Capstone Project-1 Submission

Play Store App Review Analysis

Manjeet Srivastava

Vivek Kumar

Rishabh Srivastav

Shubham Bharti

Durga Shankar

Data science trainees

AlmaBetter, Bangalore

----------***----------

Manjeet Srivastava- srivastavamanjeet01@gmail.com

Vivek Kumar-viku35@gmail.com

Rishabh Srivastava- rishabhsri999@gmail.com

Shubham Bharti- shubhambharti731@gmail.com

Durga Shankar- dspathak50@gmail.com

GitHub Link –

Manjeet Srivastava - https://github.com/manjeetsrivastava/play-store-app-reveiw-analysis

Vivek Kumar - https://github.com/endlessstory35/google-play-store

Rishabh Srivastav - https://github.com/RishabhSrivastav-1994/play-store-app-review-analysis

Shubham Bharti -https://github.com/Shubh-am1999/play_store_app_analysis.git

Durga Shankar - https://github.com/bigdeepak/play-store-app-reveiw-analysis

Abstract-

This paper presents an exploratory data analysis of the Google Play store app data using Python. The data set used consists of 10,000+ apps and their respective details such as ratings, size, category, etc. The analysis begins with basic data cleaning and wrangling techniques, followed by visualizing the data to gain insights. The analysis focuses on understanding the most popular app categories, the rating distribution of apps, the app size distribution, the free and paid apps and the correlation between app ratings and their sizes. The results of the analysis reveal that the most popular category is Games, followed by Tools and then Entertainment. The ratings distribution of apps is heavily skewed towards the higher ratings, indicating that people are mostly satisfied with the apps. The size of the apps is also heavily skewed towards the lower size, indicating that most apps are not too large. Finally, the correlation between app ratings and sizes is weak and insignificant.

***Key Words*:** Google Play Store Apps, Ratings Prediction, Exploratory Data Analysis, Machine Learning.

1. PROBLEM STATEMENT

Exploratory Data Analysis (EDA) on the Google Play Store dataset involves analyzing the data to identify patterns, trends and correlations among the different features of the dataset. This can be done by visualizing the data in various forms such as histograms, box plots, and scatter plots. Additionally, descriptive statistics such as maximum, minimum, mean, median, and standard deviation can be calculated to gain insights on the data. The data can also be grouped by different features such as category, size, and ratings to evaluate how different apps perform in comparison to each other. By doing this, we can gain valuable insights and make predictions about the success of a newly launched app.

2. INTRODUCTION

This project is an exploratory data analysis project focusing on the Google Play Store app reviews. We will use Python programming language to analyze the reviews and extract useful insights from them. The data set we will use here is a collection of user reviews, ratings, and other related information from the Google Play Store. The aim of this project is to gain a better understanding of the user feedback for a Google Play Store app and to identify the most important aspects of the app that users consider when giving their reviews and ratings. We will also be looking into the sentiment of the reviews to determine the overall opinion of the users. In addition, we will explore how the reviews and ratings differ across different apps, their categories, and other factors that might influence the reviews. Finally, we will investigate which apps have the most satisfied users and how this can be used to improve the user experience.

2.1 GOOGLE PLAY STORE AND USER REVIEW ANALYSIS

Data analysis is an important part of the modern business landscape. By analyzing data, businesses can gain insights into customer behavior, understand market trends, and make better decisions about their products and services. This is especially true for the Google Play Store, where data analysis can help developers understand user preferences and improve the performance of their apps. Data analysis of the Google Play Store dataset can provide valuable insights to developers. For example, it can show which apps have the highest ratings and reviews, which have the most downloads, and which are most popular within certain demographics. This information can be used to inform marketing strategies and product development, as well as to identify potential areas for improvement. Additionally, data analysis can show which apps are most profitable, helping developers

to focus their efforts on generating more revenue. Finally, data analysis of the Google Play Store can be used to help developers understand user behavior and preferences. By analyzing user reviews, developers can gain insights into how users interact with their apps and what features they find most valuable. This can help developers to prioritize features and make improvements accordingly. Data analysis can also be used to identify user trends and preferences, allowing developers to tailor their apps to better meet customer needs.

2.2 GOOGLE PLAY STORE DATASET

The Google Play Store dataset provided by Almabetter is an invaluable resource for Exploratory Data Analysis. It contains over 10,000 apps from the Google Play Store, with detailed analytics on each app, including number of downloads, ratings, reviews, and more. With this data, analysts can easily explore the trends, correlations, and insights of the apps in the Google Play Store. The dataset is provided in CSV format and can be easily imported into various software tools for analysis. It is also highly organized, with each column representing a different aspect of the app, such as its title, category, rating, etc. This makes it easy to quickly filter and analyze the dataset. Additionally, the dataset also contains links to the app's page in the Google Play Store, which allows further exploration of the app. Overall, Almabetter's Google Play Store dataset is a great tool for Exploratory Data Analysis.

The data set contains the following columns:

- **App:** This Column contains the name of the app
- **Category:** This contains the category to which the app belongs. The category column contains 33 unique values.
- **Rating:** This column contains the average value of the individual rating the app has received on the play store. Individual rating values can vary between 0 to 5.
- **Reviews:** This column contains the number of people that have given their feedback for the app.
- **Size:** This column contains the size of the app i.e. The memory space that the app occupies on the device after installation.
- **Installs:** This column indicates the number of time that the app has been downloaded from the play store, these are approximate values and not absolute values.
- **Type:** This column contains only two values- free and paid. They indicate whether the user must pay money to install the app on their device or not.
- **Price:** For paid apps this column contains the price of the app, for free apps it contains the value 0.
- **Content Rating:** It indicates the targeted audience of the app and their age group.
- **Genre:** This column contains to which genre the app belongs to, genre can be considered as a sub division of Category.
- **Last updated:** This column contains the info about the date on which the last update for the app was launched.
- **Current version:** Contains information about the current version of the app available on the play store.
- **Android version:** Contains information about the version of the android OS on which the app can be installed.

2.3 USER REVIEW DATASET

User reviews data frame has 64295 rows and 5 columns. The 5 columns are identified as follows:

- **App:** Contains the name of the app with a short description (optional).
- **Translated Review:** It contains the English translation of the review dropped by the user of the app.
- **Sentiment:** It gives the attitude/emotion of the writer. It can be 'Positive', 'Negative', or 'Neutral'.
- **Sentiment Polarity:** It gives the polarity of the review. Its range is [-1,1], where 1 means 'Positive statement' and -1 means a 'Negative statement'.

- **Sentiment Subjectivity:** This value gives how close a reviewer's opinion is to the opinion of the general public. Its range is [0,1]. Higher the subjectivity, closer is the reviewer's opinion to the opinion of the general public, and lower subjectivity indicates the review is more of a factual information.

2.4 PYTHON

Python is an extremely versatile language that is becoming increasingly popular in the field of data science. It is a powerful and dynamic language that is easy to learn and provides a wide range of features and libraries to facilitate data analysis. One of the main advantages of Python is its ability to quickly and easily manipulate data. Its syntactical structure is simple and intuitive, allowing developers to quickly write data-driven scripts and programs. In addition, Python's libraries and packages offer a wealth of features and functions to assist in data analysis. These include powerful algorithms, data visualization tools, and machine learning tools. Python also provides excellent support for numerical computation, making it well suited for data analysis tasks such as statistics and linear algebra.

Another advantage of Python for data science is its wide range of libraries and packages available. Python is an open-source language, so developers can easily access a range of powerful libraries and packages to assist in the development of data science applications. These include the popular NumPy, SciPy, and Panda's libraries, which provide powerful tools for numerical computation and data manipulation. Python also provides excellent support for data visualization, allowing developers to quickly create beautiful and insightful visualizations from their data. Finally, Python's machine learning libraries, such as Scikit-learn, offer powerful tools for data mining and predictive modeling. With its range of features and libraries, Python provides a powerful and versatile platform for data science.

2.5 DATA CLEANING AND PREPARATION

Data cleaning and preparation is a crucial step in exploratory data analysis of the Google Play Store dataset. Data cleaning is a process that involves identifying and removing inaccurate or incomplete data. This process may also involve fixing incorrect data values, filling in missing values, and removing outliers. Data preparation is the process of transforming raw data into a form that can be used for analysis. This includes merging datasets, selecting and transforming variables, and creating new features.

The Google Play Store dataset contains data on millions of apps and is highly unstructured. The first step in preparing this dataset for exploratory data analysis is to clean the data by removing any inaccurate and incomplete records. This can be done by checking for missing values, incorrect values, and outliers. After data cleaning, the next step is to prepare the dataset for analysis by selecting and transforming the relevant variables and creating any new features that may be useful for the analysis. This may include creating dummy variables for categorical data, creating new variables based on existing variables, and transforming numerical variables. After data preparation, the dataset can then be used for exploratory data analysis.

- **Step1**: We write a function play storeinfo (), that will display 5 attributes about all the columns: Data type, Count of non-null values, Count of null values, number of unique values in that column and percentage of null value in that columns in the play store dataset.

- **Step2**: we start off with the column 'Type' we can see that it has one null value. We checked this row and found out from the play store that it is a free app. We use filllna() function of the pandas library to fill this value.

- **Step 3**: We drop the columns 'Current Ver', 'Android Ver' and 'last updated' from our dataset using the drop() function of the pandas library.

- **Step 4**: We can see that the 'Rating' column has 1474 null values. Due to low variations in the rating values and a lot of repeated values the 'median' would be a suitable statistical indicator to replace the null values with. We calculate the mode of the column using the median () aggregate method, and fill this value in place of null values using the fillna() function.

- **Step 5:** We can see that the 'Reviews' column despite being a numerical indicator is of the 'object' data type, we will convert this to 'int' data type using the as type(int) function.

- **Step 6:** We can see that the size column, which should be numeric, is of the data type 'object', it also has characters 'k' and 'M' in the values which stand for kilobytes and Megabytes, we will replace the 'k' with 1000 and 'M' with 1000000. Some values also have '+' sign in them, which will be removed. Next, we will convert this column into 'int' datatype.

- **Step 7:** The 'Installs' column values contain the characters '+' and ',' which are going to prevent us from converting this column into a numeric datatype. We will get rid of these using the strip() and replace() functions.

- **Step 8:** The values in the column 'Price' might have the '$' sign in some values and the column is of the datatype 'object'. We will first remove the '$' sign using the **strip()** function and then convert the column into 'int' datatype.

- **Step 9:** Handling the duplicates in the App column we drop the no of duplicate rows that are present in the App columns.

- **Step 10:** We write a function Ur info (), that will display 5 attributes about all the columns: Data type, count of non-null values, Count of null values, number of unique values in that column and percentage of null value in that column in the User review dataset.

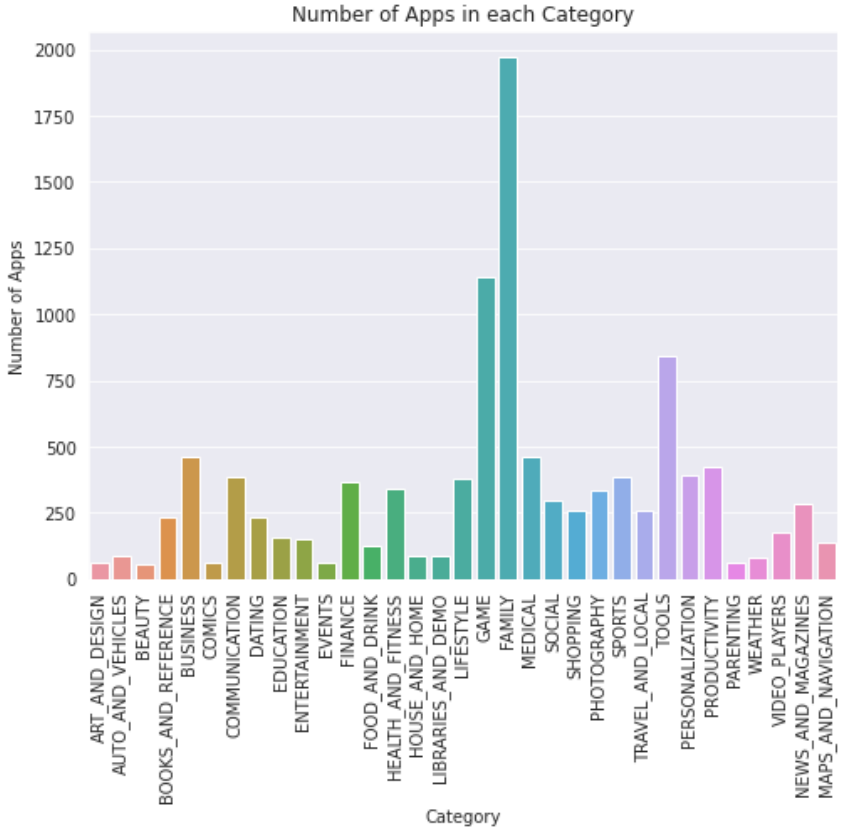
- **Step11:** In the User review dataset the columns are App, Translated Review, Sentiment, Sentiment Polarity, Sentiment Subjectivity in this total 26863 NaN value are present so we drop them using dropna() function.

3. EXPLORATORY DATA ANALYSIS

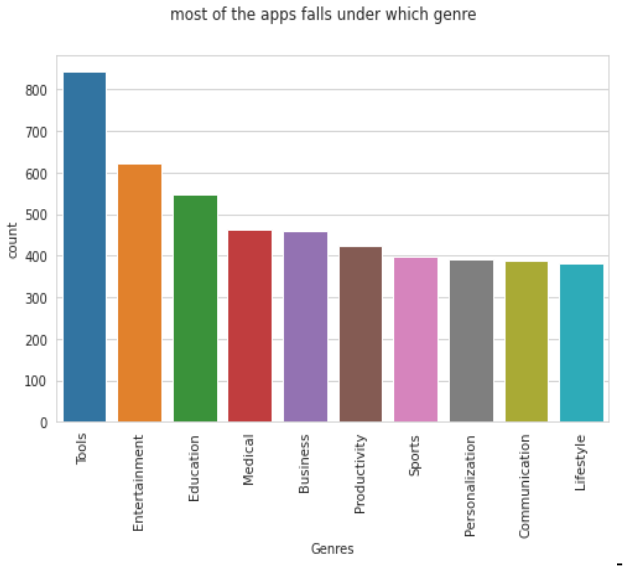
Exploratory Data Analysis (EDA) is an important step in the data analysis process. It involves looking at data in order to get a better understanding of the underlying patterns and relationships it contains. It is a great way to gain insights into the data and to identify any potential problems that may require further investigation. EDA involves a variety of techniques such as data visualization, statistical analysis, and data mining. Data visualization can help to reveal trends, outliers, and other patterns in the data that would otherwise be difficult to detect. Statistical analysis can help to identify correlations between different variables and to identify areas where more data is needed. Data mining can help to uncover hidden patterns in the data that could be used to make predictions or to inform decision-making.

EDA is a crucial step in the data analysis process and should be done before beginning any other analysis. It can provide valuable insights into the data and can help to identify problems and areas for further investigation. It can also help to identify potential relationships between variables that could be used to make predictions or to inform decision-making. EDA should always be done before any other type of analysis in order to ensure that the data is properly understood and the results are reliable.

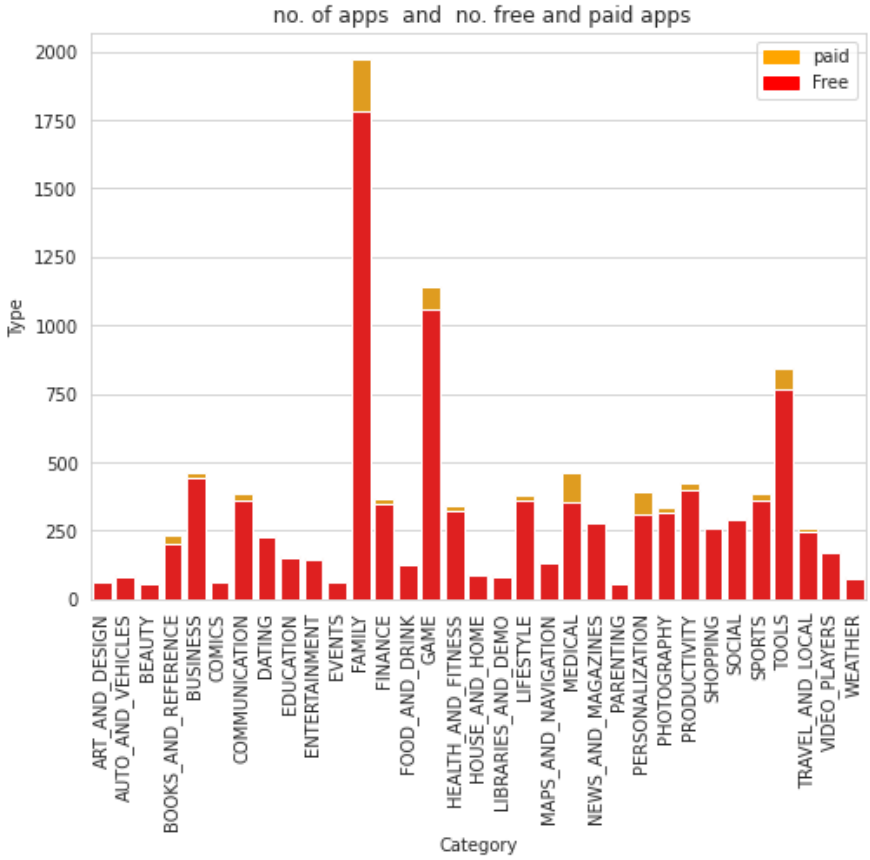
1. **NUMBER OF APPS IN EACH CATEGORY**



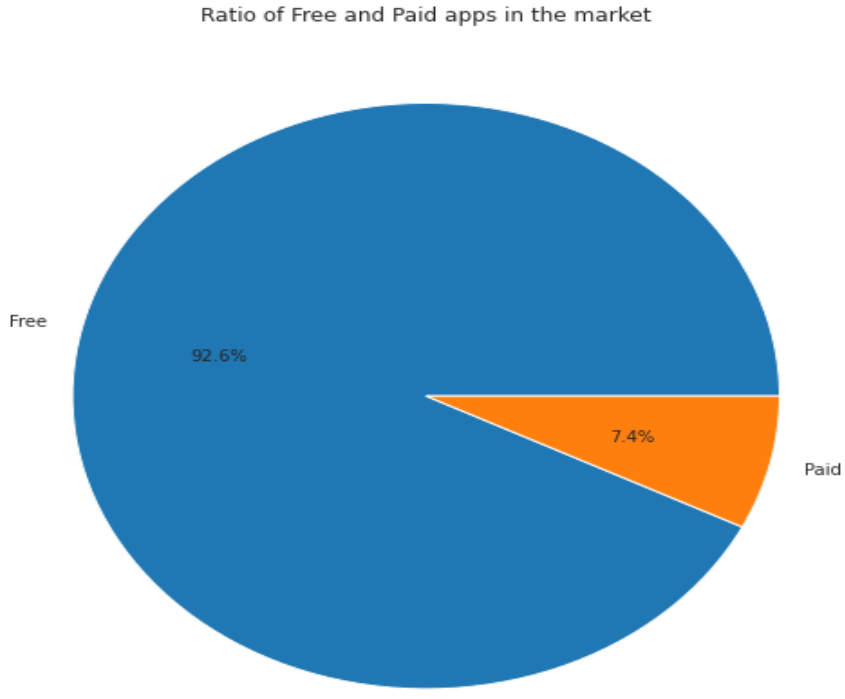
2. **MOST OF THE FALLS UNDER WHICH GENRE**



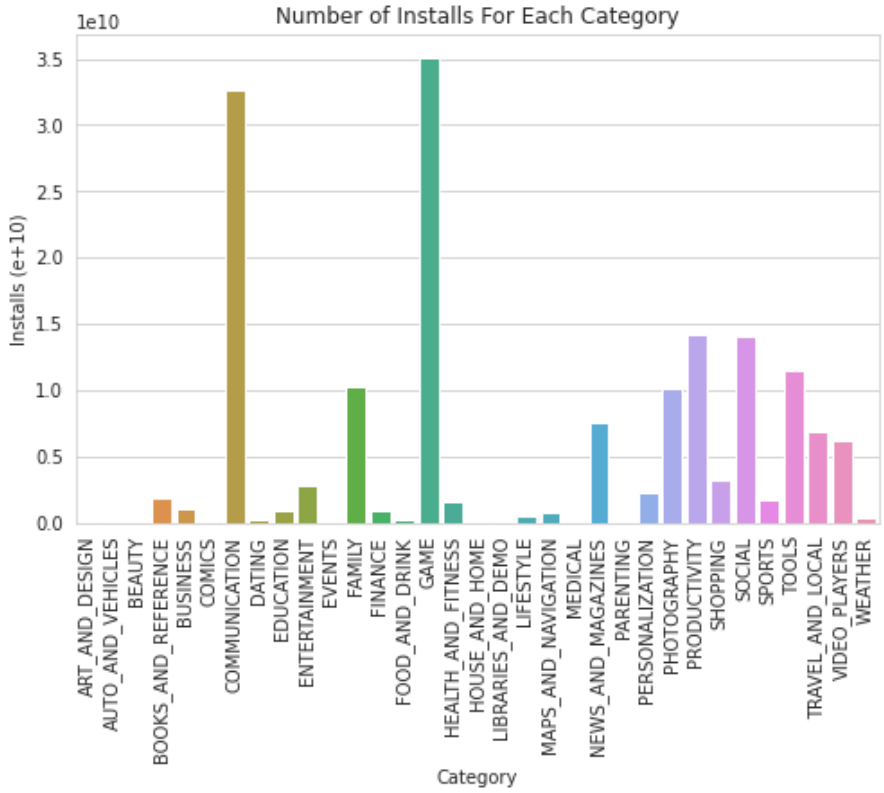
3. NUMBER OF FREE AND PAID APPS IN EACH CAEGORY



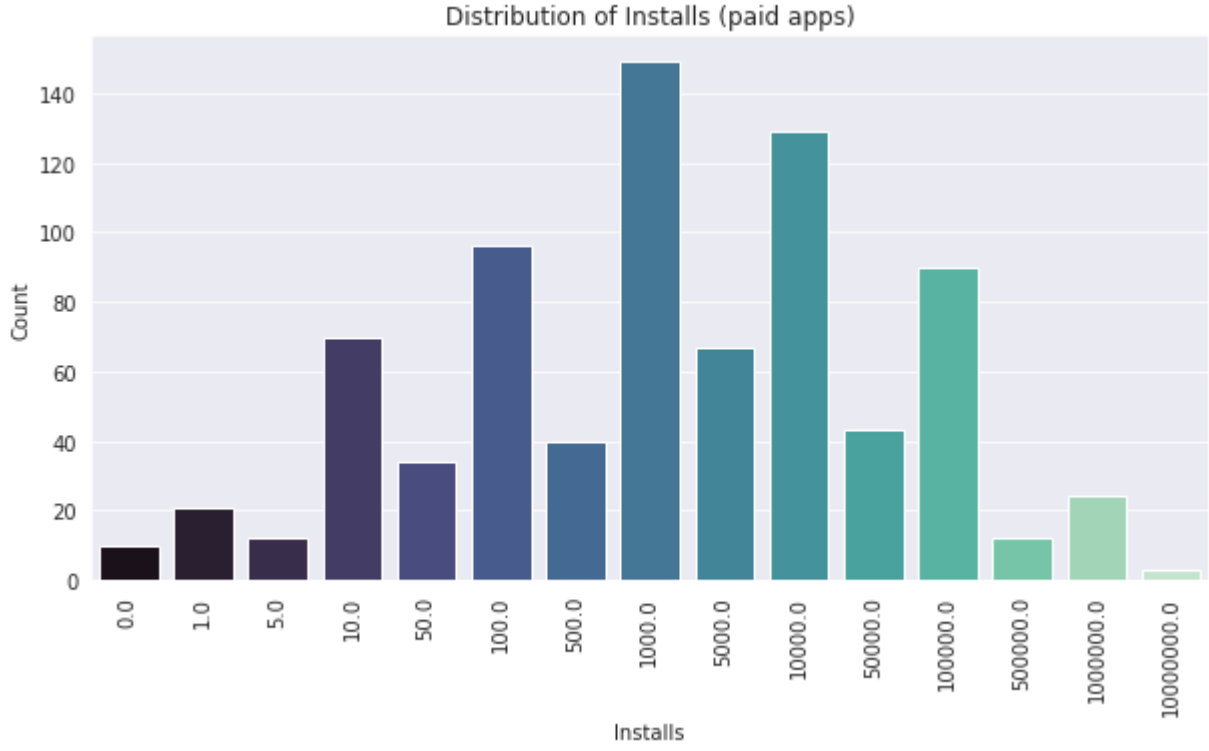
4. RATIO OF FREE AND PAID APPS



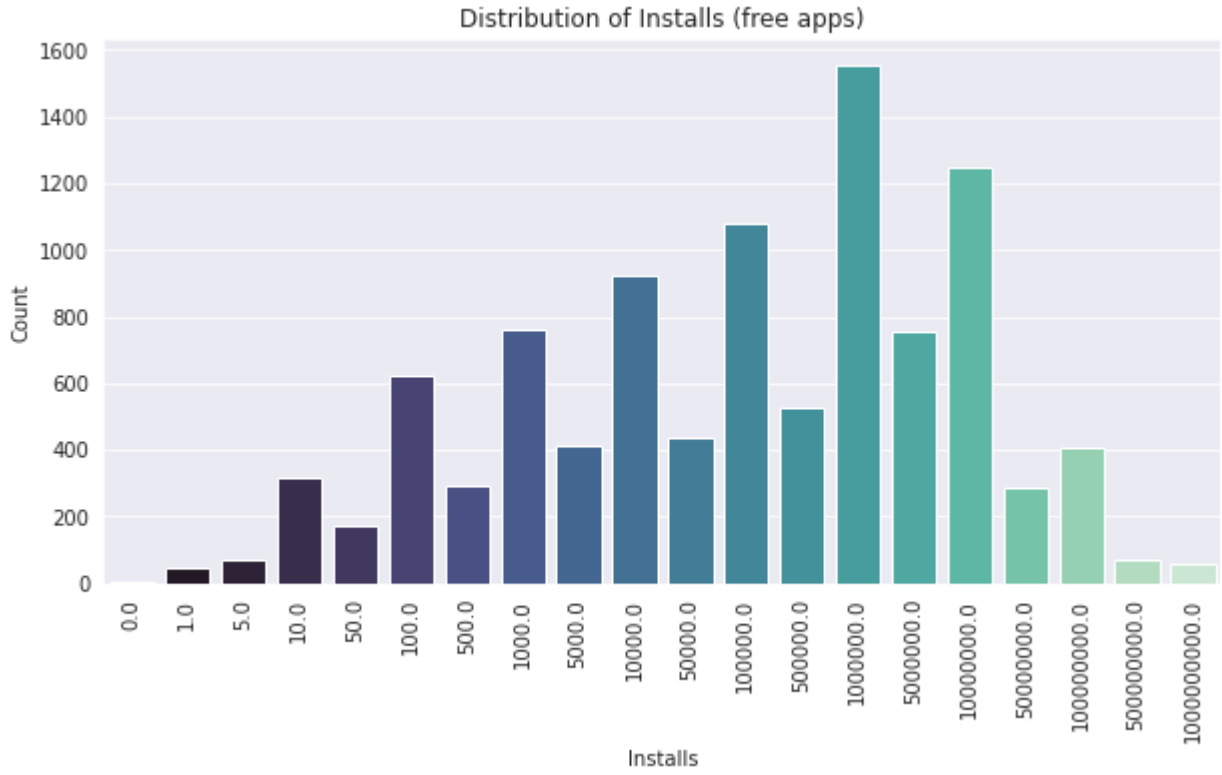
5. NUMBER OF APPS INSTALLS IN EACH CATEGORY



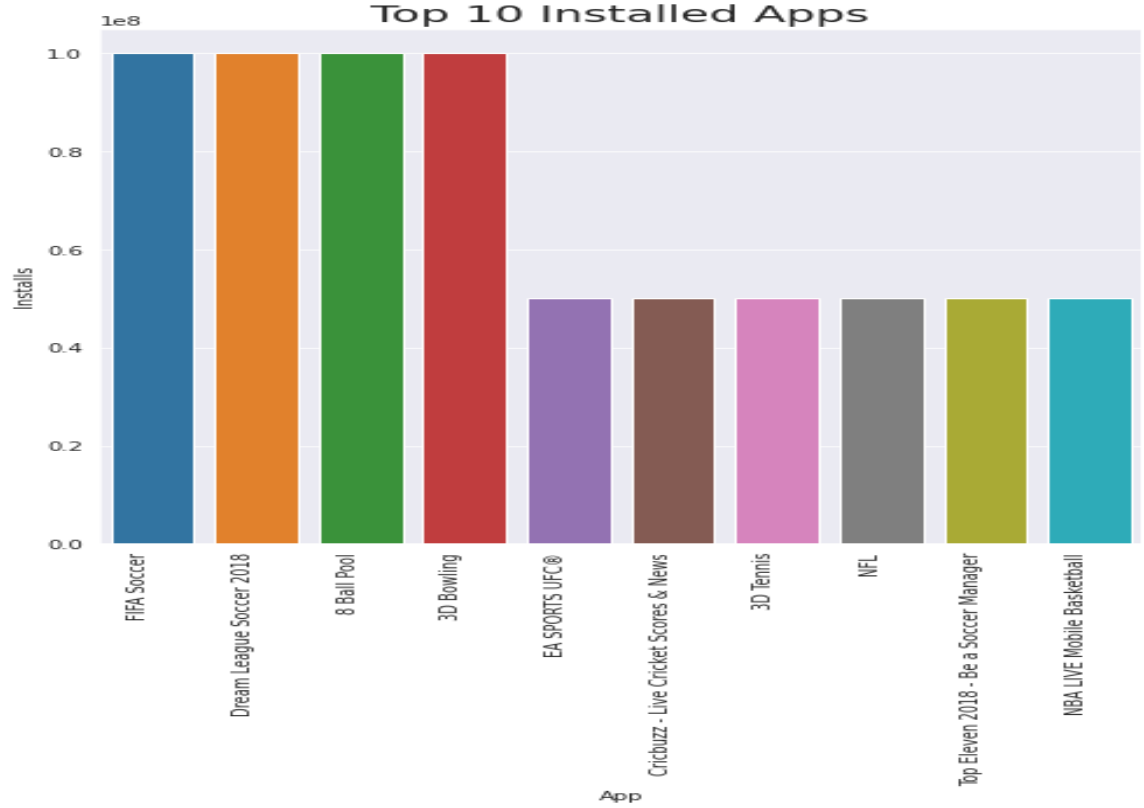
6. DISTRIBUTION OF PAID APPS



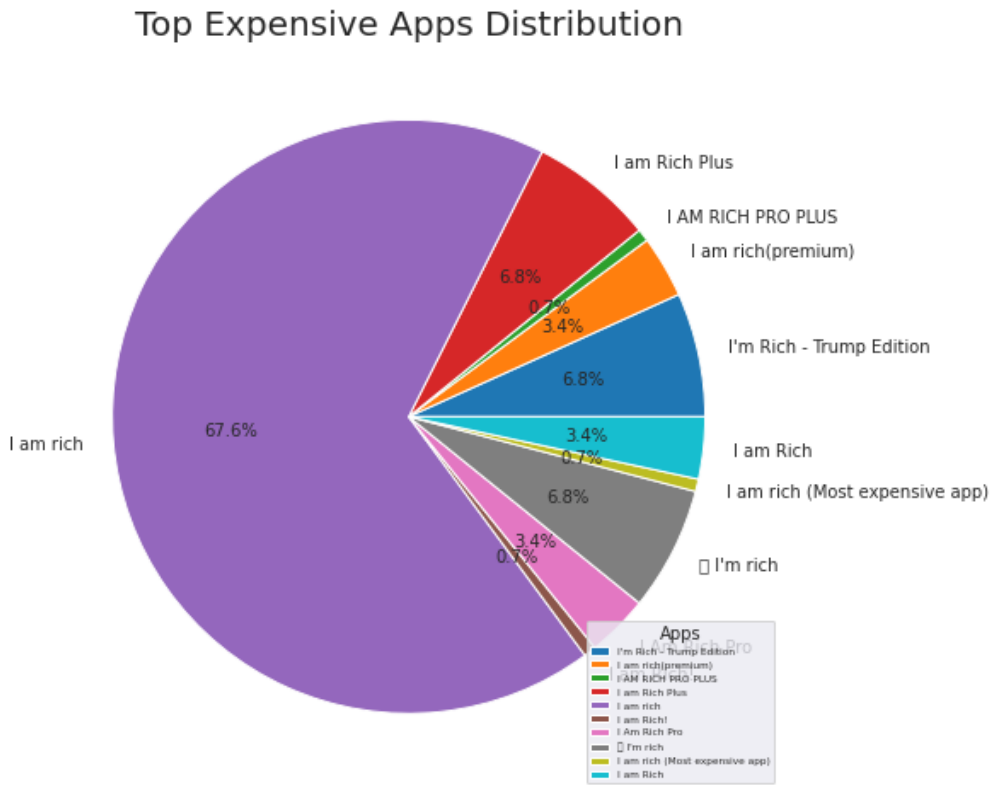
7.DISTRIBUTION OF FREE APPS



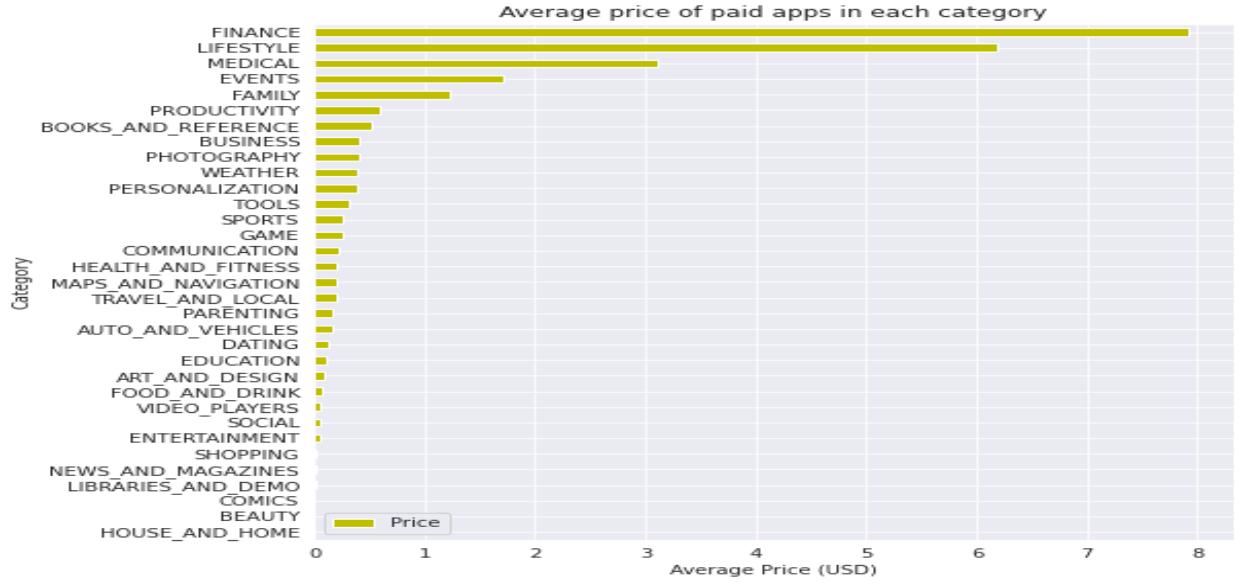
8. TOP 10 INSTALLED APP IN SPORTS CATEGORY



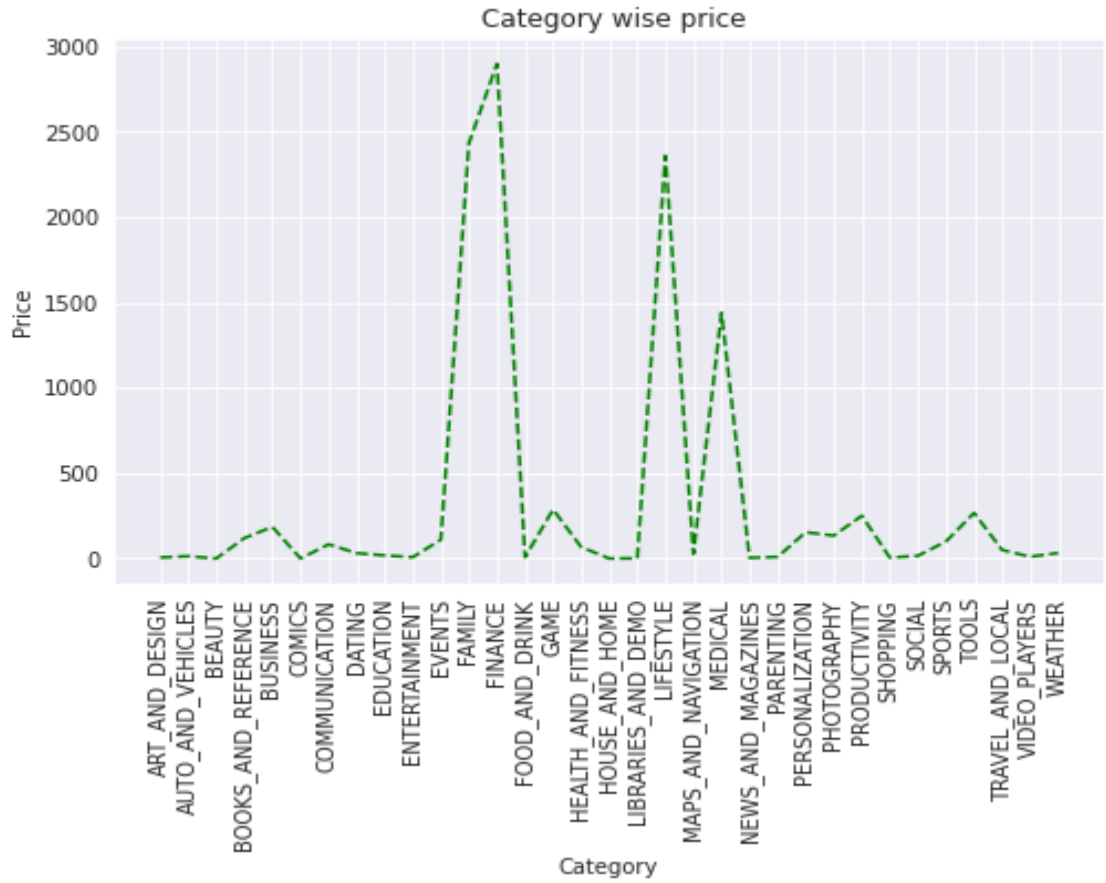
9. TOP 10 PAID APPS



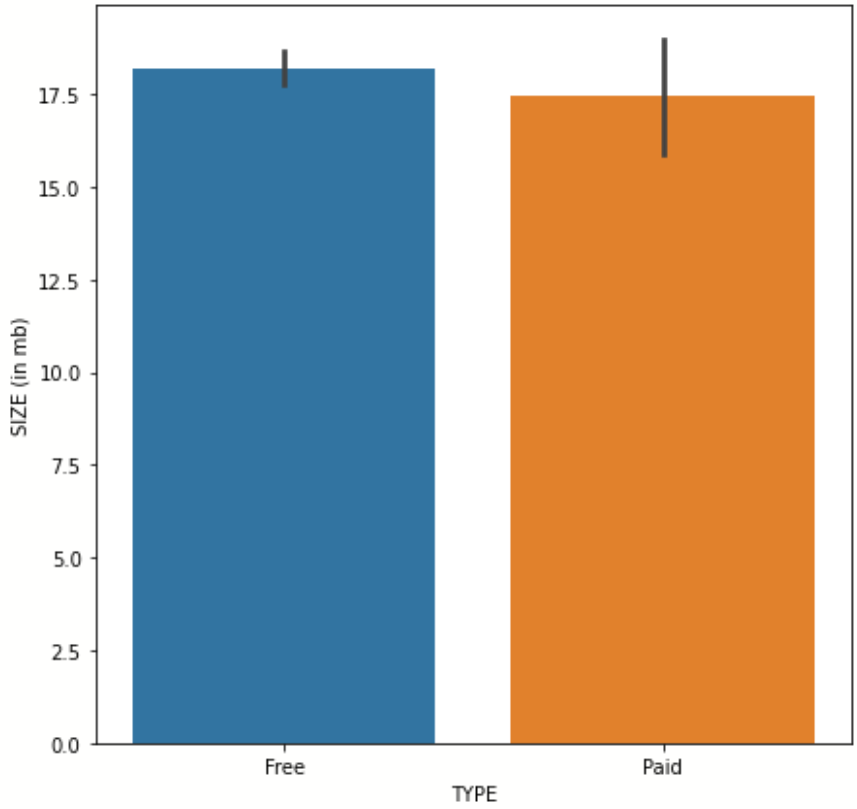
10. AVERAGE PRICE OF APPS IN EACH CATEGORY(IN USD)



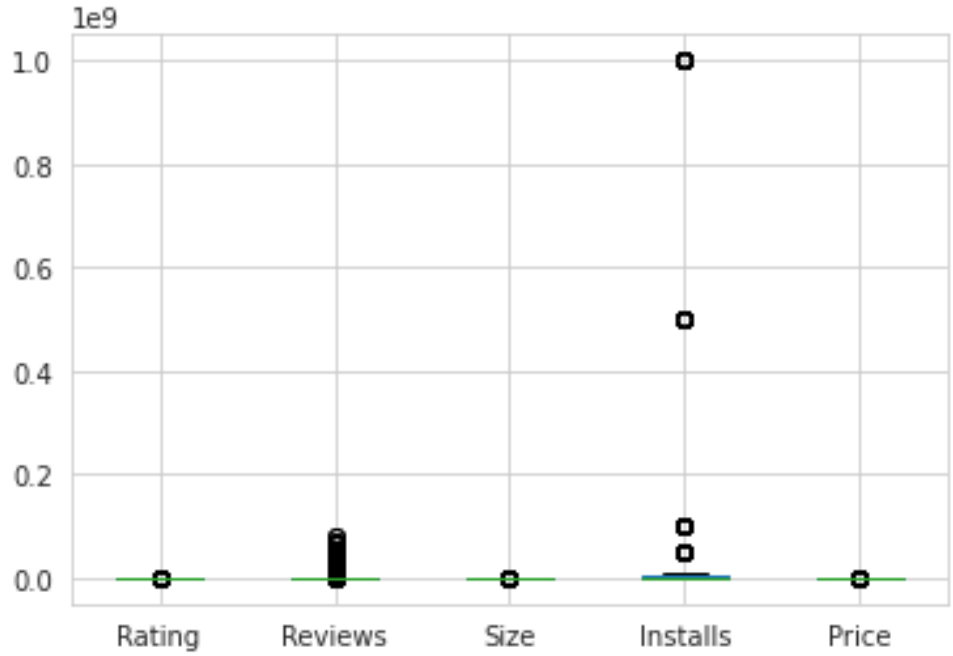
11. TOTAL PRICE IN EACH CATEGORY(SUM)



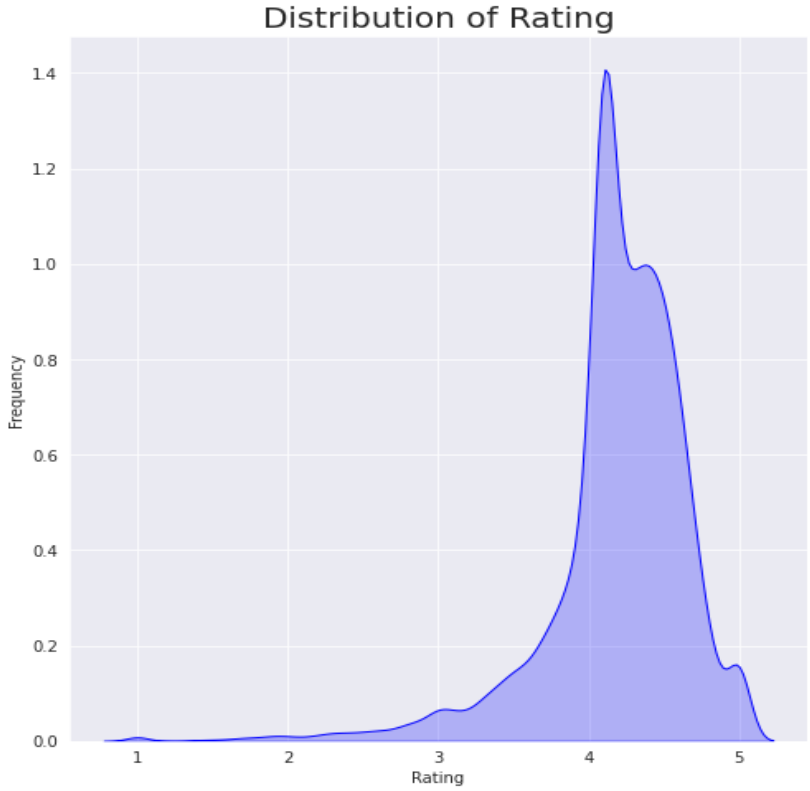
12. WHAT IS THE FREE AND PAID APPS SIZE (IN MB)



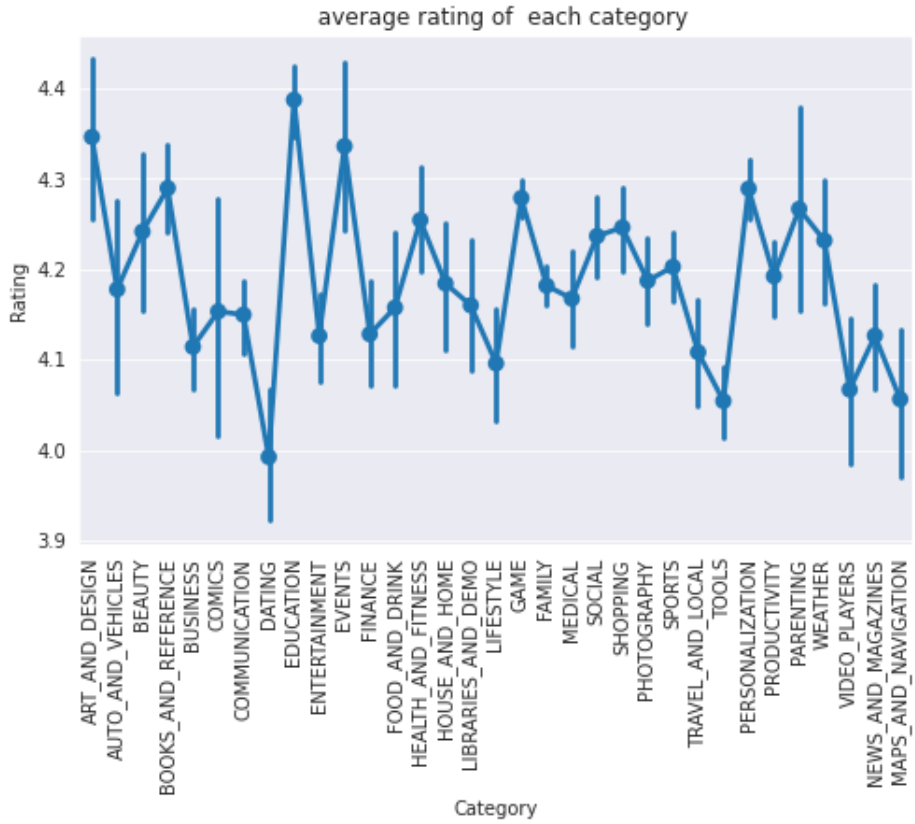
13. DATA SHAPE



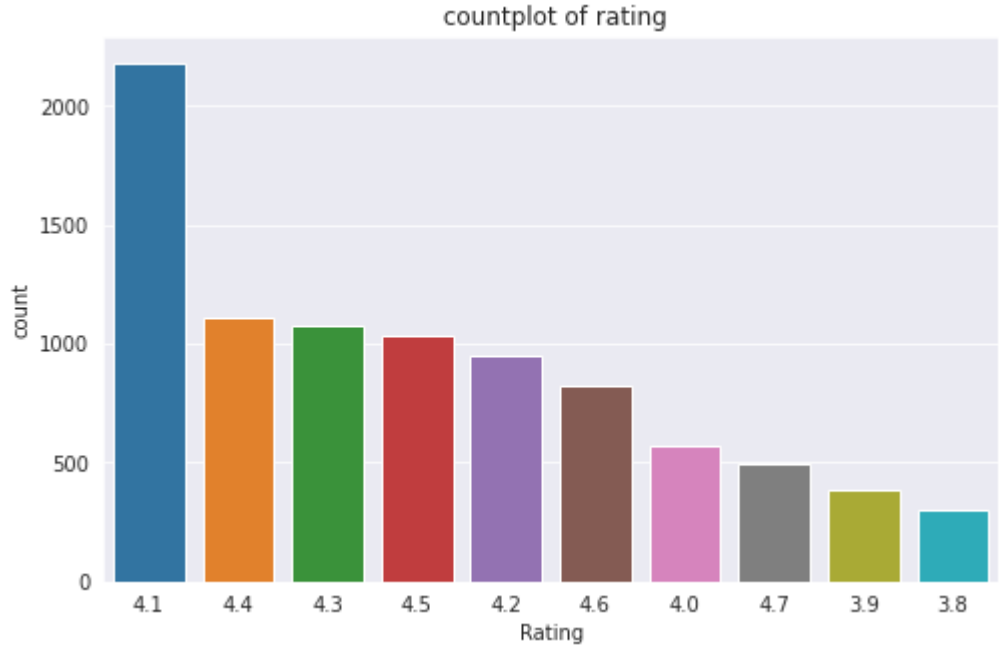
14. DISTRIBUTION OF RATING



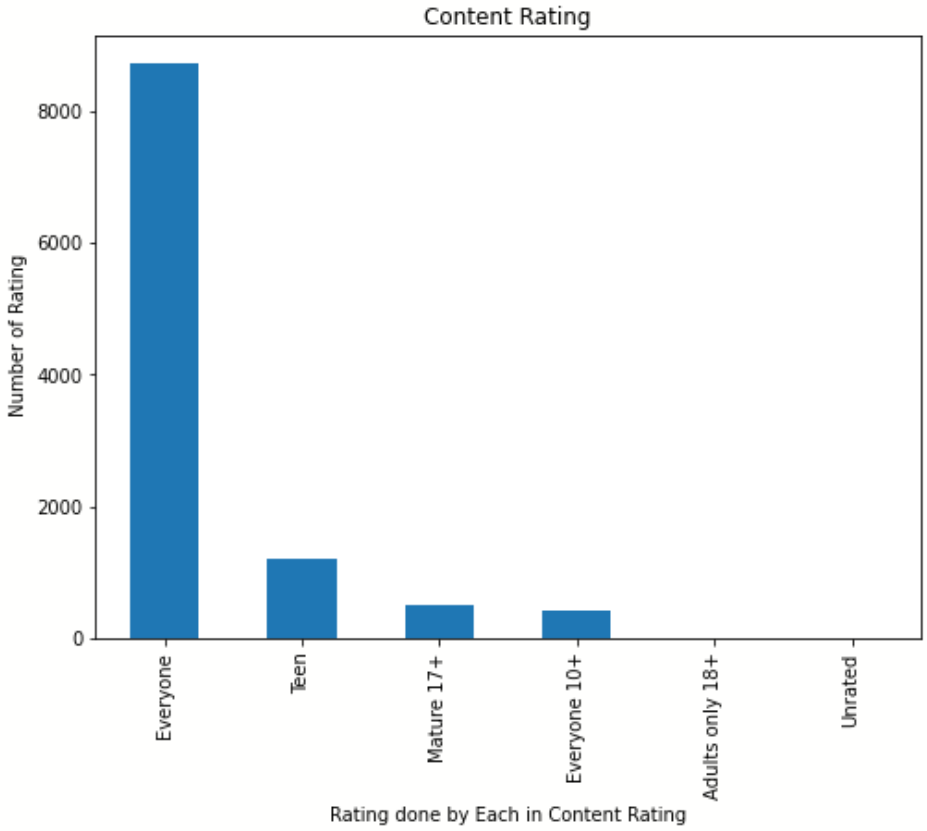
15. POINT PLOT OF AVERAGE RATING IN EACH CATEGORY



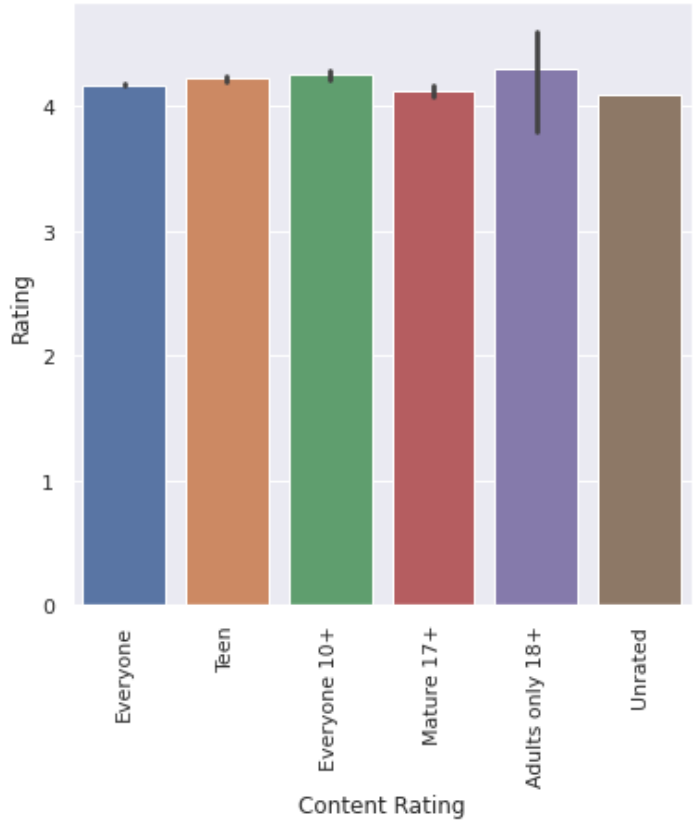
16. COUNTPLOT OF RATING(TOP10)



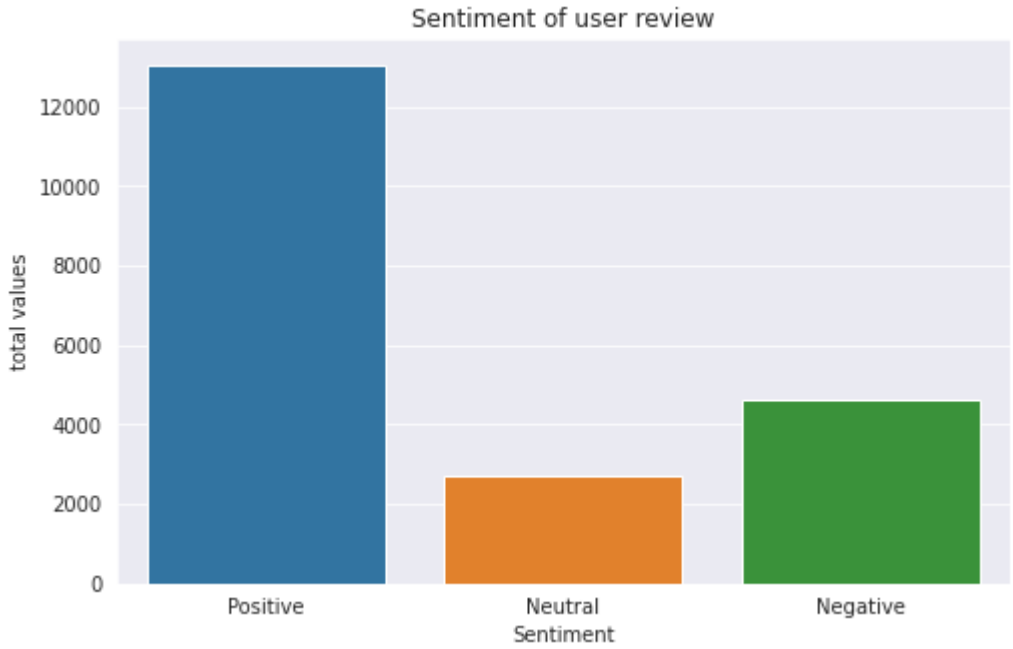
17. PLOT THE BARGRAPH FOR CONTENT RATING



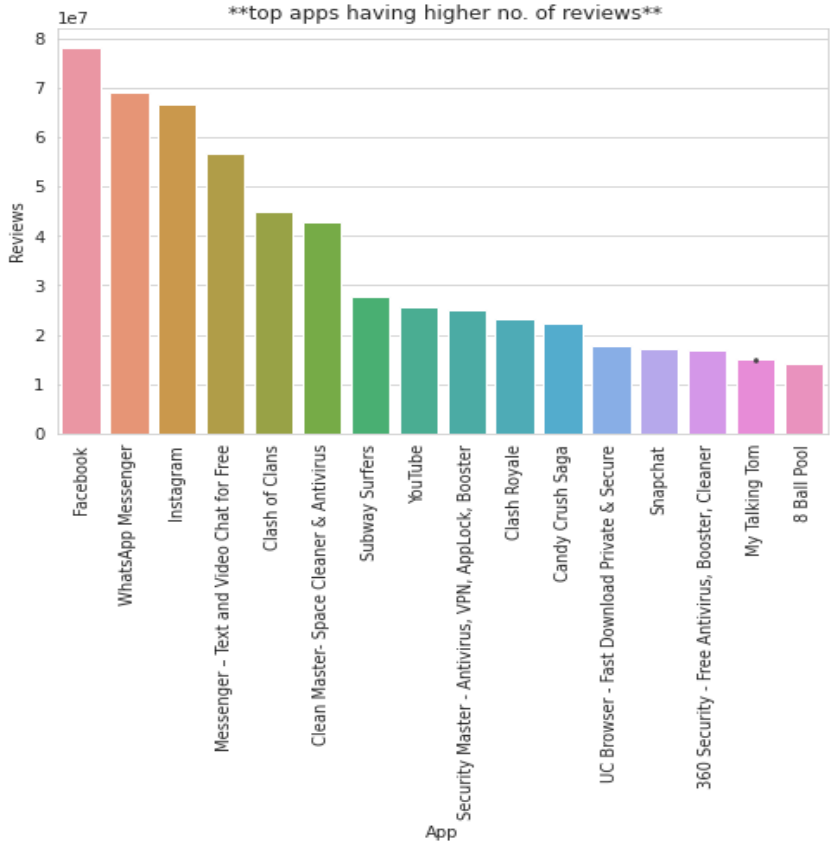
18.BARPLOT DISPLAYING THE RATING FOR EACH CONTENT RATING



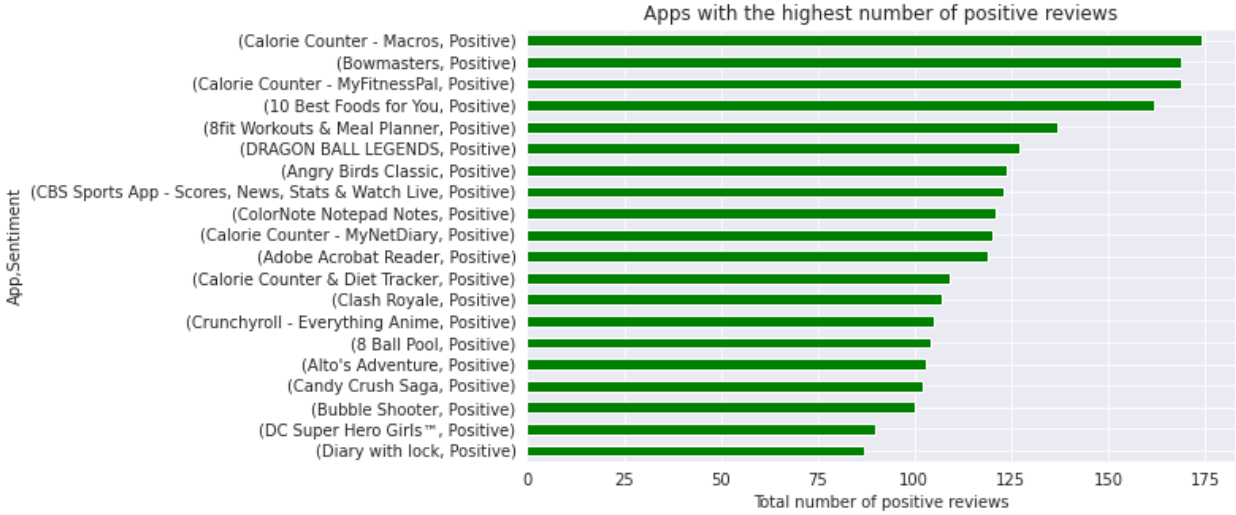
19. PLOTTING THE OVERALL SENTIMENT OF REVIEWS



20. TOP APPS HAVING HIGHER NUMBER OF REVIEWS

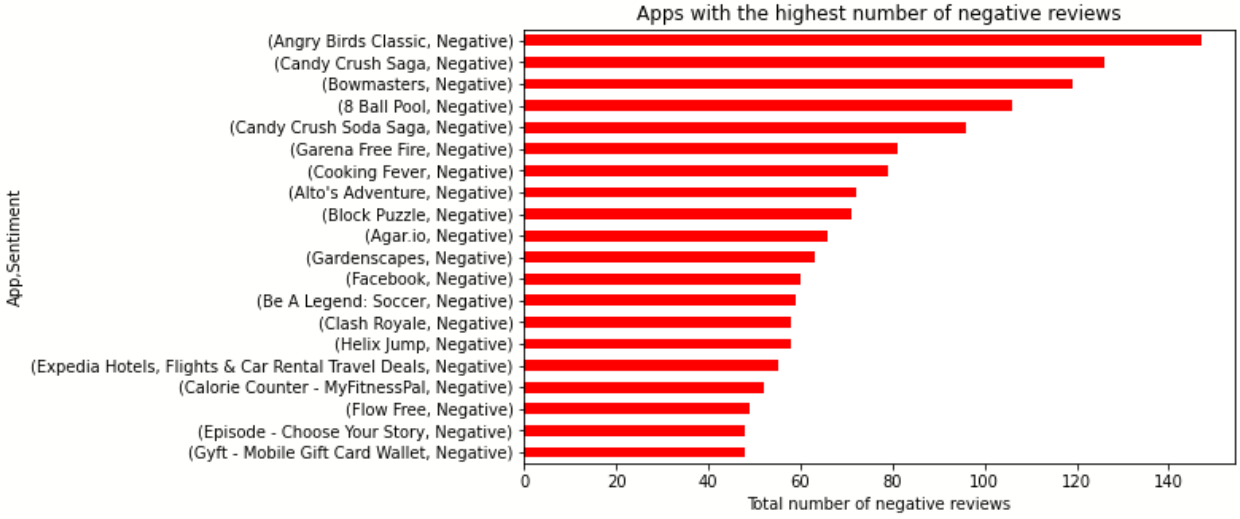


21. PLOTTING OF POSITIVE REVIEWS



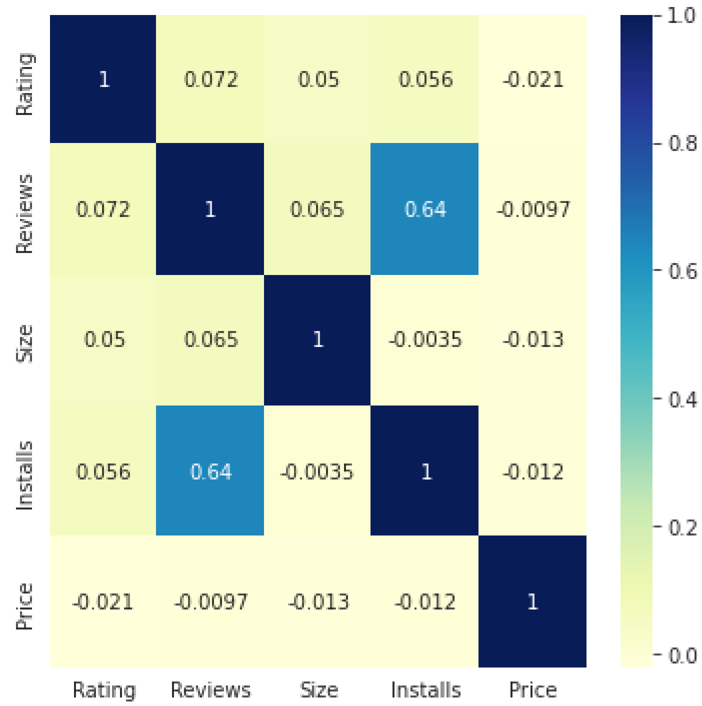
22. WORD CLOUD FOR POSITIVE REVIEWS

23. PLOTTING OF NEGATIVE REVIEWS



24. WORD CLOUD OF NEGATIVE REVIEWS

25. CORELATION BETWEEN RATING,REVIEWS,SIZE AND INSTALLS



Thank You