Capstone Project

Play Store App Review Analysis

By

Krushna Chaure

Data Science Trainee

AlmaBetter,Bangalore

Topic for Discussion

1. Introduction
2. What is Exploratory Data Analysis
3. Understand Play Store and User Review Data
4. Data Cleaning and Transforming
5. Count of Apps in each Category
6. Category of Apps have most number of Installed
7. Percentage of Free vs Paid Apps
8. Apps accessible according age group
9. Average rating of different Category of Apps
10. Correlation between Apps(Free vs Paid) and Size
11. Correlation between multiple variable
12. Percentage of review Sentiments
13. Correlation between merged Dataframe
14. Customer Sentiment Subjectivity
15. Challenges Faced
16. Conclusion

Introduction

- In today's scenario we can see that mobile apps playing an important role in any individual's life. The Google play store is one of the largest and most popular Android app stores. It has an enormous amount of data that can be used to make an optimal model. In today's scenario we can see that market has increased to over 3.5 million Apps and around 3000+ apps are being added per day as per a Google survey report. Thus, the market, in turn, led to around 5 billion users downloading all over the world. therefore enormous datasets & variety of insights can be concluded for business improvements. There are various key factors that play a major role in the success & engagement from the user's end. In our data set we have two csv files for data analysis: Play Store data User Reviews At first, we analysis the play store data and in the play store data we have 10841 rows and 13 columns & in the user review data we have 64295 rows and 5 columns of data. We have to take the maximum outcomes from the data which help us to analysis the which type of app is most preferable and comparisons between different insights. Our goal is to filter and make plots accordingly for a better EDA with respect to the final data. We need to explore and analyze the data to discover key factors responsible for app engagement and success.

What is Exploratory Data Analysis

- Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets for patterns, and anomalies (outliers), and form hypotheses based on our understanding of the dataset and summarize their main characteristics, often employing data visualization methods. It is an important step in any Data Analysis or Data Science project. It helps determine how best to manipulate data sources to get the answers you need.
- EDA involves generating summary statistics for numerical data in the dataset and creating various graphical representations to understand the data better and make it more attractive and appealing.

Understand Play Store Data

- **App:** It contains the name of the app with a short description (optional).
- **Category:** This section gives the category to which an app belongs. In this dataset, the apps are divided among 33 categories.
- **Size:** The disk space required to install the respective app.
- **Rating:** The average rating given by the users for the respective app. It can be in between 1 and 5.
- **Reviews:** The number of users that have dropped a review for the respective app.
- **Installs:** The approximate number of times the respective app was installed.
- **Type:** It states whether an app is free to use or paid.
- **Price:** It gives the price payable to install the app. For free type apps, the price is zero.
- **Content rating:** It states which age group is suitable to consume the content of the respective app.
- **Genres:** It gives the genre(s) to which the respective app belongs.
- **Last updated:** It gives the day in which the latest update for the respective app was released.
- **Current Ver:** It gives the current version of the respective app.
- **Android Ver:** It gives the android version of the respective app.

Understand User Review Data

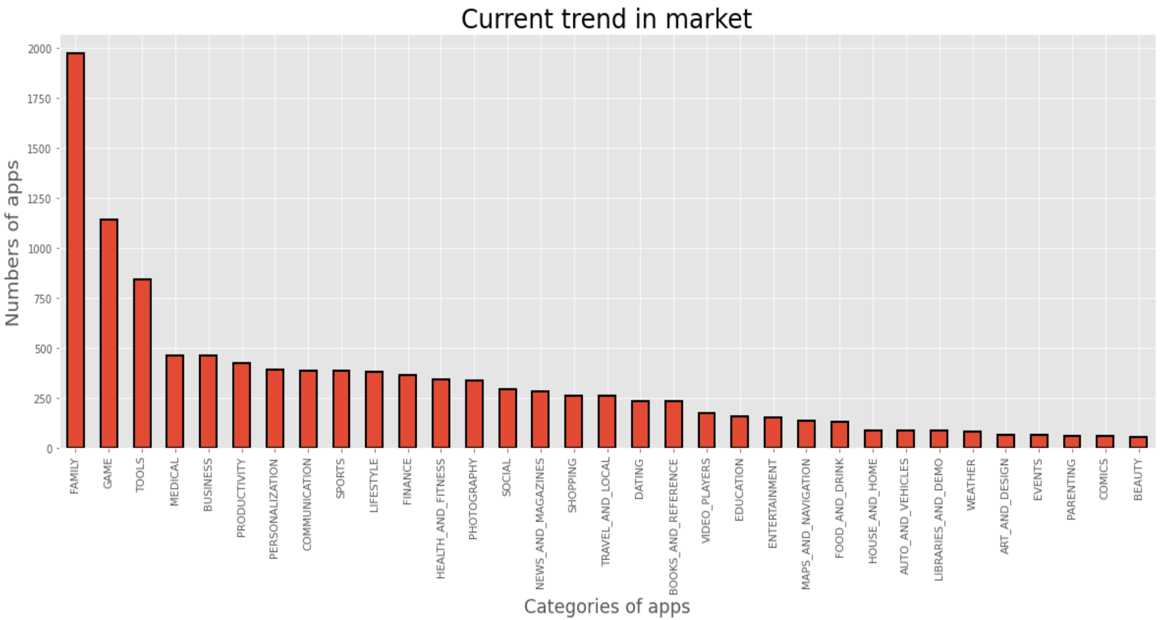
- **App:** It contains the name of the app with a short description (optional).
- **Translated_Review:** It contains the English translation of the review dropped by the user of the app.
- **Sentiment:** It gives the attitude/emotion of the writer. It can be ‘Positive’, ‘Negative’, or ‘Neutral’.
- **Sentiment_Polarity:** 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.

Data Cleaning and transforming

- So during running the code in row number 10472 found mismatch data on that row so we remove this row.
- Some of the columns have a smaller number of null values, so we replace the null values in these columns with the mode value of that particular column.
- From the information of data frame, we can see that all the columns except rating have the object data type but some of the columns like, reviews, size, installs and price have the numerical value. So, we have to transform them in proper data type and also remove the unwanted values from the numerical columns like ‘+’ and ‘,’ from installs and ‘$’ from price also 'M' & 'k' '+' ',’ from size column.

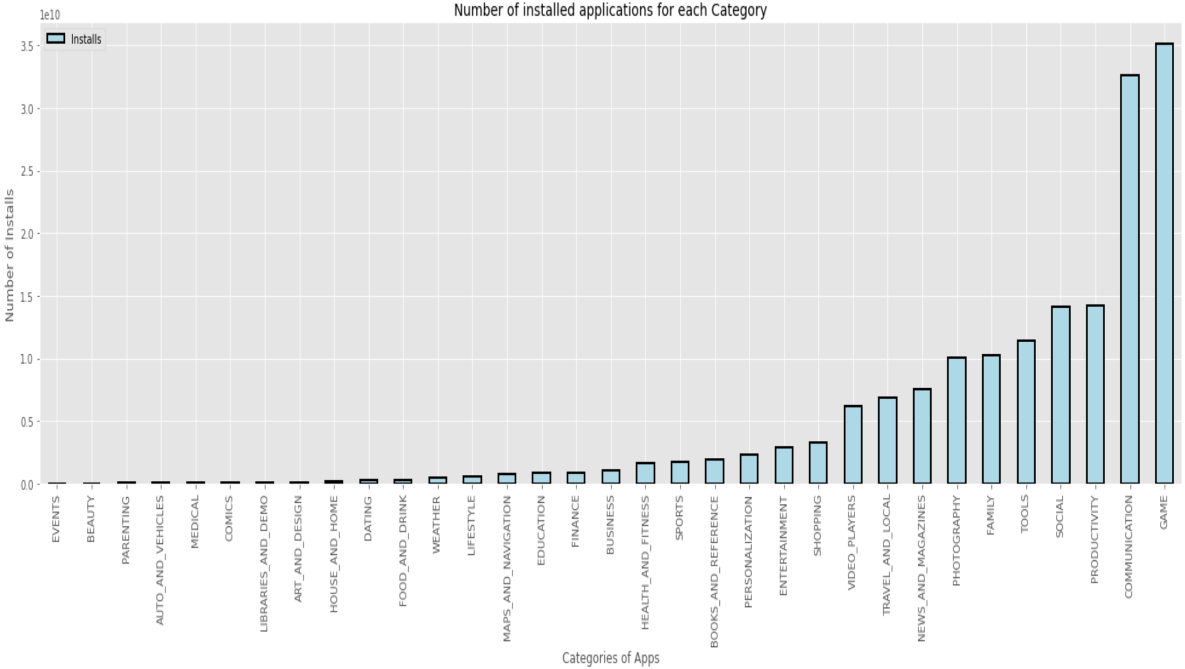
Count of Apps in each Category

From this Bar plot we analyze and finding some insights where Family and Game have maximum number of apps available in play store and Comics & Beauty has less number of apps available in play store.



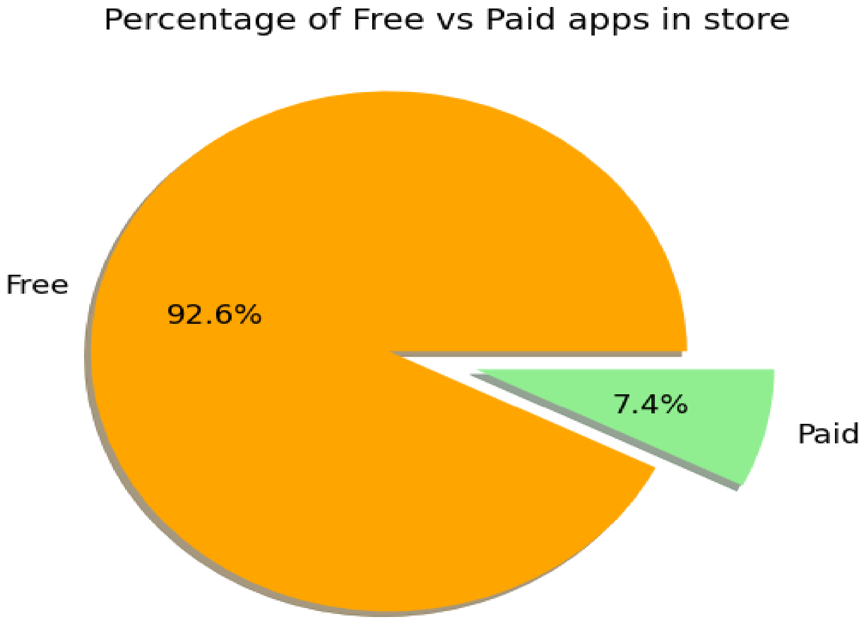
Category of Apps have most number of Installed

From this Bar plot we analyze and finding some insights where gaming and communication type of apps installed top in this list.



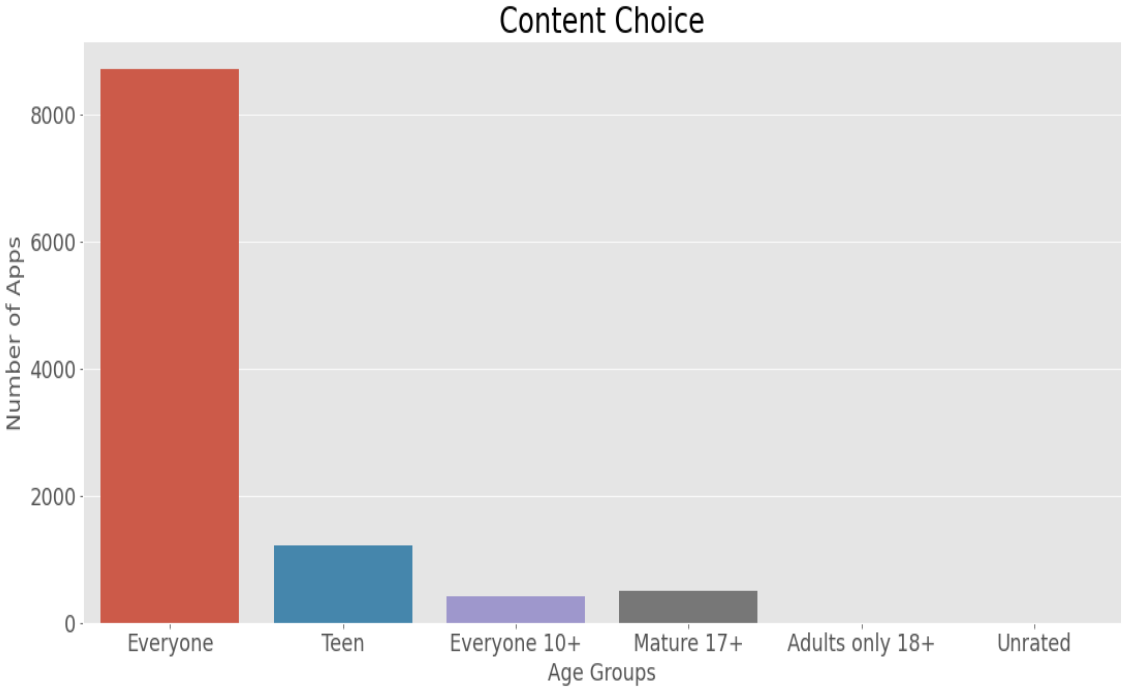
Percentage of Free vs Paid Apps

From this Pie plot we analyze and finding some insights where most of the apps which is 92.6% are free and only 7.4% apps are paid.



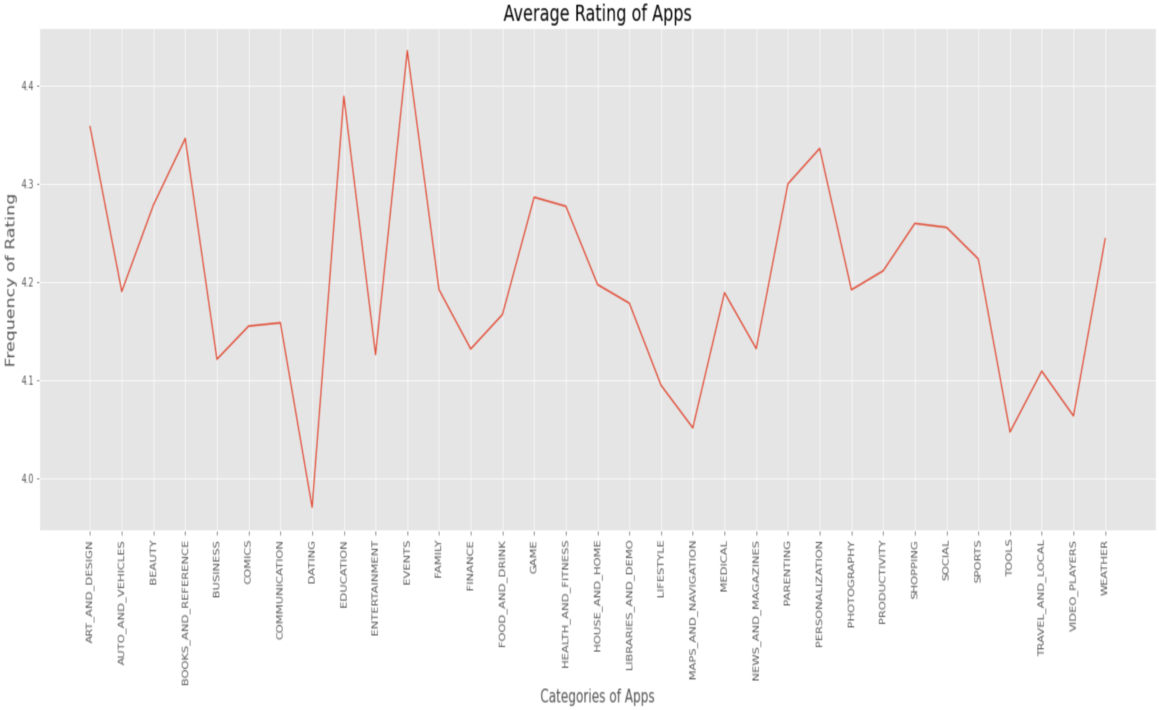
Apps accessible according age group

From this Count plot we analyze and finding some insights where most of the apps in play store are accessible for everyone, there is no any restriction to use this apps and there min number of apps accessible only Teen, Everyone10+, Mature 17+.



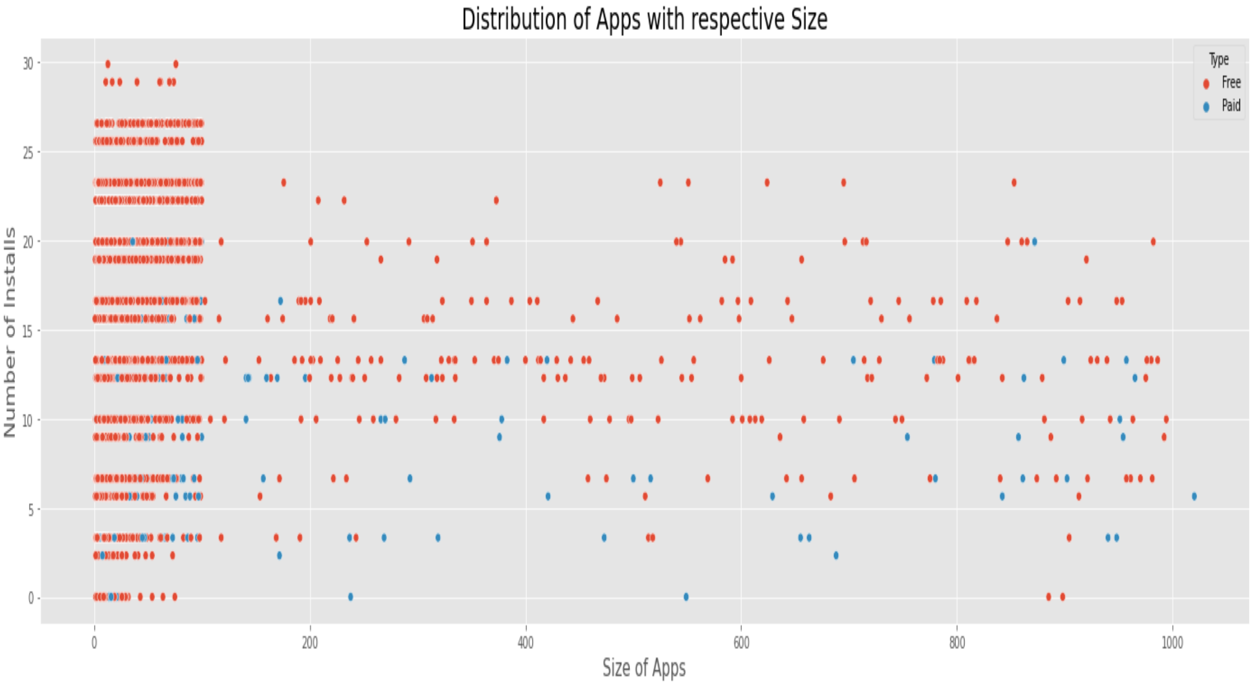
Average rating of different Category of Apps

From this Line chart we analyze and finding some insights where most of the apps rating are between in 4 to 4.5 and education base & event base apps top in this list.



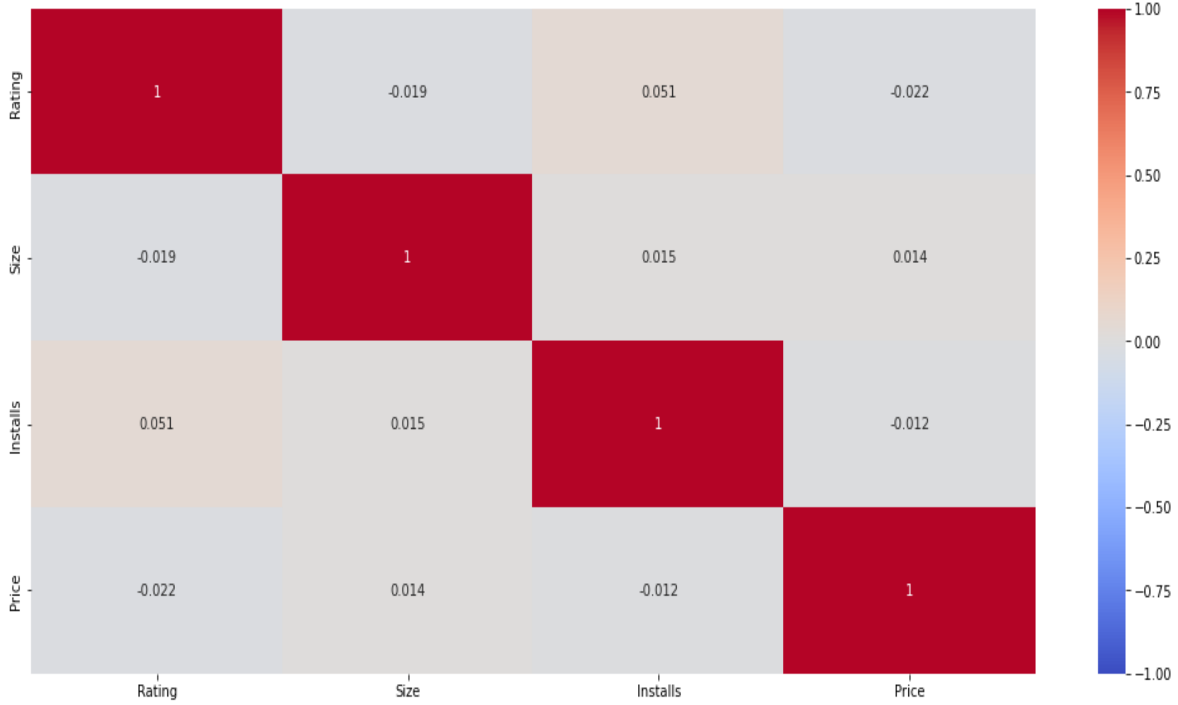
Correlation between Apps(Free vs Paid) and Size

From this Scatter plot we analyze and finding some insights where size of apps smaller then it will more install and size of apps less it will less install also free apps installed maximum time compare to paid apps.



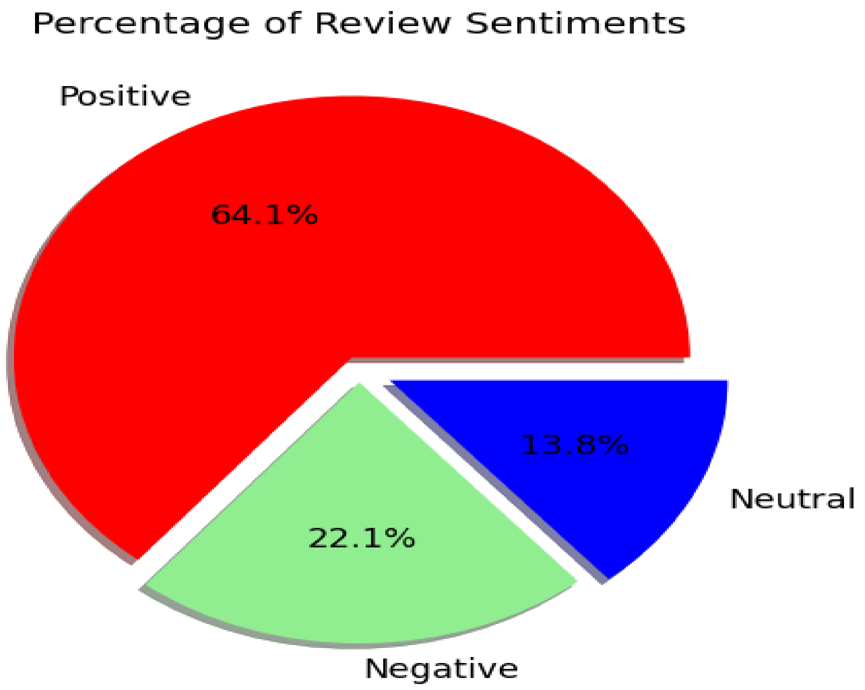
Correlation between multiple variable

From this Heatmap we analyze and finding some insights where positive rating between installs and rating & there negative rating between size and rating.



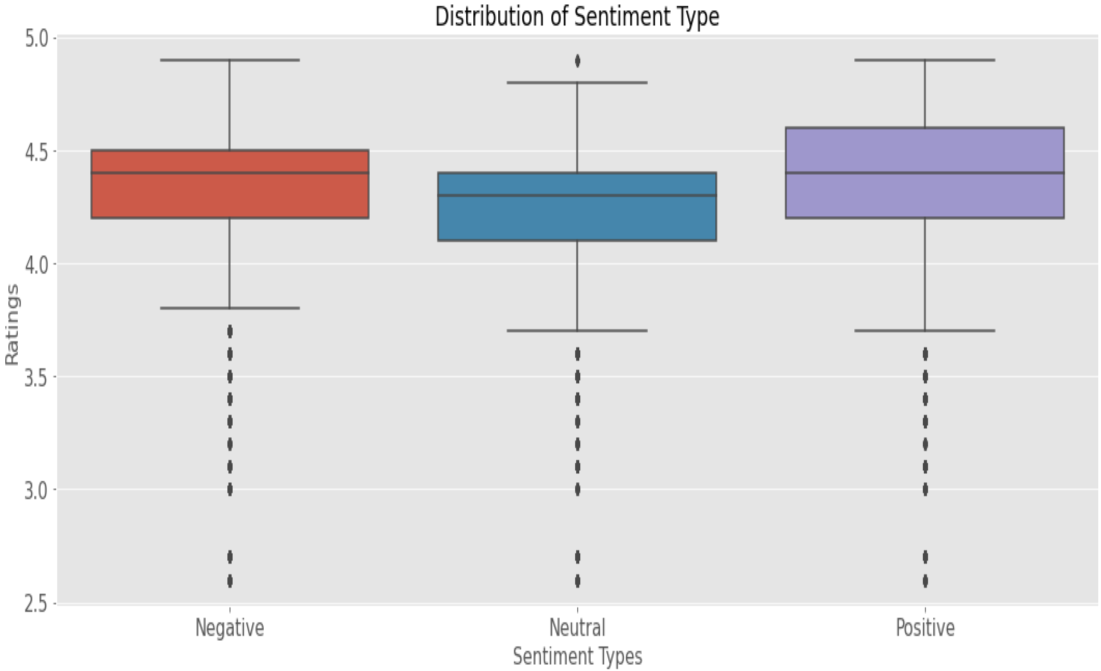
Percentage of review Sentiments

From this Pie plot we analyze and finding some insights where positive sentiment is more which is 64.1%, negative sentiment is 22.1%, neutral sentiment is 13.8%.



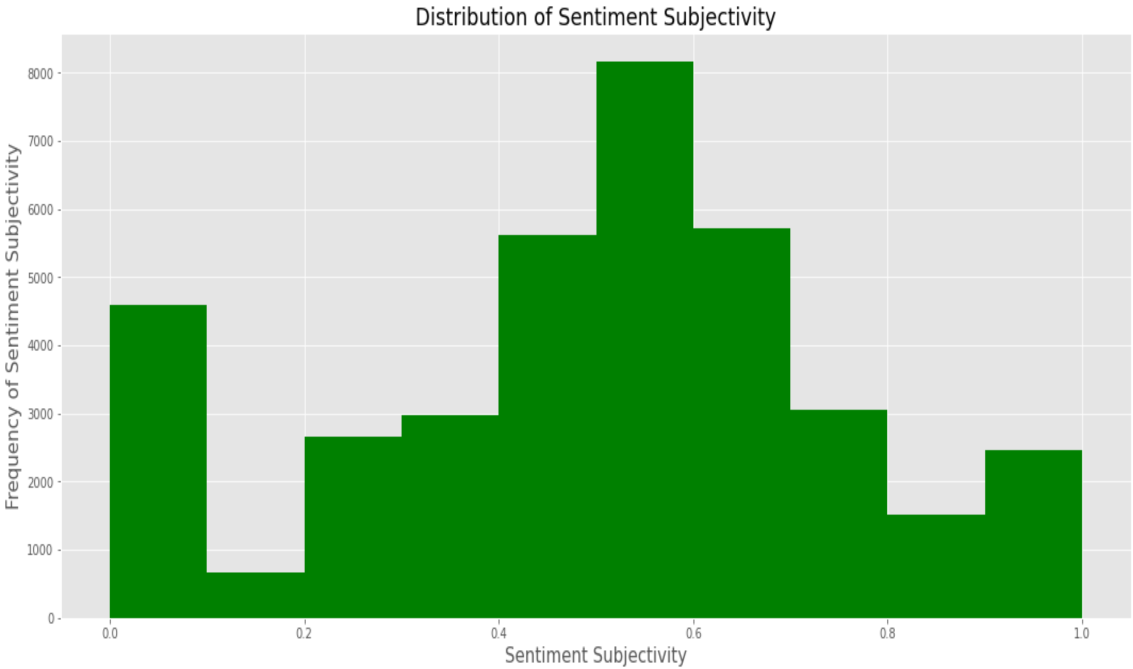
Correlation between merged Dataframes

From this Box plot we analyze and finding some insights where that even its positive, negative or neutral the median(50%) remains near 4.4.



Customer Sentiment Subjectivity

From this Histogram we analyze and finding some insights where maximum number of customers give review to the apps according to their experience.



Challenges Faced

- Reading the dataset and comprehending the problem statement. Our major challenge was data cleaning.
- Handling the error, duplicate and NaN values in the dataset.
- 13.80% of reviews were NaN values, and even after merging both the Dataframes, we could not infer much in order to fill them. Thus we had to drop them.
- User Reviews had 42% of NaN values, which could have been used for developing an understanding of the category wise sentiments.
- Machine learning can help us to deploy more insights by developing models which can help us interpret even more better. We have left this as future work as this is something where we can work on.
- Designing multiple visualizations to summarize the information in the dataset and successfully communicate the results and trends to the reader.

Conclusion

1. Family and Game have maximum number of apps available in play store and Comics & Beauty has less number of apps in play store.
2. Gaming and Communication type of apps installed top in this list.
3. Most of the apps which is 92.6% are Free and only 7.4% apps are Paid.
4. Most of the apps in play store are accessible for Everyone, their is any restriction to use this apps and there min number of apps accessible only Teen, Everyone10+, Mature 17+.
5. Most of the apps rating are between in 4 to 4.5 and Education base & Event base apps top in this list.
6. Size of apps smaller then it will more install and size of apps less it will less install also Free apps installed maximum time compare to Paid apps.
7. Positive rating between Installs and Rating & there negative rating between Size and Rating.
8. Positive sentiment is more which is 64.1%, Negative sentiment is 22.1%, Neutral sentiment is 13.8%.
9. That even its Positive, Negative or Neutral the median(50%) remains near 4.4.
10. Maximum number of customers give review to the apps according to their experience.

Thank You