



# Abstract

This EDA project focuses on utilizing two raw datasets from Kaggle, which contain information about the Google Play Store. The datasets include attributes of Play Store apps and user reviews. By conducting exploratory data analysis (EDA), we aim to extract valuable insights, identify trends, and anticipate future challenges.

The Google Play Store is renowned as one of the largest and most popular Android app stores, making its data a valuable resource for developing optimal models and understanding market dynamics. EDA serves as the crucial first step in any data science project, allowing us to uncover patterns and extract meaningful information through visual analysis.

Our EDA process involved cleaning the datasets and then leveraging Python libraries such as matplotlib and seaborn to create various graphs. These visualizations helped us explore the relationships between different attributes and derive insightful findings. The outcomes of our analysis can provide valuable guidance to industries seeking to capture the Android market.

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

# 1. Introduction

The Google Play Store is a widely used website and app that allows users to download and purchase various digital content such as apps, books, games, movies, and music for Android devices. With Android being the dominant mobile operating system, the Play Store has become the largest and most popular app store for Android.

Originally launched as the "Android Market" in 2008 alongside the first Android devices, its primary purpose was to distribute apps and games. Initially, it did not support paid apps and games, but as the Android platform grew, so did the Android Market. By 2012, it featured over 450,000 Android apps and games.

This article aims to explore the dynamics of the Google Play Store and demonstrate how different features from its dataset can be utilized for prediction purposes. By conducting sentiment analysis and data visualization, we can analyze customer preferences and provide developers with insights on the best app development strategies.

Understanding user download behavior across various app categories in the Google Play Store is crucial as the mobile app industry continues to expand. This knowledge helps developers stay competitive and ensures that the value of their apps in the store remains high. The Google Play Store serves as a digital distribution service, offering not only apps but also music, digital media, books, movies, and TV programs.

By analyzing and researching relevant data, developers can determine the success rate of apps and make informed decisions about adding or modifying features based on current trends and future predictions. Factors such as price, app size, user reviews, and ratings influence users' motivation to download an app. This project work aims to provide valuable insights for developers, allowing them to understand what types of apps are likely to attract more users and adapt their strategies accordingly.

# Analysis Methodology

Our analysis is structured into four distinct phases: data extraction, data cleaning, data visualization, and interpretation of results. We conclude with a summary and discuss future research directions.

In the data extraction phase, we collect the necessary data from the provided dataset, which includes the Play Store data and review data. These datasets are typically available in CSV format. Following data extraction, we proceed with the data cleaning phase. Here, our objective is to reduce errors in the dataset by addressing issues such as null values and duplicates. By performing data cleaning, we ensure that the dataset is reliable and ready for analysis.

Once the dataset is cleaned, we move on to the data visualization phase. Here, we utilize various plotting techniques and visualizations to gain insights and understand the relationships between different variables. This step helps us identify patterns, trends, and correlations within the dataset. After analyzing the dataset through visualization, we move to the interpretation of results. In this phase, we draw conclusions and make observations based on the patterns and relationships we have identified. This step helps us extract meaningful insights and answer research questions or address specific objectives.

Finally, we provide a comprehensive conclusion that summarizes the analysis results and their implications. We may also discuss future research directions and potential areas for further investigation. This allows us to provide a clear vision of the relationships and findings derived from the analysis, and offer suggestions for future work in the field.



# Google Play Store Dataset

1. **App**- It tells us about the name of the application.
2. **Category** - It tells us about the category to which an application belongs.
3. **Rating**- It tells us about the ratings given by the users for a specific application.
4. **Reviews** - It tells us about the total number of users who have given a review for the application.
5. **Size**- It tells us about the size of the application on the mobile phone.
6. **Installs** - It tells us about the total number of installs/downloads for an application.
7. **Type** - It tells us whether the application is free or a paid one.
8. **Price**- It tells us about the price of the application.
9. **Content\_Rating**- It tells us about the target audience for the application.
10. **Genres** - It tells us about the various other categories to which an application can belong.
11. **Last\_Updated** - It tells us about when the application was updated.
12. **Current\_Ver** - It tells us about the current version of the application.
13. **Android\_Ver** - It tells us about the android version which can support the application on its platform.

## Problem Definition:

The objective of this project was to perform a comprehensive analysis of the Google Play Store, addressing various problem statements to gain insights into app categories, ratings, content ratings, sentiments, genres, pricing, size, and correlations between different variables. By analyzing these aspects, the project aimed to provide valuable insights for app developers and stakeholders to make informed decisions in the highly competitive app marketplace.

**App Category Share:** The project aimed to identify the category with the highest share of apps in the market, providing an understanding of the app landscape and the most dominant categories.

**App Rating Distribution:** The distribution of app ratings was analyzed to gain insights into the satisfaction levels of users and understand the overall quality of apps available on the Google Play Store.

**Content Rating Count:** The project sought to determine the content rating category that had the highest count, shedding light on the types of apps that are most prevalent in terms of content appropriateness.

**Sentiment Count:** The analysis focused on identifying the sentiment with the highest count, providing an understanding of the predominant user sentiment towards apps.

**Genre Analysis:** The project aimed to identify the genres with the highest number of apps, offering insights into the popular types of apps available on the Google Play Store.

**Percentage of Free and Paid Apps:** The analysis aimed to determine the percentage of free and paid apps in the market, helping stakeholders understand the pricing dynamics and revenue models prevalent in the app ecosystem.

**Frequency Distribution:** The project aimed to perform frequency distribution analysis on rating, price, sentiment polarity, and sentiment subjectivity, revealing patterns and distributions within these variables.

**App Size and its Impact:** The relationship between app size, ratings, and number of installs was explored to understand how the size of an app affects its popularity and user perception.

**Correlation between Sentiment Polarity and Subjectivity:** The project aimed to determine the correlation between sentiment polarity and sentiment subjectivity, providing insights into the relationship between the emotional tone and subjectivity of user reviews.

**App Category and Price Relation:** The analysis focused on examining the relationship between app category and app price, offering insights into pricing strategies and dynamics across different categories.

**Distribution of Ratings and Price:** The distribution of ratings and price across the dataset was explored, providing an understanding of the overall range and variation in ratings and app pricing.

**Correlation between Ratings and Price:** The project aimed to establish the correlation between app ratings and prices, determining whether higher-priced apps tend to have higher ratings or vice versa.

**Correlation with Year of Last Update and Rating:** The correlation between the year of the last update and the app ratings was examined, helping understand if the recency of updates has an impact on user ratings.

**Correlation between Size, Reviews, Installs, and Price:** The analysis aimed to explore the correlations between app size, reviews, number of installs, and price, providing insights into the interplay of these variables and their impact on app performance.

**Correlation Heatmap:** A correlation heatmap was constructed using the merged dataset, presenting an overview of the correlations between various attributes, helping identify significant relationships within the data.

## 2. Data Cleaning and Preparation

Data cleaning and preprocessing play a crucial role in data analysis. They involve transforming raw data into a more comprehensible, valuable, and optimized format.

## 2.1 Why is Data Cleaning and Preprocessing required?

Data often contains noise in the form of outliers, null values, or data points that deviate significantly from the overall trend. Prior to conducting exploratory data analysis, it is essential to address and treat these noisy data points.

## 2.2 Available Dataset:

Before proceeding with any analysis, it is necessary to explore and clean the raw dataset of Play Store Data and User Reviews. This initial step is crucial to make the data usable and suitable for exploratory data analysis.

## 2.3 Steps Performed:

In **Step 1**, we created a function called "getinfo()" to gather information about the datasets. This function displays various attributes for each column, including the data type, total number of records, count of null values, count of non-null values, percentage of null values, and the number of unique values. We applied this function to both the Play Store dataset and the User Reviews dataset to understand their characteristics.

In **Step 2**, we noticed that the "Rating" column had a significant number of null values, accounting for approximately 15% of the total records. Instead of dropping these null values, we decided to replace them with the mode of the Rating column to maintain data integrity.

Moving on to **Step 3**, we encountered special characters such as "$" and "+" in the "Installs" column, and "M" and "K" in the "Size" column (representing MB and KB, respectively). To ensure consistency, we removed these special characters using a for loop. Additionally, we converted KB values to MB by implementing a function. We also dropped any remaining rows with null values, as they were very few and had a negligible impact on the dataset.

In **Step 4**, we observed that the "Reviews" column, despite being a numerical indicator, was of the "object" data type. To rectify this, we converted the data type of the "Reviews" column to "int" using the "astype(int)" function.

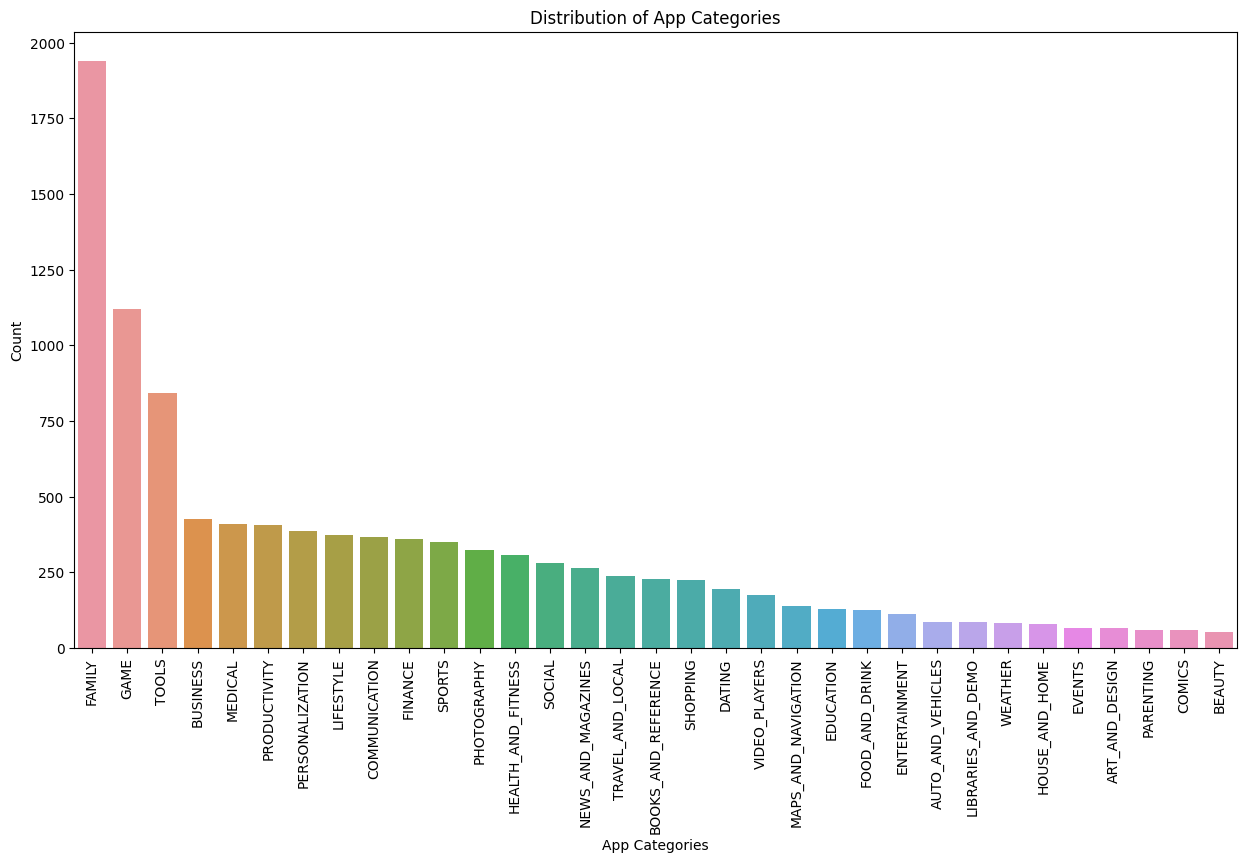
Finally, in **Step 5**, we utilized the "getinfo()" function again and confirmed that there were no remaining null values. However, we did identify duplicate records in both datasets. To address this issue, we eliminated the duplicates from both the Play Store dataset and the User Reviews dataset.

## 3. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a critical initial step in any Data Science project. Once the raw data has been cleaned, EDA involves examining the dataset to uncover patterns, gain insights, and address future challenges. It encompasses the process of analyzing the dataset to identify trends, outliers, and forming hypotheses based on our understanding of the data.

During EDA, statistical measures are created for numerical data, correlations between numerical values are explored, and various graphical representations and plots are generated using visual methods. These techniques help in gaining a concise and improved understanding of the data. In this article, we will demonstrate EDA using a Google Play Store dataset. By leveraging Python and its essential libraries, we aim to derive meaningful insights and extract valuable information from the dataset.

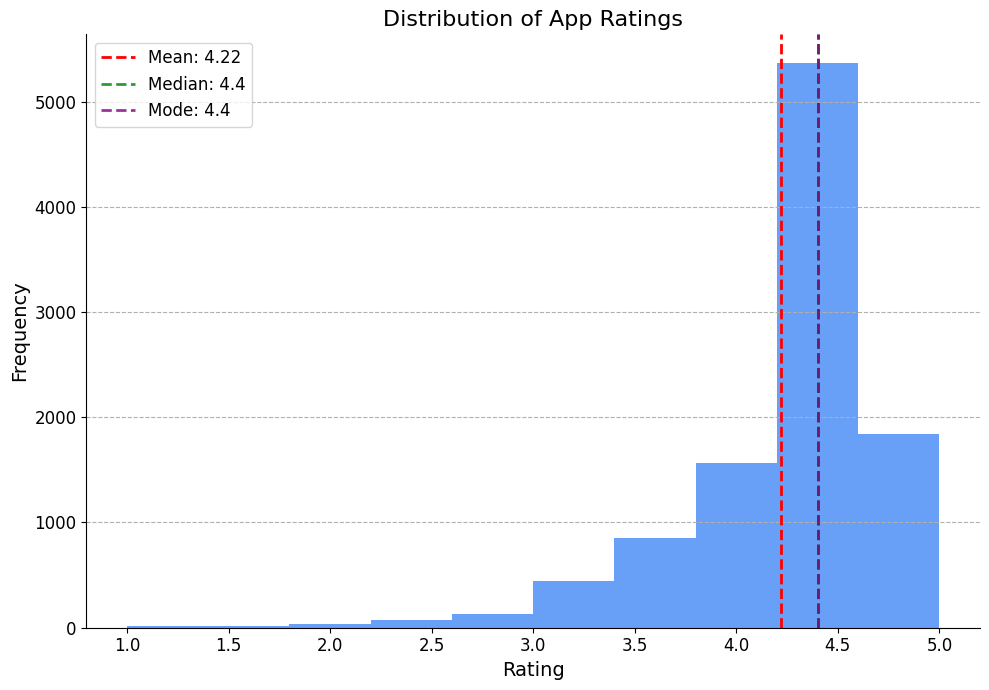
## 3.1 Top Categories in play store

Based on our analysis, we have observed that apps on the Google Play Store are categorized into different categories. Through our analysis, we have determined that the "Family" category has the highest number of applications compared to other categories.

From the bar graph we plotted, it is evident that out of the 33 categories, the "Family" category stands out as the most prominent, while the category with the least number of applications is "Beauty". This conclusion is drawn based on the data and the visualization of the bar graph.

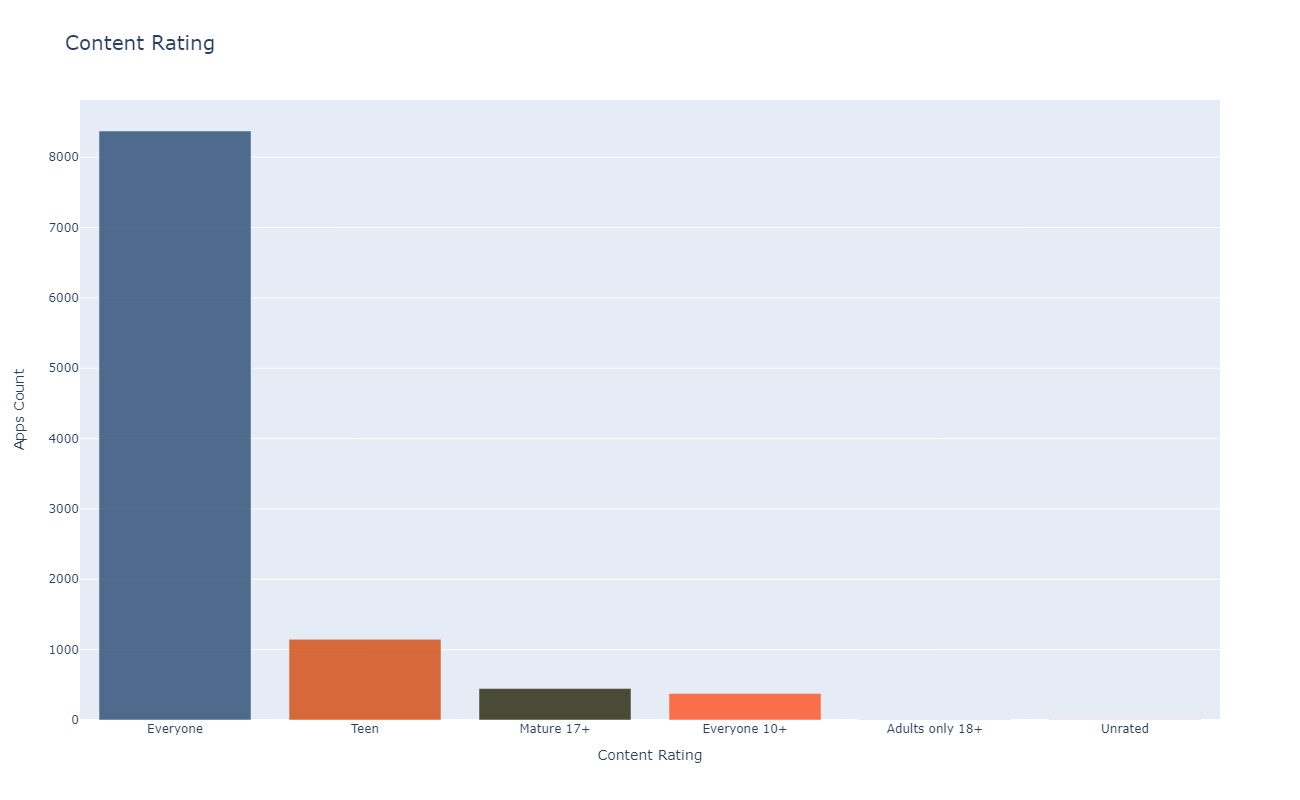
## 3.2 What Is Distribution Of Apps Rating?

From our observation, we found that the average volume of ratings across all app categories is 4.217. In the histogram plot majority of the apps are highly rated with only few exceptions in the low-rated apps.



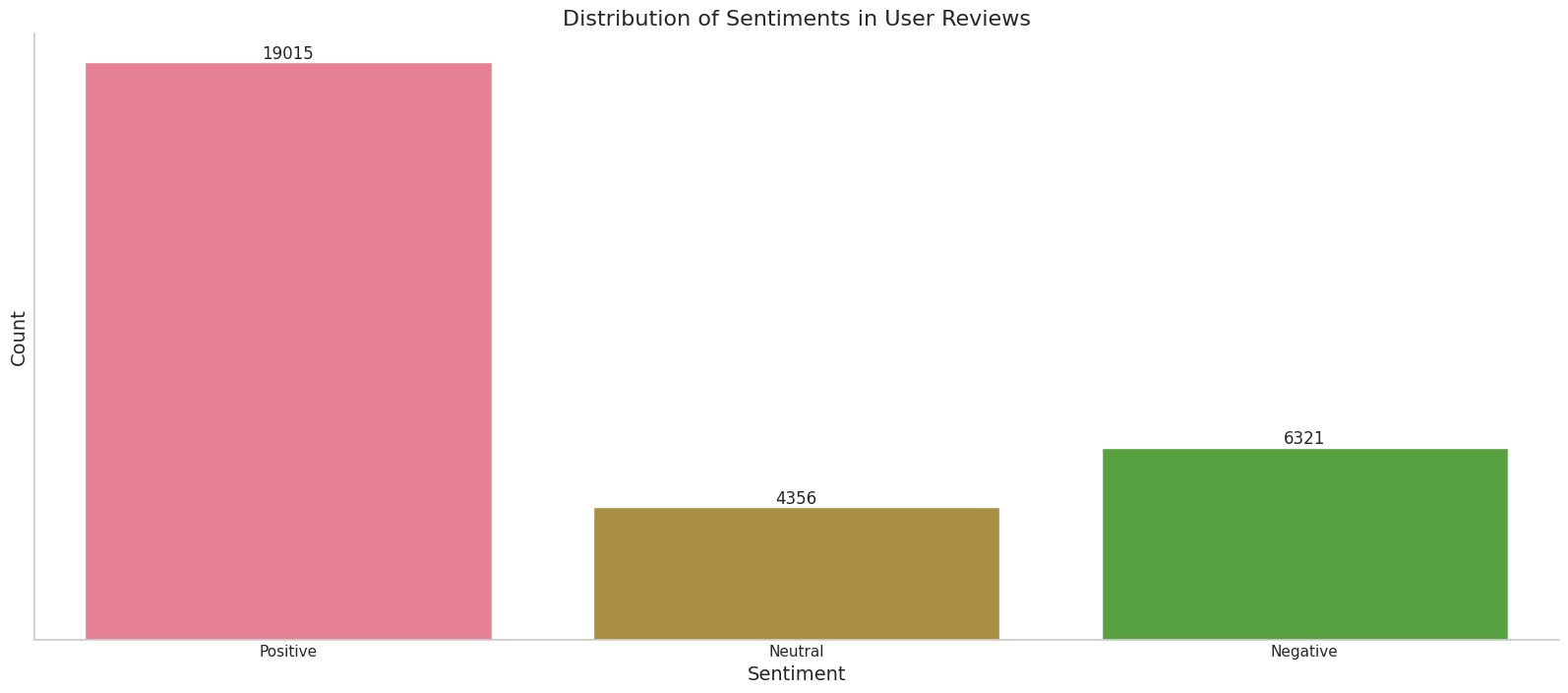
## 3.3 Which Content Rating Has Highest Count?

The bar plot displaying the app count by content rating reveals that the "Everyone" rating has the highest count, followed by "Teen" and "Mature 17+". This suggests that a significant number of apps in the Google Play Store cater to diverse audiences, reflecting the growing trend of mobile device adoption across various age groups. This insight indicates a broad appeal and potential positive business impact for app developers targeting a wide range of users.



## 3.4 Which Sentiment Has Highest Count?

Based on the presented pie chart, it can be observed that the majority of apps on the Google Play Store have received positive reviews, accounting for approximately 64% of the total. Additionally, a smaller proportion of apps have received negative reviews, representing approximately 15% of the total.



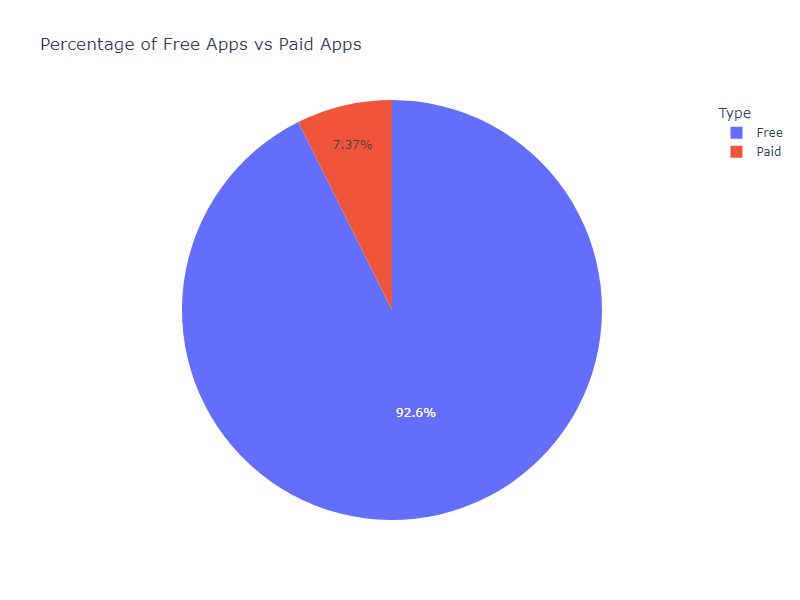
## 3.5 Which Genres have most number of apps?

The "Tools" category exhibits intense competition with the highest number of apps, suggesting that developers should prioritize creating distinctive and innovative features to differentiate themselves. Moreover, the significant presence of apps in the "Entertainment" and "Education" categories implies a demand for apps in these areas, presenting developers with potential growth prospects if they invest in these categories.

### 

## 3.6 What Is Percentage Of Free And Paid Apps?

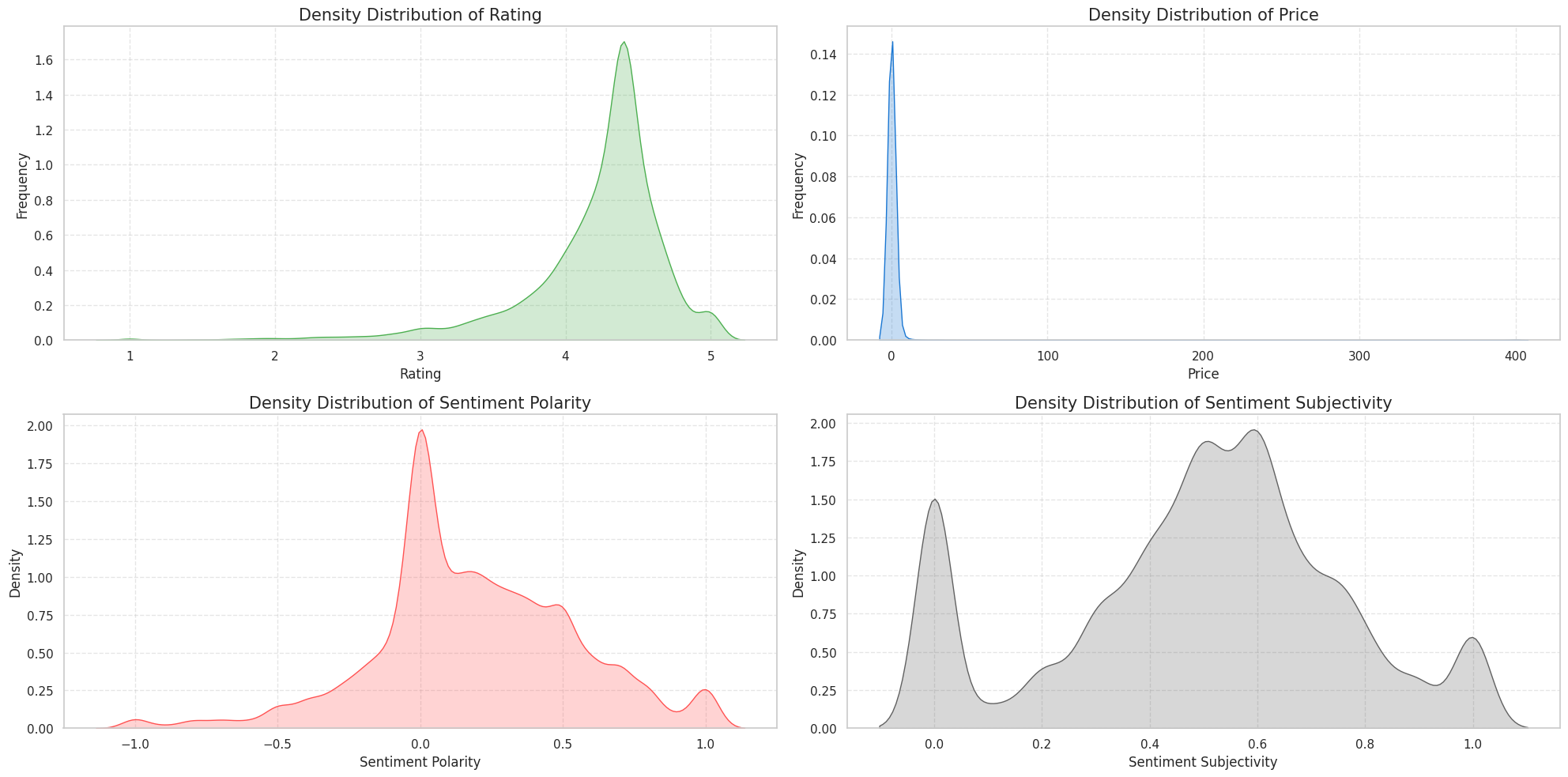
Based on the pie chart analysis, it can be inferred that the majority of apps on the Google Play Store are free, constituting approximately 92% of the total.



## 3.7 Checking Frequency Distribution Of Rating, Price, Sentiment\_Polarity and Sentiment\_subjectivity.

From the above density distribution plots we can say that:

* Average rating of given dataset is 4.2178813 on play store.
* The Mean price of our dataset is 1.0315609 which means that the majority of the apps are free (0$) on google play store.
* The polarity score lies in the range of[-1,1]. Anything below a score of -0.05 we tag as negative and anything above 0.05 we tag as positive Sentiment score. Here we can see from our calculations and the Sentiment Polarity Desnsity distribution graph that the Mean Sentiment Polarity Score is 0.18886801 which is resembles a good average sentiment score(Majority of the users liking the apps).
* The subjectivity is a float within the range [0.0, 1.0] where 0.0 is very objective and 1.0 is very subjective. As per our analysis and plotted graph the Mean Sentiment Subjectivity Score is 0.49093045. That means around 50% users are sharing personal opinion while others 50% are just sharing the factual information in reviews.



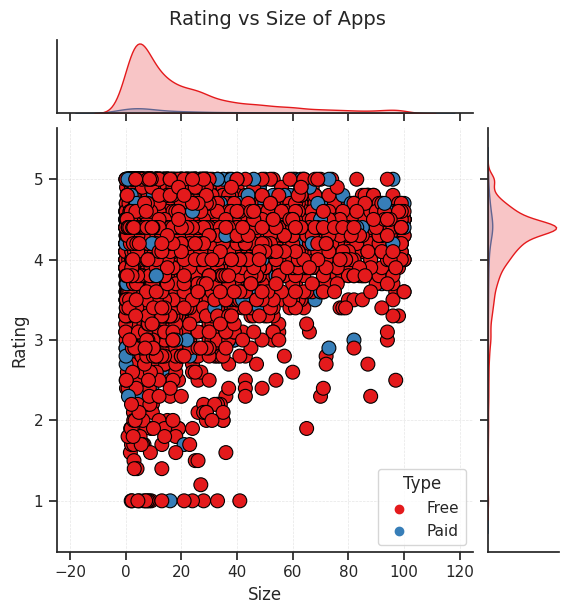
## 3.8 How does the size of an app affect the ratings and number of installs

The analysis reveals that a majority of users prefer free apps with smaller sizes, as indicated by the limited presence of large-sized paid apps with high ratings. This suggests a preference for apps that are both free and compact.

An intriguing finding is that despite the scarcity of large-sized apps, those that do exist tend to have favorable ratings. This phenomenon may be attributed to their zero cost, which potentially offsets the impact of their larger size.

Furthermore, it is noteworthy that certain paid apps exhibit lower ratings. This suggests a weak correlation between price and ratings, indicating that users' assessments of value and satisfaction are not solely determined by the app's cost.

Overall, these insights emphasize the importance of considering app size and pricing strategies in addition to ratings to meet user expectations and optimize user experience.

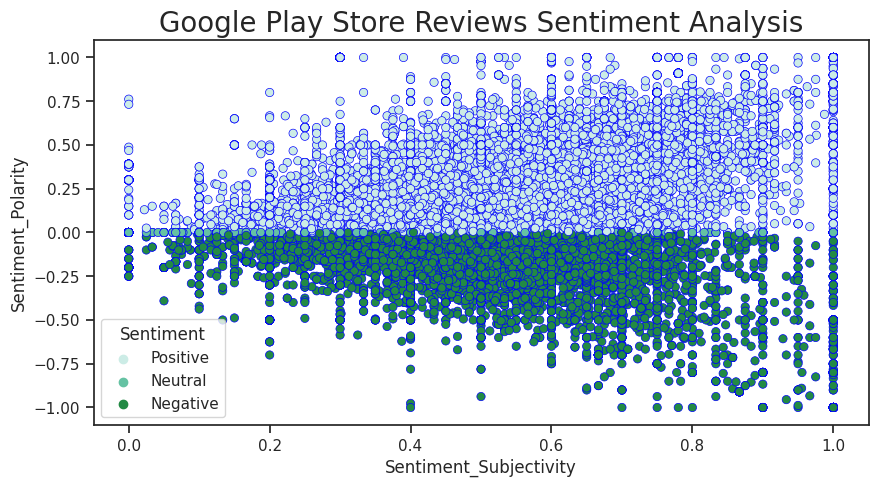


## 3.9 Correlation between Sentiment\_Polarity and Sentiment\_Subjectivity.

The analysis of the scatter plot reveals that the sentiment subjectivity of the reviews predominantly falls within the range of 0.5 to 0.8. This indicates that users tend to provide reviews based on opinions and personal experiences rather than solely factual information.

Furthermore, the sentiment subjectivity values are widely distributed between -0.5 and 0.75. This suggests that polarity (sentiment polarity) is not consistently proportional to sentiment subjectivity. However, it is observed that in a majority of cases, there is a proportional relationship between the two variables.

These findings highlight the nuanced nature of user sentiments expressed in the Google Play Store reviews. The predominant presence of subjective opinions and experiences, along with the varying relationship between sentiment subjectivity and polarity, emphasizes the importance of considering both subjective and objective aspects when analyzing user feedback.

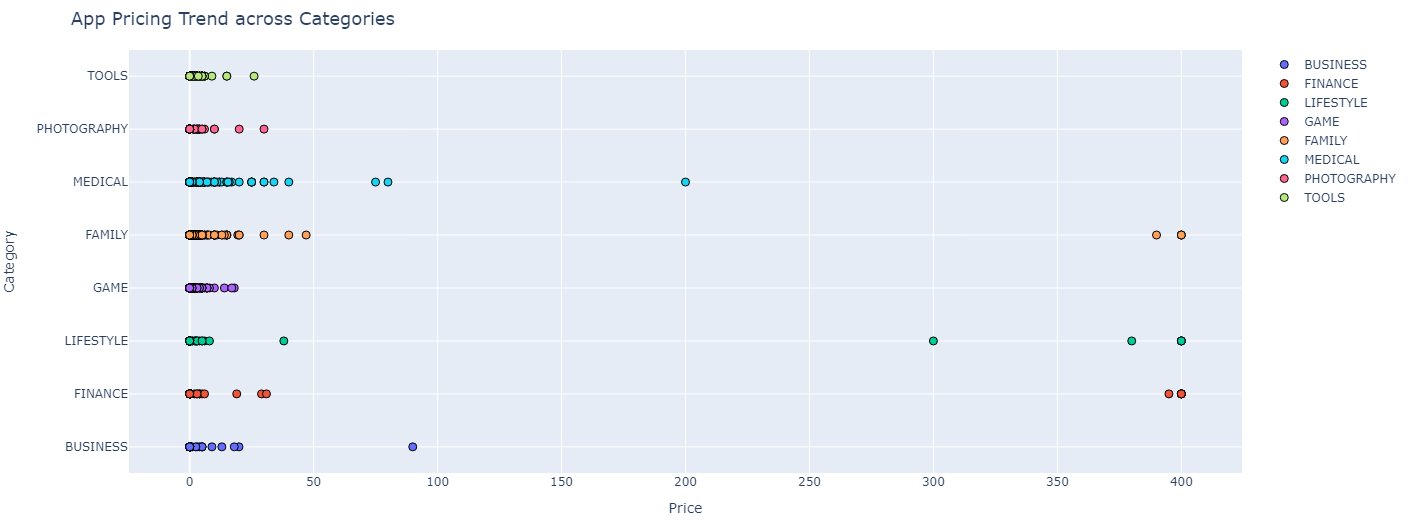


## 3.10 what is the Relation between app category and app price

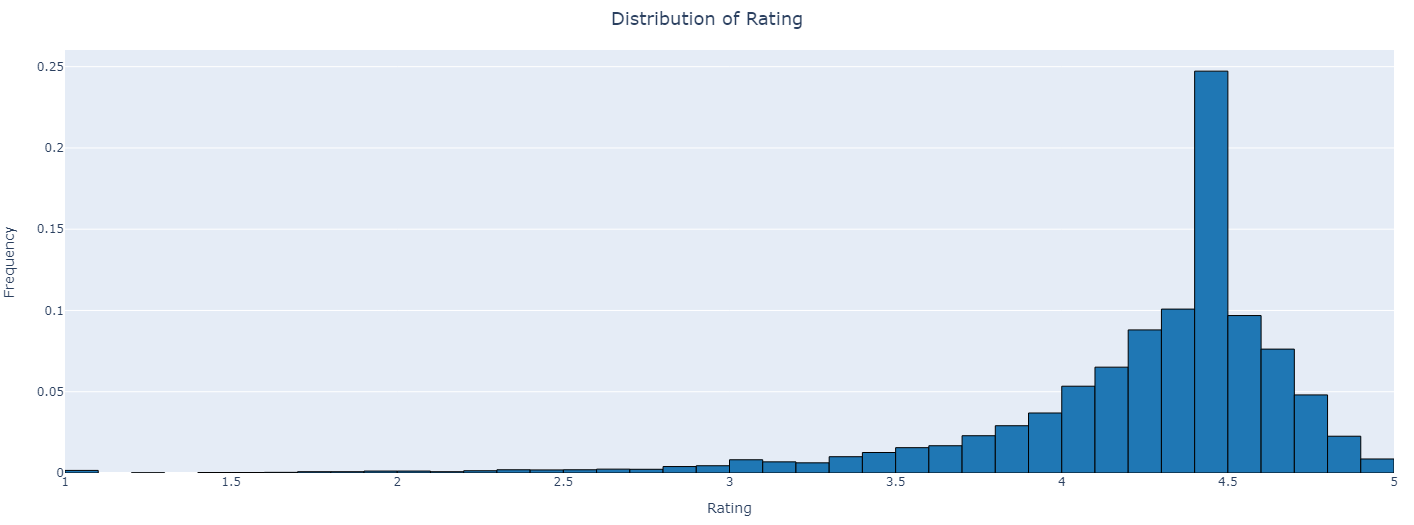
When determining the appropriate pricing strategy for a mobile app, various factors come into play. It is crucial to reassess the app's price before entering the market to ensure its worthiness.

Analyzing the data, we observe that different app categories exhibit varying price ranges. While simpler apps are often available for free, categories such as FAMILY, LIFESTYLE, FINANCE, and MEDICAL tend to have higher price points.

Interestingly, game apps generally have lower prices. This could explain their higher download numbers, as we have previously observed.

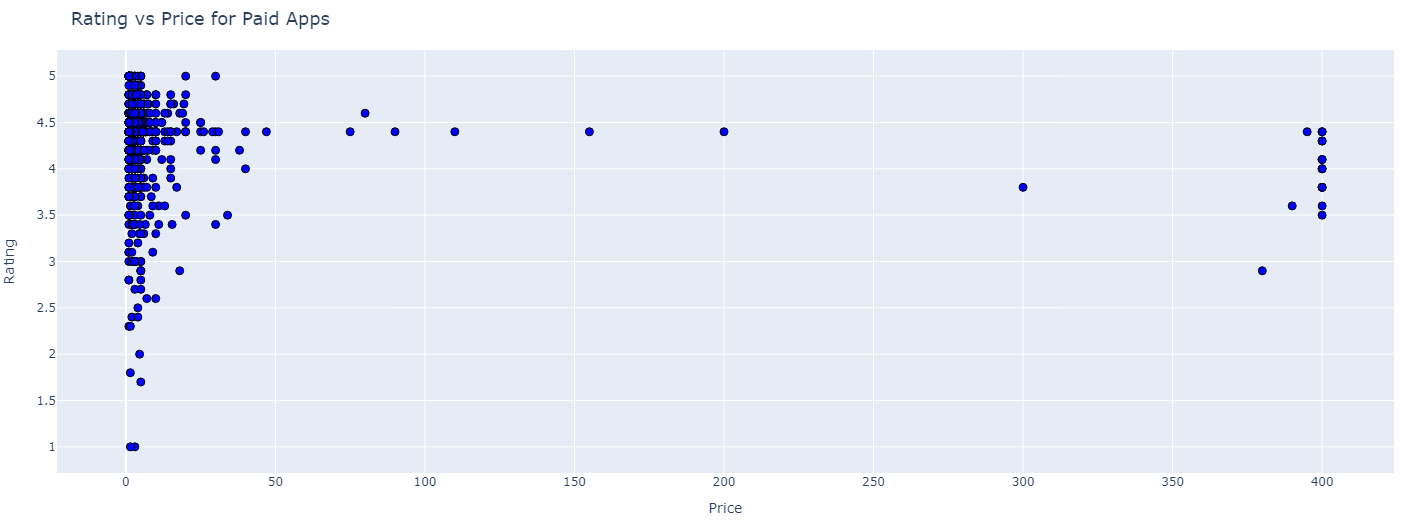


## 3.11 Below is the distribution of Ratings and price over the dataframe.



The analysis of the histogram reveals a predominantly positive average rating for apps, typically exceeding 4. This rating is found to be interconnected with the app's price and size, as observed earlier. It is noteworthy that a minor peak occurs at a rating of 1, indicating a small number of poorly rated apps that consequently experience lower installation numbers.

## 3.12 Now, let's have a look at correlation between Ratings and price.

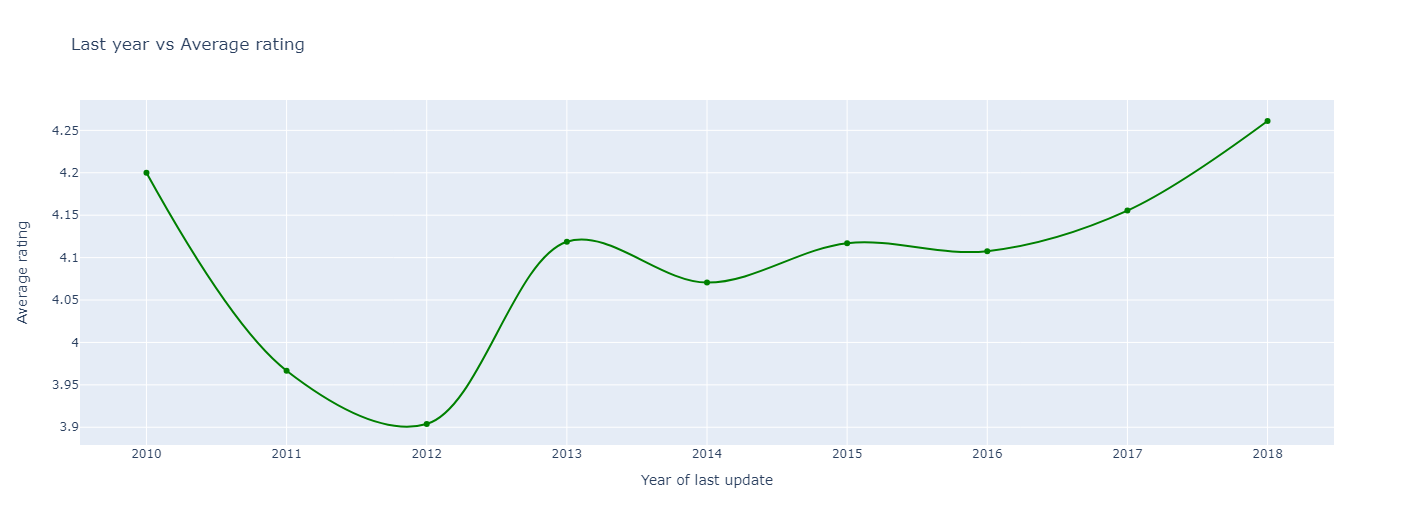


The analysis of the scatter plot reveals that there is a higher number of apps with lower prices, and these apps generally have good ratings. On the other hand, apps with higher prices are relatively fewer in number but tend to have ratings above 3.5. This indicates that these apps, particularly in the FAMILY and HEALTH categories, offer valuable services to users.

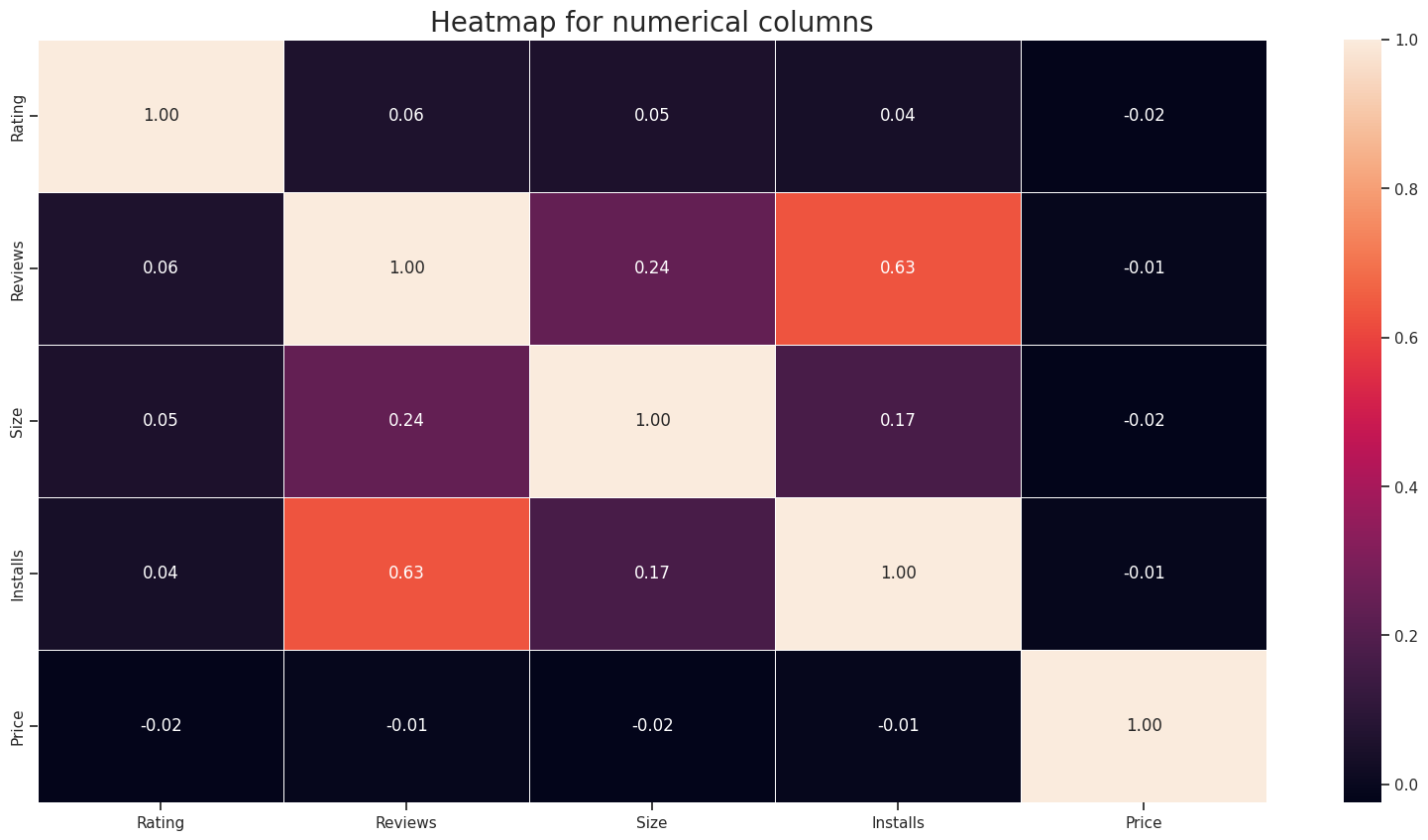
## 3.13 Correlation between Year of last Update and Rating.

The analysis of the line plot reveals a gradual increase in the average app rating from 2016 onwards. This indicates that apps with regular updates tend to receive higher average ratings, suggesting a positive correlation between update frequency and user satisfaction.

Interestingly, the graph shows a decline in average rating between 2010 and 2012. This could be attributed to factors such as poor user experience, lack of regular updates, or other reasons that led to users giving lower ratings during that period.



## 3.14 How Size, Reviews, Installs and Price of apps are correlated.



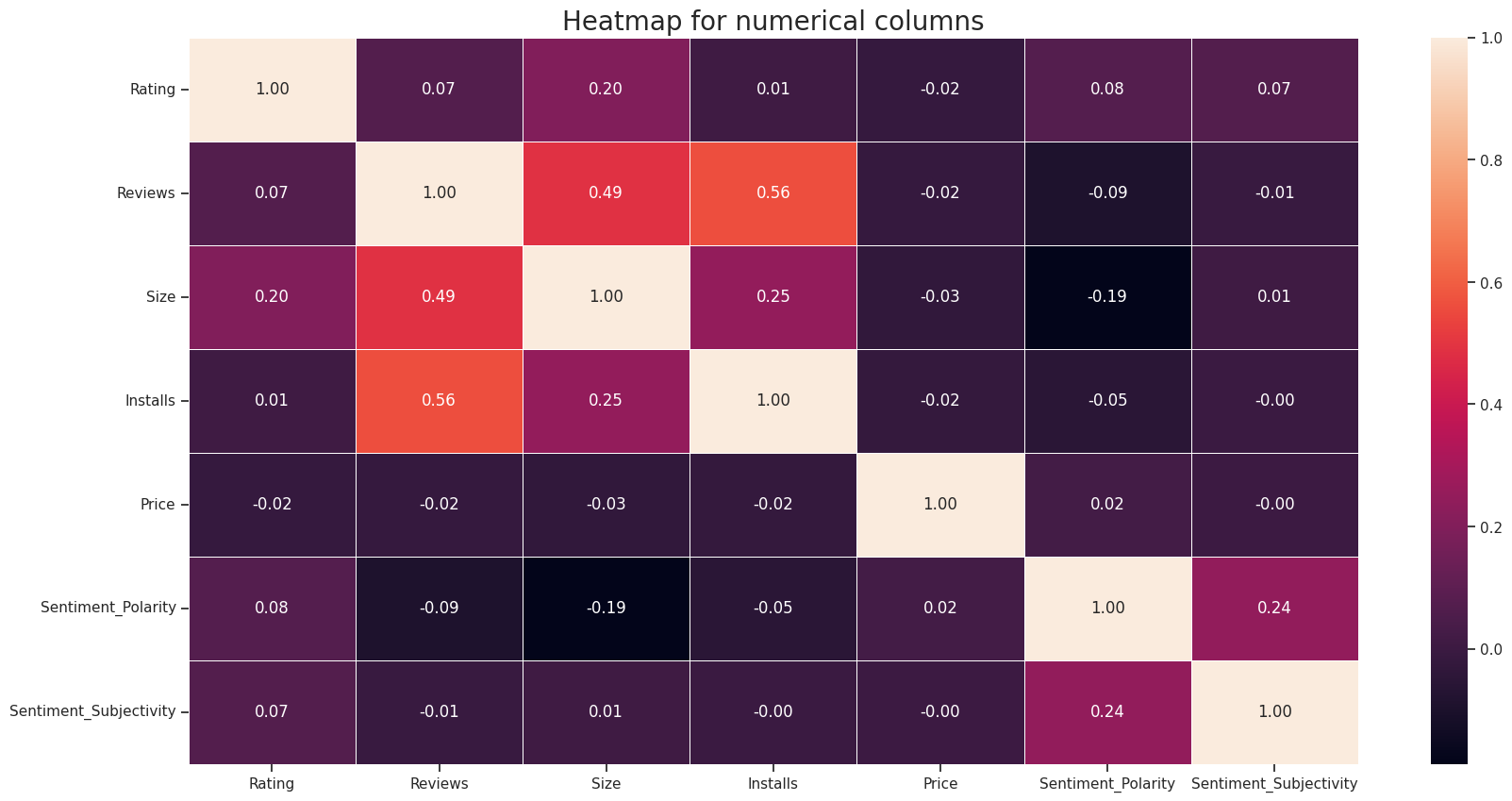
The heatmap analysis reveals several correlations within the Playstore data. Firstly, there is a positive correlation of 0.63 between the "Reviews" and "Installs" columns. This suggests that a higher number of reviews is associated with a greater number of app downloads, indicating that customer behavior is influenced by reviews.

Additionally, the "Price" column shows a slightly negative correlation with the "Rating", "Reviews", and "Installs" columns. This implies that as the price of an app increases, the average rating, number of reviews, and number of installs tend to decrease. This highlights the importance of pricing strategies in attracting users and maximizing app performance.

Furthermore, the "Rating" column exhibits a slight positive correlation with the "Installs" and "Reviews" columns. This indicates that higher app ratings are associated with increased downloads and reviews. Therefore, improving the rating and review count can potentially enhance the market share of an app.

To capture the market more rapidly, it is advisable to initially launch the app at a lower price and gradually increase it over time. This strategy can help attract more users and gain a larger market share.

## 3.15 Correlation Heatmap on merged dataframe.



The analysis of the heatmap reveals a negative correlation of -0.19 between the "Size" and "Sentiment Polarity" columns. This suggests that as the size of an app increases, there is a tendency for people to express more negative sentiment in their reviews. This could be attributed to factors such as increased storage consumption, higher RAM usage, and the need for a faster internet connection, which may lead to user dissatisfaction.

Furthermore, there is a positive correlation of 0.56 between the "Reviews" and "Installs" columns. This indicates that as the number of reviews increases, more people take notice of the app and are more likely to install it. This underscores the impact of reviews on user perception and app popularity.

Additionally, a slight positive correlation of 0.24 is observed between the "Sentiment Polarity" and "Sentiment Subjectivity" columns. This suggests that when users share positive reviews (higher sentiment polarity), there is a greater likelihood that they are expressing their personal opinions rather than providing purely factual information (higher sentiment subjectivity).

## Conclusions:

* There is a positive correlation between the number of reviews and the number of installs, meaning that as the number of installs increases, it is more likely that users will leave reviews. Conversely, there is a negative correlation between price and rating, indicating that higher-priced apps tend to have lower ratings.
* Art and design have the most number of installs.
* Developing apps in the Family and Lifestyle categories can be a strategic choice for maximizing profitability and generating high revenue.
* Approximately 61% of people expressed positive sentiments, while a relatively low percentage of around 15% reacted negatively. The majority of the remaining respondents had a neutral sentiment.
* Out of the total number of apps, approximately 92.12% are free, while around 7.81% are paid, indicating a significant majority of apps being offered for free compared to paid alternatives.
* The "Everyone" content rating, which encompasses all age groups, has the highest percentage of apps, accounting for approximately 81.80% of the total.
* The Family, Game, and Tools categories have the highest number of apps, indicating that these categories contain a significant portion of the total app offerings.
* Game category presents a promising opportunity for developers as it demonstrates a high demand with a large number of installs, suggesting that it is an area with potential for growth and where competition may not be as saturated.
* Users have a strong inclination towards downloading apps from genres such as Tools, Entertainment, Education, Business, and Medical, indicating a high level of interest and demand for apps in these categories.
* The average rating of apps available on the Play Store is a commendable 4.17, indicating a generally positive reception from users. When it comes to purchasing apps, users tend to show a preference for lightweight applications, indicating a willingness to pay for apps that have a smaller file size or consume less device storage.
* Paid apps with larger file sizes may face challenges in the market and may not achieve optimal performance.The presence of a large number of user reviews positively influences the download frequency of a given app, indicating that users are more likely to download an app that has received a substantial number of reviews.
* Users tend to provide more critical reviews for paid apps.
* There is a positive relationship between the number of installs and the app rating.
* Developing an app that achieves a high rating requires regularly updating it with the latest version while ensuring it remains optimally sized.
* Developing a free app with a content rating suitable for all age groups is a favorable approach.

## References:

1. GeeksforGeeks
2. Analytics Vidhya
3. AlmaBetter Class material
4. Pandas and Numpy libraries
5. Stack overflow
6. YouTube
7. Researchgate.net
8. W3schools.com