GOOGLE PLAY STORE DATA ANALYSIS PROJECT REPORT

ABSTRACT

This report provides an in-depth analysis of Google Play Store data, aimed at identifying key factors that influence app popularity, user satisfaction, and revenue potential.

By: Saniya Randive

Google Play Store Data Analysis Project Report

Agenda

1. Introduction

   - Overview and objectives of Google Play Store data analysis project.

2. Dataset Description

   - Details of datasets used, including key features and preprocessing steps.

3. Methodology

   - Data cleaning, exploratory data analysis, sentiment analysis, and revenue insights.

4. Results and Insights

   - Findings on app categories, free vs. paid apps, and user sentiment.

5. Challenges and Solutions

   - Issues encountered and methods used to address them.

6. Conclusion

   - Summary of findings, developer recommendations, and impact.

7. Future Scope

   - Potential improvements and predictive modeling opportunities.

8. Appendix

   - GitHub link and visual output screenshot.

1. Introduction

- ## Project Overview

This project offers a detailed exploration of the Google Play Store, which hosts millions of Android apps covering various categories such as Games, Productivity, Lifestyle, and Education. With a growing demand for mobile apps, app developers and marketers face intense competition, making it essential to understand the factors that drive app success. By analyzing patterns in app ratings, installation counts, pricing strategies, and user reviews, this project aims to highlight significant trends and actionable insights.

Using data science techniques, the analysis dives into understanding user preferences, examining both quantitative metrics (e.g., downloads, ratings, and revenue) and qualitative feedback (e.g., user sentiments). The insights derived provide developers and marketers with guidance on optimizing app engagement, enhancing user experience, and achieving competitive advantages in the Google Play Store ecosystem.

- ## Objectives

- Popularity Analysis: Identify the main factors that contribute to an app's popularity, including its category, type, and user ratings.

- **Comparison of Free and Paid Apps:** Investigate differences in performance and user satisfaction between free and paid apps, providing insights into user expectations and monetization strategies.

- **Sentiment Analysis**: Examine sentiment trends in user reviews to assess how they influence overall app ratings and user perception.

- **Revenue Potential**: Analyze how pricing strategies and install volumes impact revenue generation for paid apps, offering insights into profitable app development strategies.

By focusing on these objectives, this project provides a comprehensive view of the Google Play Store's dynamics, equipping stakeholders with the information needed to make data-driven decisions.

2. Dataset Description

- **Data Sources**: The dataset for this analysis was acquired from Google Play Store and includes two main files—Google Play Store App data and Google Play Store User Reviews.
- **Dataset Details**:
  - **Google Play Store Dataset**: Contains 10,841 records of apps with features such as app name, category, rating, reviews, size, installs, type (free/paid), price, and last updated.
  - **User Reviews Dataset**: Contains 64,295 records with user sentiments labeled as positive, negative, or neutral, along with sentiment polarity and subjectivity scores.
- **Key Features**:
  - **App**: Name of the app.
  - **Category**: Genre/category to which the app belongs.
  - **Rating**: Average user rating of the app.
  - **Reviews**: Number of reviews the app has received.
  - **Installs**: Number of times the app has been downloaded.
  - **Price**: Cost of the app (0 for free apps).
  - **Sentiment Polarity**: Numeric sentiment score from user reviews (-1 to 1, where -1 is negative, 1 is positive).
- **Data Cleaning and Preprocessing**:
  - Handling missing values (e.g., filling NaNs in ratings).
  - Encoding categorical variables (e.g., converting 'Type' into numeric values).
  - Converting data types for consistency (e.g., parsing the 'Installs' field to remove special characters).

3. Methodology

- **Data Preprocessing**: Essential preprocessing tasks included handling missing values, converting data types, and scaling numeric data. For example, columns with currency symbols or special characters were parsed to ensure consistency across numerical features.
- **Exploratory Data Analysis (EDA)**:
  - Distribution of apps across categories to identify which app categories dominate in the Play Store.
  - Analysis of app ratings and installs for insights into app popularity and user satisfaction.
  - Free vs. paid app analysis to understand differences in performance metrics such as installs and ratings.
- **Sentiment Analysis**:
  - Sentiment analysis was conducted on user reviews, using sentiment polarity scores to assess overall user satisfaction.
  - A comparative analysis of sentiment between free and paid apps provided additional insights into user expectations and perceptions.
- **Revenue and Pricing Analysis**:
  - Revenue analysis was performed by calculating estimated revenue for paid apps based on price and install numbers.
  - Comparison of app prices within categories helped to determine market trends in pricing.

4. Results and Insights

- **Category Insights**:
    - Categories like **Games** and **Family** have the highest number of apps and downloads, indicating high user engagement and demand. These categories present lucrative opportunities for developers.
- **Free vs. Paid Apps**:
    - Free apps dominate the market in terms of downloads, but paid apps have higher ratings on average. This suggests that users expect higher quality from paid apps and are willing to rate them positively if they deliver.
- **Sentiment Analysis Findings**:
    - User sentiment analysis reveals that free apps receive more mixed reviews, potentially due to high user expectations and a larger user base. Paid apps receive fewer negative comments, which could indicate higher user satisfaction.
- **Revenue Analysis**:
    - Apps with higher prices and downloads, such as **Minecraft**, are significant revenue generators. Categories like **Lifestyle** and **Productivity** also show high average revenues per app, suggesting user willingness to pay for specialized functionality

5. Challenges and Solutions

- **Handling Missing Data**

  The dataset presented a significant challenge due to the presence of missing values, primarily within the 'Rating' column. To address this issue, we implemented imputation techniques. For numeric fields, including 'Rating', we utilized the median value to fill in the gaps. This approach ensured that the missing values were replaced with a representative value from the dataset, minimizing the impact on subsequent analysis.

- **Class Imbalance**

  A notable challenge arose from the class imbalance within certain categories, where some app types were significantly more prevalent than others. This imbalance could potentially skew the analysis and lead to biased results. To mitigate this issue, we employed a combination of techniques, including downsampling. By reducing the number of instances in overrepresented categories, we achieved a more balanced distribution, allowing for a fairer and more accurate assessment.

- **Outliers in Revenue and Install Data**

  The dataset contained outliers, particularly in the 'Revenue' and 'Installs' columns, which were attributed to high-priced or "junk" apps with minimal downloads. These outliers had the potential to distort the analysis and skew the results. To address this, we implemented a filtering process to remove these extreme values. By focusing on a more representative subset of data, we ensured that the analysis was not unduly influenced by these anomalous data points.

- **Data Transformation**

  Several fields, including 'Installs' and 'Price', required careful parsing and cleaning to ensure data consistency and facilitate accurate analysis. The raw data often contained inconsistencies and variations in formatting, which could hinder the effectiveness of subsequent data processing and modeling techniques. By standardizing the data formats and removing inconsistencies, we prepared the dataset for robust analysis.

6. Conclusion

- **Key Findings**: The analysis highlighted that regular app updates, a positive sentiment from user reviews, and category selection are essential factors for an app's success. Popular categories like Games and Family continue to lead, while paid apps maintain high ratings with more niche audiences.
- **Impact for Developers**:
  - Developers should focus on user satisfaction, especially for free apps, to retain engagement.
  - Regular updates are recommended to improve user retention and ratings.
  - Paid apps have significant revenue potential, especially if targeted at specialized categories.
- **Potential for Business Decisions**: Insights gained from this analysis can help app developers make data-driven decisions on category focus, pricing strategies, and maintaining user engagement through regular updates.

7. Future Scope

While this analysis provides valuable insights into the Google Play Store ecosystem, several avenues for future research and development remain:

Advanced Feature Engineering:

- **Sentiment Analysis Granularity:** Delve deeper into user reviews to identify specific sentiment themes (e.g., performance issues, feature requests).
- **Textual Analysis:** Employ natural language processing techniques to extract keywords and topics from reviews.
- **Time-Series Analysis:** Analyze trends in app ratings, installs, and revenue over time.

Predictive Modeling:

- **Rating Prediction:** Develop models to predict app ratings based on factors like category, features, and user reviews.
- **Installation Prediction:** Build models to forecast app installations, considering marketing campaigns and app updates.
- **Churn Prediction:** Identify factors influencing user churn and develop models to predict user retention.

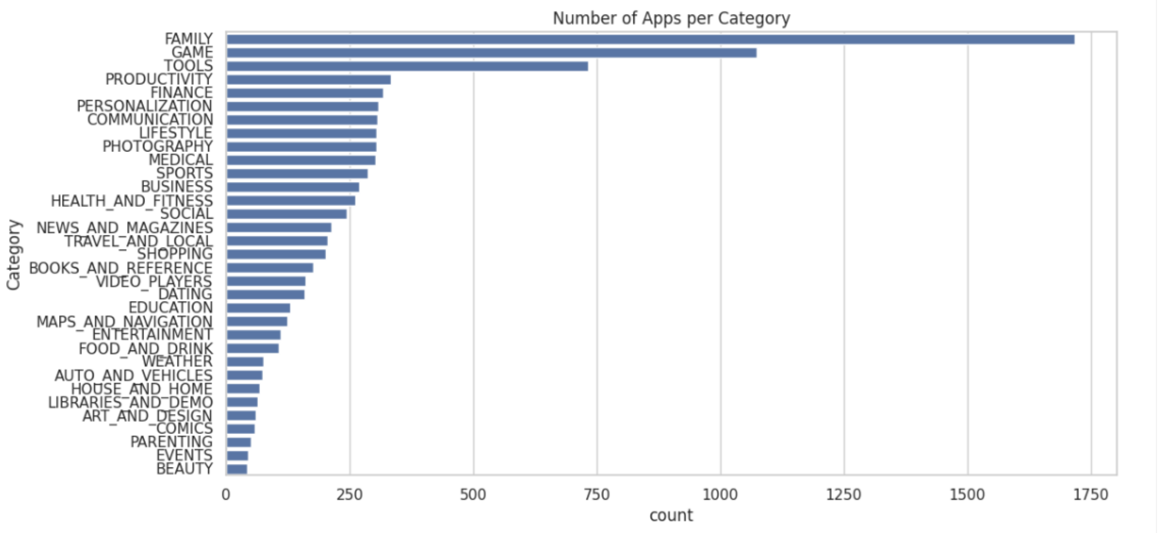
Incorporating Additional Data:

- **User Demographics:** Analyze user demographics (age, gender, location) to understand target audiences.
- **Device Information:** Consider device-specific factors (OS version, screen size) to optimize app performance.
- **App Store Optimization (ASO) Data:** Analyze ASO metrics (keyword rankings, search visibility) to improve app discoverability.

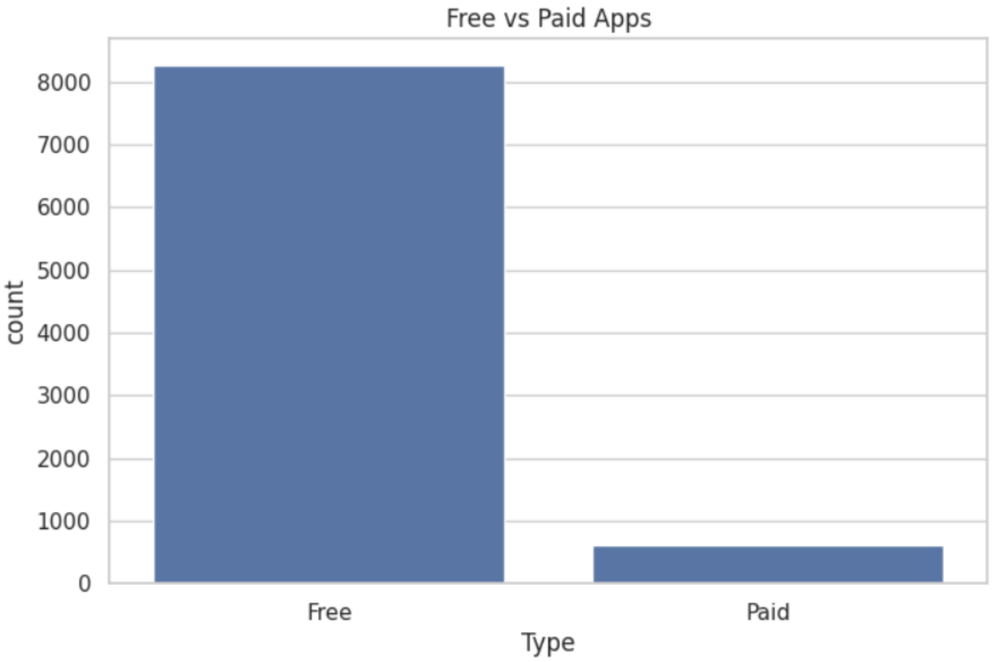
8. Appendix

- **GitHub Repository**: https://github.com/Saniya6112003/GooglePlayStoreAnalysis
- **Screenshots of Key Outputs**:

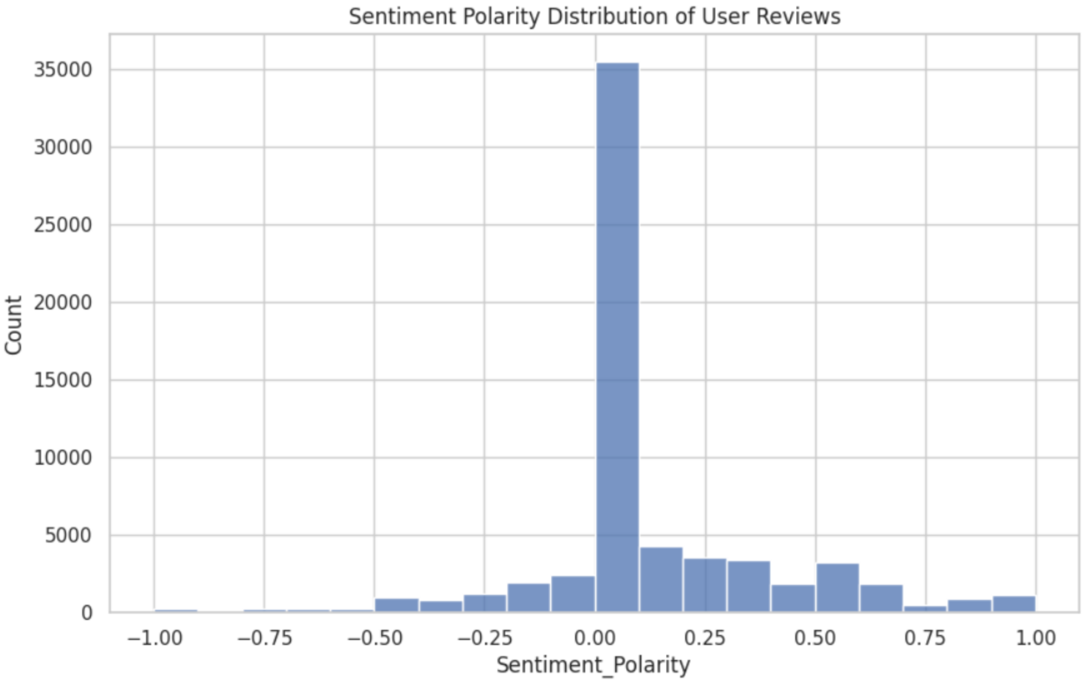
Category Distribution: A bar plot showing the number of apps per category.



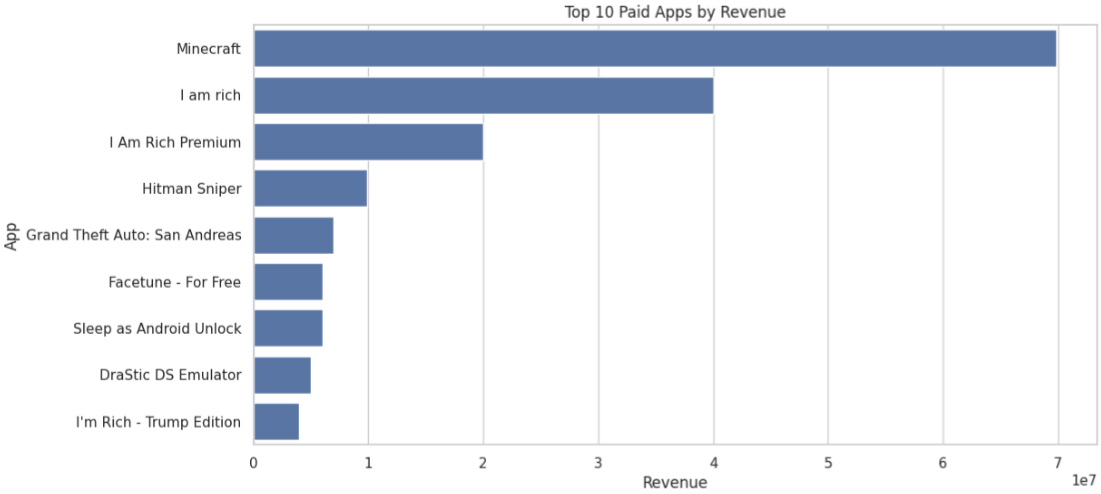
Free vs Paid Apps: Visual comparison of installs and ratings for free vs. paid apps.



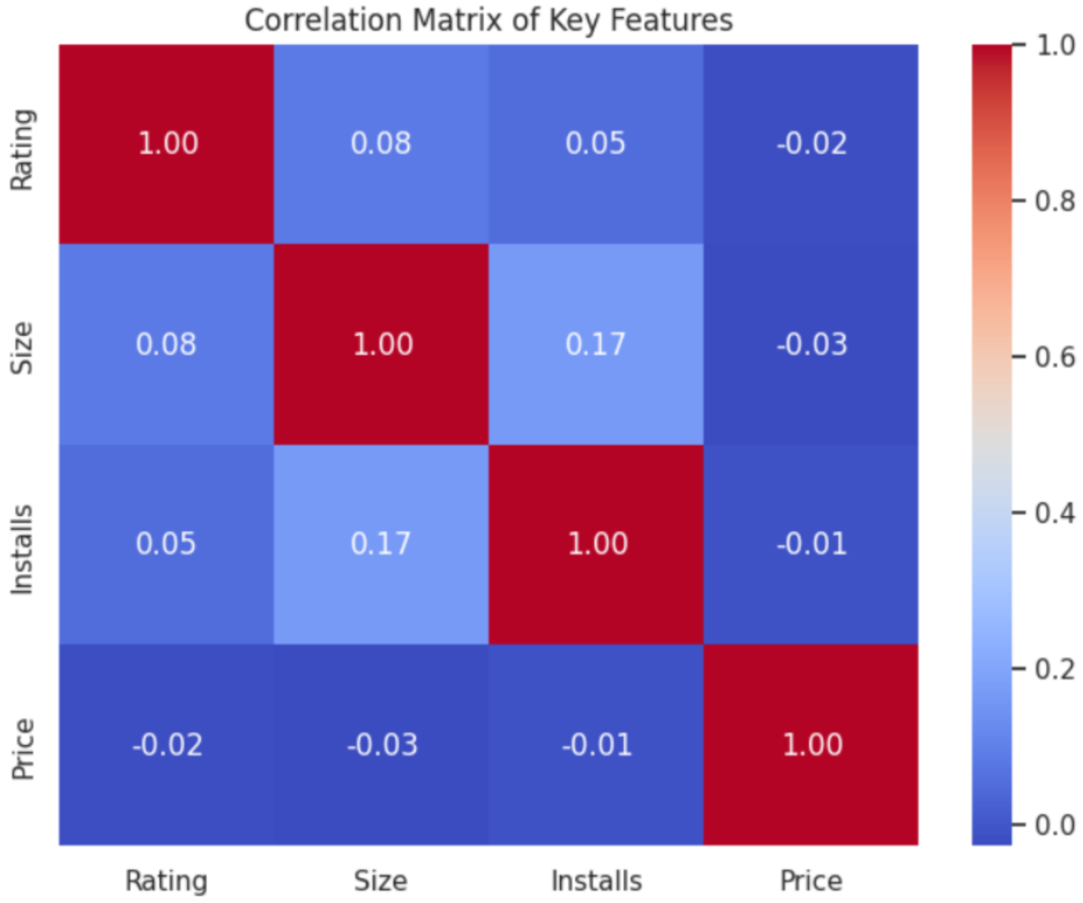
Sentiment Analysis Results: Histogram of sentiment polarity scores.



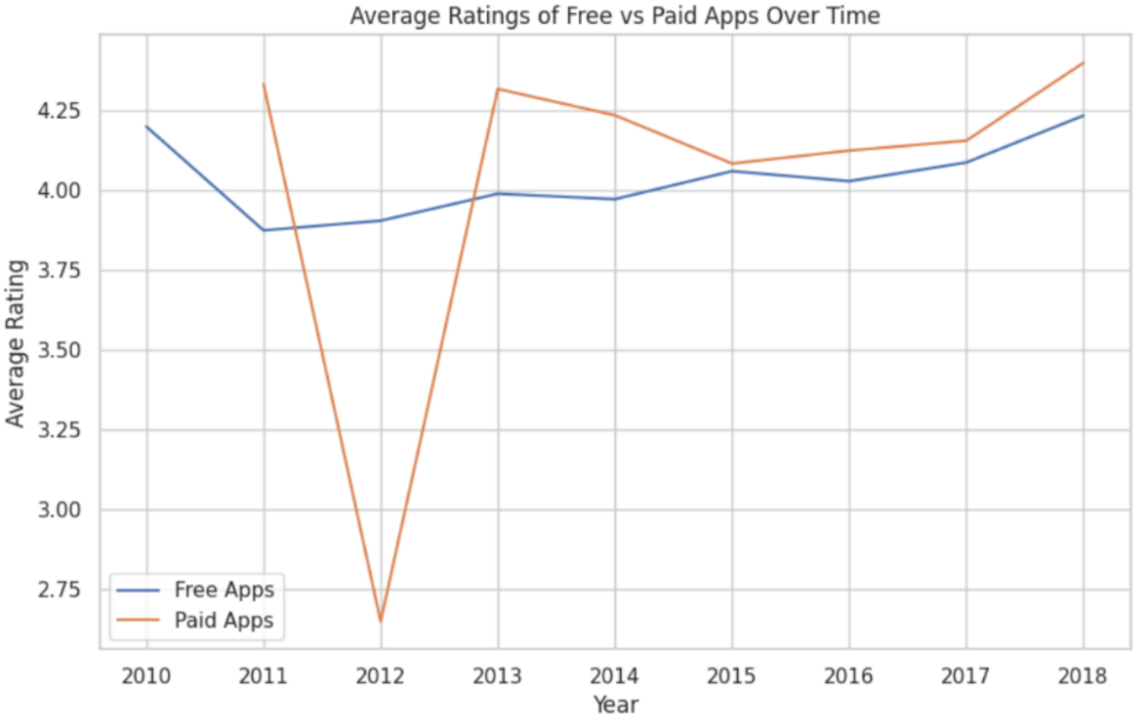
Revenue Insights: Bar plot of top revenue-generating paid apps.



Correlation Heatmap: Showing relationships between numerical variables (e.g., installs, rating, price).



Paid vs free apps: Performance of paid vs free apps over time.


Average Ratings of Free vs Paid Apps Over Time
Average Rating
Free Apps
Paid Apps
Year