# Technical Documentation

## Play Store App Review Analysis

### CAPSTONE PROJECT 1


**Submitted By:**

**TEAM- Scraper Nerds**

Navneet Sawhney
Tinu J Pooppally
Biswanath Das
Manoj Chetry
Shubham V Jadav


# Contents

## ABSTRACT

*Nowadays, becoming digital is essential for any organization to prosper, and developing an app for your company can help it generate more revenue. Understanding the factors and trends connected to app creation from a business perspective is crucial given the steadily increasing quantity of apps. In this project, we performed an exploratory data analysis on the Play Store datasets to delve further into the data and identify patterns that may be useful to various business sectors. The datasets were thoroughly cleaned before being combined to form a single dataset.*

*An overview of the present state of the Play Store and its apps was conducted after this. We paid particular attention to how the various key indicators were distributed throughout the different categories as it helps us understand the healthcare sector better. The results of the EDA provided a number of key factors to be looked into while creating a healthcare app at its initial stages. The document concludes with a discussion of potential next steps for further research using the dataset.*

***Key Words*:** Google Play Store Apps, Exploratory Data Analysis, Machine Learning, Statistics, Data Visualization

## INTRODUCTION

Google Playstore is an online platform for downloading and purchasing mobile apps and games for Android devices. It is the official app store for the Android operating system and is maintained by Google. It offers a wide range of apps and games for users to choose from, including popular titles such as WhatsApp, Facebook, and Instagram. Users can also purchase digital content such as books, movies, and music through the Playstore. With its easy-to-use interface and vast selection of apps, Google Playstore has become the go-to destination for Android users looking to enhance their mobile experience.

Google Playstore is one of the most popular app stores in the world, with over 2.9 million apps available for download as of 2021. According to a study by App Annie, Google Playstore accounted for 85% of all app downloads globally in 2020. This is primarily due to the large market share of Android devices, which make up over 75% of the global smartphone market.[1]

Research has also shown that the majority of app downloads on Google Playstore come from emerging markets such as India, Brazil, and Indonesia.

For any business to succeed these days, going digital has become paramount and getting an application for your business can lead it towards more profits. None of us can now imagine our lives without using apps in our mobile phones. Think of any problem you are facing in your life; you will find a solution of same in form of an app specially curated to fix the issue for you. As per latest Google Play stats, there are 3.6 million apps currently at the Google Play Store. The number is constantly rising as around 3,739 apps are added to the Play Store every single day [2].

Due to constant rising numbers of apps, it becomes imperative to understand the parameters and patterns related to app creation from a business point of view. The Play store datasets are quite intriguing as they consist of details like number of installations, app reviews, sentiment polarity and so on. In this project we dig deeper into the datasets by doing an Exploratory Data Analysis to see certain patterns that can be helpful to different category of businesses.

We began by in-depth cleaning of the datasets and then we merged them to create one dataset. After that we did a generalised analysis to get numerous insights. We particularly focussed on the customer behaviour and what components affect the decision of the user to install the app. With the information gathered, we further tried to see if for medical category creating an app would be beneficial. If so, what kinds of apps are more liked by the audience.

## PROBLEM STATEMENT

To perform exploratory data analysis (EDA) on the Google Playstore in order to understand the trends and patterns in app downloads, user ratings, and revenue generated by apps. Our main aim is to structure the data and then see numerous patterns and trends across different features. We will do a general study of dataset through visualizations and draw some preliminary conclusions. After that we will delve deeper into seeing the scope of medical apps and also the consumer behavior in cased of apps related to the healthcare/medical category. Also, we will make final conclusions about the probability of the success of a new app based on the gathered insights.

Additionally, the EDA will also investigate any potential biases or limitations in the data, and make recommendations for further research to improve the understanding of the app market on the Playstore.

## WHAT IS EXPLORATORY DATA ANALYSIS(EDA)?

Exploratory data analysis (EDA) is a method used to analyze and summarize a dataset in order to understand its characteristics and patterns. EDA can be used to clean and preprocess the data, as well as identify any outliers or anomalies that may be present. Some common techniques used in EDA include visualizing the data using graphs and plots, calculating summary statistics, and identifying correlations and relationships between variables.

The following are the various steps involved in the EDA process:

**Data collection-** Collecting the relevant data that is necessary for the analysis. This can include gathering data from various sources such as databases, surveys, and experiments. We took the datasets from Kaggle.

**Problem Statement** - This is the initial step on understanding the attributes of the datasets and based on that the aim of EDA is defined.

**Hypothesis** - In this step a basic a proposition made as a basis for reasoning.

**Data Cleaning** - Cleaning the data by removing any missing or corrupted values, and correcting any errors or inconsistencies.

**Data Exploration** - Exploring the data by creating visualizations, performing summary statistics, and identifying patterns and trends. This step helps to gain a better understanding of the data and identify any potential issues or outliers.

**Data Transformation -** Transforming the data to make it more suitable for analysis. This can include normalizing data, creating new variables, or removing outliers.

**Data Visualization -** Creating visual representations of the data to better understand and communicate the findings. This step can include creating charts, graphs, maps, and other visualizations to help convey the key insights of the analysis.

**Data Interpretation -** Interpreting the results of the analysis and drawing conclusions from the data. This step includes communicating the findings to stakeholders and making recommendations for further action.

**Testing Hypothesis** - We shall check if our data meets the assumptions required by most of the multivariate techniques.

## DATASET 1 - GOOGLE PLAYSTORE

The data set contains the following columns:

- **App:** This Column contains the name of the Apps
- **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.

## DATASET 2- USER REVIEW

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

- **App:** Contains the name of the App.
- **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:** A value from 0 to 1 indicating the subjectivity of the review. Lower values indicate the review is based on factual information, and higher values indicate the review is based on personal or public opinions or judgement.

## DATA CLEANING AND PREPARATION

Data cleaning is the process of identifying and correcting or removing errors, inconsistencies, and missing values in a dataset. It is an important step in the data preprocessing phase, as it can improve the quality and reliability of the data and make it more suitable for analysis.

Some common data cleaning techniques include:

**Removing duplicate records:** Identifying and removing duplicate records from the dataset.

**Handling missing values:** Identifying and handling missing values, which can be done by either removing the entire record or replacing the missing value with a suitable estimate.

**Formatting and type conversion:** Formatting data values to ensure consistency and converting them to the appropriate data type.

**Outlier detection:** Identifying and handling outliers, which are extreme values that can skew the results of the analysis.

**Normalization and scaling:** Normalizing and scaling numerical variables to ensure that they are on the same scale and can be compared more easily.

**Text cleaning:** Cleaning the text attributes like removing stop words, stemming, lemmatization.

It's important to note that data cleaning can be an iterative process, as errors and inconsistencies may be discovered during the cleaning process, and multiple rounds of cleaning may be necessary to achieve high-quality data. Additionally, data cleaning should be done with care as it could lead to loss of important information.

The following steps helped us to clean the data efficiently

- **Step 1:** First of all, we handled duplicate values, There are 483 rows with duplicate values so we need to delete them. After handling the duplicate values, the number of rows reduced from 10841 to 9660.

- **Step 2**: We wrote a function playstoreinfo (), that displays 5 attributes about all the columns: Data type, All values, Null values, number of unique values in that column and percentage of null value in that columns in the play store dataset.

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

- **Step 4**: We started off with the column 'Type' we could 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 used 'free' to replace corresponding row index.

- **Step 5**: The Rating column contains 1463 NaN values which accounts to apprximately 13.5% of the rows in the entire dataset. It is not practical to drop these rows because by doing so, we will lose a large amount of data, which may impact the final quality of the analysis. The NaN values in this case can be imputed by the aggregate (mean or median) of the remaining values in the Rating column.

- **Step 6:** We checked 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 replaced 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:** We can see that the 'Reviews' column despite being a numerical indicator is of the 'object' data type, we converted this to 'float' data type using the astype(float) function.

- **Step 8:** The ‘Installs’ column values contained the characters ‘+’ and ‘,’ that prevented us from converting this column into a numeric datatype. We got rid of these using the replace() functions and using astype(float) function..

- **Step 9:** The values in the column ‘Price’ might have the ‘$’ sign in some values and the column is of the datatype ‘object’. We first removed the ‘$’ sign using the **strip()** function and then converted the column into ‘float’ datatype.

- **Step 10:** We also changed ‘Last updated’ column type to datetime using to_datetime() function.

- **Step 11:** We wrote a function Userinfo(), that displays 5 attributes about all the columns: Data type, count of non-null values, null values ,number of unique values in that column and percentage of null value in that columns in the User review dataset

- **Step12:** In the User review dataset the columns are App, Translated Review, Sentiment, Sentiment Polarity, Sentiment Subjectivity in this total 26868 NaN value were present so we dropped them using dropna() function.

## EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis, or EDA, is an important step in any Data Analysis or Data Science project. EDA is the process of investigating the dataset to discover patterns, and anomalies (outliers), and form hypotheses based on our understanding of the dataset.

EDA involves generating summary statistics for numerical data in the dataset and creating various graphical representations to understand the data better. In this article, we will understand EDA with the help of an example dataset. We will use **Python** language (**Pandas** library) for this purpose.

- **Number Of Apps Per Category**

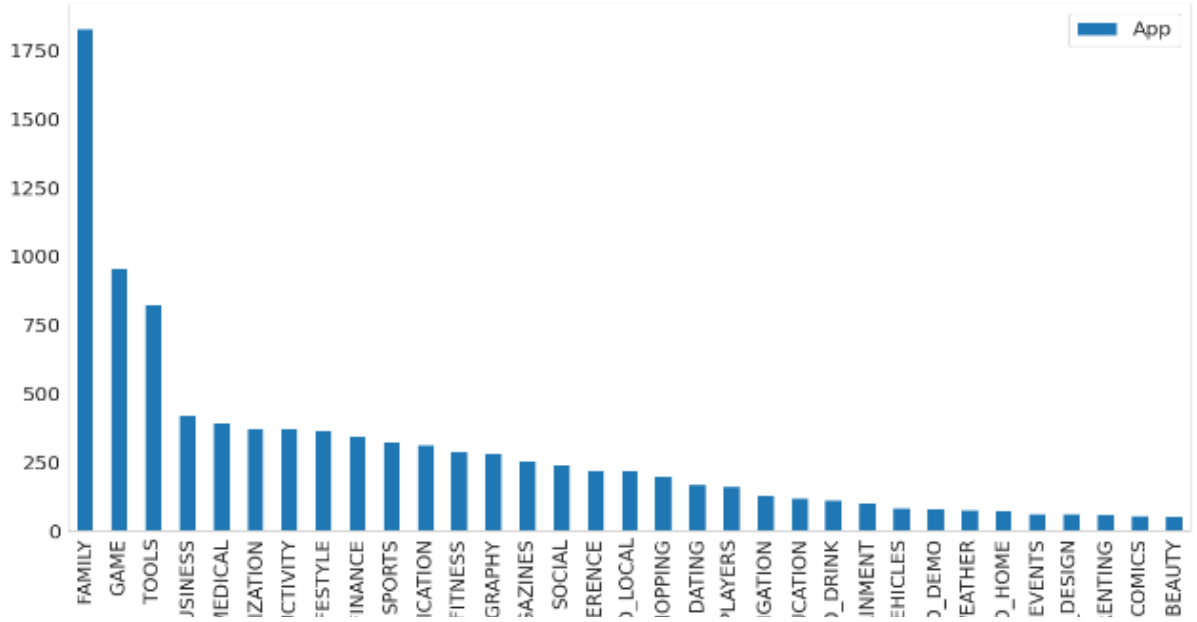


Fig -1: Max number of apps per category

The above bar graph represents the distribution of number of apps in different categories in the Play Store. It can be infered that FAMILY Category has the maximum number of Apps.

- **Rating**

In the below plot, we plotted the apps Rating

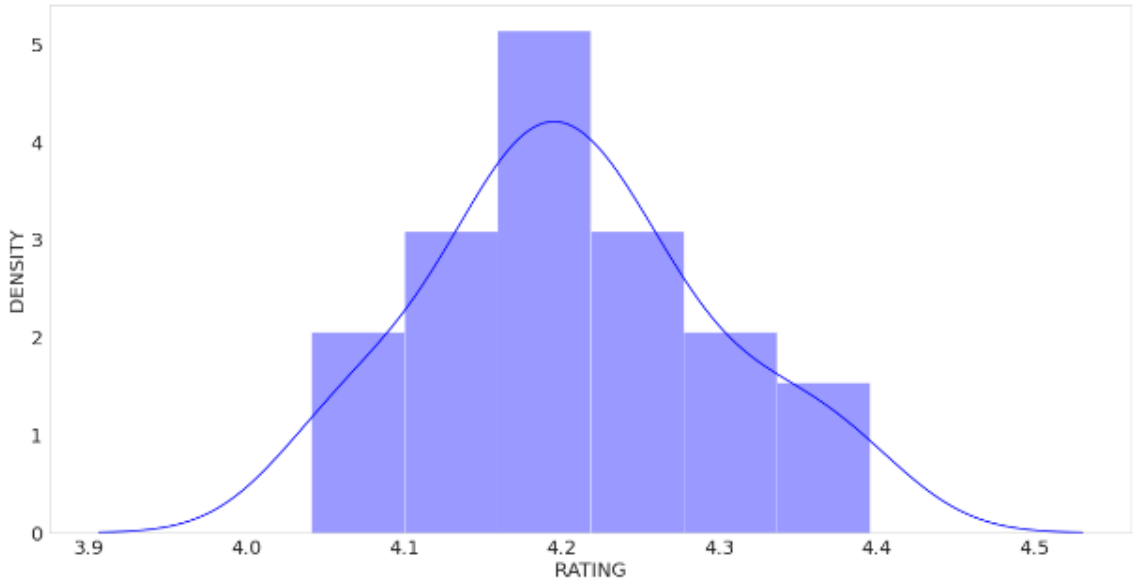


**Fig -2**: Distribution of App rating

- The mean of the average ratings (excluding the NaN values) comes to be 4.2.
- The median of the entries (excluding the NaN values) in the 'Rating' column comes to be 4.3. From this we can say that 50% of the apps have an average rating of above 4.3, and the rest below 4.3.
- From the distplot visualizations, it is clear that the ratings are left skewed.
- We know that if the variable is skewed, the mean is biased by the values at the far end of the distribution. Therefore, the median is a better representation of the majority of the values in the variable.

➢ **Average app rating**

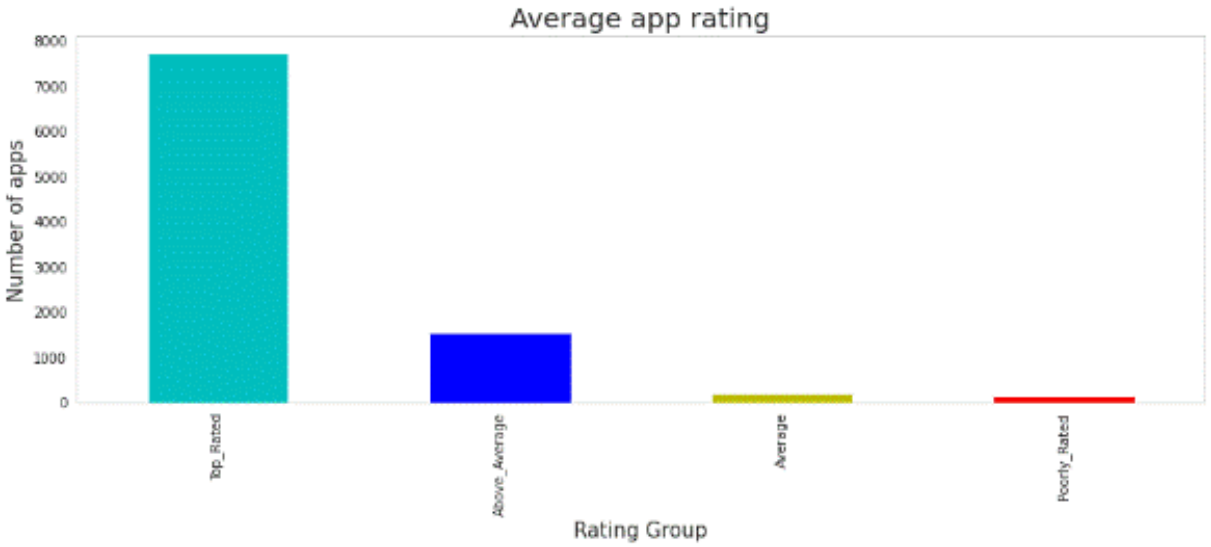


**Fig-5: Average app rating**

Nearly 8000 apps are top rated and about 2000 apps are above average.

From the above analysis we can conclude,

1. For all the categories, the average rating is above 4 stars.
2. Most of the Apps are "Top_Rated" as we can see in the above visualisation.

- **Count of app in each category based on their type**

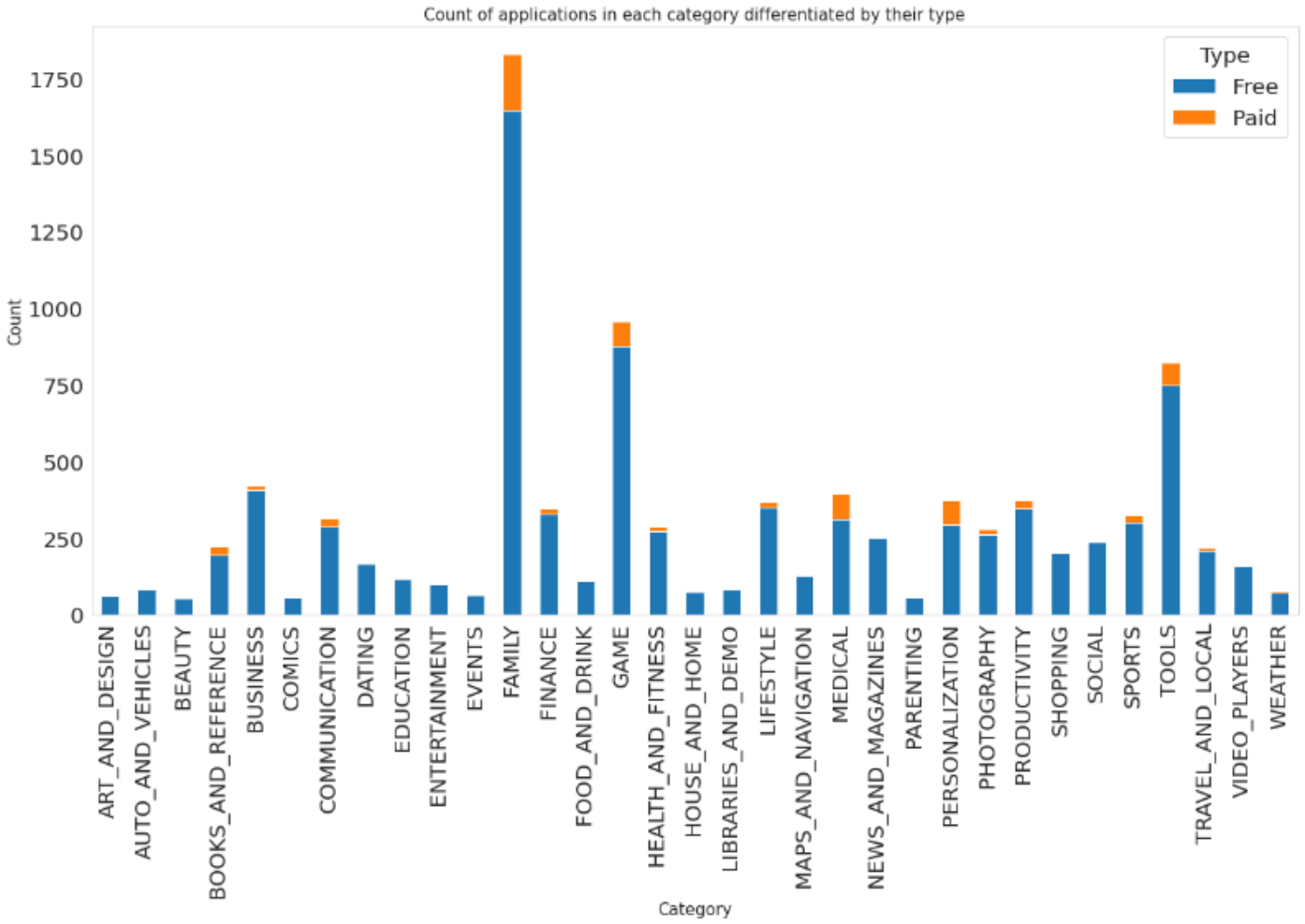


**Fig -6: count of app by type.**

It can be inferred that certain app categories have more free apps available for download than others. In our dataset, the majority of apps in Family, Games and Tools, as well as Social categories are free for users to install.

- **Top 10 apps in medical category**

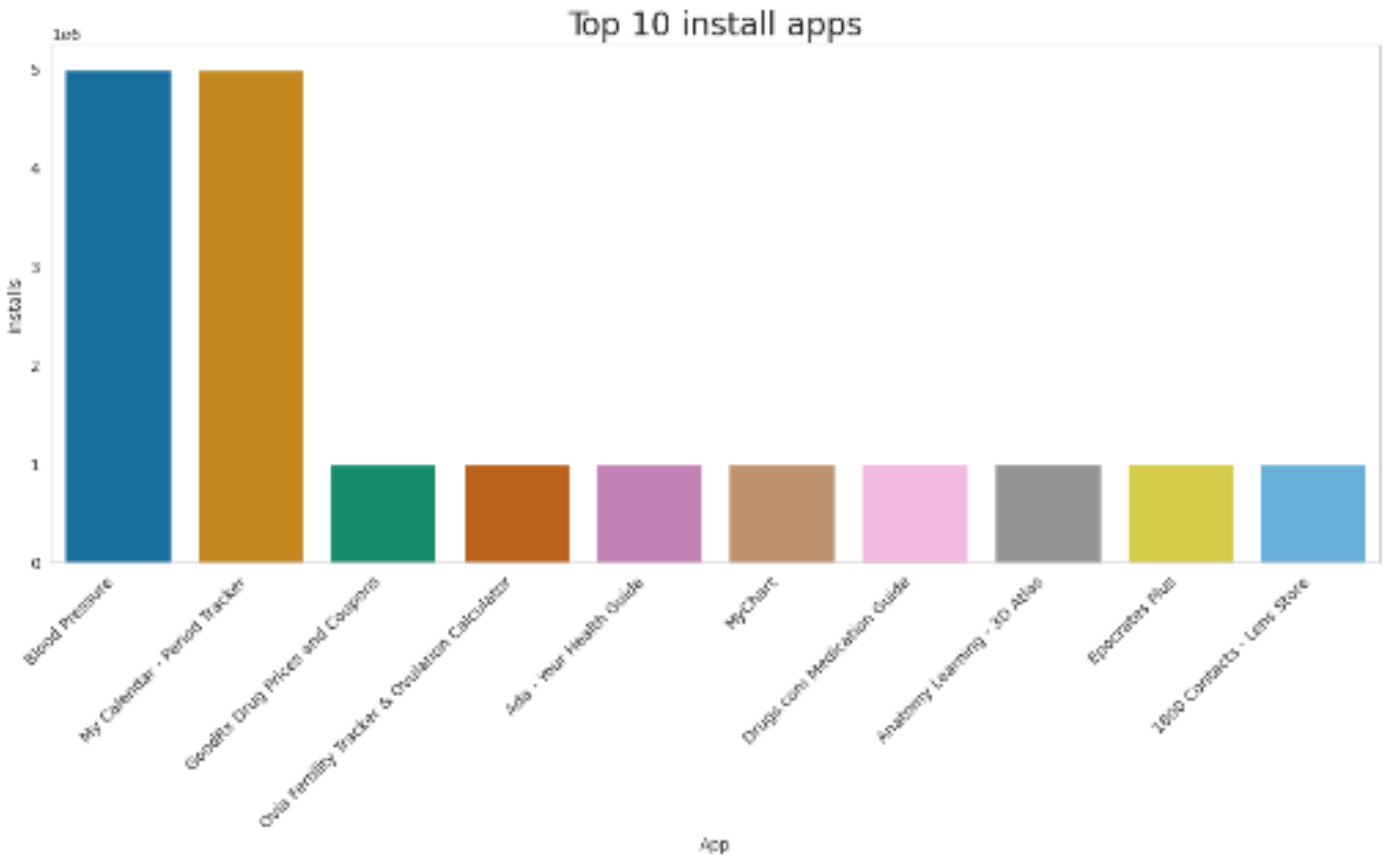


**Fig -7: top 10 medical app**

We can see the top 10 Medical Category apps in the above visualization.

- **Optimal app size**

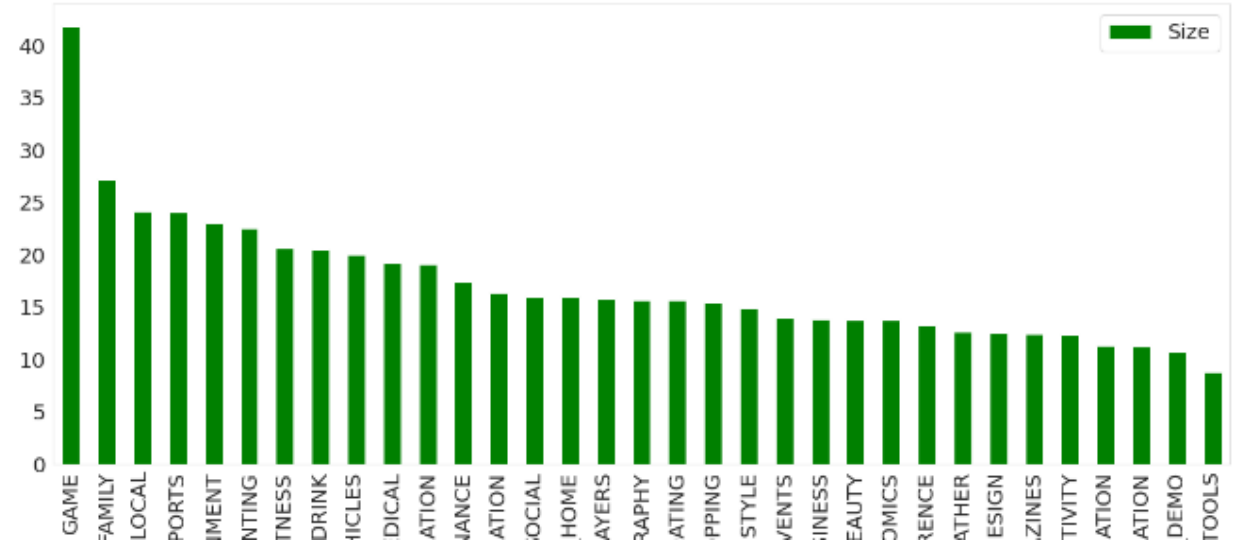


**Fig -8: Average app size by category**

This tells us the category of apps that has the maximum number of installs.

- The average App Size is in the range 10-25 Mb.
- The Game category has the maximum average App Size.
- Tools has the least average App Size.
- The average App Size of Medical category stands at 19 Mb

➢ **Distribution of Paid and Free apps**

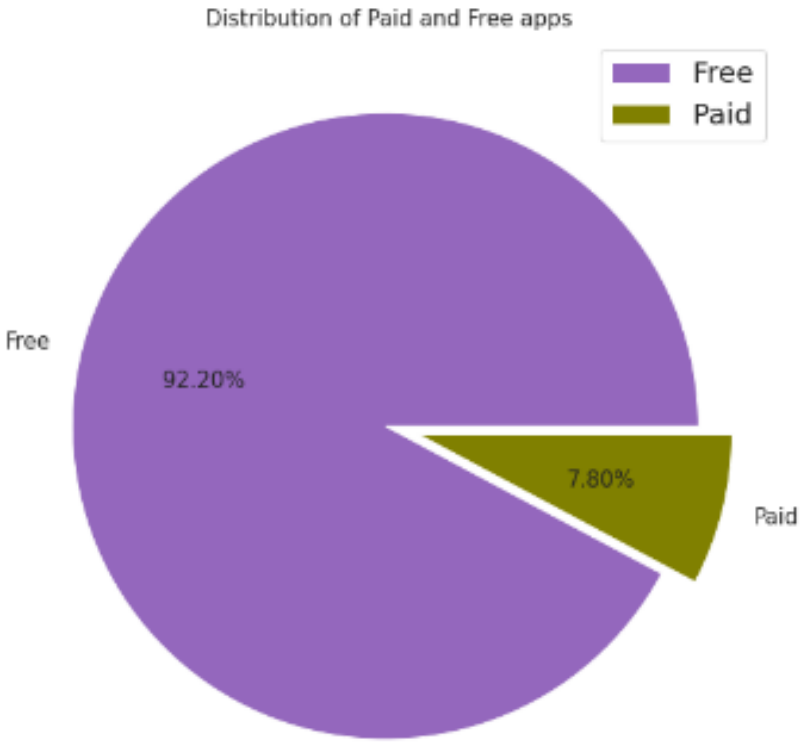


**Fig -9: Distribution of paid and free apps**

**It can be inferred that around 92% Apps are free and the 8% apps are paid.**

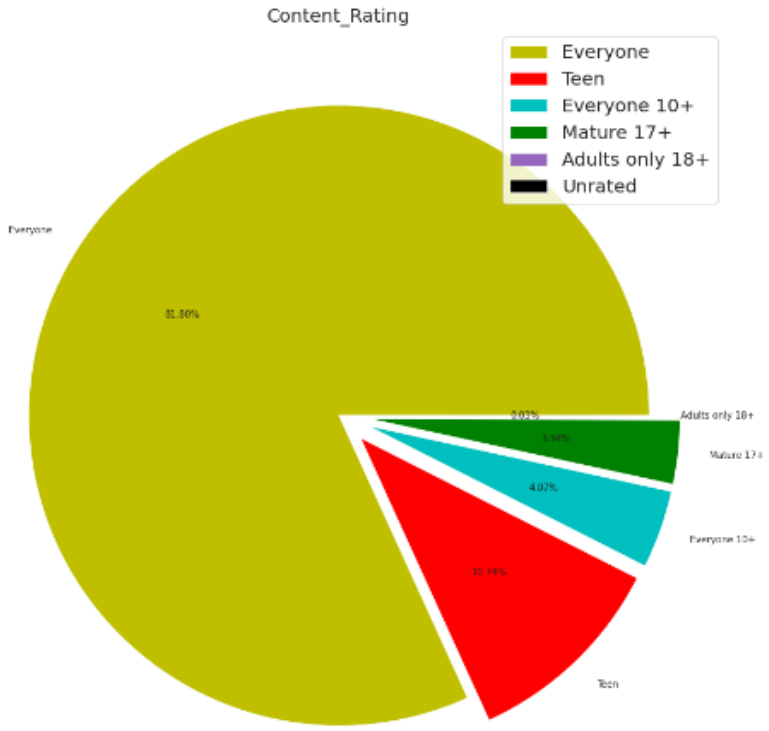


**Fig -10: Distribution of content rating.**

As per the above pie chart, around 82% Apps are created for Everyone and the least being Unrated i.e. 0.02 followed by adults only 18+ i.e. 0.03.

- **Percentage of User review Sentiments**

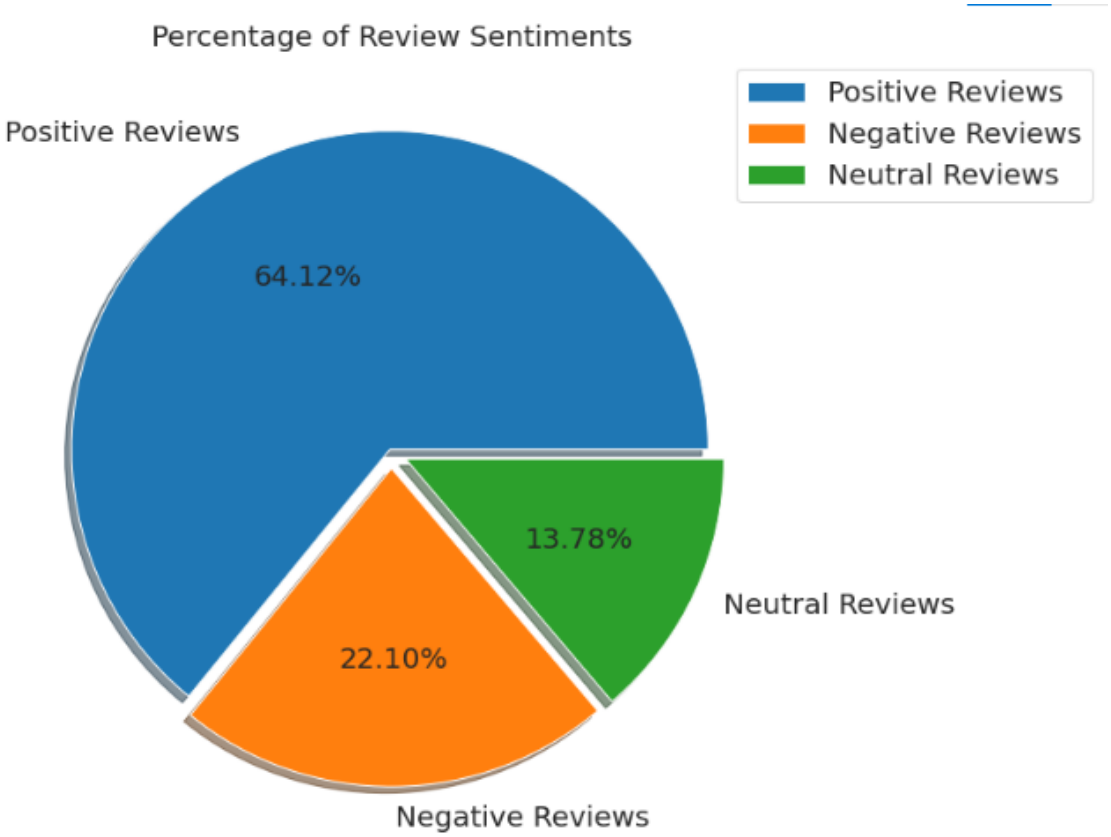

**Fig -11: Percentage of User Review Sentiments**

From the above pie chart, we can say that most of the apps that are present on the play store has received positive review by the user while there are some apps which have negative reviews as well.

- **TOP POSITIVE SENITMENT RATED FOR GENRES**

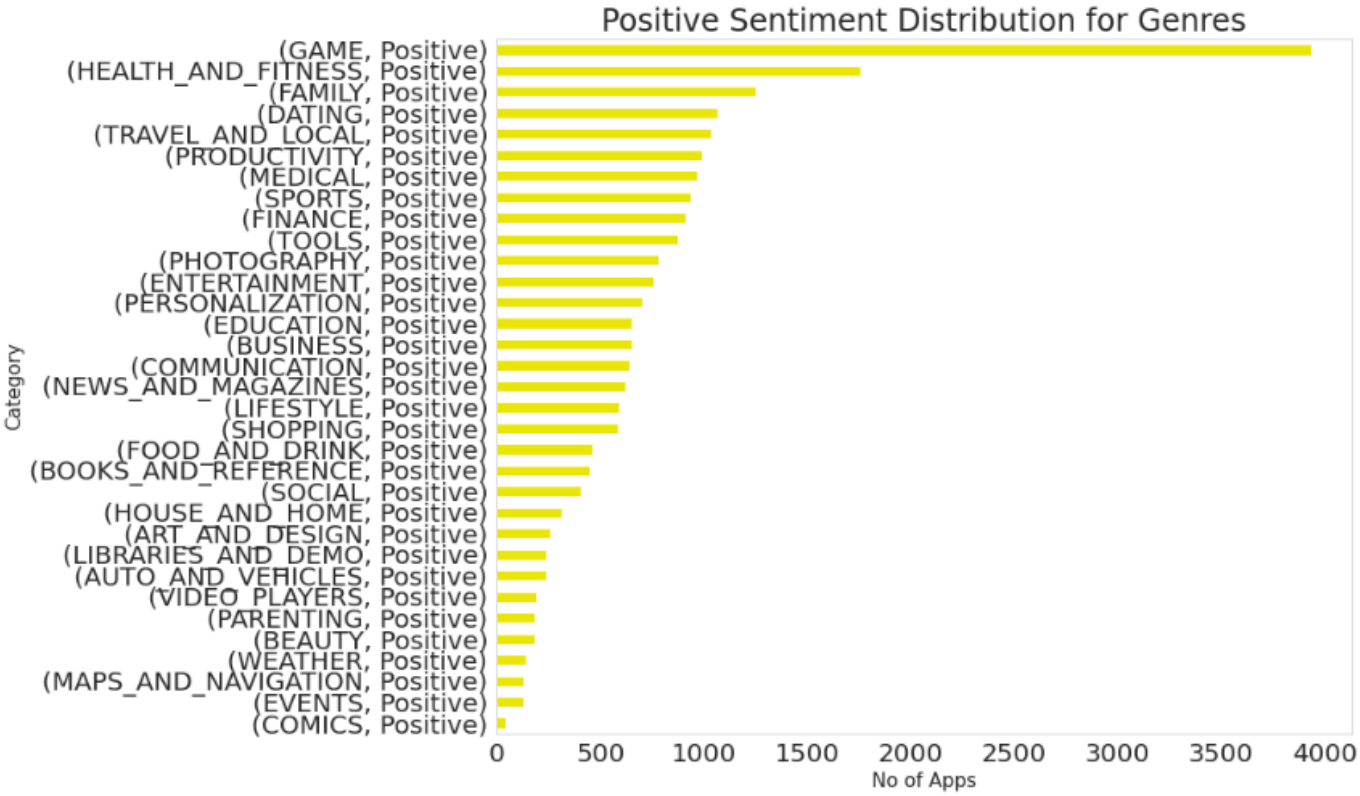


**Fig -12: Top Positive Reviewed Genres**

- **TOP NEGATIVE SENITMENT RATED FOR GENRES**

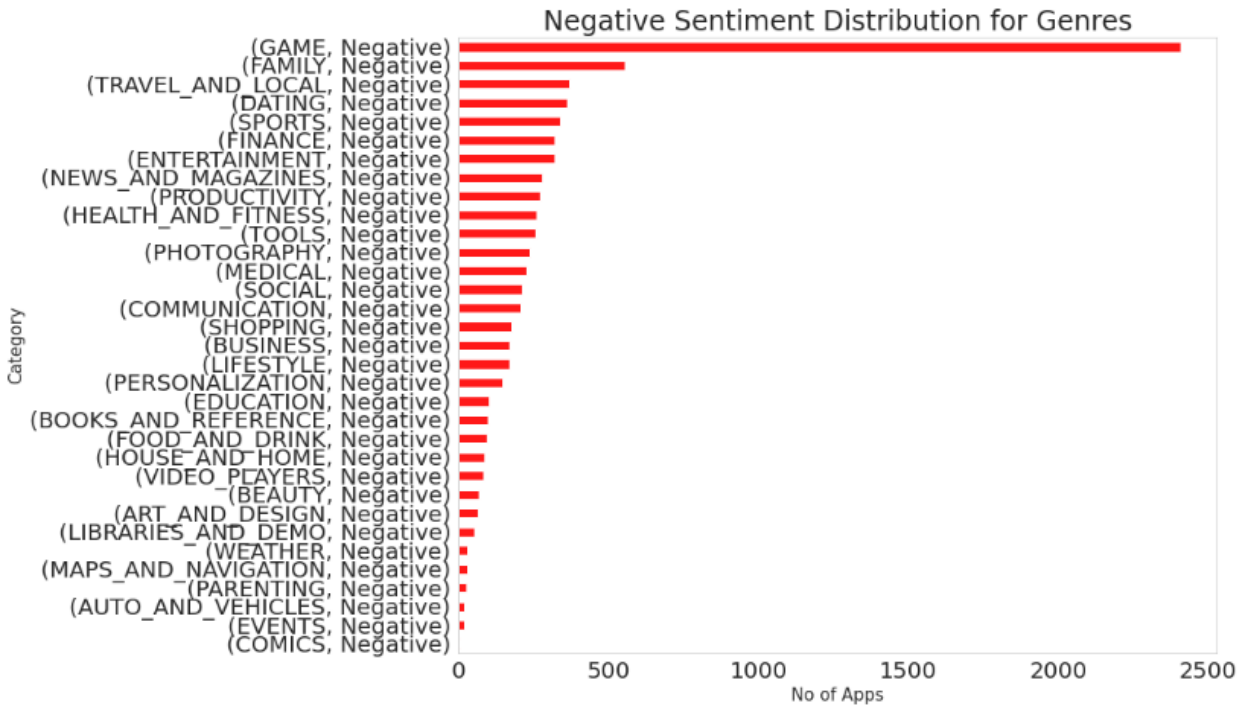

**Fig -13: Top Negative Reviewed Genres**

The graphs showcase that Games has the most number of positive and negative reviews which can be a result of it having the highest number of reviews within the category distribution. But a difference can be seen when moving to the next one, where Health and Fitness holds the second place in positive reviews followed by Family and vice versa in the negative sentiment distribution

- **Distribution of sentiment within the genres**

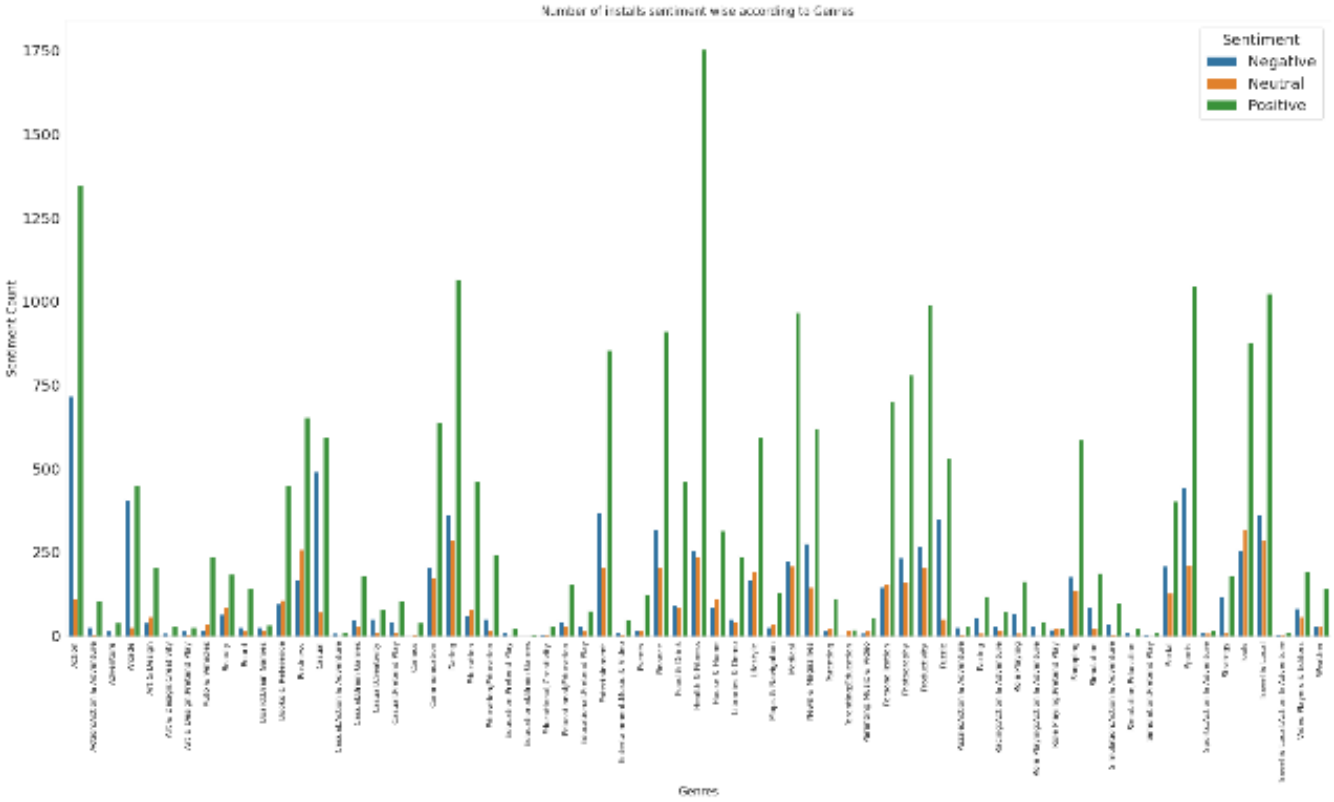


**Fig -14: App install sentiment wise by genres.**

It can be seen that Health and Fitness has the highest number of positive reviews which is followed by Action. But it is also worthwhile to note that Action has the highest number of negative reviews as well while the ratio is much less for Health and Fitness.

**Relationship between sentiment subjectivity proportional to sentiment polarity**

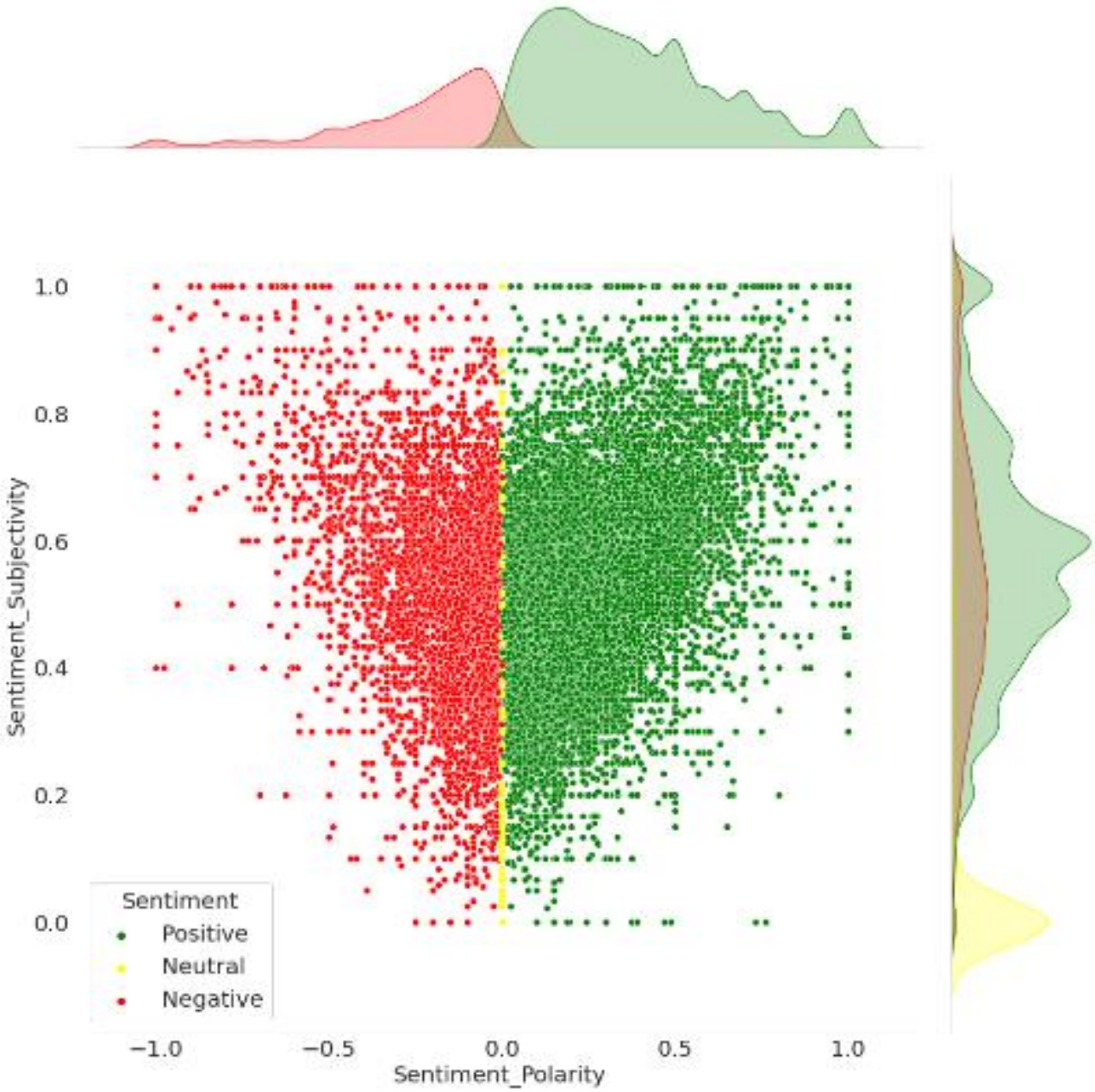


**Fig -17: Proportional check**

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, show a proportional behavior, when variance is too high or low.

- **Distribution of Subjectivity and polarity within DIFFERENT LOCATIONS categories.**

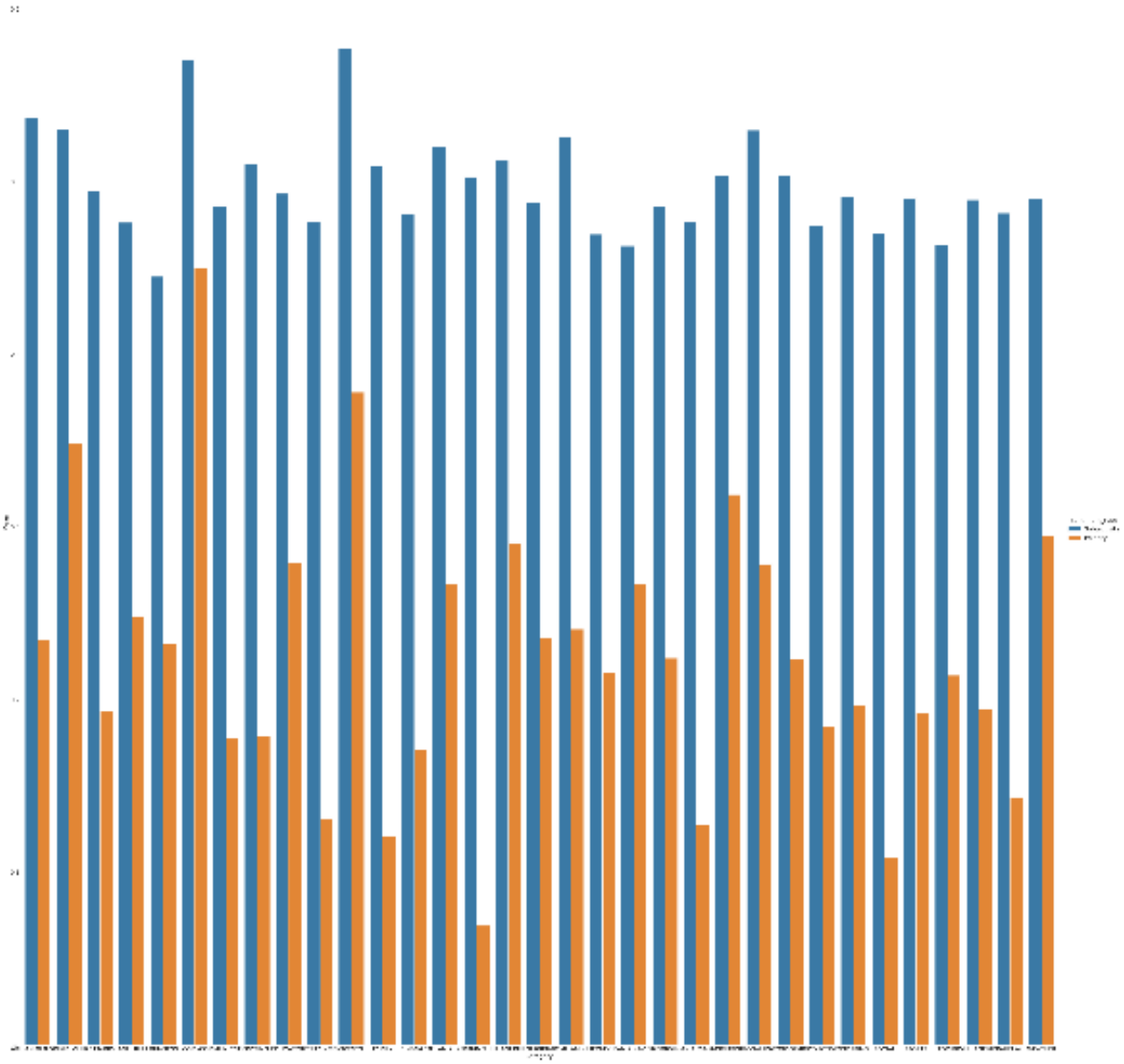

**Fig -18: Distribution of sentiment**

- **Relationship between different features of the dataset**

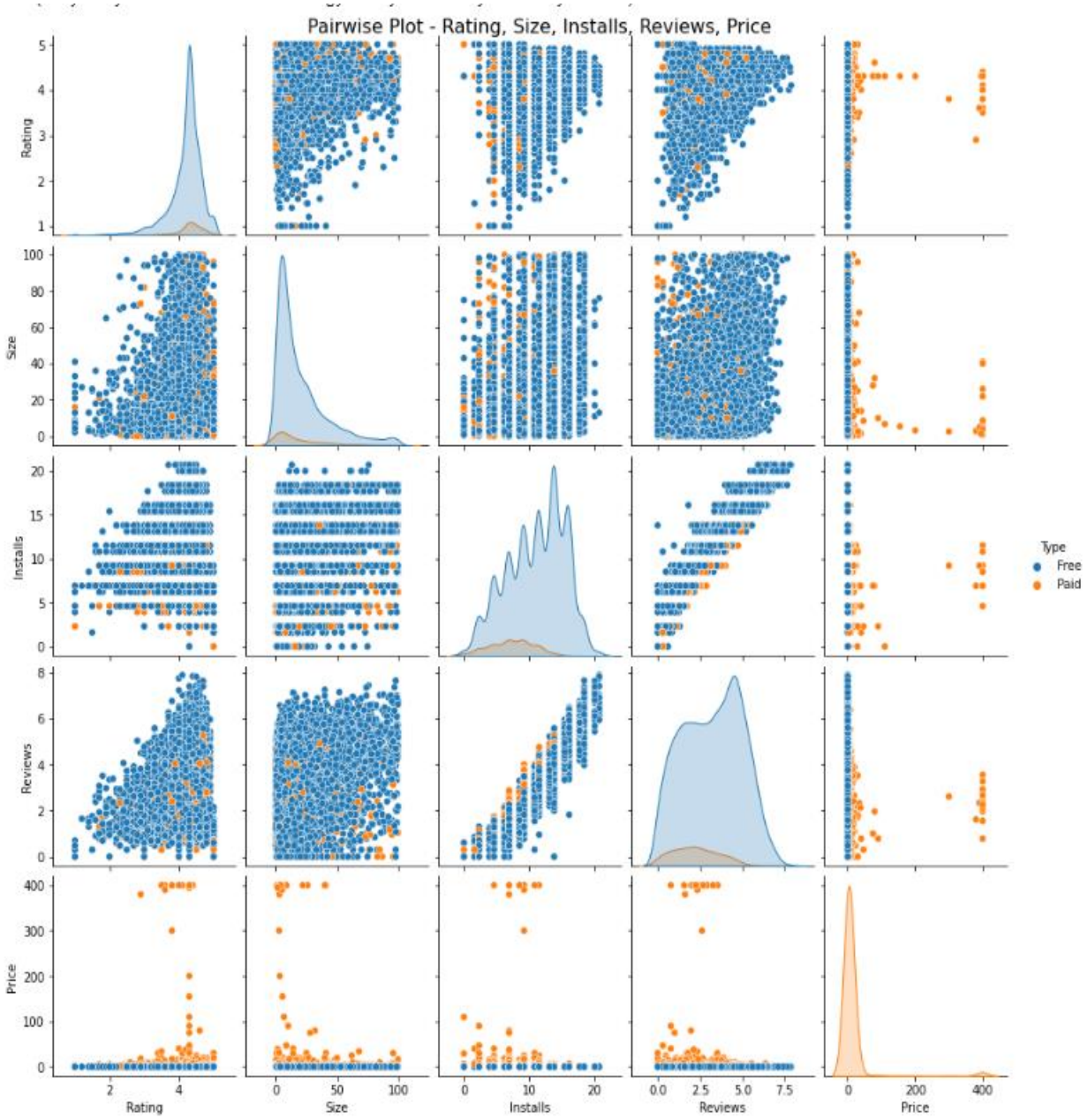


**Fig -19: Pair wise plot**

- Most of the App are Free.
- Most of the Paid Apps have Rating around 4
- As the number of installations increases the number of reviews of the particular app also increases.
- Most of the Apps are light-weighted.

## ➢ Correlation Heatmap

A correlation matrix is simply a table which displays the correlation coefficients for different variables. The matrix depicts the correlation between all the possible pairs of values in a table. It is a powerful tool to summarize a large dataset and to identify and visualize patterns in the given data.
A correlation heatmap is a graphical representation of a correlation matrix representing the correlation between different variables. The value of correlation can take any value from -1 to 1. Correlation between two random variables or bivariate data does not necessarily imply a causal relationship.

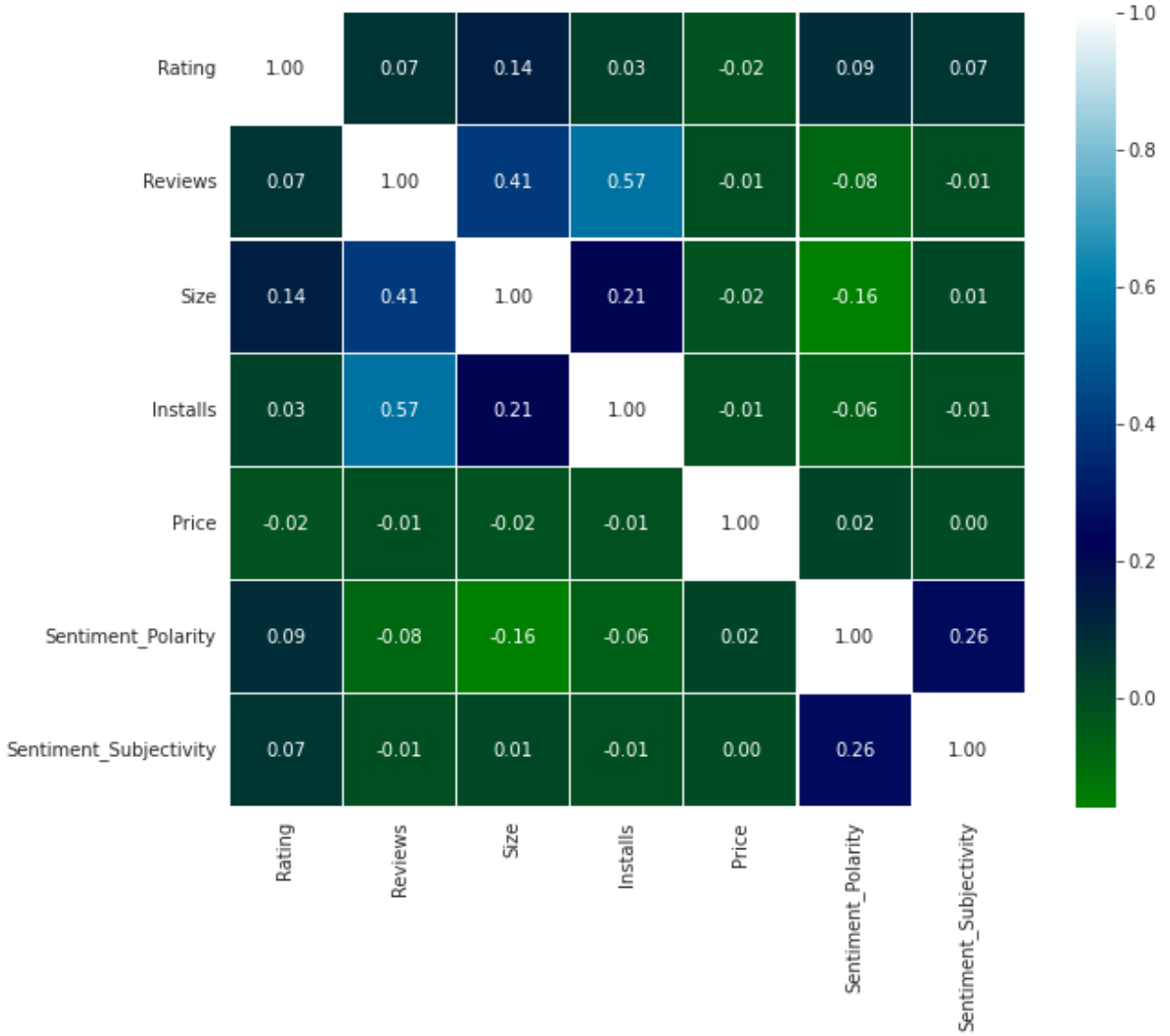


**Fig - 20:**

**Merged Data frame Heatmap**

- There is a strong positive correlation between the Reviews and Installs column. This is pretty much obvious. Higher the number of installs, higher is the user base, and higher are the total number of reviews dropped by the users.
- The Price is slightly negatively correlated with the Rating, Reviews, and Installs. This means that as the prices of the app increases, the average rating, total number of reviews and installs fall slightly.
- The Rating is slightly positively correlated with the Installs and Reviews column. This indicates that as the average user rating increases, the app installs, and number of reviews also increase.

- Sentiment Polarity is not highly correlated with Sentiment Subjectivity.

## END NOTES

**[1]** The state of mobile 2020 (2020). App Annie.*URL https://www. appannie. com/en/insights/market-data/state-ofmobile-2020/. Last accessed November*.

[2] The State of the App Economy, Q3 2021," App Annie, 2021

## REFERENCES