Abstract - *Google play store is engulfed with a few thousand new applications regularly with a progressively huge number of designers working freely or on the other hand in a group to make them successful, with enormous challenges from everywhere throughout the globe. Since most Play Store applications are free, the income model is very obscure and inaccessible regarding how the in-application buys, adverts, and memberships add to the achievement of an application. In this way, an application's prosperity is normally dictated by the quantity of installation of the application, and the client appraisals that it has gotten over its lifetime instead of the income created. Application (App) ratings are feedback provided voluntarily by users and function as important evaluation criteria for apps. However, these ratings can often be biased due to insufficient or missing votes. Additionally, significant differences are observed between numeric ratings and user reviews. This Study aims to predict the ratings of Google Play Store apps using Data Analysis. I have tried to perform Data Analysis and prediction into the Google Play store application dataset. Using Exploratory Data Analysis, I have tried to discover the relationships among various attributes present in my dataset such as which application is free or paid, the user reviews, rating of the application.*

1. PROBLEM STATEMENT

The data we have taken for analysis is the Google play store dataset. Every row contains various entries regarding a certain app. We will be doing Exploratory data analysis on this data set, which is a very important step in the data science cycle, as it helps in taking very initial business decisions and preparing the data for further modeling for use in machine learning algorithms. Our objective will be to structure the data, clean it and present certain trends that we observe that can help us draw very preliminary conclusions about the probability of success of a newly launched app.

2. INTRODUCTION

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.

3. GOOGLE PLAY STORE AND USER REVIEW ANALYSIS

In today's scenario, we can see that mobile apps play an important role in any individual's life. It has been seen that the development of mobile application advertising has an incredible effect on advanced innovation. Having said that, with the consistently developing versatile application showcase there is additionally an eminent ascent of portable application designers inevitably bringing about high as can be income by the worldwide portable application industry.

With enormous challenges from everywhere throughout the globe, it is basic for a designer to realize that he is continuing in the right heading. To hold this income and their place in the market the application designers may need to figure out how to stick to their present position. The Google Play Store is observed to be the biggest application platform. It has been seen that although it creates more than two-fold the downloads than the Apple App Store yet makes just a large portion of the cash contrasted with the App Store. In this way, I scratched information from the Play Store to direct our examination on it.

With the fast development of advanced cells, portable applications (Mobile Apps) have turned out to be basic pieces of our lives. Be that as it may, it is troublesome for us to follow along with the fact and to understand everything about the apps as new applications are entering the market each day. It accounted that the Android market achieved a large portion of a million applications in September 2011. Starting now, 0.675 million Android applications are accessible on Google Play App Store. Such a lot of applications are by all accounts an extraordinary open door for clients to purchase from a wide determination extent. We trust versatile application clients consider online application surveys as a noteworthy impact for paid applications. It is trying for a potential client to peruse all the literary remarks and ratings to settle on a choice. Additionally, application engineers experience issues in discovering how to improve the application execution dependent on evaluations alone and would profit by understanding a huge number of printed remarks.

We develop Android apps & release them on Play Store. As a Developer or say Business Perspective it’s very important to know whether users are enjoying the app or facing any issues. To know this Play Store has a Ratings & reviews section for each app released on the play store. Users can submit ratings and have the freedom to write a review for a particular app. This approach is quite lengthy to rate & review the app i.e., navigate to the Play Store to submit feedback or redirect leaving a current app workflow to open the Play Store App link using URI. We never wanted our customers to leave our application, but with this flow, we are forced to redirect the control to the Play Store app.

3.1 GOOGLE PLAY STORE DATASET:

The dataset consists of the Google play store application and is taken from Alma Better which is again taken from Kaggle, which is the world’s largest community for data scientists to explore, analyze and share data.

This dataset is for Web scratched information of 10k Play Store applications to analyze the market of Android. Here it is a downloaded dataset that a user can use to examine the Android market of different use of classifications music, camera, etc. With the assistance of this, the client can predict see whether any given application will get a lower or higher rating level. This dataset can be moreover used for future references for the proposal of any application. Additionally, the disconnected dataset is picked to choose the estimate exactly as online data gets revived all around a great part of the time. With the assistance of this dataset, I will examine various qualities like rating, free or paid, and so forth utilizing Hive and after that, I will likewise do a forecast of various traits like client surveys, ratings, etc.

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 times 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 or not the user must pay money to install the app on their device.
- **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 which genre the app belongs to, genre can be considered as a sub-division of a 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.

USER REVIEW DATASET

The 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 a ‘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 public. Its range is [0,1]. The higher the subjectivity, the closer is the reviewer's opinion to the opinion of the public, and lower subjectivity indicates the review is more factual information.

PYTHON

Most of the info scientists use Python due to the good built-in library functions and therefore the decent community. Python now has 70,000 libraries. Python is the simplest programing language to select compared to other languages. That is the reason most data scientists use Python more often, for machine learning and data processing data analyst want to use some language that's straightforward to use. That is one of the most reasons to use Python. Specifically, for data scientists the foremost popular data inbuilt open-source library is named Panda. As we have seen earlier in our previous assignment once we got to plot scatterplots, heat maps, graphs, 3and -dimensional data python built-in library comes very helpful.

DATA CLEANING AND PREPARATION

Preprocessing is important in transitioning raw data into a more desirable format. Undergoing the preprocessing process can help with completeness and comparability. For instance, you'll see if certain values were recorded or not. Also, you'll see how trustable the info is. It could also help with finding how consistent the values are. We need preprocessing because most real-world data is dirty. Data can be noisy i.e.; the data can contain outliers or simple errors generally. Data can also be incomplete i.e., there can be some missing values.

The available data is raw and unusable for Exploratory data analysis, so before we do anything with the data we will have to explore and clean it to prepare it for data analysis.

Step1: We write a function play store 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 those columns in the play store dataset.

Step 2: We drop the columns ‘Current Ver’, ‘Android Ver’, and ‘Last updated’ from our dataset using the drop() function of the panda’s library.

Step 3: Handling the duplicates in the App column we drop the no of duplicate rows that are present in the App columns.

Step 4: We can see that the ‘Rating’ column has 1474 null values. We have visually shown null values using Heatmap. 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 median of the column using the median () method and fill this value in place of null values using the fillna() function.

Step 5: 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 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. From the size column we can observe that size of an app is in two formats MB and KB. For the sake of uniformity, we need to convert KB into MB. We know that 1MB = 1024KB. so, we need to divide KB by 1024 to convert it into the format of MB. We also change the datatype from object to float.

Step 7: 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 an ‘int’ datatype.

Step 8: We have changed the Last Updated data column from data type to datetime.

Step 9: 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.

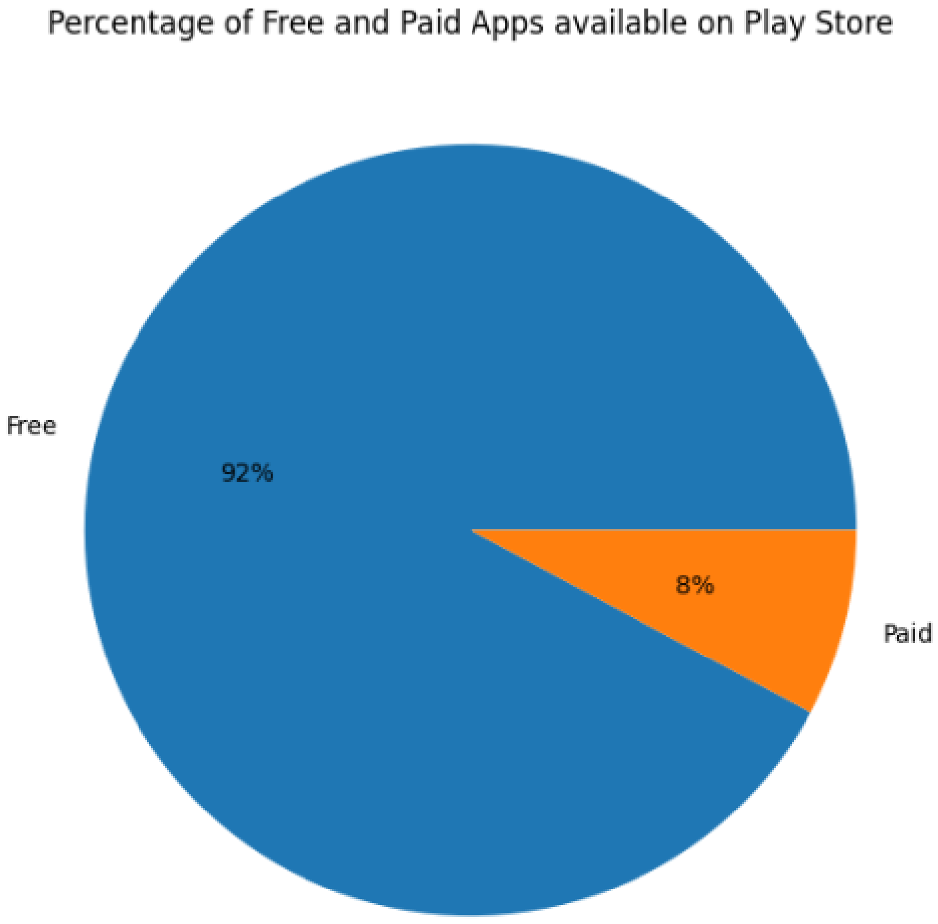
4. EXPLORATORY DATA ANALYSIS

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.

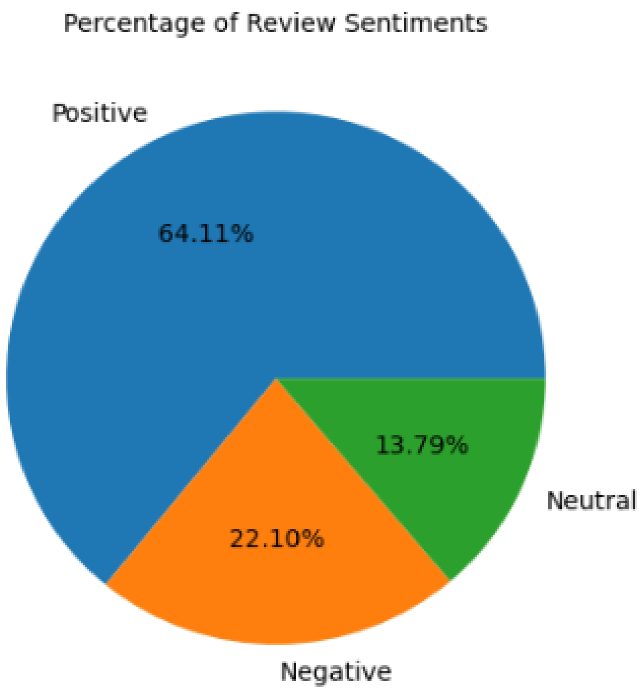
EDA involves generating summary statistics for numerical data in the dataset and creating various graphical representations to understand the data better. In this article, we will understand EDA with the help of an example dataset. We will use Python language (Pandas library) for this purpose.

4.1 Free vs Paid

From the plot below, we can observe that the majority of the apps in the Play Store are Free.

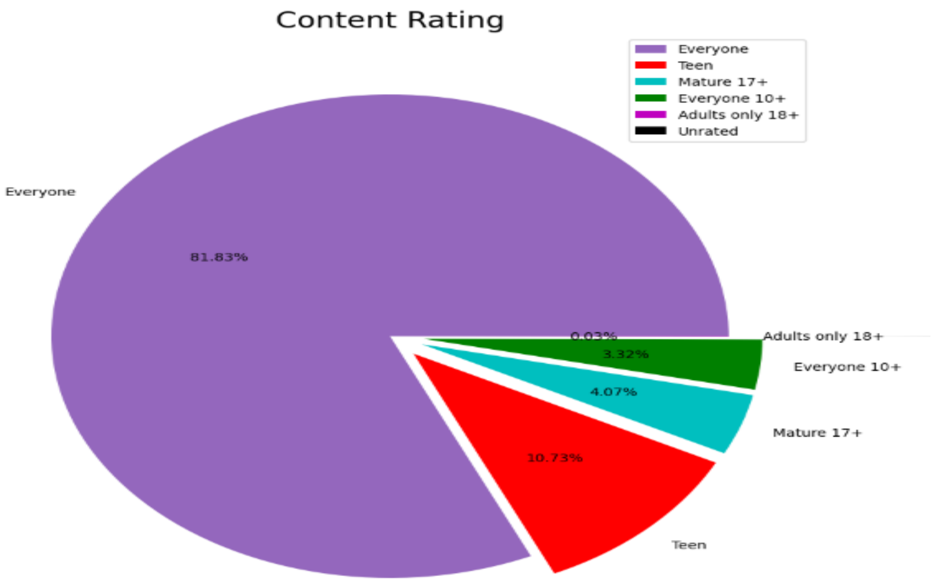


4.2 Percentage of Review Sentiments



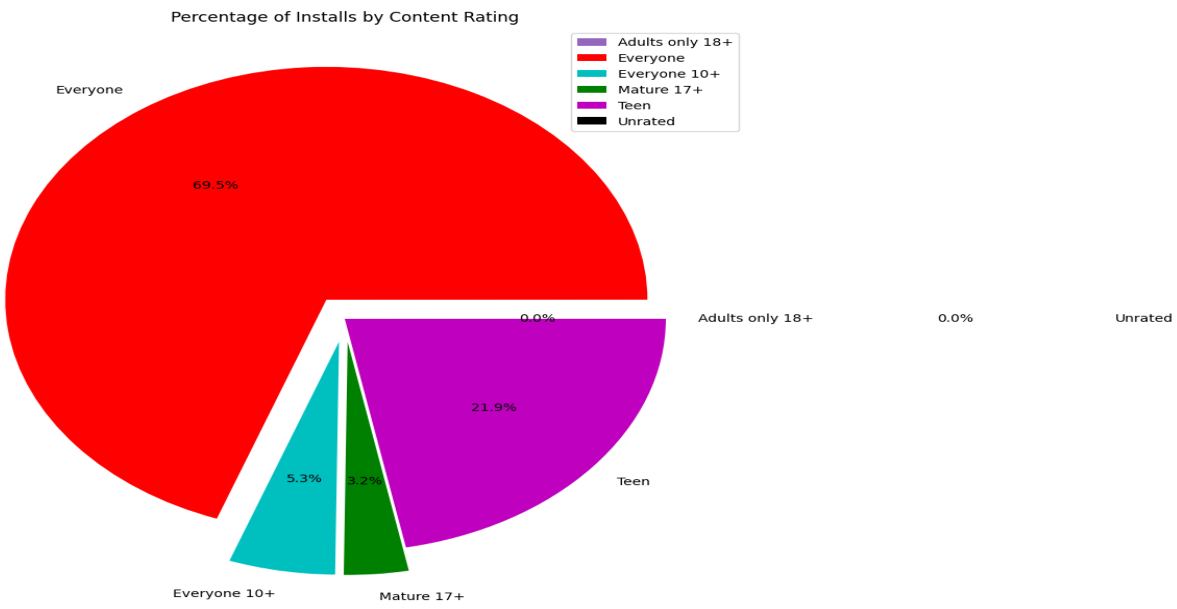
1. Positive reviews are 64.12%.
2. Negative reviews are 22.10%.
3. Neutral reviews are 13.78%.

4.3 Content Rating



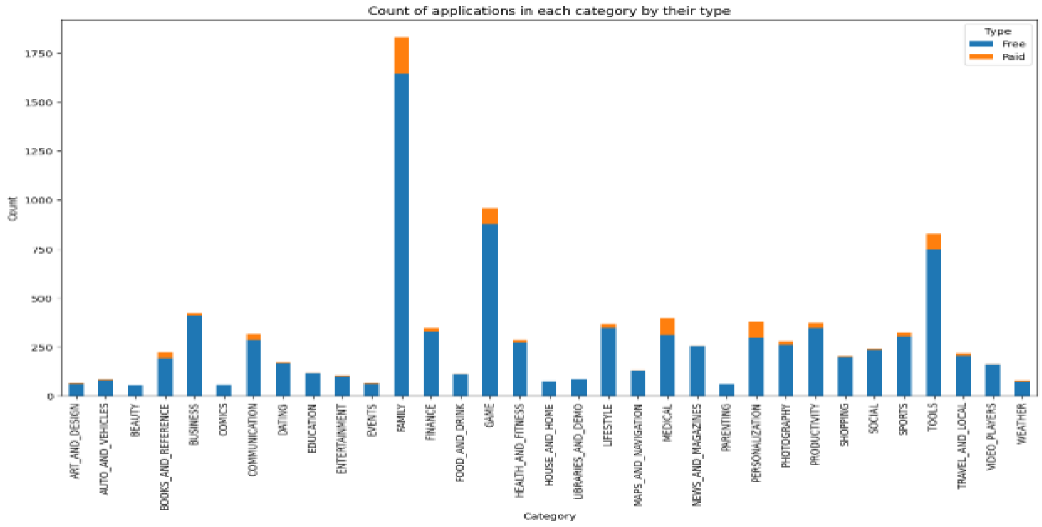
The majority of the apps can be used by Everyone.

4.4 Percentage of Installs by Content Rating



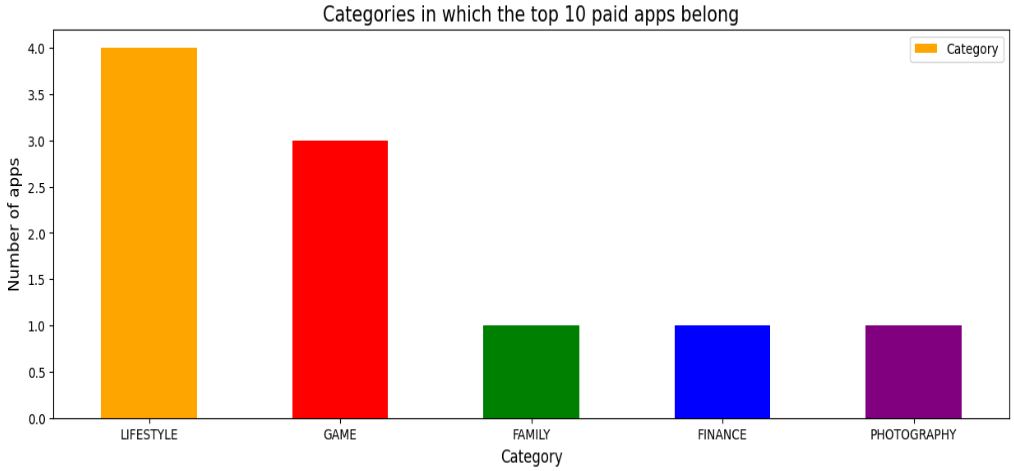
Most of the installs come from apps that can be used by everyone.

4.5 The count of applications in each category by their type



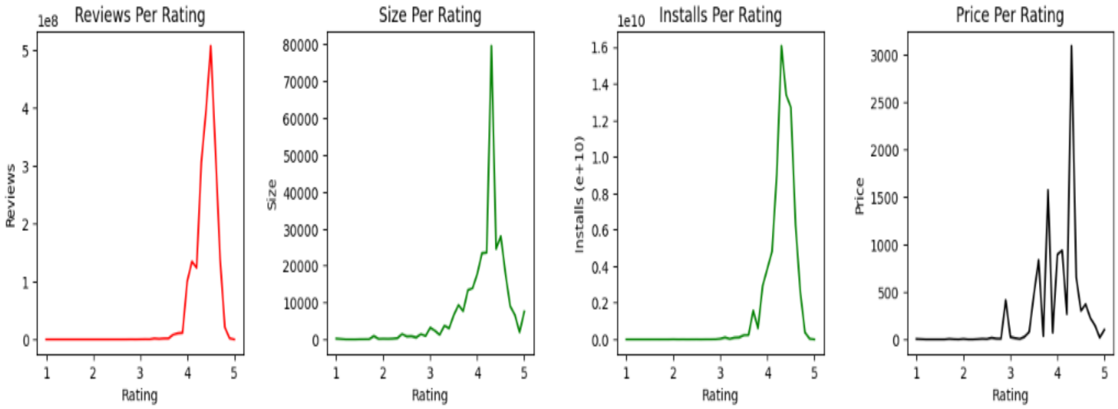
Family, Games, and Tools apps have the highest count of applications.

4.6 Categories in which the top 10 paid apps belong



Lifestyle and Game have the most paid apps.

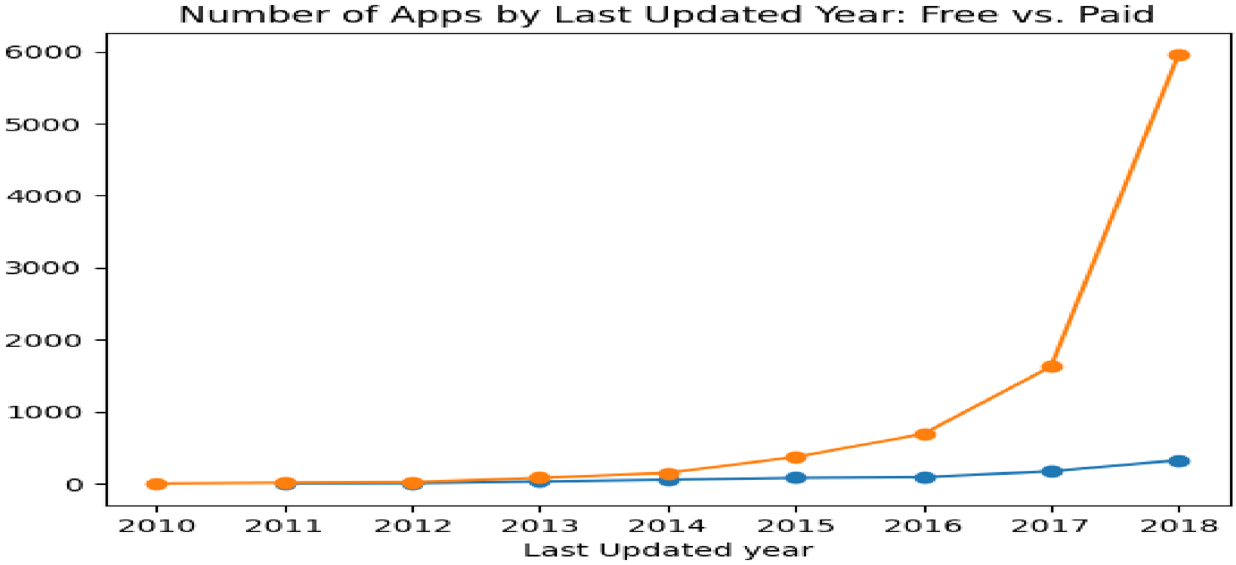
4.7 Plotting the graphs of reviews, size, installs, and price per rating



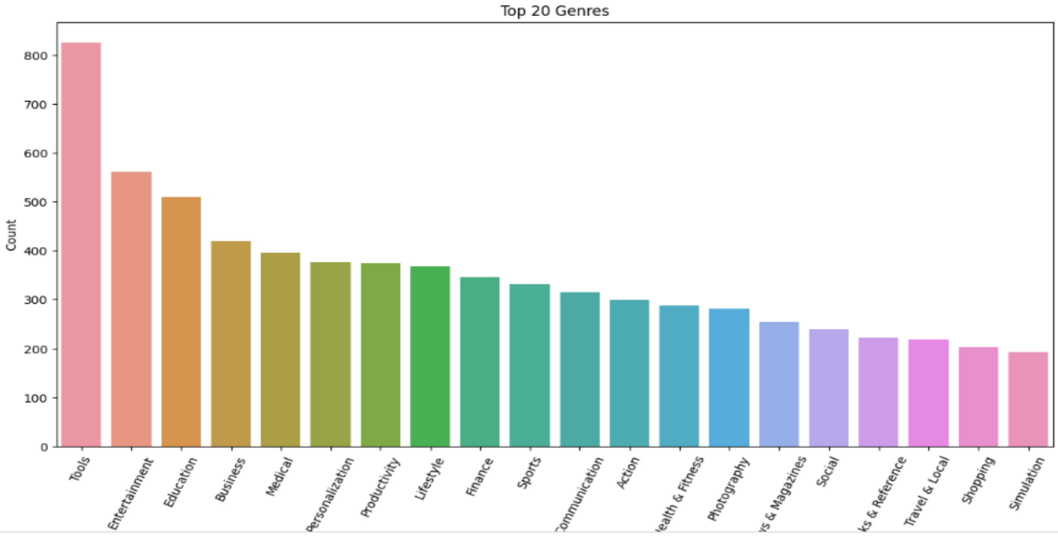
From the above plots, we can observe that most apps with a higher rating range of 4.0 - 4.7 have a high amount of reviews, sizes, and installs. In terms of price, it doesn't reflect a direct relationship with rating, as we could see a fluctuation in terms of pricing even at the range of high rating.

4.8 Number of Apps by Last Updated Year: Free vs. Paid

In the plot below, we plotted the apps updated or added over the years comparing Free vs. Paid, by observing this plot we can conclude that before 2011 there were no paid apps, but with the years passing free apps has been added more in comparison to paid apps, By comparing the apps updated or added in the year 2011 and 2018 free apps are increasing.

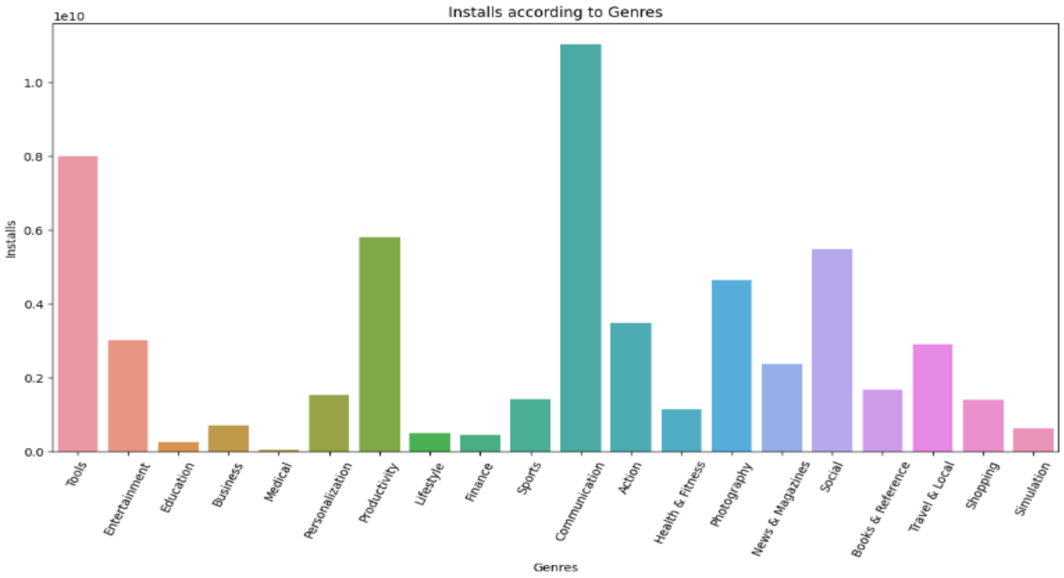


4.9 Top 20 Genres by their installs



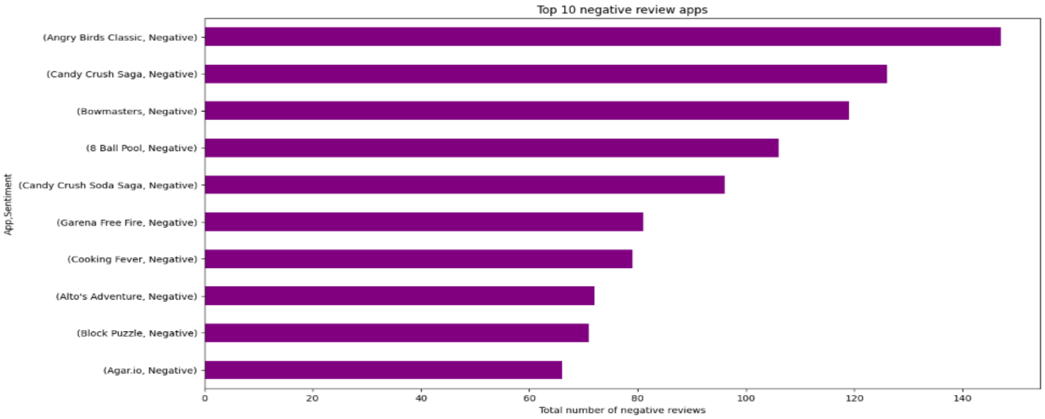
We can see from the above that the maximum number of apps present in the google play store comes under Tools, Entertainment, and Education Genres.

4.10 Installs according to Genres



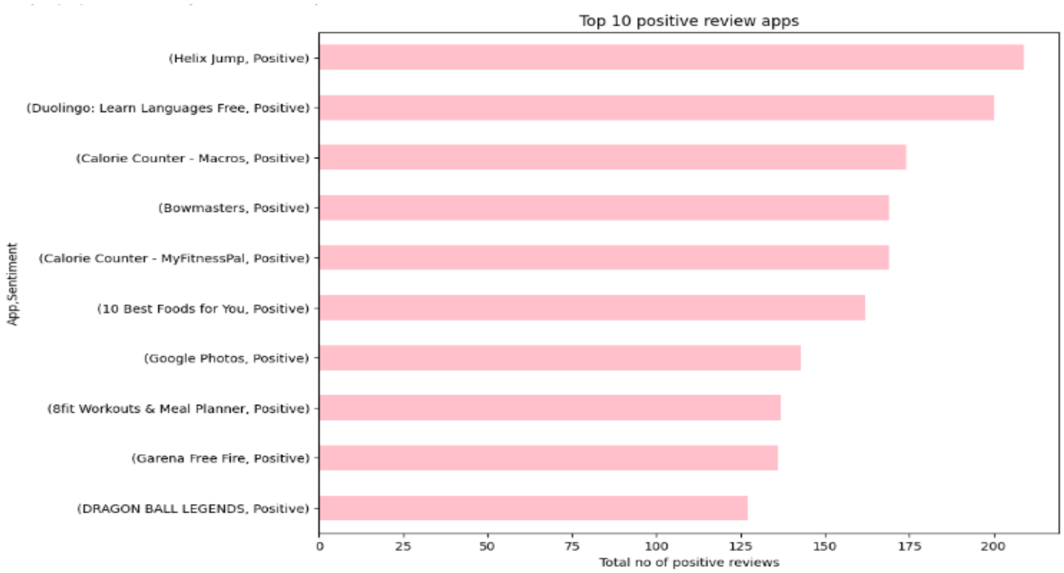
As per the installation and requirement in the market plot, the Maximum installed apps comes under the Communication, Tools, and Productivity Genres.

4.11 Top10 Negative Reviews Apps



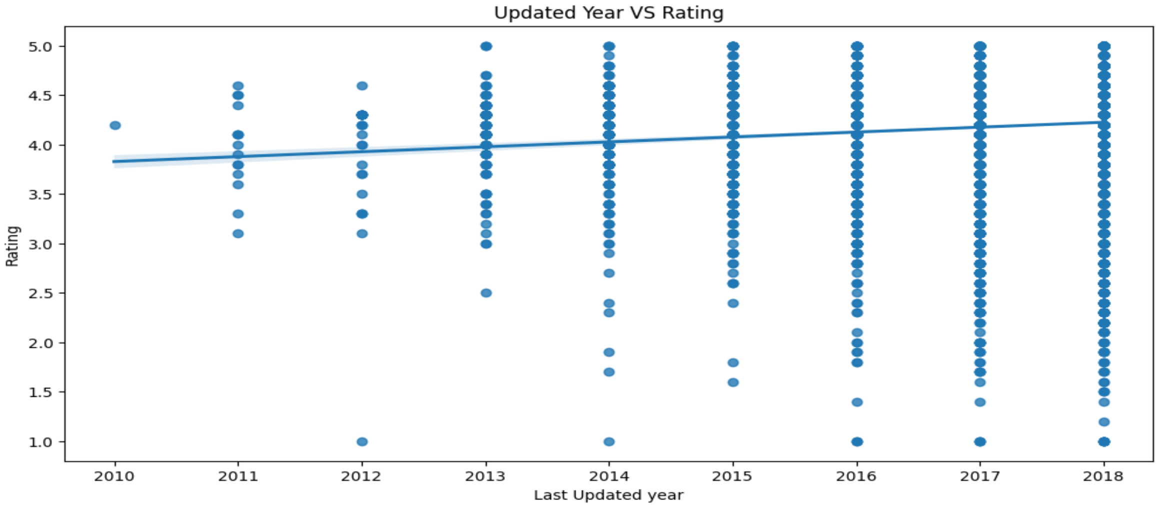
Angry Birds Classic and Candy Crush Saga has the highest number of negative reviews.

4.12 Top10 Positive Reviews Apps



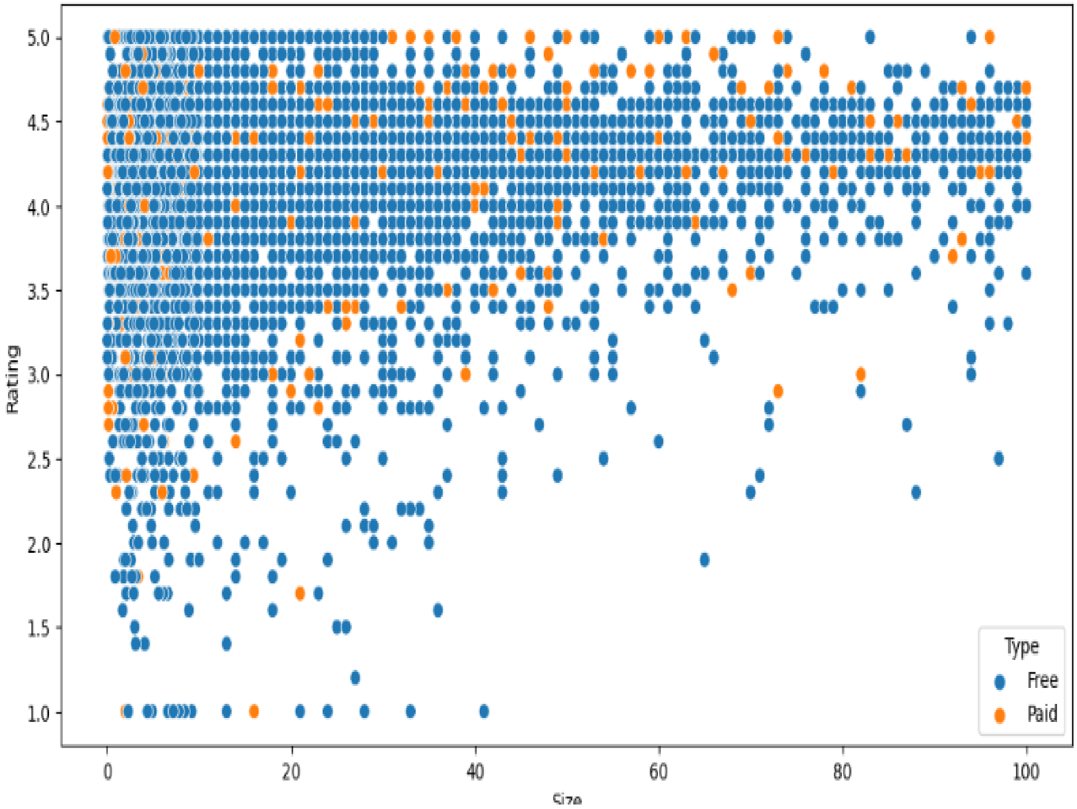
Helix Jump and Duolingo have the highest number of positive reviews

4.13 Updated Year vs Rating



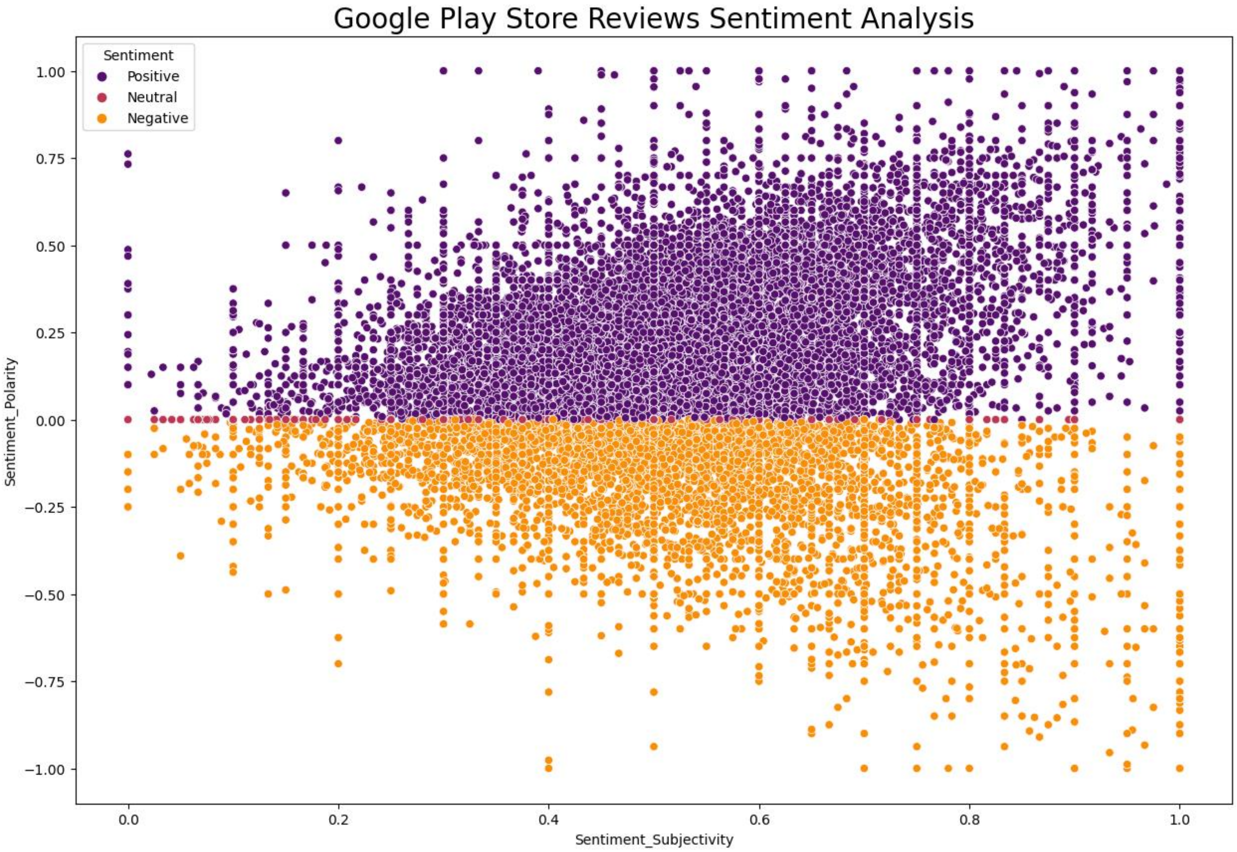
There is a positive correlation between the "Updated Year" and the "Rating" of the apps. As the update year increases, there is a tendency for the apps to have higher ratings. This suggests that apps that are more recently updated are more likely to have better ratings. Users may perceive apps that receive regular updates as more reliable, functional, and up-to-date, leading to higher ratings.

4.14 Distribution of apps in terms of their rating, size, and type



From this scatter plot, we can imply that the majority of the free apps are small in size and have high ratings. While for paid apps, we have quite equal distribution in terms of size and rating.

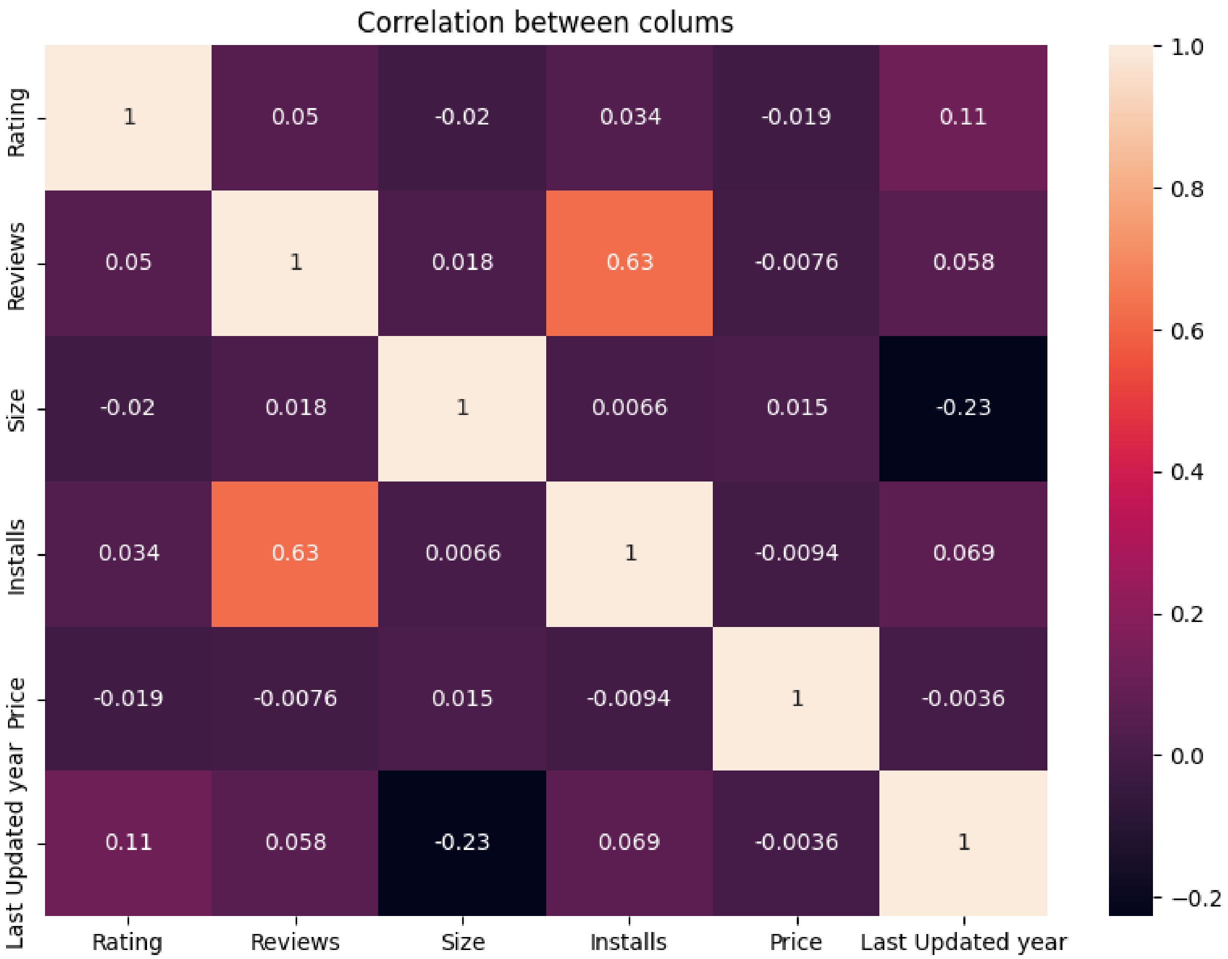
4.15 Google Play Store Reviews Sentiment Analysis



From the above scatter plot, it can be concluded that sentiment subjectivity is not always proportional to sentiment polarity but in a maximum number of cases, shows a proportional behavior, when variance is too high or low

4.16 Correlation

- There is a strong positive correlation between the Reviews and Installs columns. This is pretty much obvious. The higher the number of installs, the higher the user base, and the higher the total number of reviews dropped by the users.

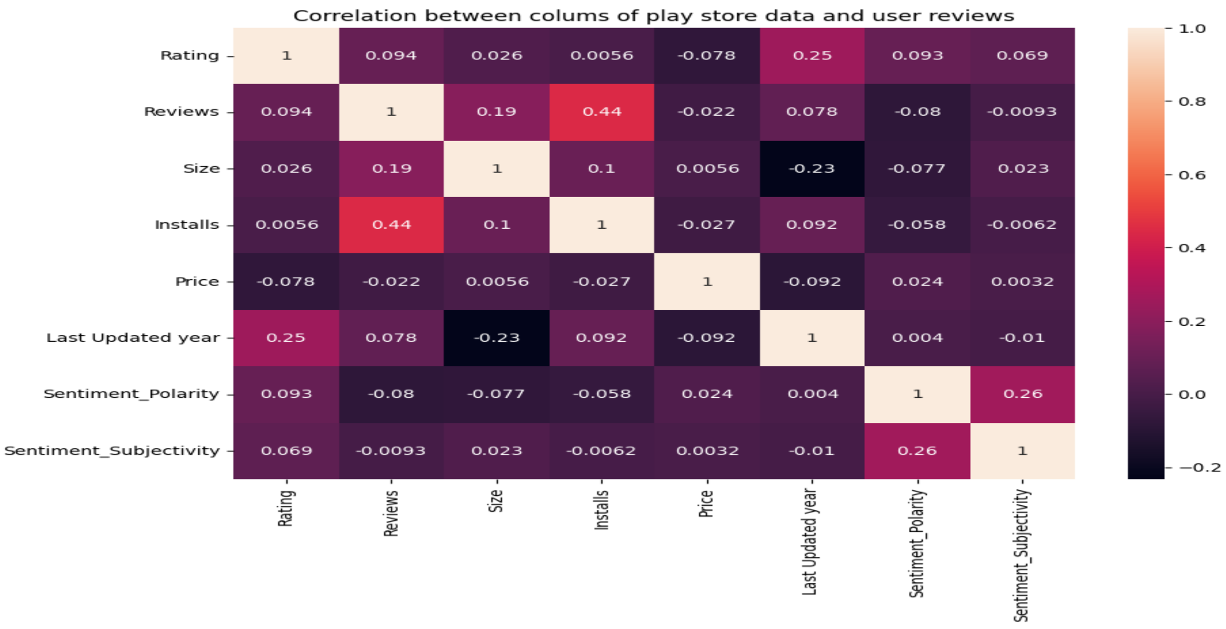


- The Price is slightly negatively correlated with the Rating, Reviews, and Installs. This means that as the prices of the app increase, the average rating, total number of reviews, and Installs fall slightly.
- The Rating is slightly positively correlated with the Installs and Reviews column. This indicates that as the average user rating increases, the app installs and the number of reviews also increases.

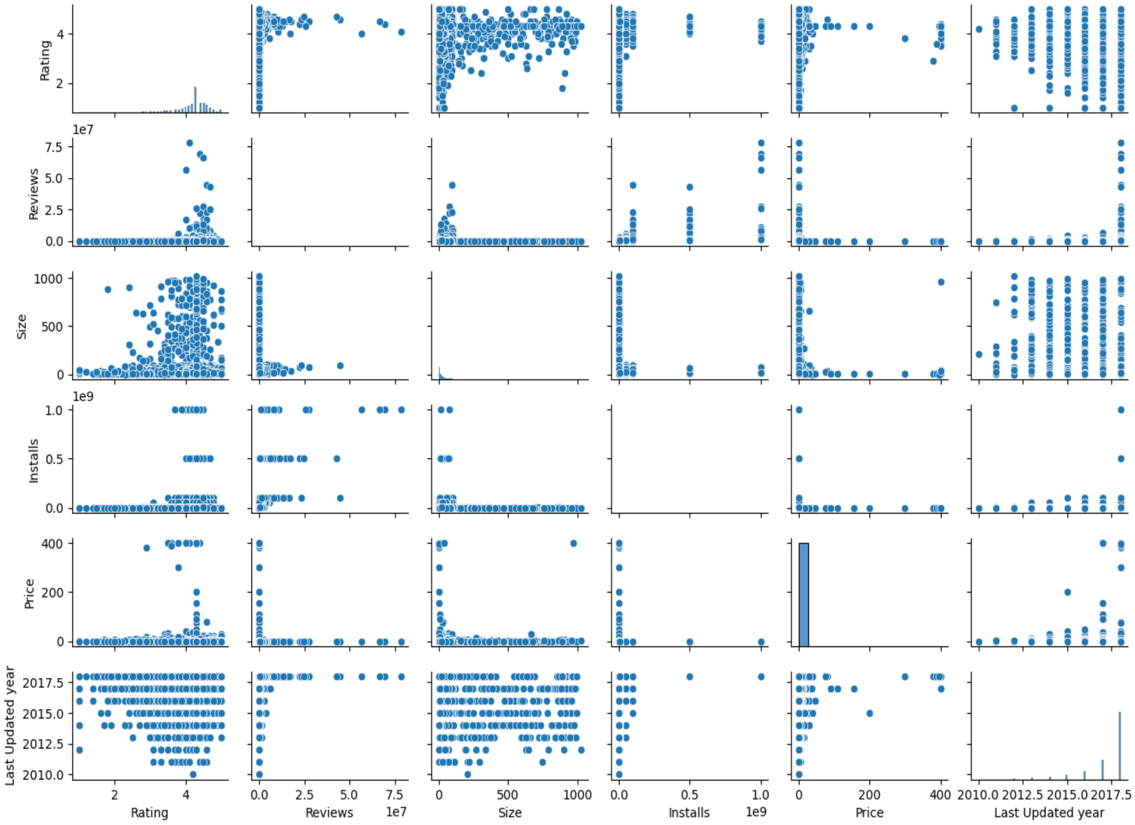
4.17 Correlation between two data frames

- Sentiment polarity shows a very weak positive correlation with sentiment subjectivity, indicating that as the polarity of sentiment (positive or negative) increases, there is a slight tendency for subjectivity (the degree of personal opinion) to also increase. However, this relationship is not very strong.

- Interestingly, sentiment polarity demonstrates a strong positive correlation with rating and price. This suggests that as the sentiment polarity becomes more positive (indicating a more positive sentiment), there is a tendency for higher ratings and higher prices.
- On the other hand, sentiment polarity exhibits a negative correlation with size and installs. This implies that as the sentiment polarity becomes more negative, there is a tendency for smaller app sizes and potentially lower numbers of installs.
- Moving on to sentiment subjectivity, it displays a strong positive correlation with price, size, and rating. This means that as the subjectivity of sentiment (the expression of personal opinion) increases, there is a tendency for higher prices, larger app sizes, and potentially higher ratings. However, the correlation between sentiment subjectivity and sentiment polarity is weak, suggesting that the subjective expression of opinion does not necessarily align strongly with the polarity of sentiment.
- Additionally, sentiment subjectivity shows a negative correlation with installs. This indicates that as the subjectivity of sentiment increases, there is a tendency for lower numbers of installs. This could imply that more subjective opinions may not resonate as strongly with app users, potentially impacting app popularity and adoption**.**



4.18 Pairplot

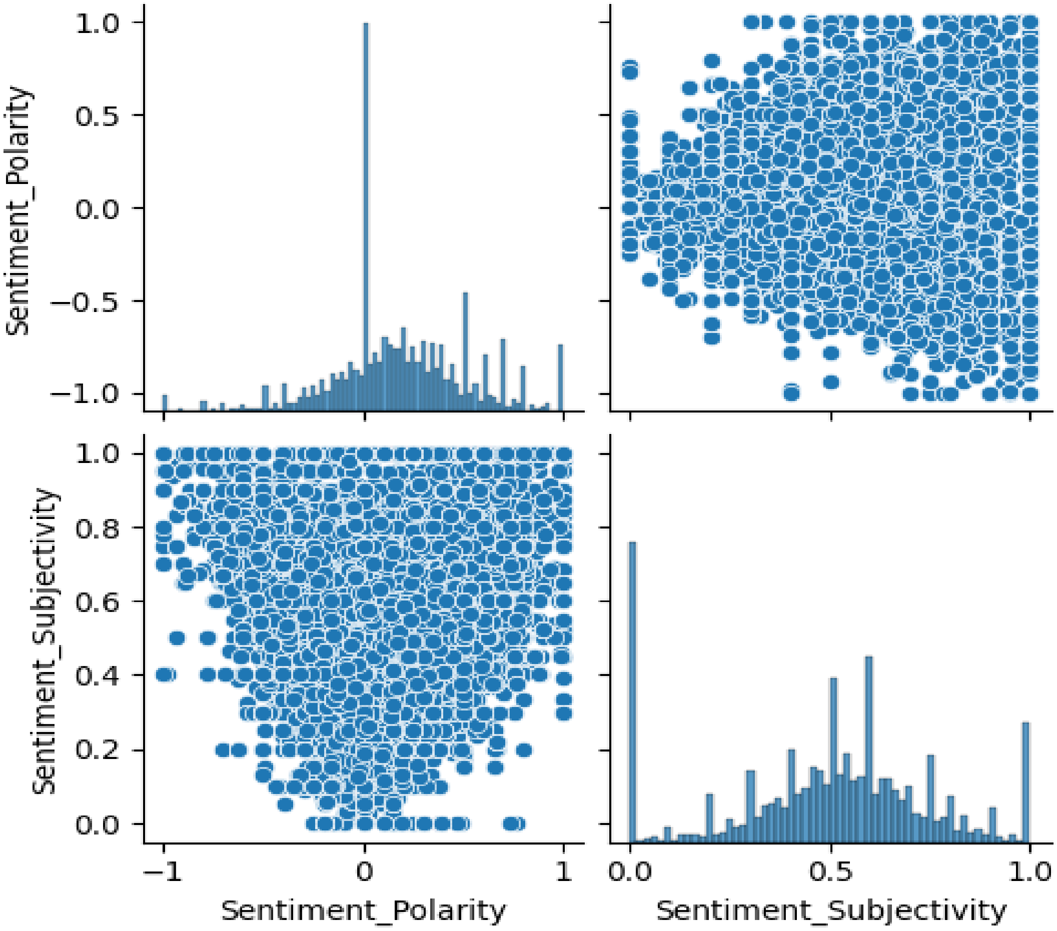


- Most of the Apps are Free.
- Most of the Paid Apps have a Rating of around 4.
- As the number of installations increases the number of reviews of the particular app also increases.
- Most of the Apps are light-weighted.

4.19 Pairplot (User reviews)

- We can see there are many outliers in Sentiment polarity and Sentiment subjectivity.

- We can observe that from rating 3 to 5, there are many Sentiment polarities and Sentiment subjectivity points.
- Free apps have a higher number of ratings and installs.
- Lesser app size has a higher number of ratings.



Project Summary and Conclusion:

This exploratory data analysis (EDA) project focuses on analyzing the Google Play Store apps data and customer reviews dataset. The goal is to derive actionable insights that can help app-making businesses succeed in the Android market.

The Google Play Store data includes information about various apps such as their names, categories, average user ratings, number of reviews, size, number of installs, pricing, content rating, genres, and more. This dataset provides a comprehensive overview of the apps available on the Play Store.

The user reviews dataset contains customer reviews for different apps. It includes the app name, the sentiment of the review (positive, neutral, or negative), sentiment polarity (numerical score indicating the review's positivity or negativity), and sentiment subjectivity (score indicating the review's objectivity or subjectivity).

By analyzing these datasets, valuable insights can be obtained. Some potential areas of analysis include:

1. App Categories: Identifying the most popular app categories can help developers understand the market demand and focus their efforts accordingly.

2. Ratings and Reviews: Examining the relationship between app ratings, the number of reviews, and sentiment can provide insights into user satisfaction and feedback. It can help developers understand the factors that contribute to positive or negative user experiences.

3. App Size and Installs: Analyzing the relationship between app size, number of installs, and user ratings can help developers optimize app sizes and understand the impact on user adoption and satisfaction.

4. Pricing Strategy: Exploring the distribution of free and paid apps, analyzing their revenue potential, and understanding the relationship between pricing, user ratings, and downloads can guide developers in formulating effective pricing strategies.

5. App Updates: Analyzing the frequency and impact of app updates on user ratings and reviews can provide insights into the importance of maintaining and improving app quality over time.

6. Sentiment Analysis: Studying sentiment polarity and subjectivity in customer reviews can help identify key strengths and weaknesses of apps and guide developers in addressing user concerns and enhancing user satisfaction.

By conducting a thorough EDA and gaining insights from the data, app developers can make informed decisions, prioritize areas for improvement, and create apps that cater to user needs and preferences.

App Distribution: The majority of apps on the Play Store (approximately 92%) are free, indicating a preference for free apps among users. This highlights the importance of considering monetization strategies such as ads or in-app purchases.

Age Restrictions: Around 81.8% of the apps in the dataset have no age restrictions, indicating a broad target audience. This provides opportunities for developers to create apps that cater to a wide range of users.

Competitive Category: The "Family" category is identified as the most competitive category, suggesting a need for developers to focus on differentiation and unique value propositions within this category.

Popular Category: The "Game" category has the highest number of app installs, highlighting the popularity of gaming apps among users. Developers can leverage this demand by creating engaging and high-quality game apps.

Top-Rated Apps: The majority of apps in the dataset are top-rated, which indicates a positive reception from users. This emphasizes the importance of delivering excellent user experiences and maintaining high-quality standards.

Category Analysis: The top three categories with the highest app count are "Family," "Game," and "Tools." Developers can consider these categories for potential business opportunities.

Genre Analysis: The top genres include "Tools," "Entertainment," "Education," "Business," and "Medical." Developers can explore these genres to target specific user needs and preferences.

Correlation: There is a strong positive correlation between the number of app reviews and app installs, suggesting that popular apps tend to have a larger user base and consequently receive more reviews.