CAPSTONE PROJECT

PLAY STORE APP REVIEW ANALYSIS

BY

ROHIT PAUL

EMAIL – rohitpaul09@gmail.com

GITHUB LINK - https://github.com/rohitpaulBA/Play-Store-Review-Data-Analysis-wih-Python

Abstract

The Google Play Store is a highly competitive market, with thousands of new apps being released every day. Developers face stiff competition from all over the world, and with most Play Store apps being free, the revenue model is complex and obscure. This makes it difficult to determine how in-app purchases, ads, and subscriptions contribute to an app's success. As a result, an app's success is typically measured by the number of installs and the user reviews it has received over its lifetime, rather than the revenue it generates. App ratings are valuable feedback provided voluntarily by users, and they serve as important evaluation criteria for apps. However, these ratings can be biased due to insufficient or missing votes. Additionally, there are often significant differences between numeric ratings and user reviews.

This study aims to analyse the Play Store apps dataset using Python to uncover key factors for app engagement and success. The findings will provide valuable insights to help optimize app performance in the Android market. The study will utilize Python libraries like NumPy and Pandas for data wrangling, and Seaborn and Matplotlib for data visualizations.

Keywords: App performance optimization, App store optimization (ASO), User engagement, App ratings and reviews, Data analysis

Problem Statement

The Play Store apps data has enormous potential to drive appmaking businesses to success. Actionable insights can be drawn for developers to work on and capture the Android market. Each app (row) has values for category, rating, size, and more. Another dataset contains customer reviews of the android apps. Explore and analyse the data to discover key factors responsible for app engagement and success.

Business Objective

Analyse the Play Store apps dataset using Python, uncovering key factors for app engagement and success, and provide valuable insights to optimize app performance in the Android market.

Datasets Overview

Descriptions for Play Store Dataset

App: The application's name and a brief description.

Category: The app's assigned category.

Rating: The average user rating.

Reviews: The total number of user reviews.

Size: The space the app occupies on a mobile phone.

Installs: The overall installations or downloads.

Type: Indicates whether the app is free or paid.

Price: The installation cost. For free apps, the price is zero.

Content Rating: Specifies if the app is suitable for all age groups.

Genres: Various categories to which an app can belong.

Last Updated: The date of the app's last update.

Current Ver: The app's current version.

Android Ver: The Android version supporting the app.

The Play Store dataset has 10,841 rows and 13 columns, with 483 instances of duplicated rows. The dataset has missing values in columns such as 'Rating' (13.60% null values), 'Type' (0.01% null values), 'Content Rating' (0.01% null values), 'Current Ver' (0.07% null values), and 'Android Ver' (0.03% null values).

Descriptions for User Reviews Dataset

App: The app's name with a brief description.

Translated_Review: English translation of the user's review.

Sentiment: The reviewer’s attitude categorized as 'Positive', 'Negative', or 'Neutral'.

Sentiment_Polarity: The review's polarity, ranging from 1 (Negative) to 1 (Positive).

Sentiment_Subjectivity: The score indicates the degree to which a reviewer’s opinion aligns with the general public’s opinion, with a range of [0, 1]. Higher scores suggest opinions closer to the general public, while lower scores indicate more factual information in the review.

The User Reviews dataset has 64,295 rows and 5 columns, with 33,616 instances of duplicated rows. The dataset has missing values in columns such as 'Translated_Review' (41.79% null value), 'Sentiment' (41.78% null value), 'Sentiment_Polarity' (41.78% null values), and 'Sentiment_Subjectivity' (41.78% null values).

Data Cleaning and Preprocessing

Data Cleaning and Preprocessing are essential steps in refining datasets for analysis. Data Cleaning involves identifying and correcting errors, handling missing or inaccurate data, and ensuring overall data quality. It aims to enhance the reliability of the dataset. Data Preprocessing focuses on transforming raw data into a format suitable for analysis. This includes tasks like normalization, scaling, and handling outliers to ensure optimal data presentation for meaningful insights.

The datasets underwent several specific actions to ensure they were analysis-ready:

Identifying Non-Numeric Reviews:
Checked and printed rows with nonnumeric characters in the 'Reviews' column.

Removing Irrelevant Row:
Dropped the row at index 10472 as it contained incorrect or irrelevant data, ensuring dataset integrity.

Converting Reviews to Integer:
Converted the 'Reviews' column to the integer data type for numerical analysis.

Converting Last Updated to Datetime:
Converted the 'Last Updated' column to datetime format for temporal analysis.

Handling Price Values:
Created a function (drop_dollar) to drop the '$' symbol and convert the 'Price' column to the float data type.

Handling Installs Values:
Created a function (drop_plus) to drop the '+' symbol and convert the 'Installs' column to the integer data type.

Converting Size Entries:
Created a function (kb_to_mb) to convert size entries to MB and handle 'k' or 'M' units.

Verifying Data Types:
Checked and printed the updated data type information after the type conversion.

Removing Duplicates:
Removed duplicate rows from both the Play Store and User Reviews datasets.

Handling Missing Values:
Filled missing values for numerical columns with the median and categorical columns with the mode. Checked and printed the updated number of missing values in both datasets.

Handling Outliers:
Visualized outliers through box plots for Reviews and Installs. Removed outliers from the data based on the quantile range (5% to 95%) for Reviews and Installs.

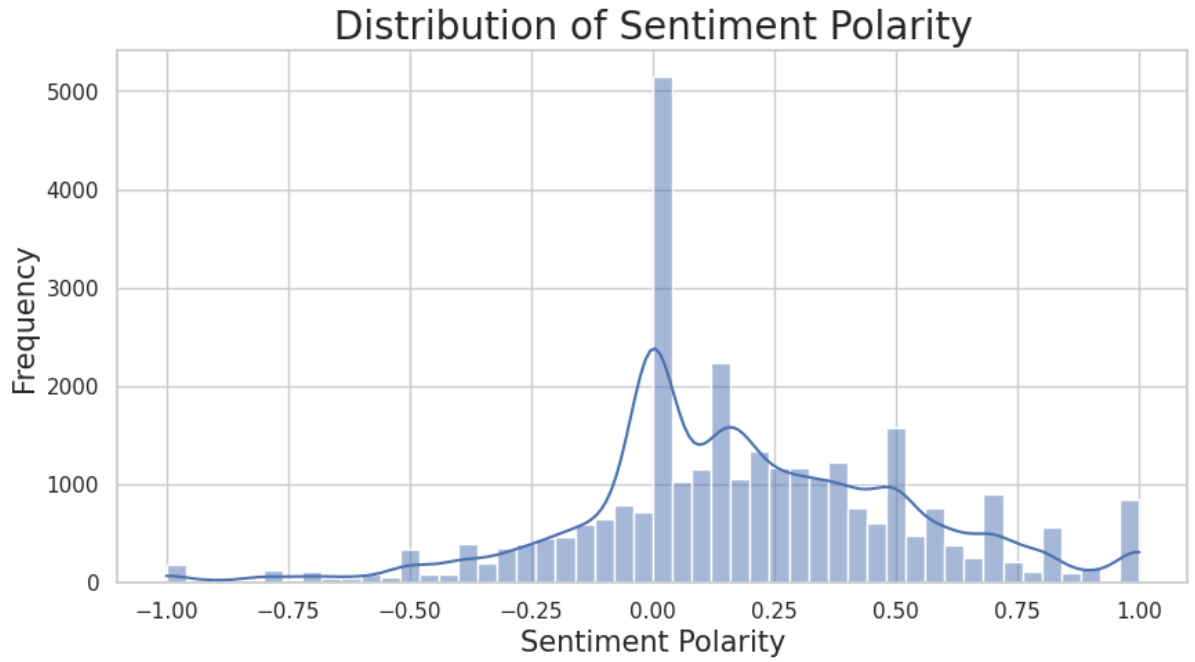
Removing Unnecessary Columns:
Certain columns were considered nonsignificant to the analysis and were subsequently dropped. Specifically, the 'Current Ver' column in the Play Store Dataset (df_psdata) and the 'Translated_Review' column in the User Reviews Dataset (df_review) were excluded.

Exploratory Data Analysis

Exploratory Data Analysis. It is a process of investigating and analysing a dataset to understand its characteristics, summarize its main features, and identify any patterns or relationships within the data. EDA is a crucial step in the data analysis process as it allows data scientists and analysts to gain insights into the data.

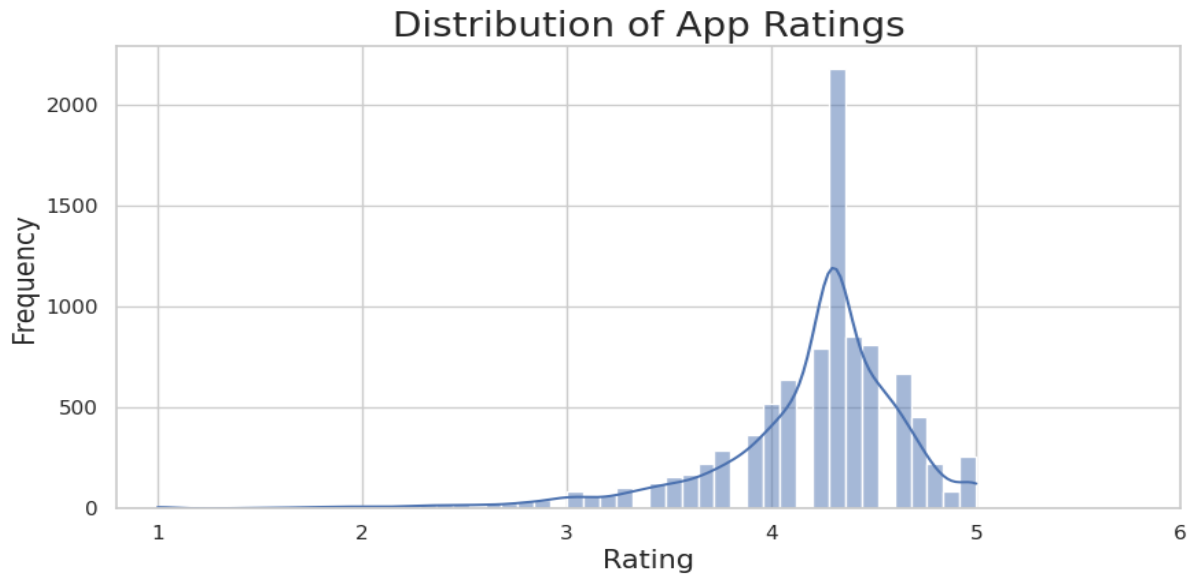
Distribution of Sentiment Polarity:

The histogram shows a clear peak towards the positive side, indicating most apps have predominantly positive sentiment amongst users. This suggests most apps are well-received and appreciated by users.



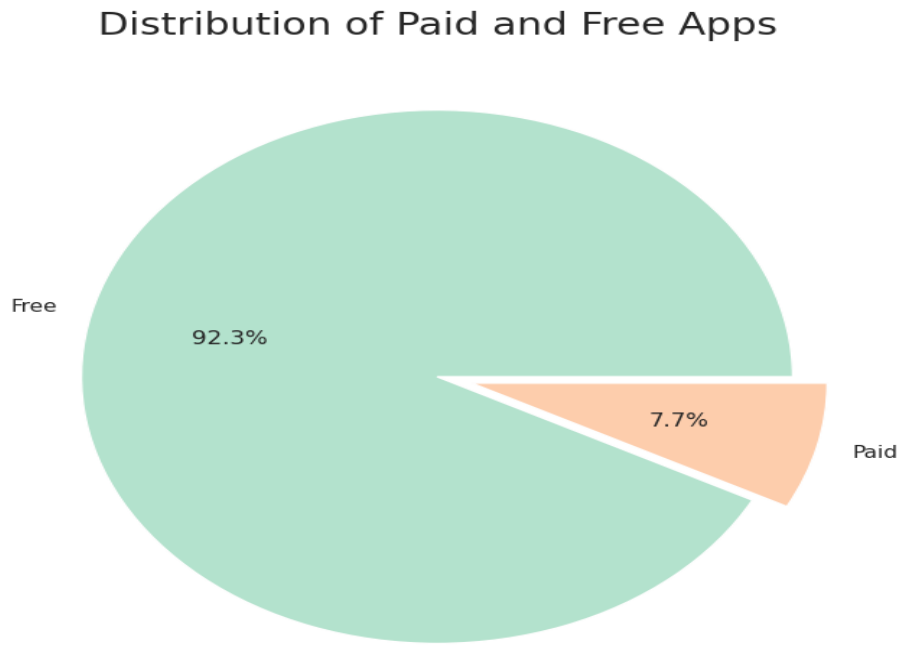
Distribution of App Ratings:

Most users are satisfied with the apps, as evidenced by the high concentration of positive ratings.



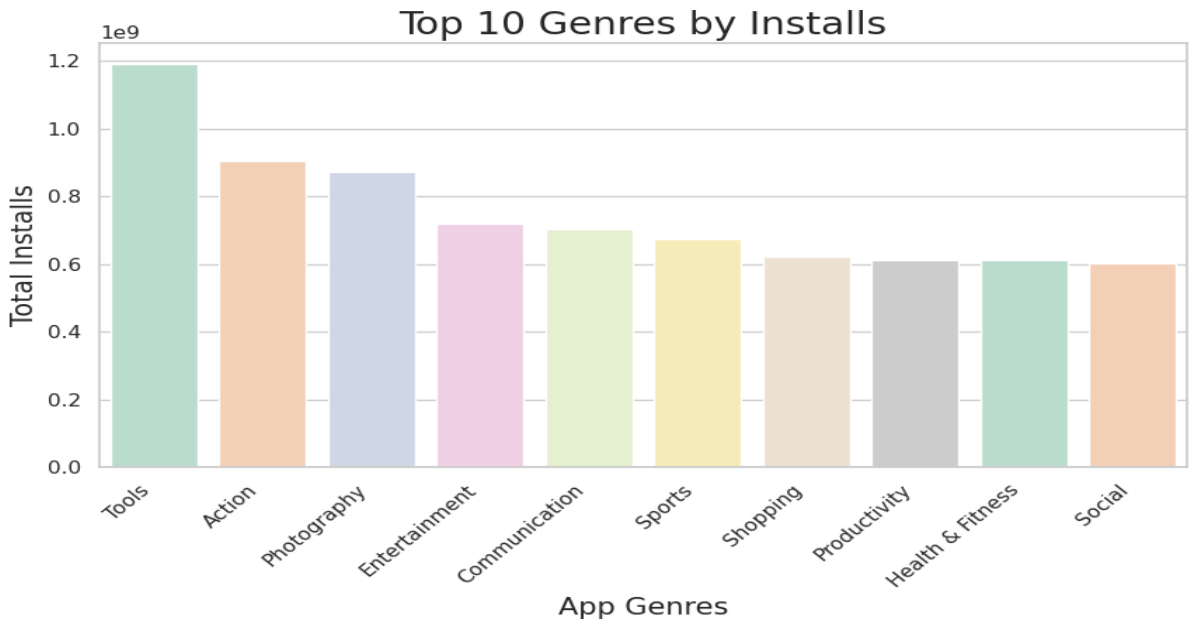
Distribution of App Types (Free and Paid):

The larger slice of the pie chart represents the free apps, with 92.3% of the total. The smaller slice representing paid apps, with 7.7% of the total. This indicates that most apps on the Google Play Store are free to download and use.



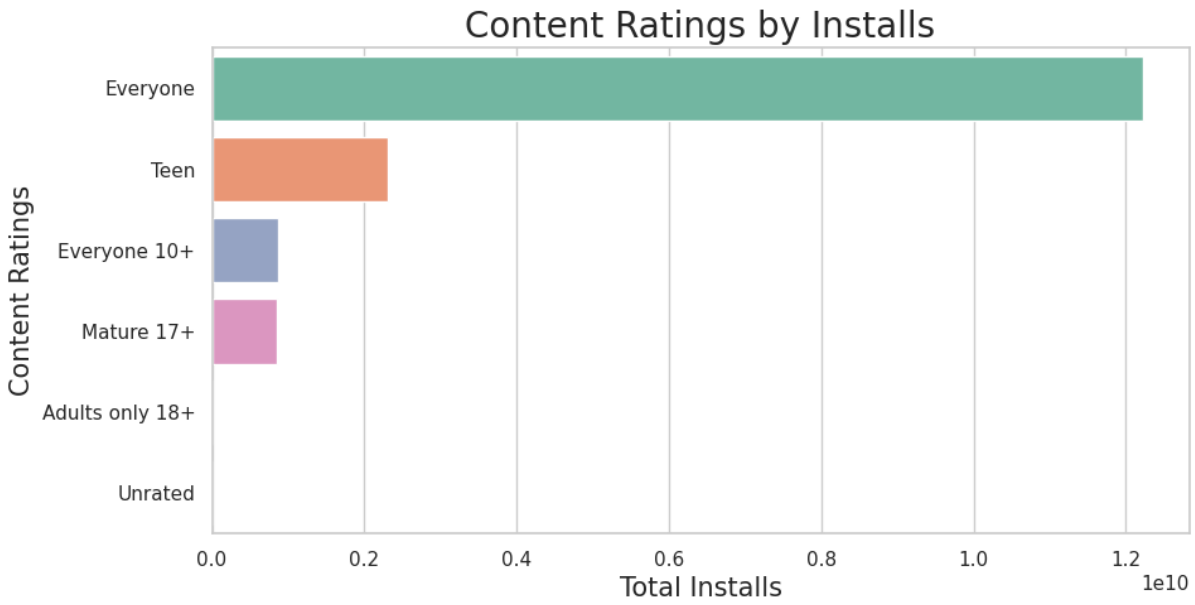
Top 10 Genres by Installs:

Tools apps lead as the most popular genre, driven by the growing reliance on smartphones and tablets for work and productivity, followed by action apps, known for their fast-paced and exciting content. Photography apps claim the third spot, reflecting the surge in smartphone photography's popularity. Entertainment apps, covering streaming services and social media, secure the fourth position, while communication apps, including messaging and video conferencing, stand as the fifth most popular genre.



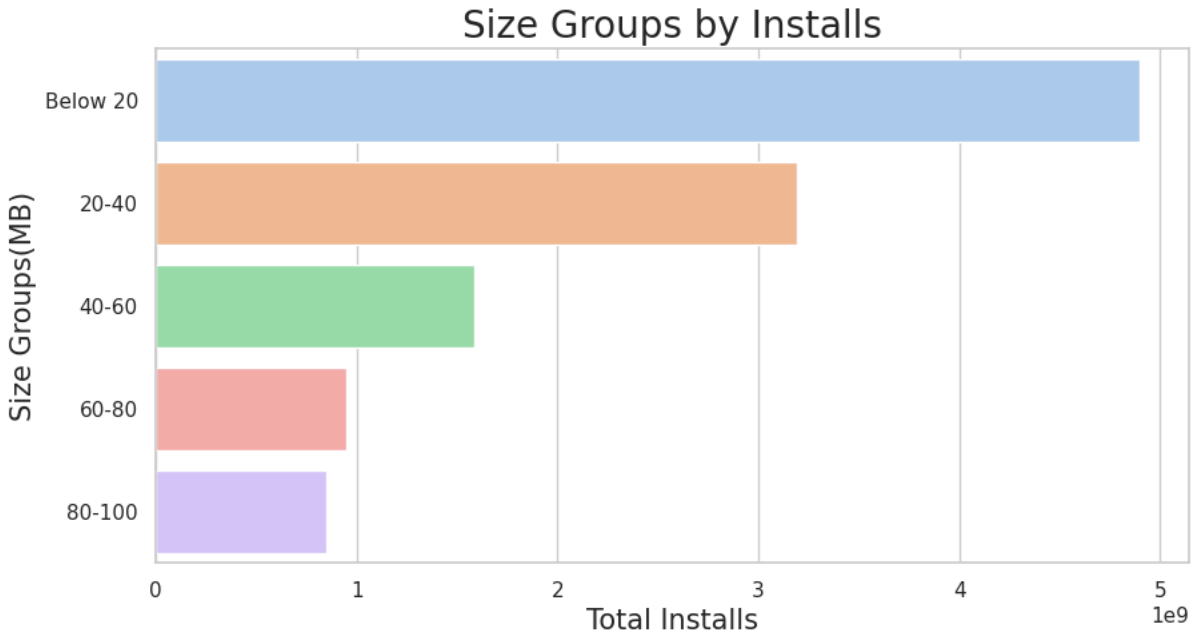
Content Ratings by Installs:

Apps for “Everyone” and “Teen” have the highest installs, indicating a preference for apps suitable for all ages or users aged 13 and above. The “Everyone 10+” category follows with the third-highest installs, indicating a preference for apps suitable for users aged 10 and above. In contrast, the “Mature 17+” and “Adults only 18+” categories show significantly fewer installs, suggesting a limited interest in apps tailored to users aged 17 or older.



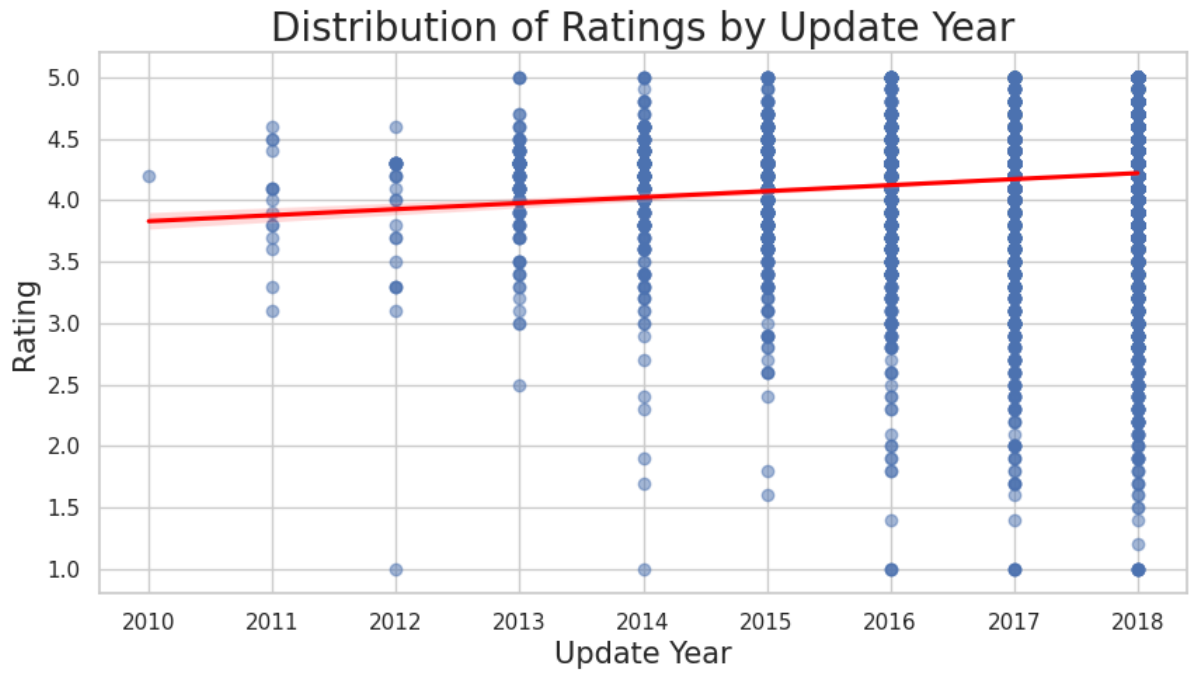
Size Groups by Installs:

Smaller apps are preferred by users, as indicated by the highest number of installs for the below-20 size group, followed by the 20-40, 40-60, and 60-80 size groups. Conversely, the 80-100 size group has noticeably fewer installs, less than a quarter of the below-20 group.



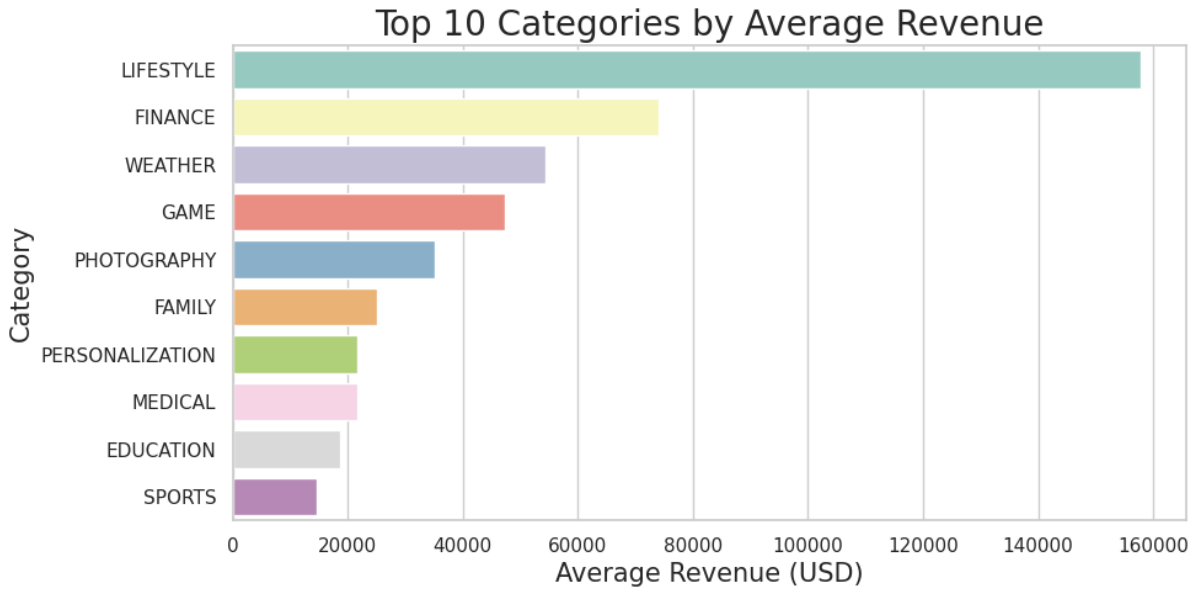
Distribution of Ratings by Update Year:

The average rating has shown an improvement, rising from approximately 3.5 in 2010 to nearly 4.5 in 2018. This indicates a general trend of increasing satisfaction among users with the product over the years.



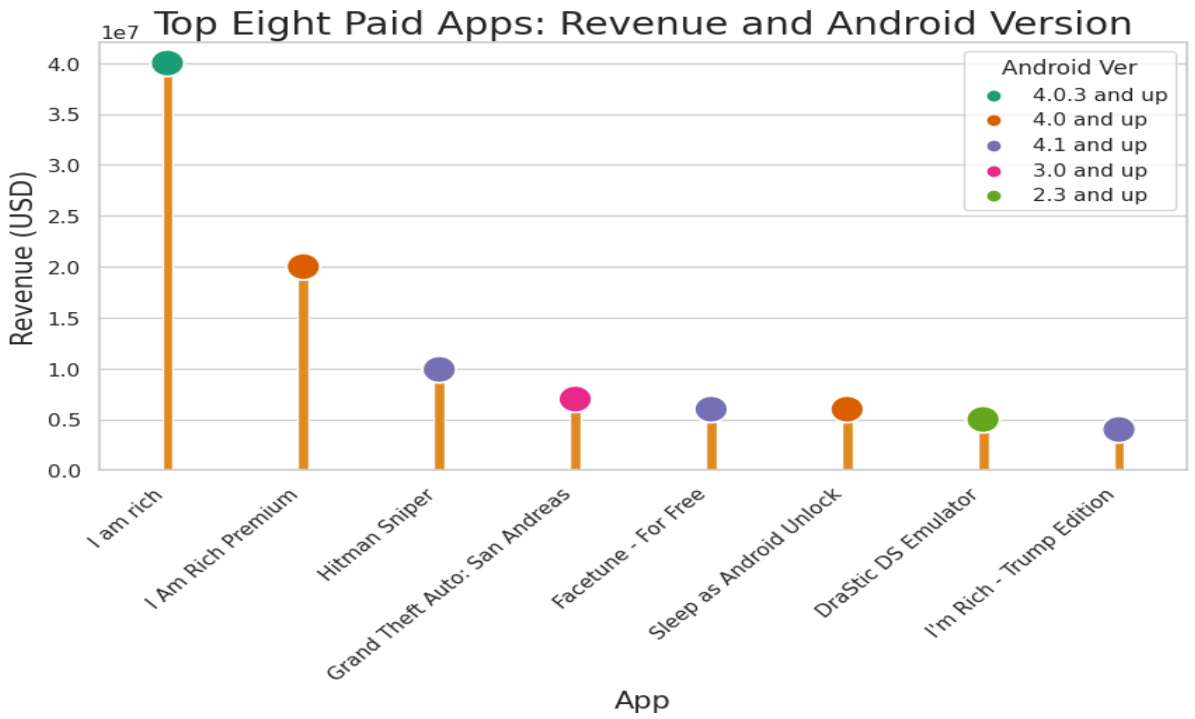
Top 10 Categories by Average Revenue:

The top revenue generating categories Lifestyle, Finance, and Weather indicate user investment in personal and financial products. Game, Photography, and Family follow in revenue, highlighting spending on leisure and entertainment. Conversely, Sports records the lowest revenue, Education closely follows, and Personalization and Medical categories show lower user interest and profitability, respectively.



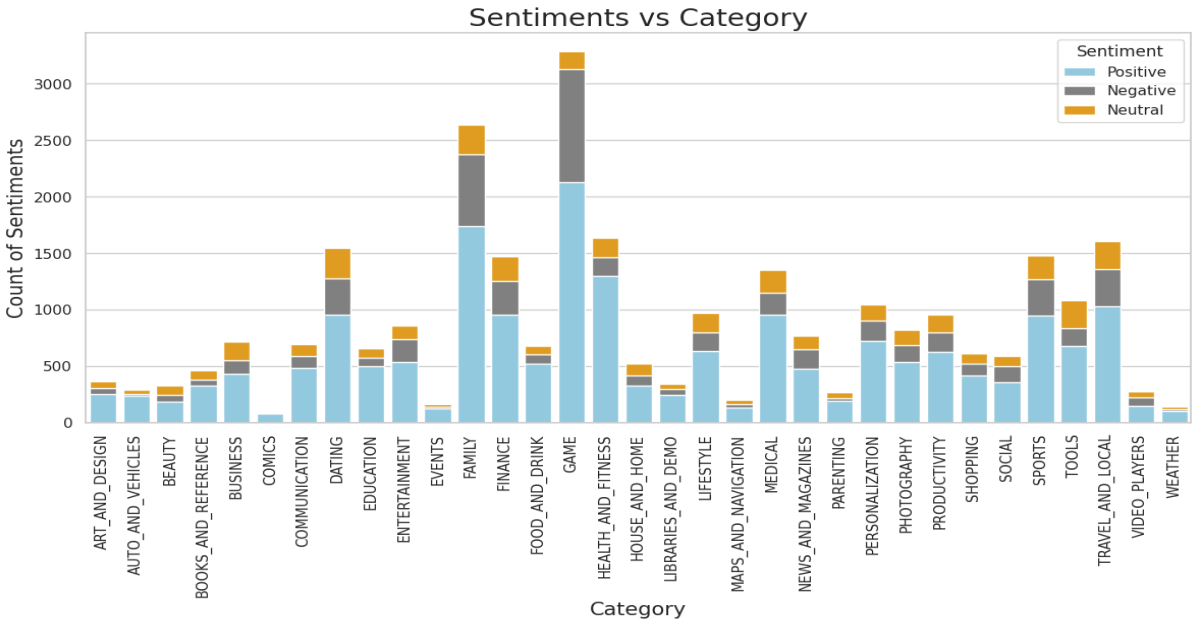
Revenue and Android Version of the Top 8 Paid Apps:

Apps designed for Android versions 4.0 and above dominate higher revenue ranks, indicating a correlation between compatibility with newer Android versions and revenue generation. Among the top 8 high revenue apps, six adhere to this compatibility, while exceptions like "Grand Theft Auto: San Andreas" (Android 3.0 and up) and "DraStic DS Emulator" (Android 2.3 and up) are on the lower end of the revenue spectrum.



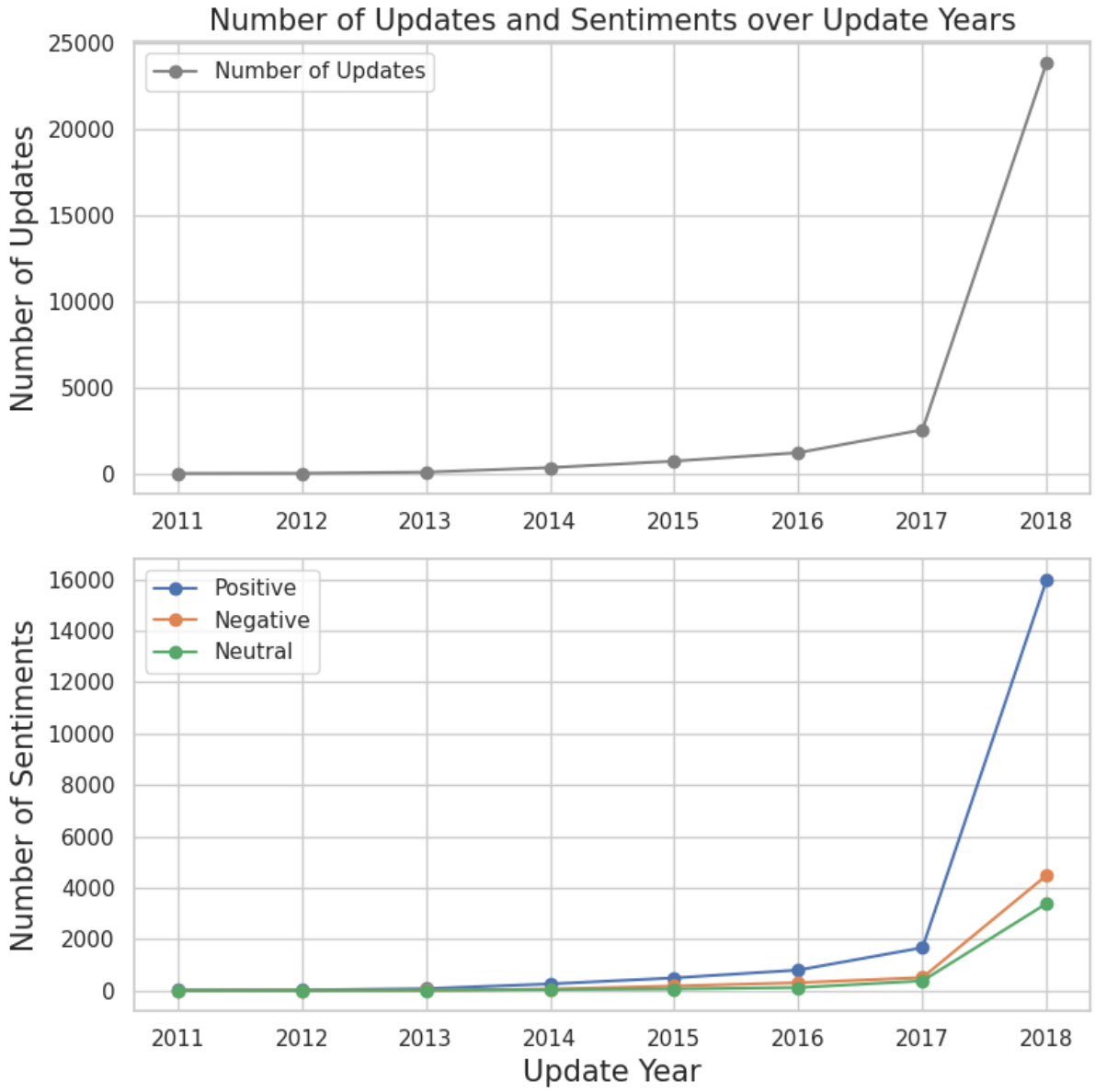
Sentiment counts across categories:

The stacked bar chart reveals a complex interplay of positive and negative sentiments across different categories. While people express positive sentiments towards categories like game, family, health and fitness, travel and local, and dating, there is also a notable presence of negative sentiment associated with most of these same categories.



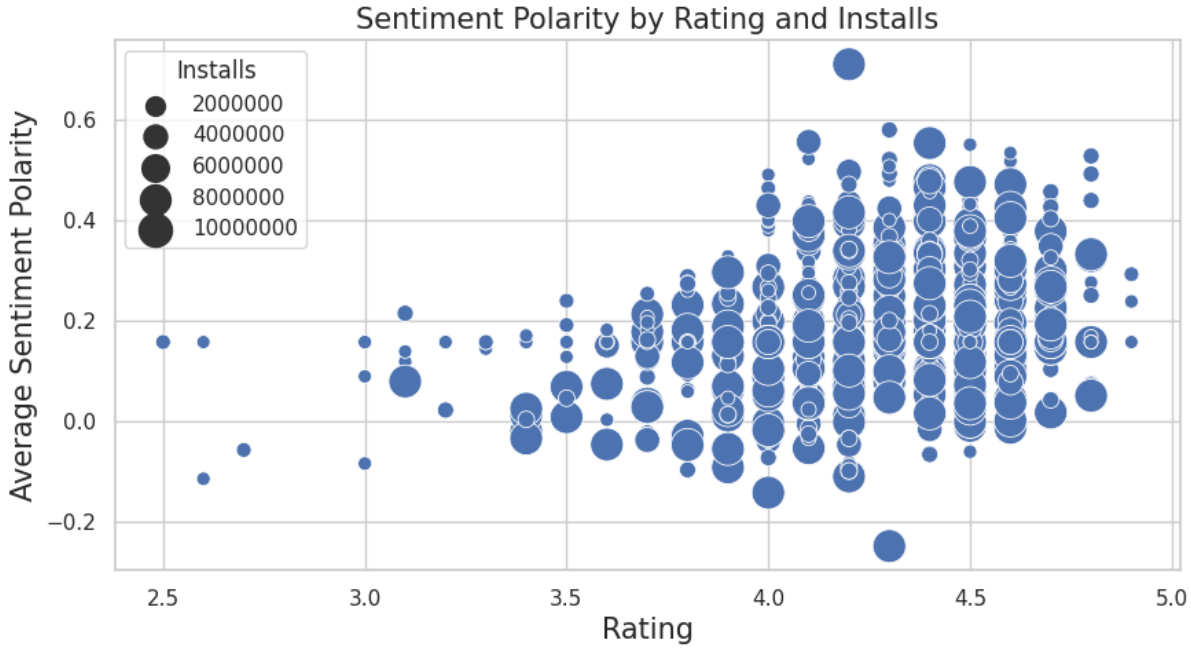
Progression of update counts and the distribution of sentiment counts over time:

A general trend of increasing positive sentiments over time indicates growing satisfaction with received updates. Simultaneously, the increasing number of updates suggests developers are releasing new updates more frequently. Despite this, the number of negative sentiments remains relatively stable, indicating overall satisfaction with the updates received.



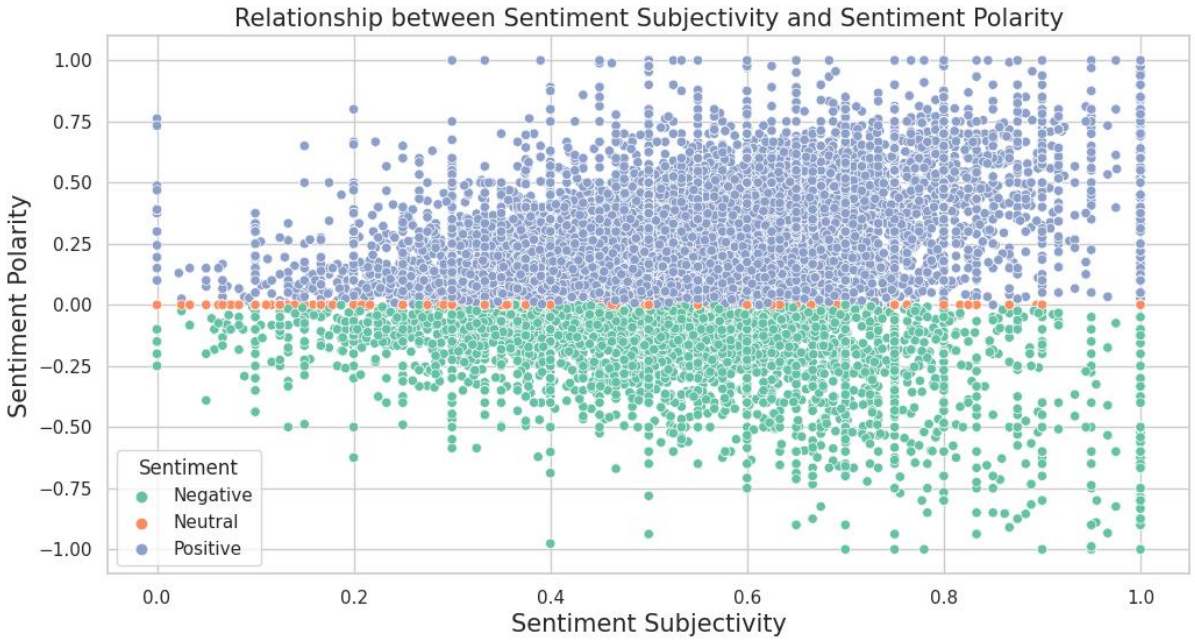
Relationship between Rating, Sentiment Polarity, and Installs:

Higher-rated apps tend to have elevated sentiment polarity, aligning with users leaving positive reviews. Despite lower install counts, some niche apps exhibit high sentiment polarity, reflecting a devoted user base. However, a disparity arises as the most installed apps show lower average sentiment polarity, indicating popularity doesn't consistently correlate with positive user sentiment.



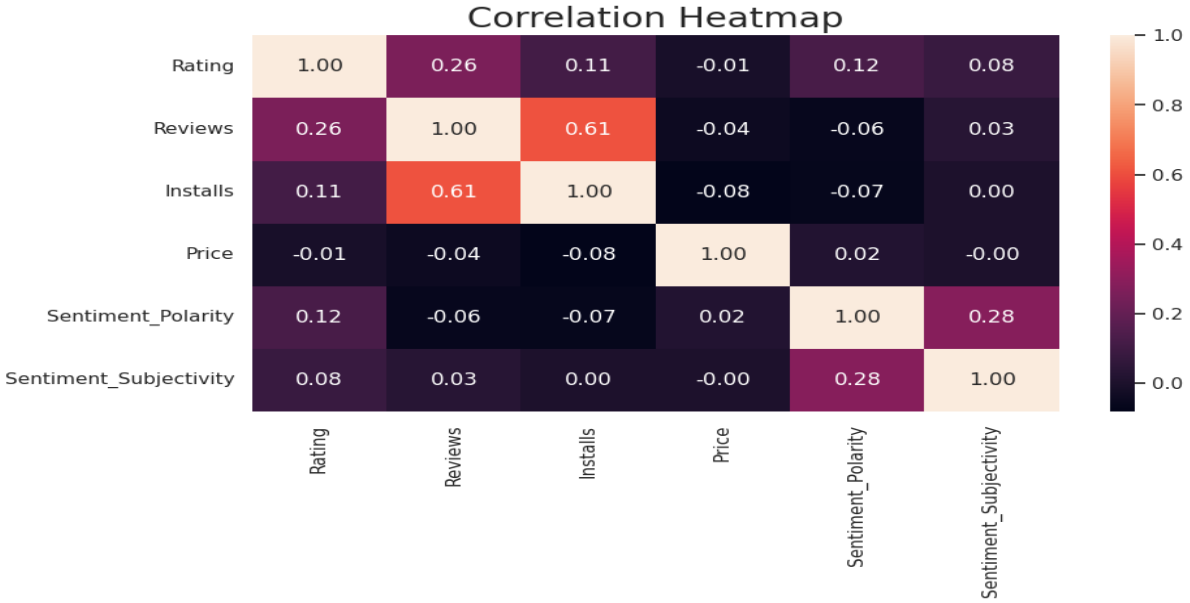
Relationship between Sentiment Subjectivity and Sentiment Polarity:

Sentiment Polarity and Sentiment Subjectivity shows a moderate positive correlation between the two variables. This means that, in general, as sentiment polarity increases, sentiment subjectivity tends to increase as well. However, the relationship is not very strong.



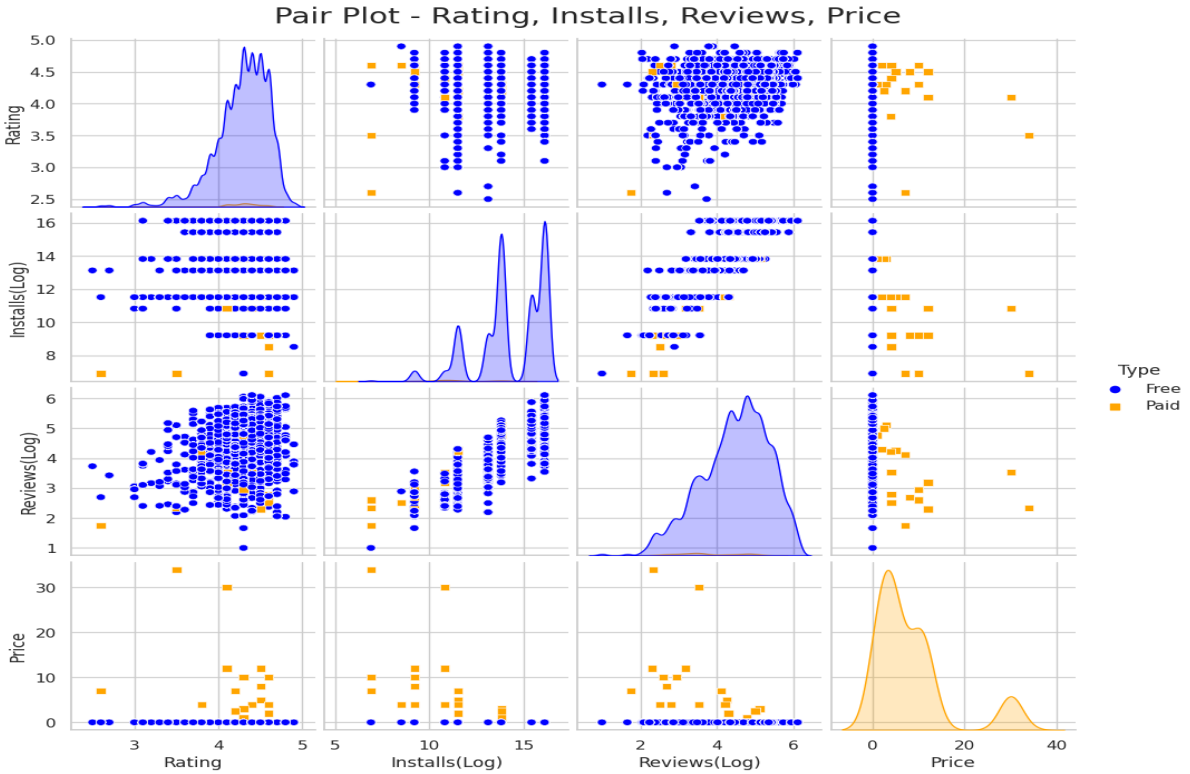
Correlation among Rating, Reviews, Installs, Price, Sentiment_Polarity, and Sentiment_Subjectivity:

Higher-rated apps tend to have more reviews, installs, and positive sentiment, while also being slightly cheaper. Strong positive correlations exist between reviews and installs. Reviews, however, show a weak negative correlation with price, indicating slightly cheaper apps. Installs exhibit a weak negative correlation with price, suggesting slightly cheaper apps. Additionally, positive sentiment in reviews correlates moderately with subjective reviews.



Pair Plot of Rating, Installs, Reviews and Price:

Higher-rated apps tend to have more installs and reviews, with a weak negative correlation between rating and price. Popular apps, indicated by higher installs, also tend to have more reviews and slightly lower prices, likely due to increased demand and developer strategies.



Recommendations

1. **Strategic Development in Popular Genres:** Acknowledge the popularity of certain genres, such as Tools, Action, Photography, and Entertainment. Consider investing in or refining apps within these genres to align with user preferences and maximize engagement.

2. **Emphasize Free Apps:** Free apps overwhelmingly dominate the market, indicating the significance of providing an engaging free version to attract users. Given that the majority of apps are free, focus on monetization strategies that complement the free model, such as in-app purchases, ads, or premium features.

3. **Optimize App Size:** Recognize the preference for smaller apps, as indicated by the popularity of the below 20 and 20-40 MB size groups. Ensure that new app developments and updates prioritize efficiency and minimal storage requirements to align with user preferences for smaller-sized apps.

4. **Tailor Content Ratings:** Understand the user preference for Everyone and Teen categories and ensure new apps align with these preferences. Be mindful of the limited interest in Mature and Adults-only categories, adjusting content accordingly.

5. **Strategic Revenue Generation:** Consider app development or improvements in categories that generate higher revenue, such as Lifestyle, Finance, and Weather. Evaluate user preferences within lower-revenue categories to identify opportunities for enhancement.

6. **Compatibility with Latest Android Versions:** Given the correlation between higher revenue and compatibility with newer Android versions, prioritize app development and updates for the latest versions of the Android operating system.

7. **User Engagement in Popular Categories:** Recognize the positive sentiments associated with popular categories like Games, Family, Health and Fitness, Travel and Local, and Dating. Strategically engage users through marketing, promotions, and feature enhancements in these categories.

8. **Continuous Improvement:** Monitor the progression of update counts and user sentiments over time. Respond to user feedback with timely updates to demonstrate a commitment to app improvement.

9. **Focus on Positive User Sentiment:** Identify the characteristics of apps with consistently positive sentiment and leverage those aspects in future app development. Encourage and amplify positive user experiences through marketing and feature enhancements.

10. **Address Negative Feedback:** Investigate apps with negative sentiment to pinpoint specific issues causing dissatisfaction. Prioritize improvements in areas highlighted by negative sentiment to enhance user satisfaction and overall app performance.

Conclusion

This project has successfully analysed the Play Store app dataset using Python, uncovering valuable insights into key factors for app engagement and success. The data visualizations and interpretations provide a comprehensive understanding of user sentiment, app ratings, genre preferences, content suitability, and the impact of updates.

Based on these insights, I have crafted actionable recommendations for the client to optimize app performance and achieve their business objectives. I advise them to focus on fostering positive sentiment, addressing user concerns promptly, catering to user preferences like smaller apps and frequent updates, targeting specific genres and content ratings strategically, and building loyal followings in niche markets.

By embracing a data-driven approach and continuously adapting to user preferences, the client can ensure long-term success in the competitive Android app market.