Statistics R Project

Google Play App Store Data Analysis

A Report Submitted to
SVKM's
Narsee Monjee Institute of Management Studies (NMIMS)

As a part of Semester III
Of
Bachelor of Science Applied Mathematical Computing
By

Diyanshi Shah (A010), SAP ID: 86022200019

Under the Guidance of

Prof. Priyanka Dangar

Introduction

In an era dominated by the proliferation of mobile applications, understanding the dynamics of the app ecosystem is paramount for developers, analysts, and stakeholders alike. The Google Play Store, a cornerstone of the Android platform, stands as a colossal repository of diverse applications catering to a multitude of interests and needs. The dataset under investigation encapsulates a wealth of information, presenting a comprehensive snapshot of these digital offerings. Comprising essential attributes such as app names, categories, ratings, reviews, and more, this dataset provides a unique opportunity to delve into the intricacies of the mobile app landscape.

The objective of this project is to conduct a thorough analysis of the Google Play Store dataset, employing a multifaceted approach to extract actionable insights. By leveraging data-driven methodologies, we aim to uncover trends, correlations, and key performance indicators that can inform app developers, marketers, and decision-makers in their pursuit of app excellence.

As we navigate through this analysis, it is imperative to bear in mind that the observations and conclusions drawn herein are predicated on the available dataset. The findings presented should be viewed as a reflection of the dataset's characteristics during the period of data collection. Moreover, the interpretations provided are subject to the limitations and assumptions inherent in any data-driven endeavor.

About The Dataset

https://www.kaggle.com/code/mdp1990/google-play-app-store
The Google Play Store dataset is a comprehensive collection of information about mobile applications available on the Android platform. It encompasses a wide range of attributes that provide valuable insights for app developers, analysts, and researchers. A brief overview of the dataset's columns:

- App: This column contains the names of the mobile applications available on the Google Play Store. Each row corresponds to a unique app.
- Category: Indicates the category or genre to which the app belongs. This helps users and developers quickly identify the app's primary purpose or niche.
- Rating: Represents the user ratings given to the app. Ratings typically range from 1 to 5, with higher values indicating better user satisfaction.
- Reviews: This column provides the total number of user reviews for the app. It's an important metric for assessing an app's popularity and user engagement.
- Size: Indicates the storage space required by the app, often represented in megabytes (MB) or with labels like "Varies with device".
- Installs: Specifies the total number of times the app has been downloaded and installed by users.
- Type: Describes whether the app is available for free or comes with a price tag (Paid).
- Price: Denotes the cost of the app, if it is a paid application. It is given in the respective currency
- Content Rating: Provides information about the target audience or age group for which the app is deemed suitable.

- Genres: Offers additional categorization of the app based on its characteristics or features. This can provide more specific information than the main category.
- Last Updated: Indicates the date when the app was last updated in the Google Play Store
- Current Ver: Represents the version number of the app that is currently available for download.
- Android Ver: Specifies the minimum Android version required for the app to run smoothly on a user's device.

Data Cleaning

Using the unique function, we extracted the unique data in the Category and the checked for null values using 'is.na'. Further the "Reviews" column was changed to numeric data type for analysis. Later the data was sorted with respect to "Reviews" and "Apps". This ensured the integrity and reliability of our dataset which is paramount to producing accurate and meaningful insights with addressing any anomalies, eliminating any missing values, or discrepancies that may be present.

Data Exploration

In this section, we're going to give you a snapshot of some important numbers from our Google Play Store data. These numbers help us understand things like how popular apps are, how they're priced, and more. Imagine it like a quick preview before we dive deeper into the details. We'll look at things like averages, which tell us what's typical, and other numbers that show how spread out the information is.
These basic stats help us get a sense of the big picture, and they're like our starting point for the more detailed investigations we'll do later.

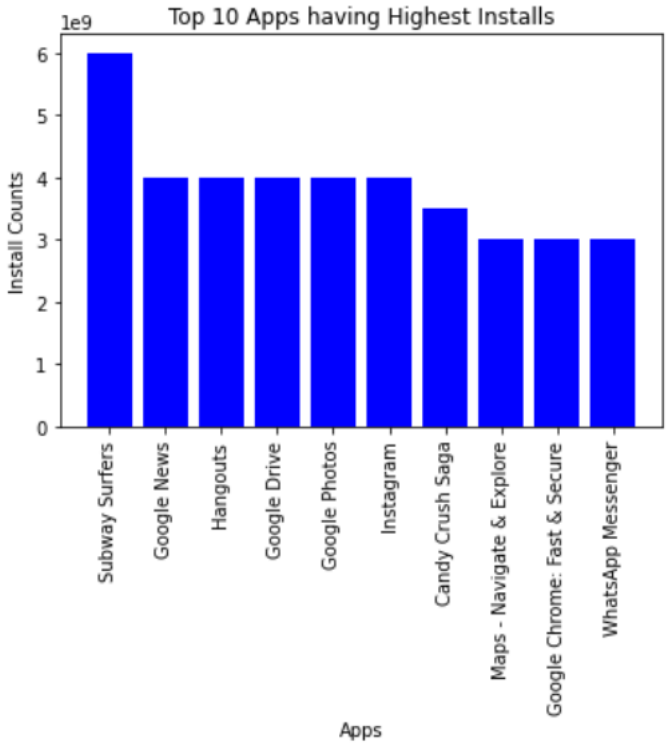
1. App with large number of reviews
   Using the sort function, we determine the app with the largest number of reviews

   - Facebook

2. Count of Paid v/s Free Apps
   Using the sum function under the Type column, we extract the number of Free and Paid apps each.

   - Free Apps 10039
   - Paid Apps 800

3. App with the largest number of installs.



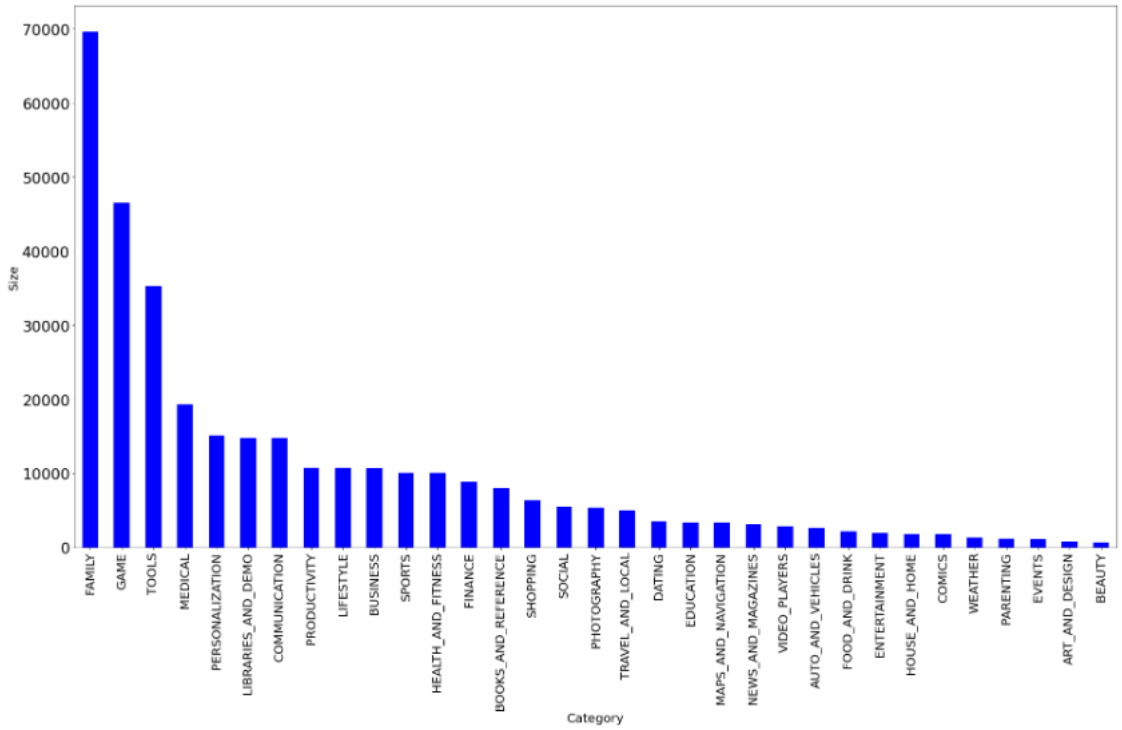
According to the Bar Chart **"Subway Surfers"** has the highest number of installs. We do this using the barplot command in R.

4. App with the largest size

To find this, we first remove rows where Size is 'Varies with device, then remove 'M', 'k', and '+' from 'Size' column and convert to numeric, Sort the data by 'Size' in descending order and finally Get the app with the largest size which according to the process is

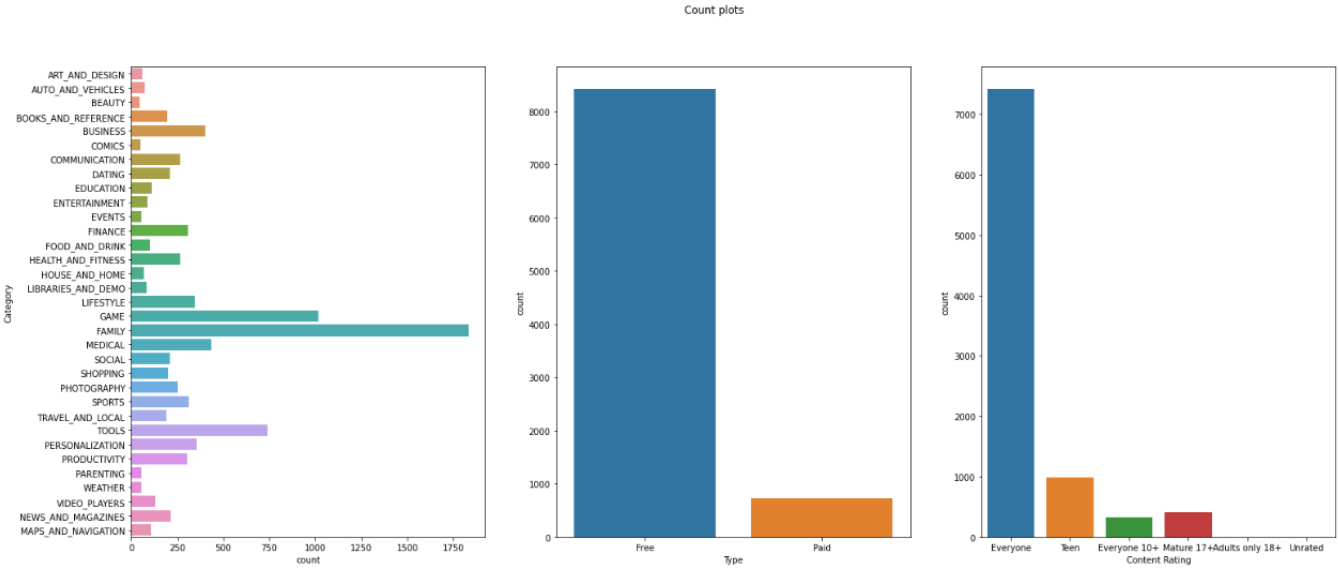
- Word Search Tab 1 FR

5. Most Popular Category

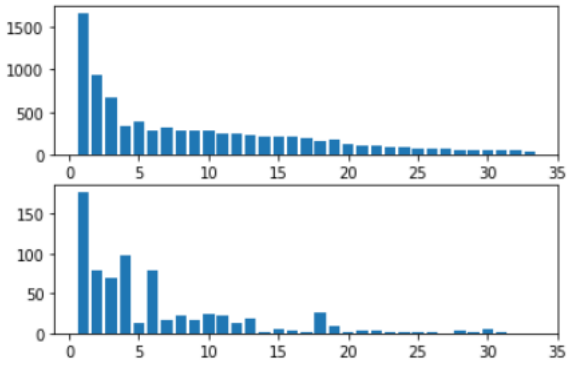


According to the graph “Family” is the most popular category.

Data Visualization



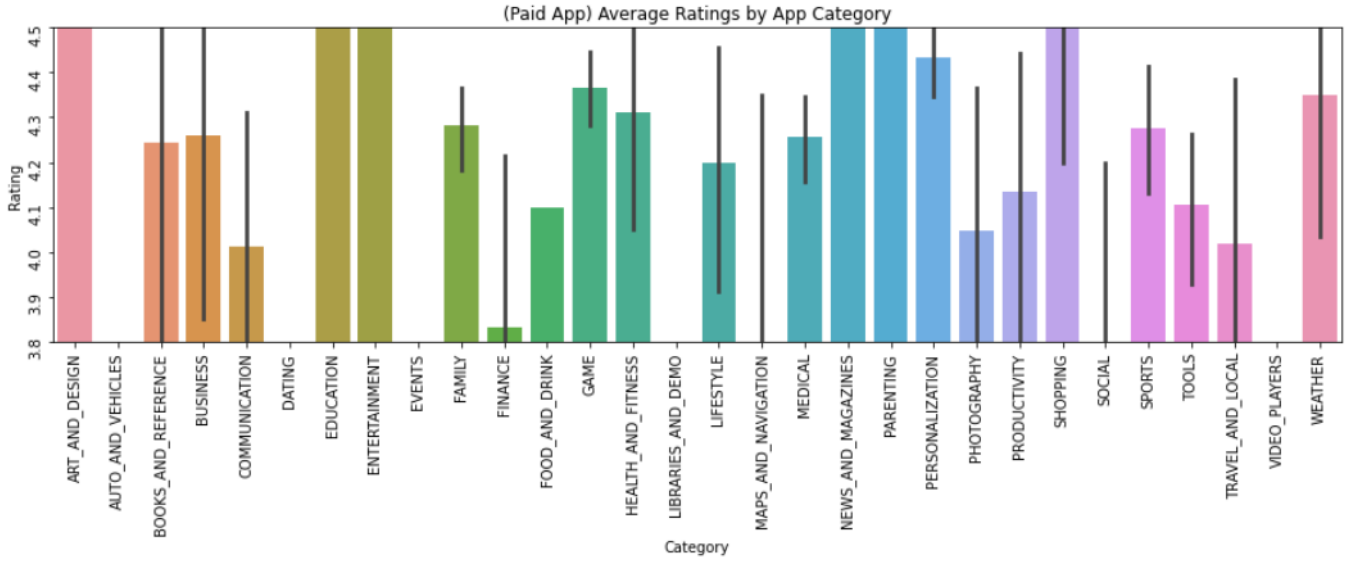
Insights from the graph: Using Count plot to summarize and visualize the data.

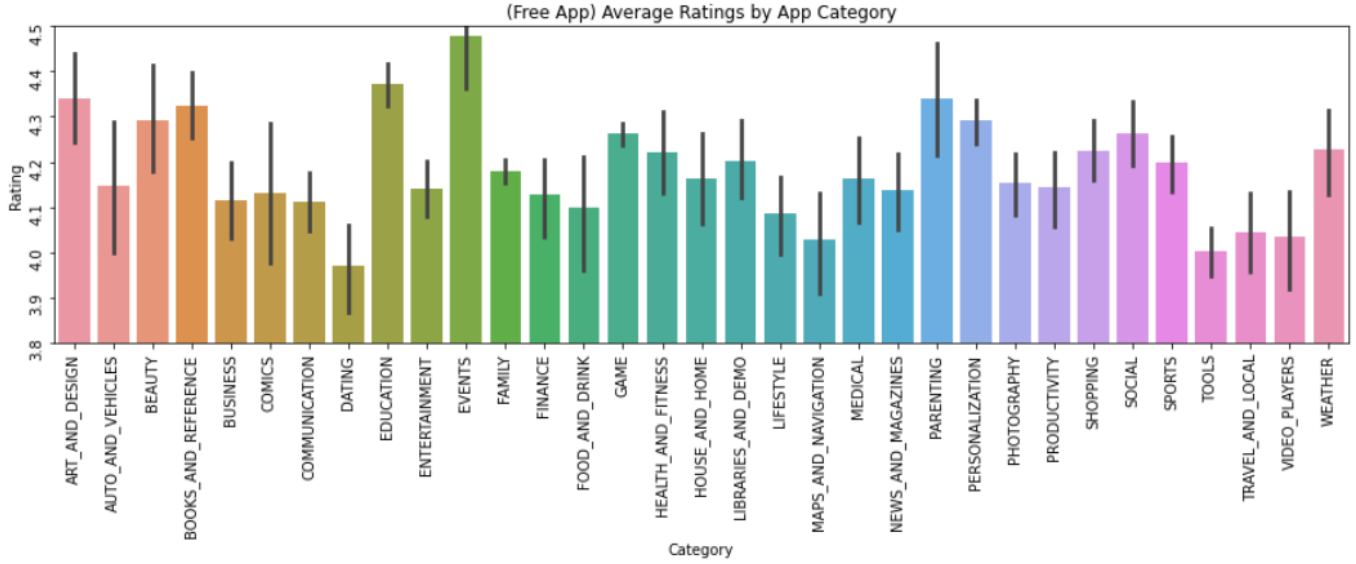


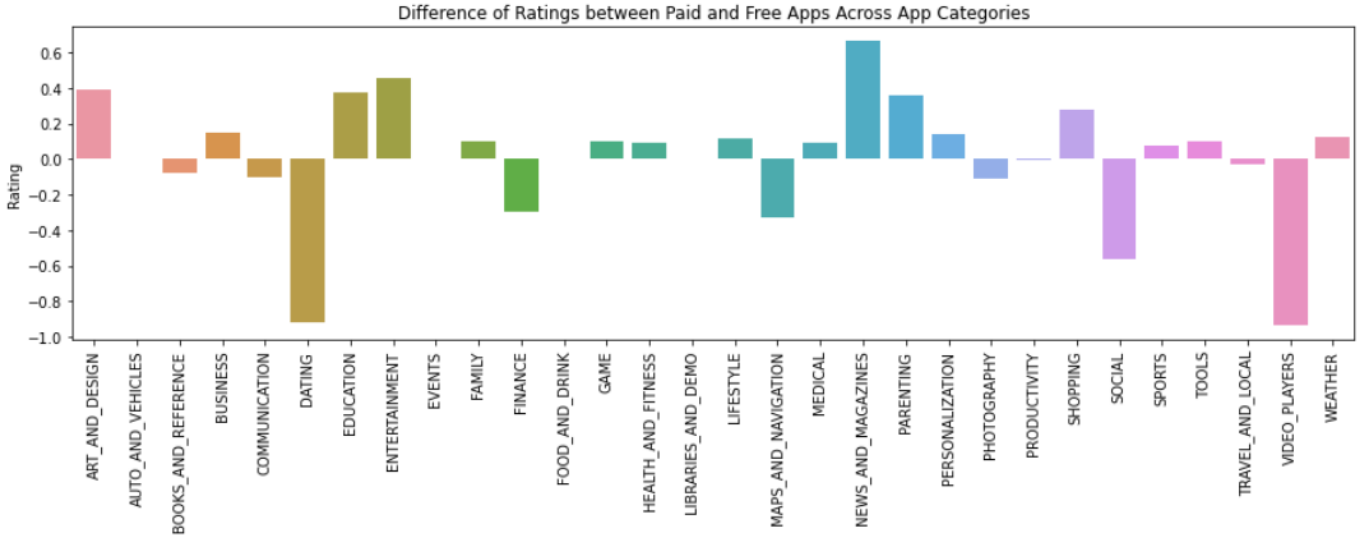
The above graph describes the Rating dataframe.

- Rating Distribution based on Free and Paid Apps

We use the subplots function to summarize the above statement




(Free App) Average Ratings by App Category
Rating
3.8
3.9
4.0
4.1
4.2
4.3
4.4
4.5
ART_AND_DESIGN
AUTO_AND_VEHICLES
BEAUTY
BOOKS_AND_REFERENCE
BUSINESS
COMICS
COMMUNICATION
DATING
EDUCATION
ENTERTAINMENT
EVENTS
FAMILY
FINANCE
FOOD_AND_DRINK
GAME
HEALTH_AND_FITNESS
HOUSE_AND_HOME
LIBRARIES_AND_DEMO
LIFESTYLE
MAPS_AND_NAVIGATION
MEDICAL
NEWS_AND_MAGAZINES
PARENTING
PERSONALIZATION
PHOTOGRAPHY
PRODUCTIVITY
SHOPPING
SOCIAL
SPORTS
TOOLS
TRAVEL_AND_LOCAL
VIDEO_PLAYERS
WEATHER
Category


Difference of Ratings between Paid and Free Apps Across App Categories
Rating
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
ART_AND_DESIGN
AUTO_AND_VEHICLES
BOOKS_AND_REFERENCE
BUSINESS
COMMUNICATION
DATING
EDUCATION
ENTERTAINMENT
EVENTS
FAMILY
FINANCE
FOOD_AND_DRINK
GAME
HEALTH_AND_FITNESS
LIBRARIES_AND_DEMO
LIFESTYLE
MAPS_AND_NAVIGATION
MEDICAL
NEWS_AND_MAGAZINES
PARENTING
PERSONALIZATION
PHOTOGRAPHY
PRODUCTIVITY
SHOPPING
SOCIAL
SPORTS
TOOLS
TRAVEL_AND_LOCAL
VIDEO_PLAYERS
WEATHER

Hypothesis Testing

Objective 1: The average user ratings of paid apps are higher than the average user ratings of free apps.

Null Hypothesis (H0): The average user ratings of paid apps are the same as the average user ratings of free apps.
- H0: μ_paid = μ_free

Alternative Hypothesis (H1): The average user ratings of paid apps are higher than the average user ratings of free apps.
- H1: μ_paid > μ_free 3.

Significance Level: 0.05

Using t – test while checking if p_value < alpha

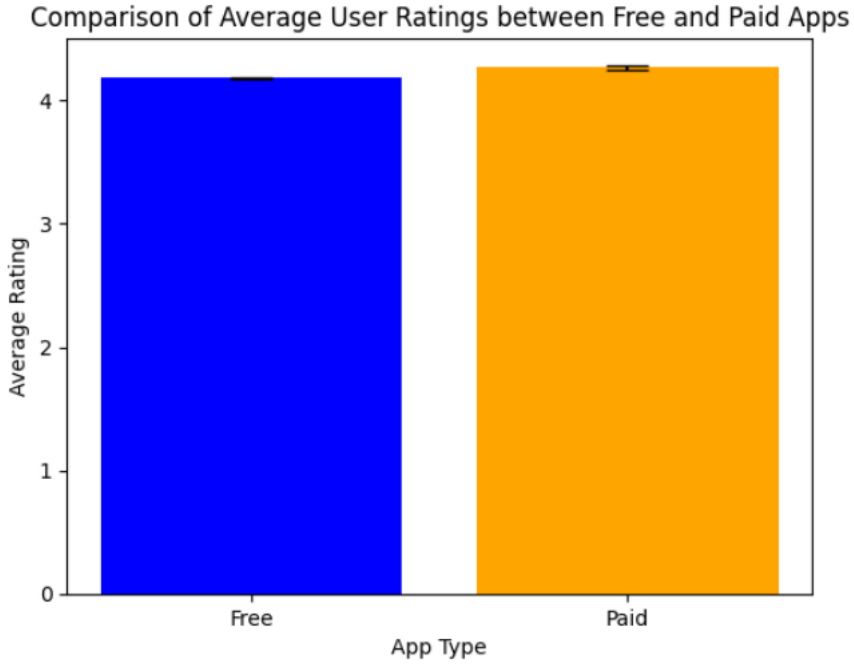
T-statistic: -3.8332
p-value: 0.0001

The first test is a two-sample t-test, which compares the means of two groups (Free and Paid apps) to determine if they are significantly different.

The second test is the Mann-Whitney U test, which is a non-parametric test for comparing two independent samples. It is used when the assumptions of the t-test are not met.

Interpretation: The results indicate a statistically significant difference in average ratings between paid and free apps. This suggests that paid apps tend to have higher average ratings compared to free apps.

Conclusion: Based on the results, we reject the null hypothesis at the 5% significance level, indicating a significant difference in average ratings between paid and free apps. This implies that users tend to rate paid apps higher than free apps.



Conclusion:

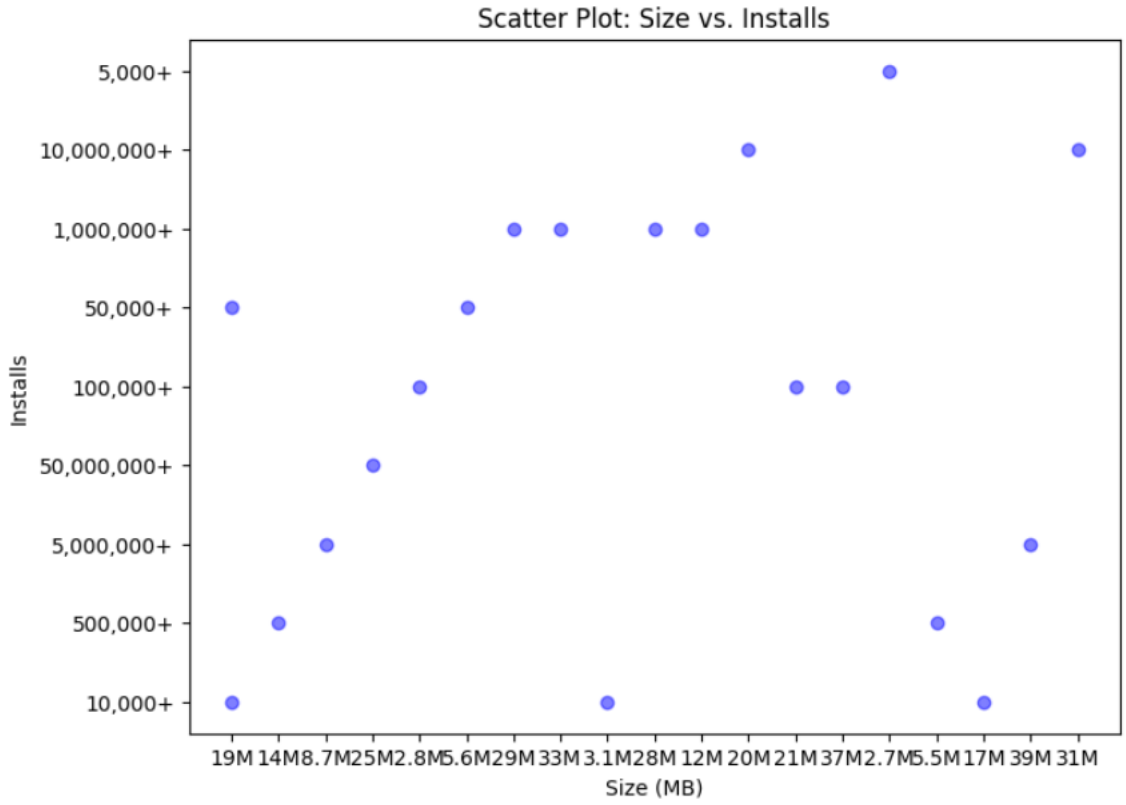
Reject null hypothesis. The average user ratings of paid apps are higher than free apps.

Objective 2: The size of an app (in MB) is correlated with the number of installations.

Null Hypothesis (H0): There is no correlation between the size of an app and the number of installations.
- H0: $\rho = 0$ (where ρ is the correlation coefficient)

Alternative Hypothesis (H1): There is a correlation between the size of an app and the number of installations.
- H1: $\rho \neq 0$ 3.

Significance Level: 0.05



Conclusion:

After conducting the Pearson's correlation coefficient test, we obtained a correlation coefficient. Since the p-value is less than the chosen significance level (α=0.05), we reject the null hypothesis.

Objective 3: Checking installations of different categories of apps.

Null Hypothesis (H0): The mean installations of "Productivity" app category is equal to the mean installations of "Lifestyle" app category.

H0: $\mu_{productivity} = \mu_{lifestyle}$

Alternative Hypothesis (H1): The mean installations of "Productivity" app category is greater than the mean installations of "Lifestyle" app category.

H1: $\mu_{productivity} > \mu_{lifestyle}$

Conclusion:

The p-value is 4.79e-08 which is much less than the chosen level of significance (0.05)
Hence we reject H0.

Frequency Table:

```
Reviews Frequency Table:
0         596
1         272
2         214
3         175
4         137
         ...
342912      1
4272        1
5517        1
4057        1
398307      1
Name: Reviews, Length: 6002
```

'Reviews' quantifies user feedback and interaction with apps. The table demonstrates that 398307 unique review counts exist, with ‘0’ being the most prevalent.

	Var1	Freq
1	Free	10039
2	Paid	800

'Type' denotes whether an app is offered for free or as a paid download. The frequency table highlights that 10039 apps are free, while 800 apps require payment.

Conclusion & Findings

In this comprehensive analysis of the Google Play Store dataset, we embarked on a journey to unravel key insight. Through meticulous examination of various attributes, we uncovered valuable information that can guide decision-making and provide a deeper understanding of user preferences.

1. App Categories and Popularity:
   - The frequency table for app categories highlighted the diverse range of classifications available. We observed that **FAMILY** emerged as the most prevalent, indicating a significant user interest in this category.

2. User Ratings and Preferences:
   - User ratings revealed a generally positive sentiment towards apps, with an average rating of **4.19**. This suggests a high level of user satisfaction within the Google Play Store ecosystem.

3. Content Rating and Audience Appeal:
   - The frequency table for content ratings illustrated a broad spectrum of content suitability for different age groups. Notably, **'EVERYONE'** emerged as the most prevalent, indicating a widespread appeal.

4. App Versions and Compatibility:
   - The analysis of app versions and Android compatibility demonstrated a diverse environment catering to various Android versions. `4.1 and up` emerged as the most prevalent, indicating a focus on compatibility with this version.