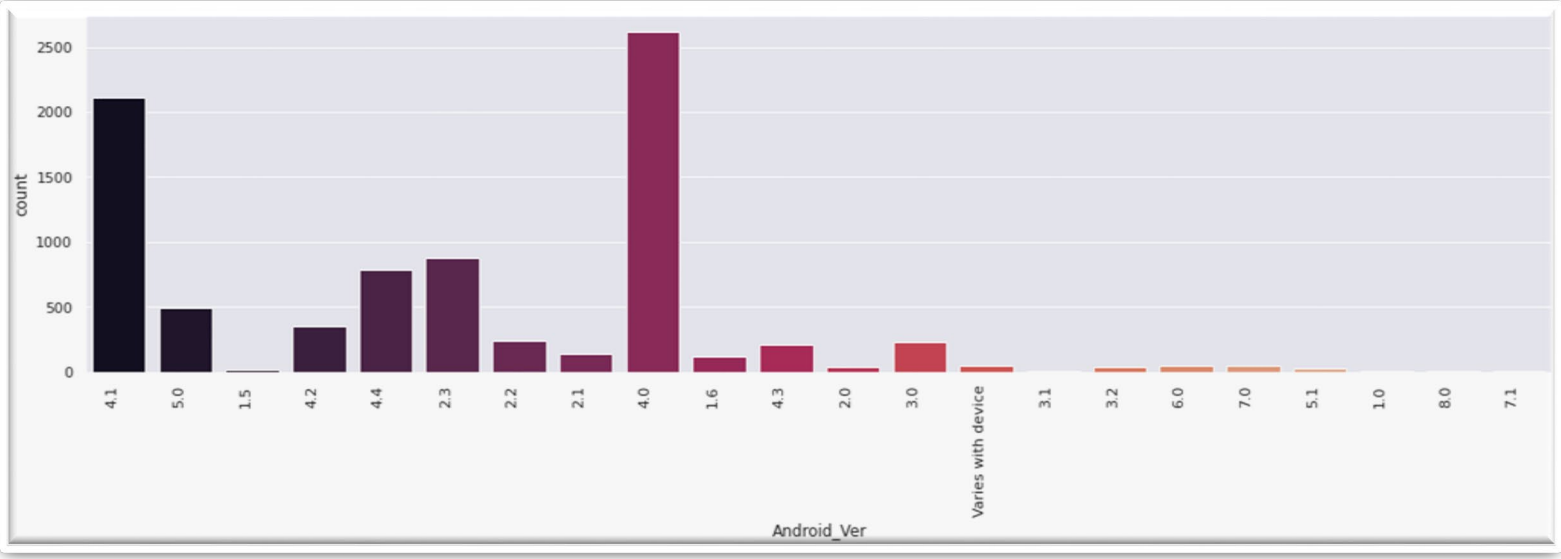
- Apps with the highest rating i.e., between 4.0 to 4.5 are having sizes between 0-20 MB
- We can see Gaming and Family categorical apps are bulky.

CONTENT RATING COUNT PLOT



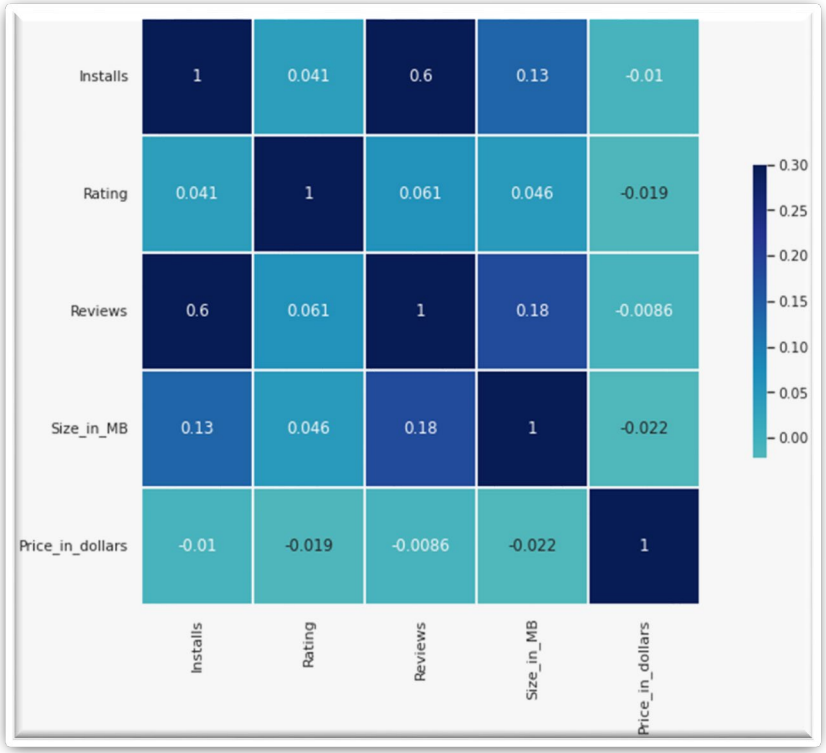
- **Most of the apps are not age-restricted. Every age group can access the apps.**

Minimum Android Version



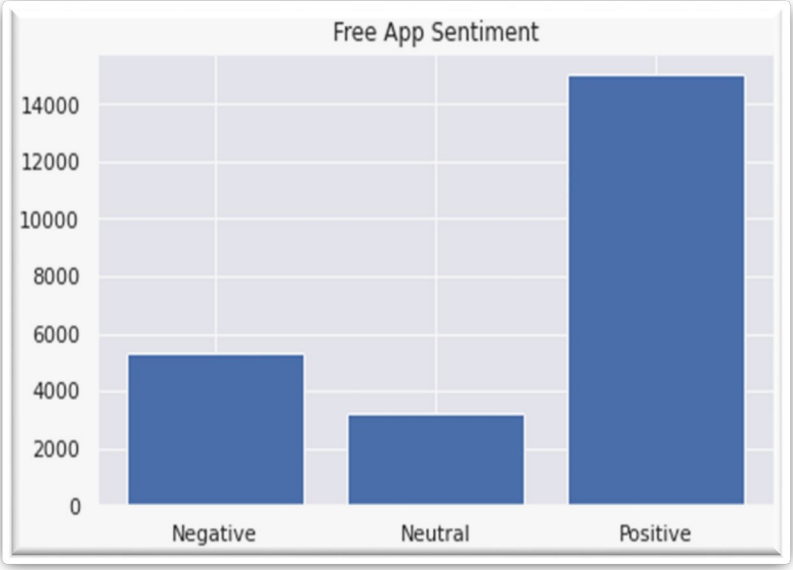
- Many apps are having lowest Android Ver preference as 4.0

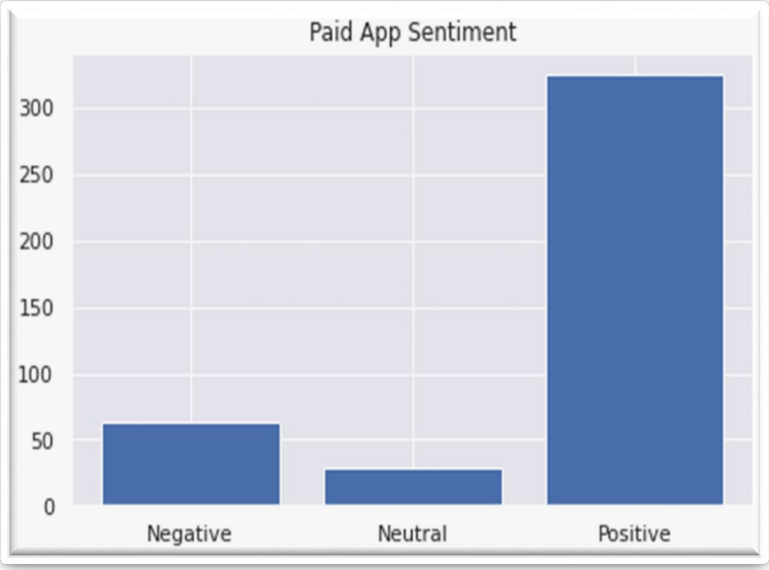
HEAT MAP FOR PLAY STORE DATASET



- To determine the pairwise correlation of all columns in the data frame, Pandas dataframe.corr() is used.
- By default Pearson correlation is used.
- The number of installs is greater if reviews are higher.
- A positive correlation exists between Installs and reviews i.e., 0.6.

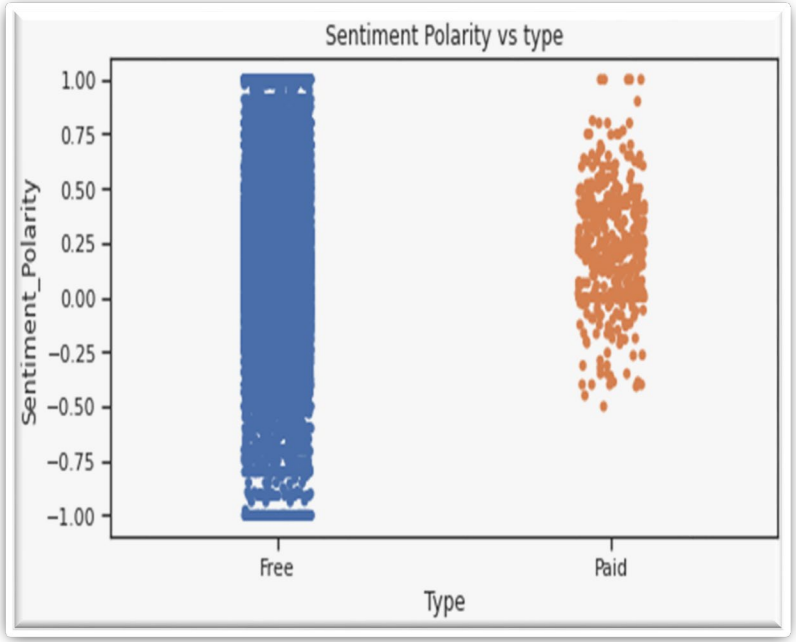
Type Vs Sentiment





- In the free app, we are having positive responses more whereas negative and neutral are very less compared to positive responses.
- In the paid app, we are having positive responses more whereas negative and neutral are very less compared to positive responses.

Sentiment Polarity Vs Type



- The polarity of sentiment measures how negative or positive the context is.
- In the data that we have, the polarity ranges from -1 to +1. In the strip plot, we can see that for free apps the sentiment polarity value count is very high compared to paid apps.
- Free apps are having Positive, Neutral, as well as positive sentiments whereas Paid Positive sentiments, are more

Word Cloud of Free and Paid Apps

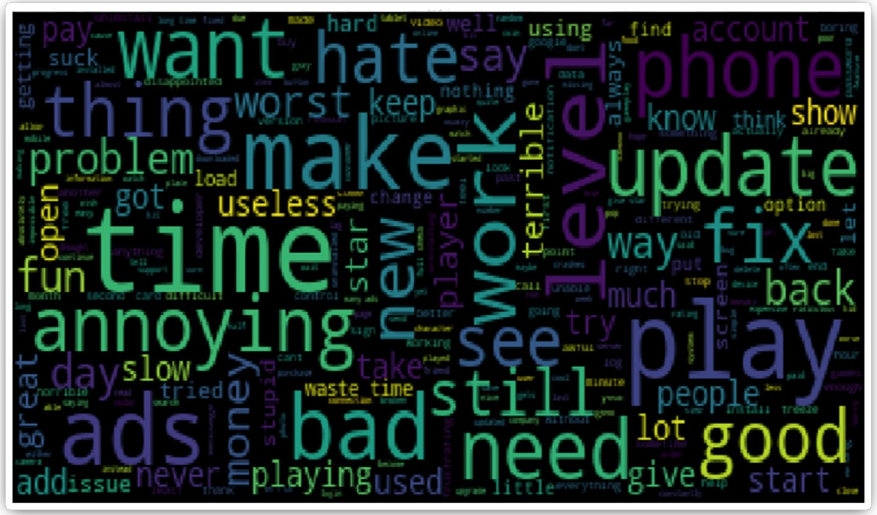
- **Free App:**

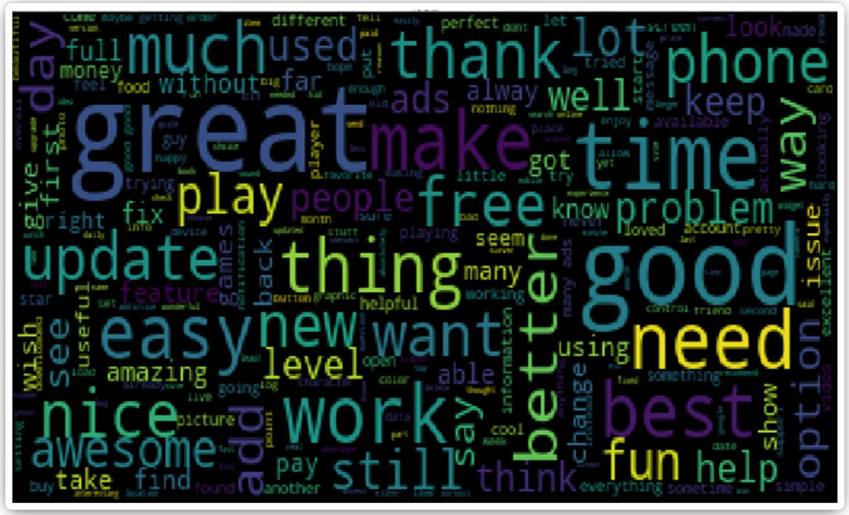
The most frequently used words in free apps are great, game, good, work, etc.

- **Paid App:**

The most frequently used words in paid apps are great, game, good, time, good, etc.

Word Cloud of Positive and Negative Reviews





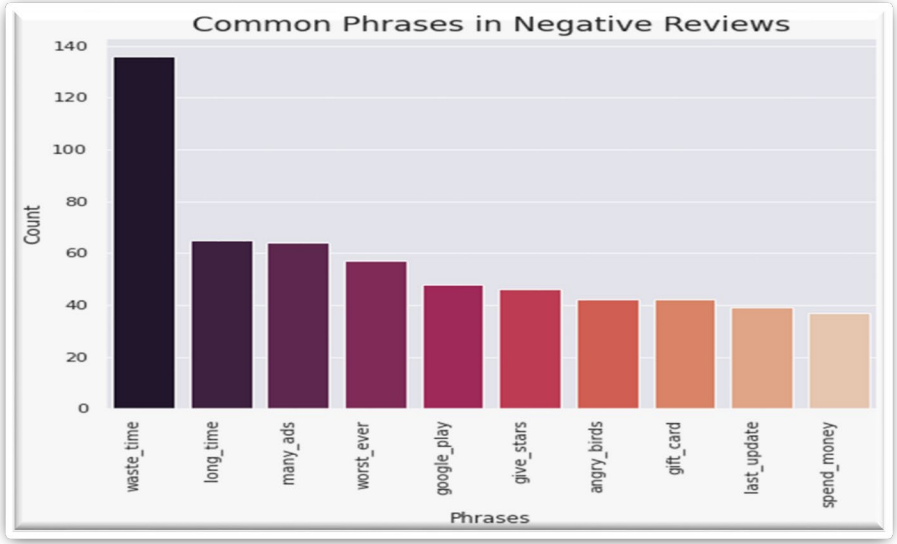
- **Negative Reviews:**

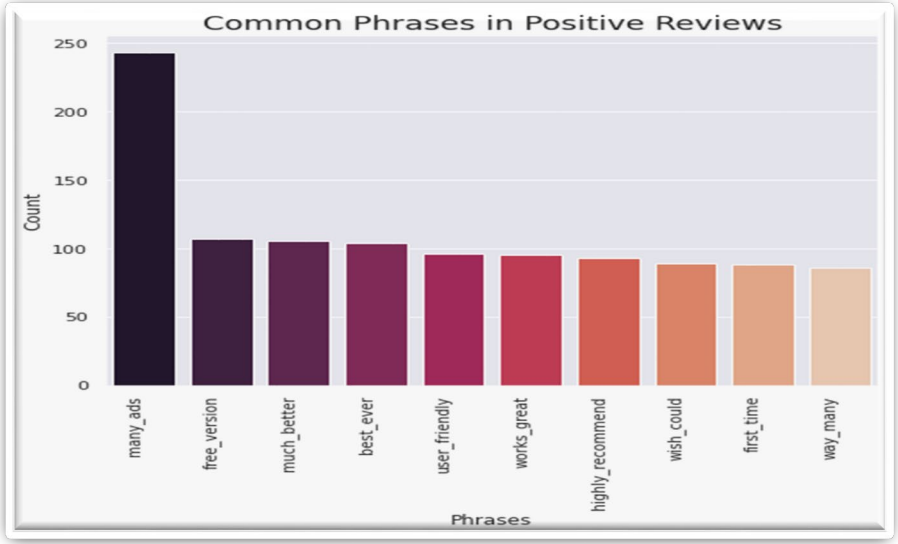
Most frequently negative review words are time, update, bad, ads, etc.

- **Positive Reviews:**

Most frequently positive review words are great, good, work, awesome, fun, etc.

COMMON PHRASES USED IN REVIEWS





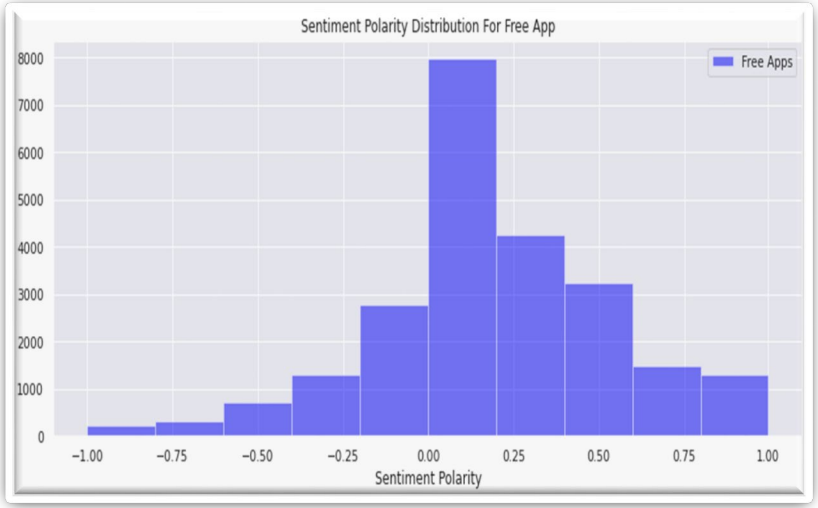
- **Common Phrases in Negative Reviews:**

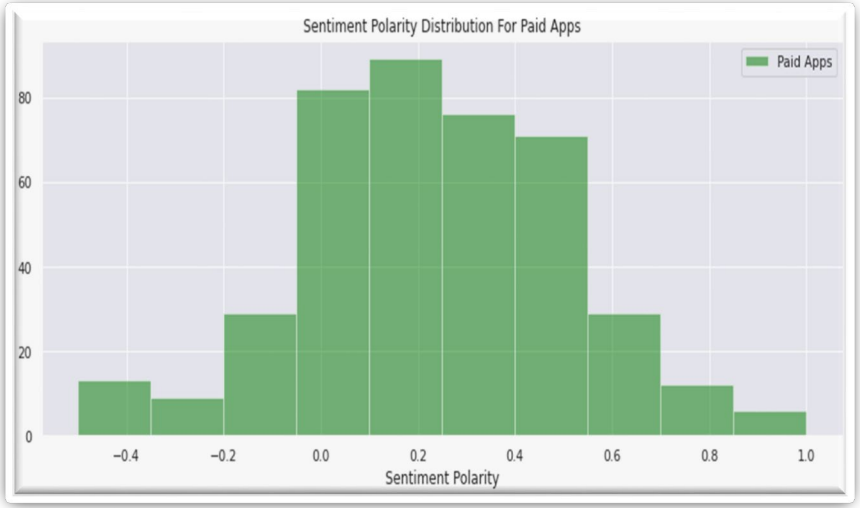
waste_time is most used phrase in negative reviews.

- **Common Phrases in Positive Reviews:**

many_ads is a most used phrase for positive reviews.

Sentiment Analysis for Free and Paid Apps





- The most sentiment polarity is in between 0.00 to 0.25 for free apps.

- The most sentiment polarity is in between 0.1 to 0.2 for paid apps.

CONCLUSIONS

1. The average rating for apps is between 4.0 and 4.5. There are 2000+ apps with ratings between 4.0 to 4.5.
2. Lifestyle, Family and Finance categories are having the highest price. Most of the categories are having prices less than or equal to 50 $.
3. Many apps are having lowest Android Ver preference as 4.0.
4. Family category is having the highest count. Game is the second-highest category.
5. The number of installs is greater if reviews are higher.
6. Apps with the highest rating i.e., is between 4.0 to 4.5 is having sizes 0-20 MB. Gaming and Family categorical apps are bulky but for the Gaming category, most people download apps even if it is bulky.
7. Travel and Local is the category which is having the highest number of ratings as 5.
8. "I'm Rich - Trump Edition" this app is having the highest price i.e., 400 $ but is not Installed by many users.

CONCLUSIONS

9. “Clash of Clans” is the app with the highest review which comes under the game category which is free of cost. The first 4 apps with reviews are of GAME category Clash of Clans, Subway Surfers, Clash Royale, Candy Crush Saga
10. Clash of Clans app has the greatest number of reviews. While Subway Surfers is the greatest number of install app
11. When it comes to free apps, users are more pessimistic and harsher than when it comes to paid apps.
12. 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

- While data cleaning there were columns with a lot of unnecessary data. For ex. In the Current Ver column, there were hyphens and it was also containing random strings at the end.
- In some app names emojis were there we need to clean those as well.

REFERENCE

- https://www.analyticsvidhya.com
- https://www.kaggle.com
- https://www.geeksforgeeks.org/overview-of-data-science/
- https://towardsdatascience.com/exploratory-data-analysis-in-python-c9a77dfa39ce
- https://techvidvan.com/tutorials/python-sentiment-analysis/
- stack overflow


THANK-YOU