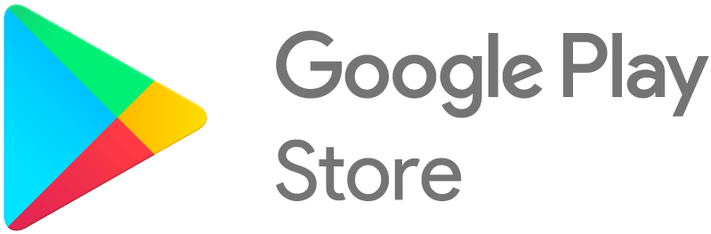
CAPSTONE PROJECT

PLAY STORE APP ANALYSIS



By

RAHUL KUMAR

Email - [rahulkumar.19k8@gmail.com](mailto:rahulkumar.19k8@gmail.com)

GitHub Link-

Abstract

The Google Play Store represents one of the most competitive digital marketplaces, with thousands of new apps released daily. Developers face fierce global competition, and with the majority of Play Store apps offered for free, revenue models based on in-app purchases, ads, and subscriptions remain complex and opaque. This creates challenges in accurately assessing how these monetization strategies contribute to app success. Consequently, an app’s performance is often gauged through user engagement metrics such as install counts and lifetime user reviews, rather than direct revenue. User ratings, a key metric, provide critical voluntary feedback, but they can often be skewed by limited or missing reviews, leading to potential biases. Moreover, notable discrepancies frequently emerge between numerical ratings and qualitative user reviews.

This study aims to perform an in-depth analysis of the Play Store apps dataset using Python, identifying key drivers of app engagement and success. By leveraging powerful data wrangling tools such as NumPy and Pandas, combined with rich visualizations from Seaborn and Matplotlib, this research seeks to uncover actionable insights for optimizing app performance in the Android ecosystem. The results will offer critical guidance for app developers and marketers aiming to enhance user engagement and streamline app store optimization (ASO) strategies.

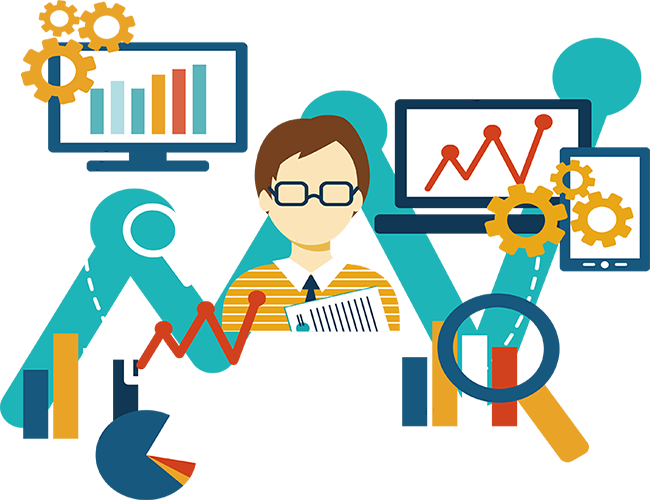
Keywords: App performance optimization, App store optimization (ASO), User engagement, App ratings and reviews, Data analysis, Play Store insights

**Problem Statement**

The Play Store apps dataset holds significant potential to propel app-making businesses towards greater success. By analyzing this data, developers can uncover actionable insights to effectively capture and expand their presence in the Android market. Each app entry provides valuable metrics, such as category, rating, size, and other attributes, while an additional dataset offers customer reviews. Analyzing and exploring these datasets will reveal the critical factors driving app engagement and success, enabling developers to make informed decisions.

**Business Objective**

Utilize Python to perform an in-depth analysis of the Play Store apps dataset to uncover critical factors that influence app engagement, user satisfaction, and revenue generation. The objective is to provide actionable insights that will not only optimize app performance but also enhance user experience and maximize revenue potential, ultimately improving the app's market positioning and success in the Android ecosystem.



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:

**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 critical steps in preparing datasets for effective analysis. Data cleaning involves systematically identifying and correcting errors, addressing missing or inaccurate information, and maintaining overall data quality. This process is crucial for enhancing the dataset's reliability and ensuring that the insights derived are based on accurate and trustworthy data. Data preprocessing goes a step further by transforming raw data into a format that is suitable for analysis. This includes operations such as normalization, scaling, and outlier management, which collectively optimize data presentation and facilitate the extraction of meaningful insights. Together, these steps are vital for ensuring that the data is both clean and well-structured, ultimately leading to more robust and actionable analytical outcomes.

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\_symbol) to drop the '$' symbol and convert the 'Price' column to the float data type.

**Handling Installs Values:**

Created a function (drop\_addition\_symbol) to drop the '+' symbol and convert the 'Installs' column to the integer data type.

**Converting Size Entries:**

Created a function (convert\_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\_apps) and the 'Translated\_Review' column in the User Reviews Dataset (df\_reviews) were excluded.

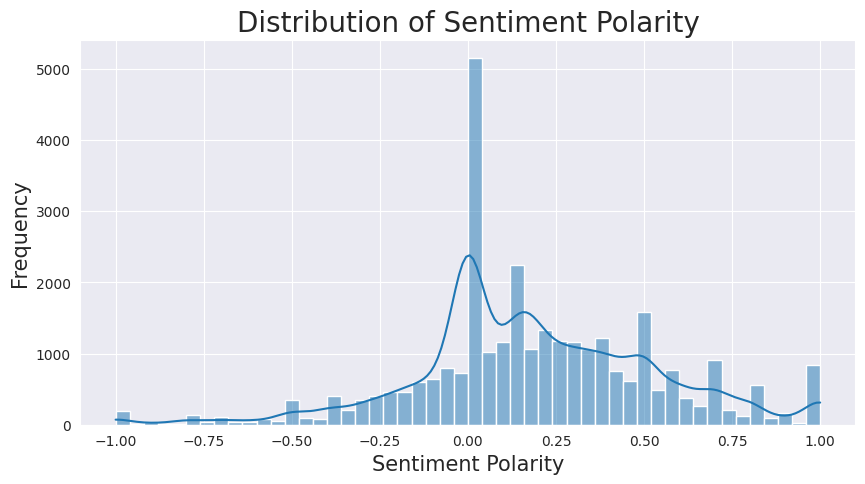
Exploratory Data Analysis

**Exploratory Data Analysis (EDA):**

It is a fundamental process for investigating and understanding a dataset's underlying characteristics. This involves summarizing the dataset's main features, uncovering patterns, and identifying relationships within the data. EDA is a critical step in the data analysis workflow as it enables data scientists and analysts to gain valuable insights into the data's structure and nuances. By visualizing and analyzing data distributions, trends, and correlations, EDA helps in forming hypotheses, guiding further analysis, and making informed decisions based on a comprehensive understanding of the dataset.

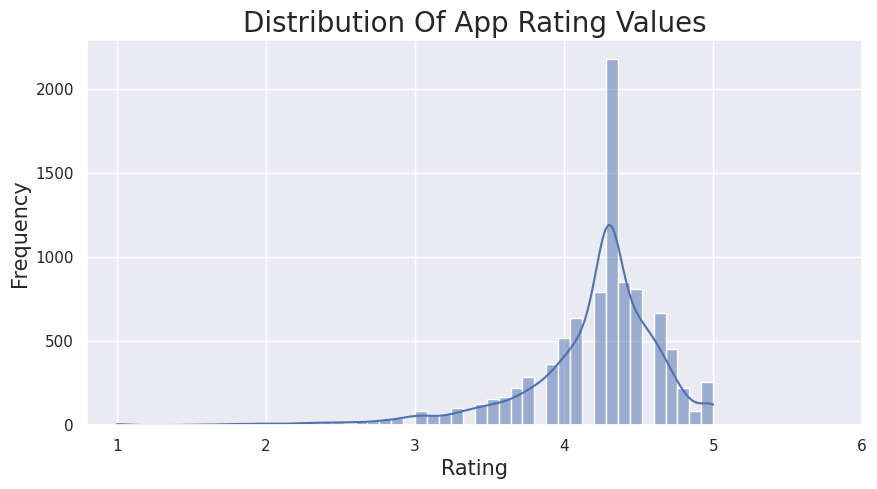
Distribution of Sentiment Polarity:

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.



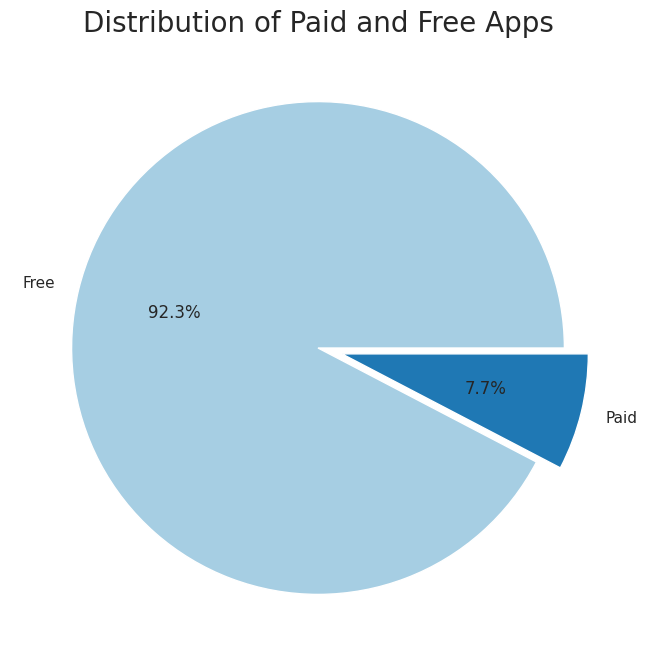
Distribution of App Ratings:

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



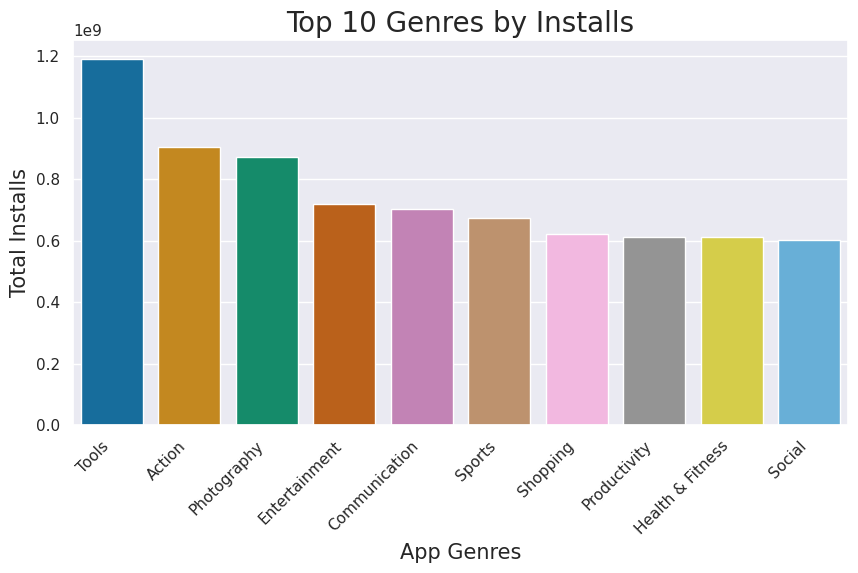
Distribution of App Types (Free and Paid):

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.



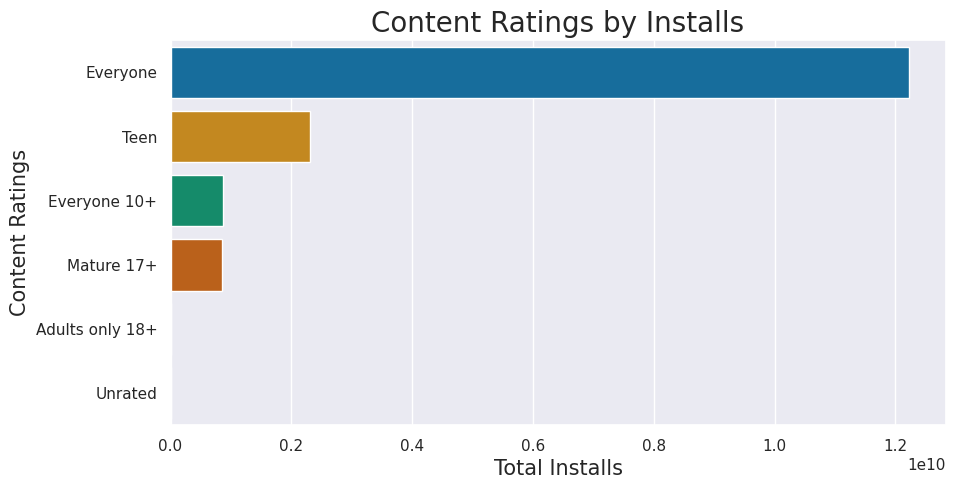
Top 10 Genres by Installs:

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.



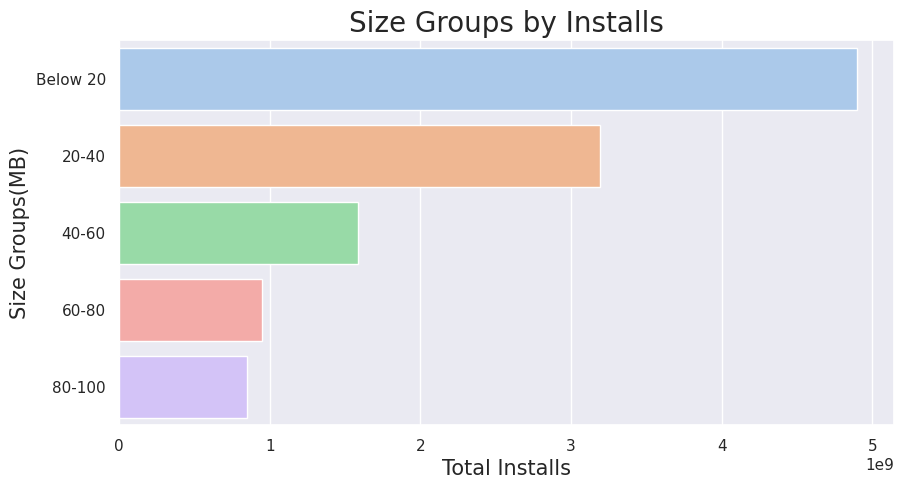
Content Ratings by Installs:

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.



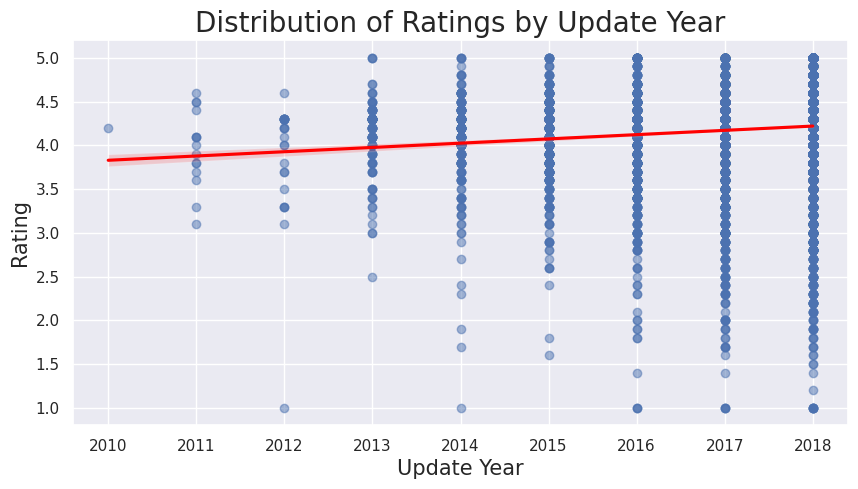
Size Groups by Installs:

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.



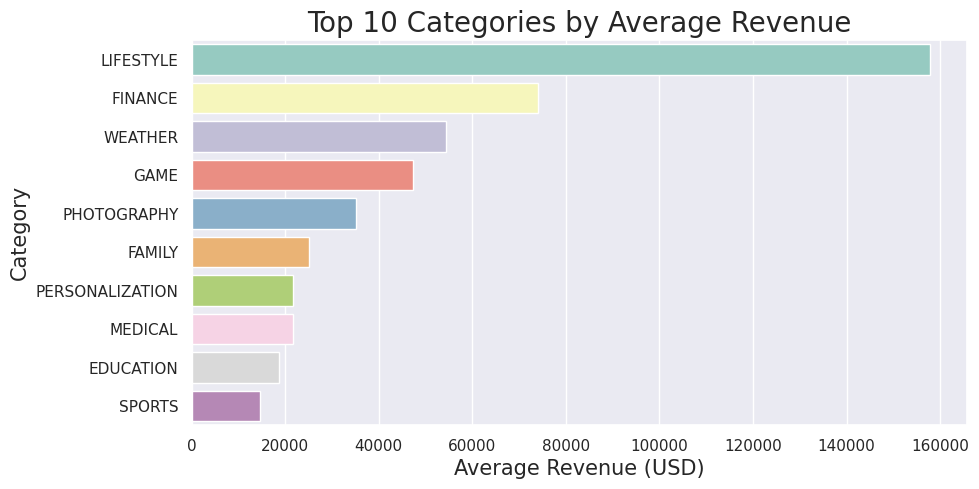
Distribution of Ratings by Update Year:

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.

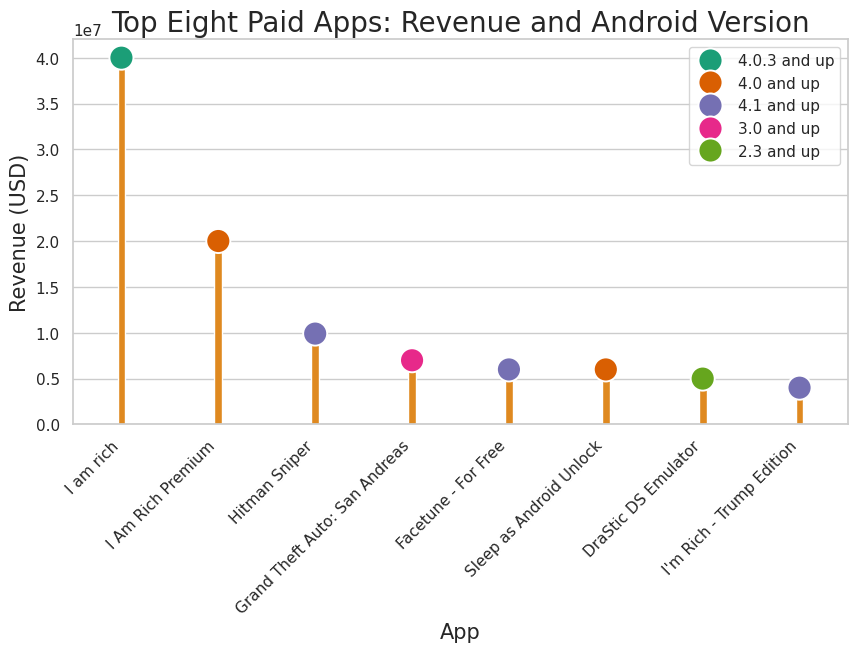


Top 10 Categories by Average Revenue:

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.

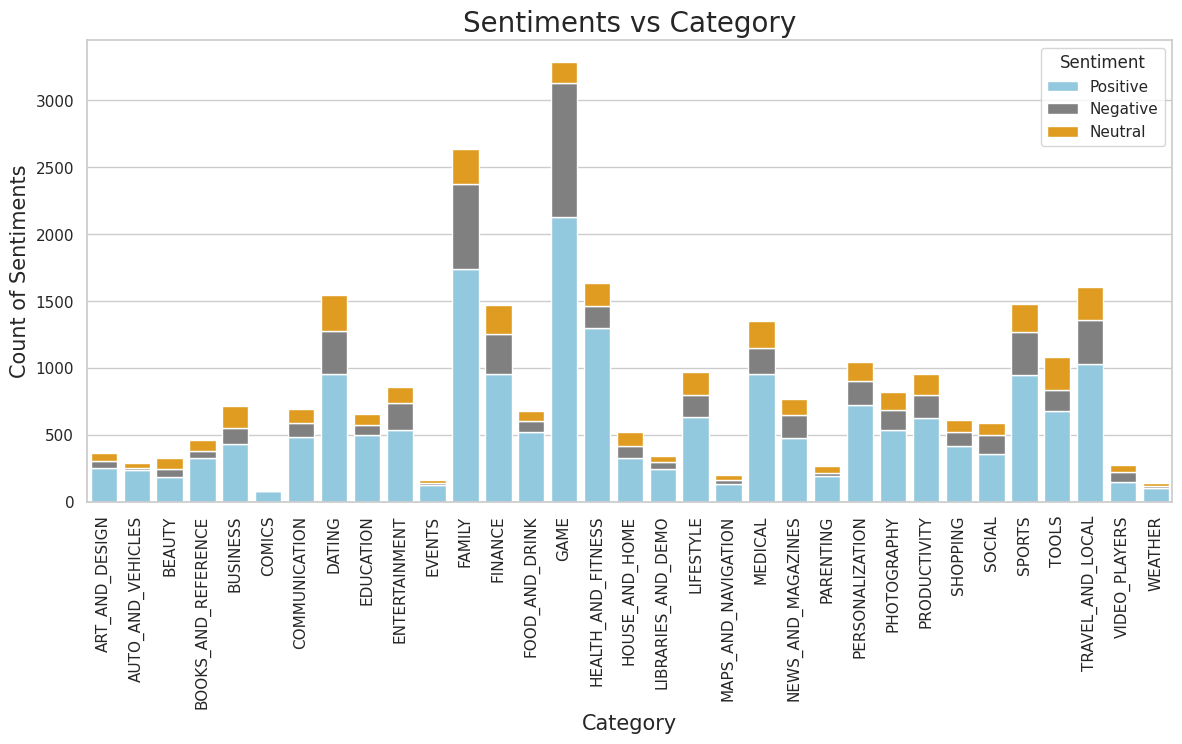


Revenue and Android Version of the Top 8 Paid Apps:

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. 

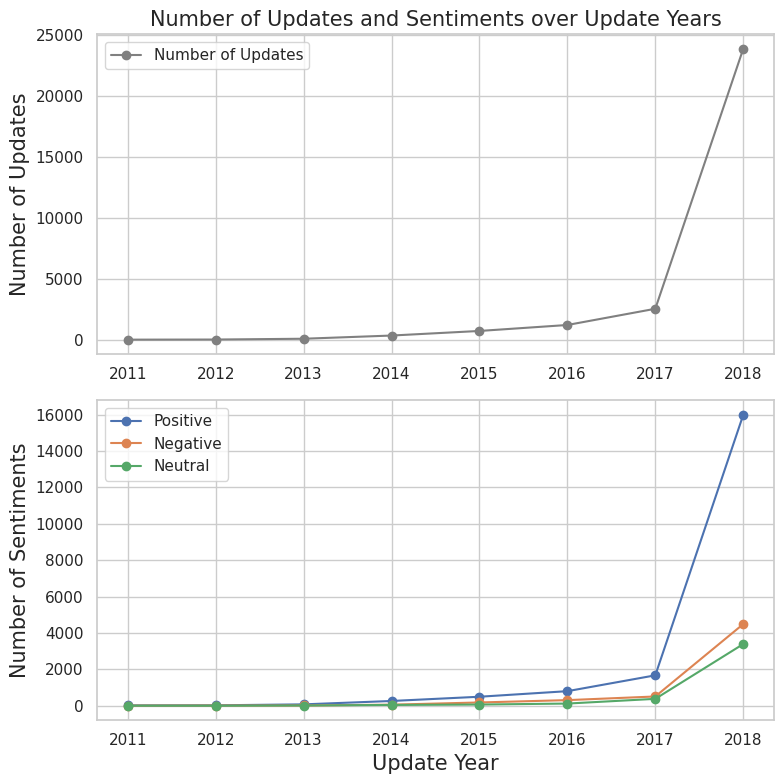
Sentiment counts across categories:

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.



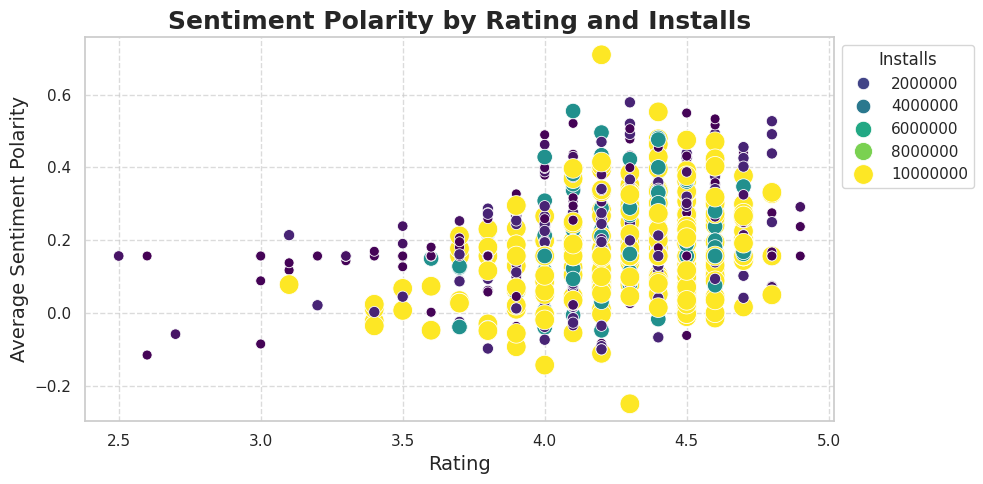
Progression of update counts and the distribution of sentiment counts over time:

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.



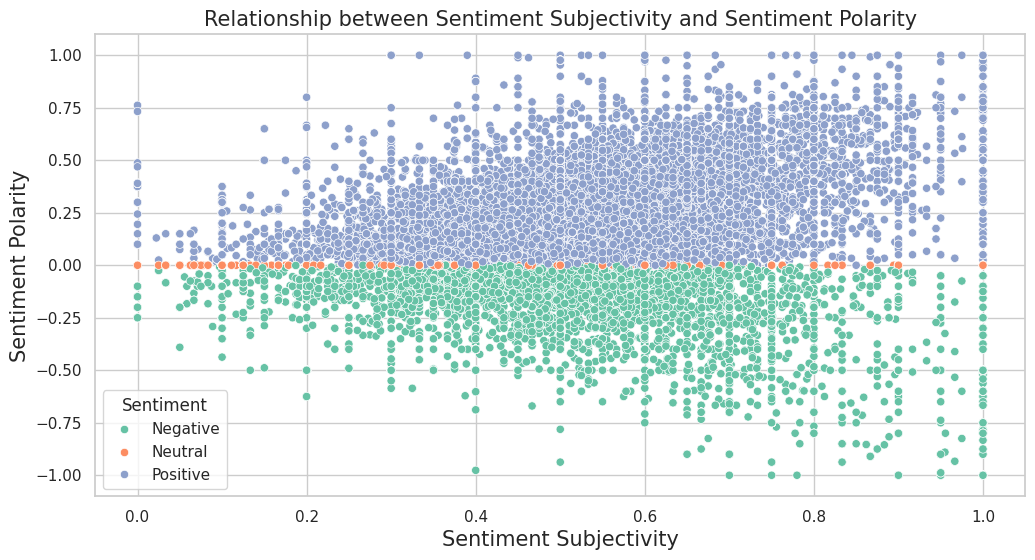
Relationship between Rating, Sentiment Polarity, and Installs:

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.



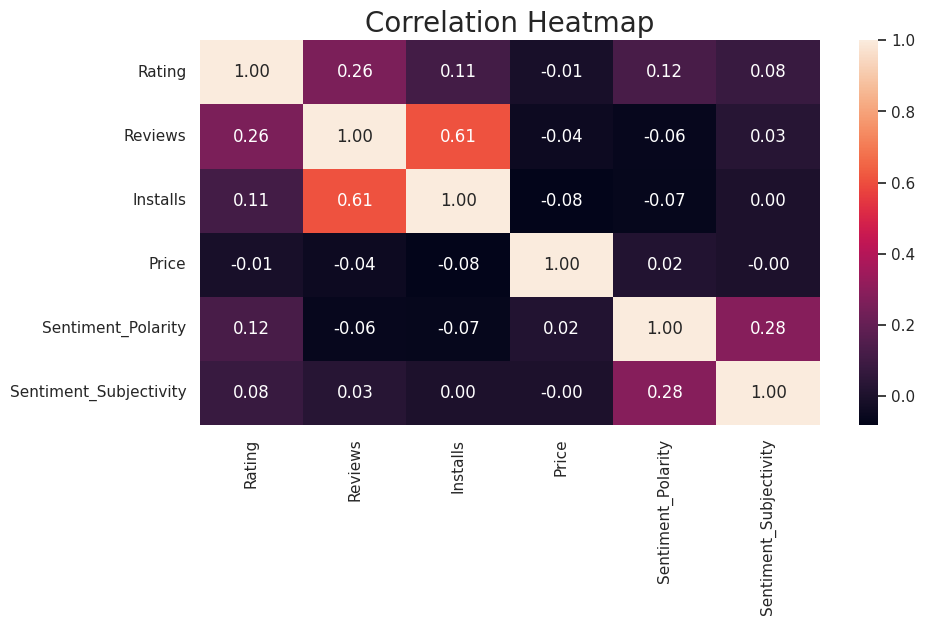
Relationship between Sentiment Subjectivity and Sentiment Polarity:

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.

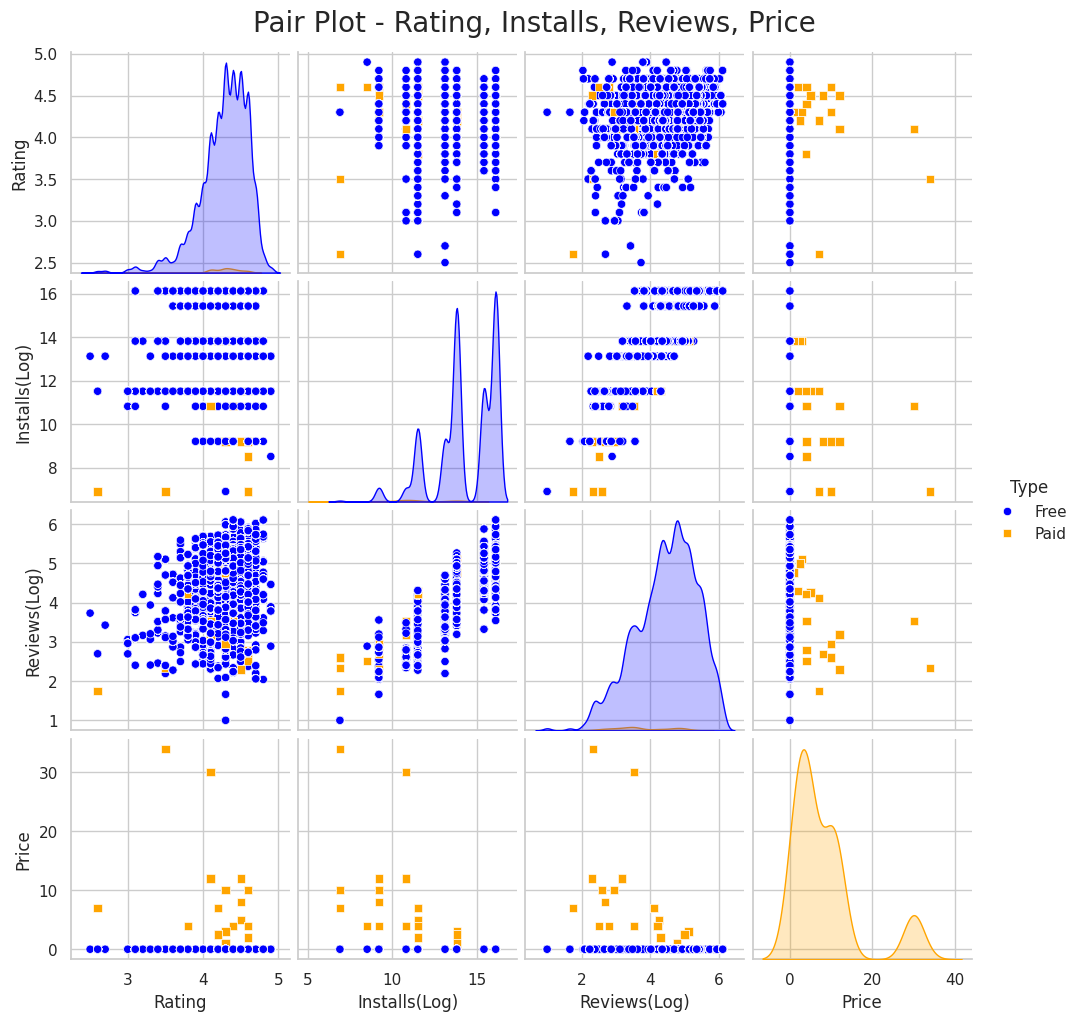


Correlation among Rating, Reviews, Installs, Price, Sentiment\_Polarity, and Sentiment\_Subjectivity:

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 Investment in Popular Genres:**

Leverage the popularity of genres such as Tools, Action, Photography, and Entertainment by investing in innovative features, performance improvements, and user-centric design. Conduct market research to identify gaps in these genres, and focus on building apps that cater to unmet user needs or introduce novel functionalities. This will help maximize user engagement and loyalty within these high-demand areas.

**2. Emphasize Free App Models with Smart Monetization:**

With free apps dominating the market, it’s crucial to offer a compelling free version to attract a broad user base. Invest in \*\*in-app purchases (IAPs), targeted advertisements\*\*, and \*\*freemium models\*\* that offer value without overwhelming users. Provide a seamless transition between free and paid features, ensuring that users feel the premium version is worth the investment while maintaining a strong free experience to retain engagement.

**3.Optimize for Small App Size and Efficiency:**

Users show a clear preference for smaller apps (particularly under 20MB and 20-40MB categories). Focus on \*\*app efficiency\*\*, using data compression techniques, cloud-based services, and optimizing code to reduce bloat. This not only aligns with user preferences but also improves app performance on lower-end devices, expanding your potential user base globally.

**4. Tailor Content Ratings to User Demographics:**

Apps in the Everyone and Teen rating categories receive significantly more engagement than those targeted at more mature audiences. Prioritize content and features that appeal to these groups, such as family-friendly games, educational tools, or teen-centered social apps. Avoid excessive focus on Mature and Adults-only content unless there is a specific strategic reason, as these categories have limited interest from broader user bases.

**5. Maximize Revenue by Targeting High-Earning Categories**:

Categories such as Lifestyle, Finance, and Weather consistently generate higher revenue. Analyze user behavior within these categories to develop apps that cater to specific user needs, whether it’s personal finance management tools, lifestyle-based subscription services, or weather forecasting enhancements. Identify opportunities in low-revenue categories and innovate to turn them into profitable ventures by adding unique value or niche services.

6. **Prioritize Compatibility with Latest Android Versions:**

Ensure that your apps are fully compatible with the latest Android OS versions, as there’s a clear link between revenue generation and compatibility with new platforms. Regularly update your apps to optimize performance on modern devices and take advantage of new Android features to attract tech-savvy users and enhance user experience.

**7. Enhance User Engagement in Popular Categories:**

Categories like Games, Family, Health and Fitness, Travel and Local, and Dating enjoy strong user engagement and positive sentiment. Boost engagement through targeted marketing campaigns, seasonal promotions, and consistent feature updates. Invest in loyalty programs, user-generated content, and community-building features to maintain long-term engagement and increase lifetime value (LTV) for these popular categories.

**8. Implement a Robust Feedback Loop for Continuous Improvement:**

Regularly monitor user feedback, app store reviews, and update cycles to identify areas for improvement. Maintain a transparent communication channel with your users, addressing concerns through timely updates and feature enhancements. Responding to feedback not only increases user satisfaction but also boosts your app’s credibility and reputation in the marketplace.

**9. Leverage Positive User Sentiment for Growth:**

Identify common characteristics of apps with consistently positive sentiment, such as user-friendly interfaces, fast load times, and unique features. Amplify these aspects in your marketing campaigns, focusing on the qualities that drive strong user experiences. Encourage user reviews and testimonials to build trust and social proof, while continuously improving the app based on what resonates most with users.

**10. Proactively Address Negative Sentiment:**

Investigate apps or features receiving negative feedback to pinpoint the root causes of dissatisfaction, whether it's performance issues, confusing interfaces, or missing functionality. Implement corrective measures such as performance improvements, UX redesigns, or bug fixes, prioritizing the most pressing concerns to reduce churn and improve app ratings. Use data analytics to track improvement after addressing these issues.

**Conclusion**

This project has successfully analyzed the Play Store app dataset using Python, uncovering key insights into the factors driving app engagement and success. Through detailed data visualizations and interpretations, we have gained a comprehensive understanding of user sentiment, app ratings, genre preferences, content suitability, and the impact of app updates.

Based on these findings, I have developed actionable recommendations to help the client optimize app performance and meet their business objectives. These include fostering positive user sentiment by promptly addressing feedback, prioritizing smaller app sizes and regular updates, strategically targeting high-demand genres and content ratings, and building loyal user bases in niche markets.

By adopting a data-driven approach and continuously adapting to evolving user preferences, the client is well-positioned to achieve sustained growth and success in the highly competitive Android app market.

