Google Play Store Data Analysis

Project Report

Prepared by: Dhrutidipa Kabi

Date: August 08, 2025

Introduction

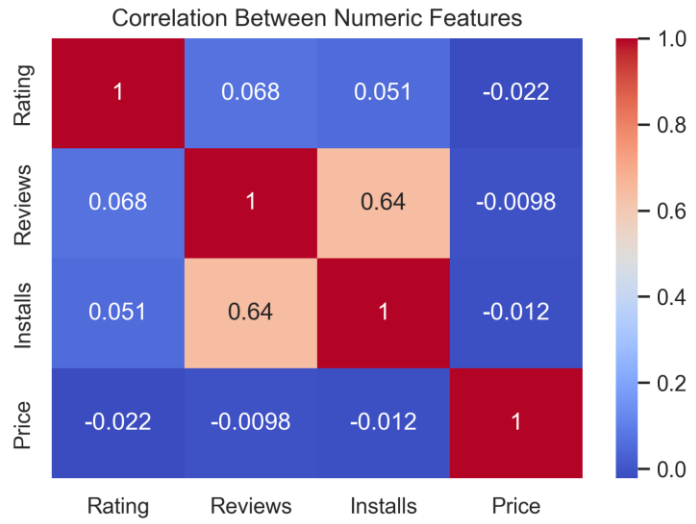
The Google Play Store is the largest platform for Android applications, hosting millions of apps across diverse categories such as Games, Education, Productivity, and Health. With billions of global users, it generates massive amounts of data in the form of app metadata, user reviews, ratings, download counts, and pricing information.

Analyzing this data is essential for several reasons:
• Understanding Market Trends – Identifying which categories are most popular and how trends shift over time.
• Evaluating User Satisfaction – Using ratings and sentiment from reviews to gauge app quality and reception.
• Optimizing App Performance – Discovering the relationship between app size, price, installs, and ratings to improve adoption.
• Strategic Decision-Making – Helping developers decide whether to offer apps for free or paid, and in which categories to compete.

This project systematically processes and analyzes Google Play Store datasets from Kaggle, covering both app listings and user reviews. The workflow includes:
1. Data Cleaning – Removing duplicates, handling missing values, and converting formats for numerical analysis.
2. Exploratory Data Analysis (EDA) – Examining distributions, relationships, and category dominance.
3. Visualization – Creating charts to visually represent patterns in installs, ratings, and app types.
4. Sentiment Analysis – Applying Natural Language Processing (NLP) techniques (VADER) to classify user feedback as positive, negative, or neutral.
5. Predictive Modeling – Building a simple regression model to estimate ratings based on app features and sentiment.

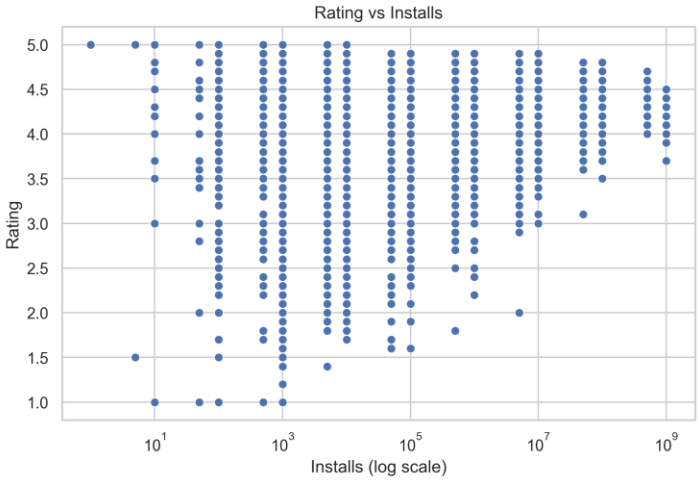
The insights generated from this analysis can assist app developers, marketers, and business strategists in making data-driven decisions to maximize user engagement, improve ratings, and stay competitive in the rapidly evolving mobile application market.

Figure 1: Correlation Heatmap of Numeric Features



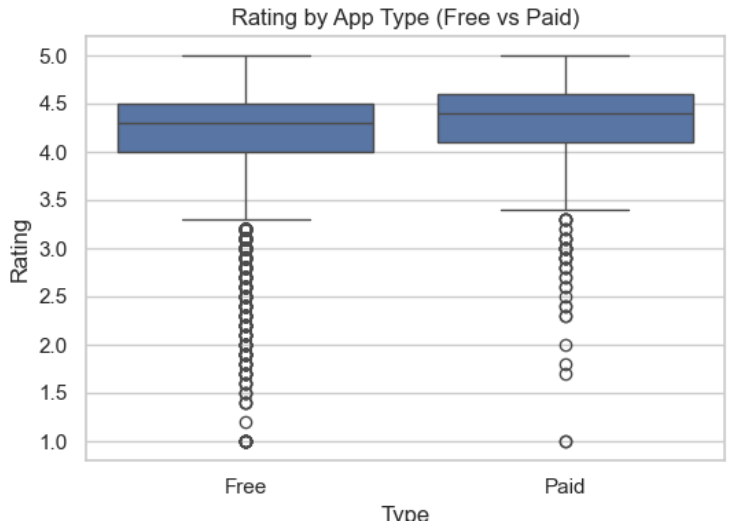
This heatmap shows the correlation between key numerical variables in the dataset, such as installs, reviews, price, and ratings. A strong positive correlation between installs and reviews suggests that popular apps often receive more feedback. Understanding these relationships helps identify which features most influence app success.

Figure 2: Installs vs. Ratings (Log Scale)



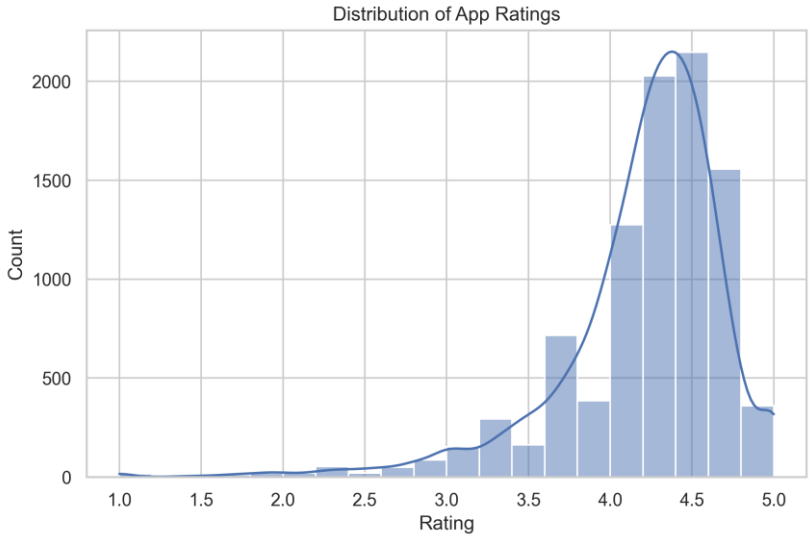
This scatter plot compares the number of installs (log scale) against app ratings. Highly installed apps generally maintain ratings above 4.0, although outliers exist. This indicates that high downloads do not always guarantee user satisfaction.

Figure 3: Ratings by App Type (Free vs Paid)



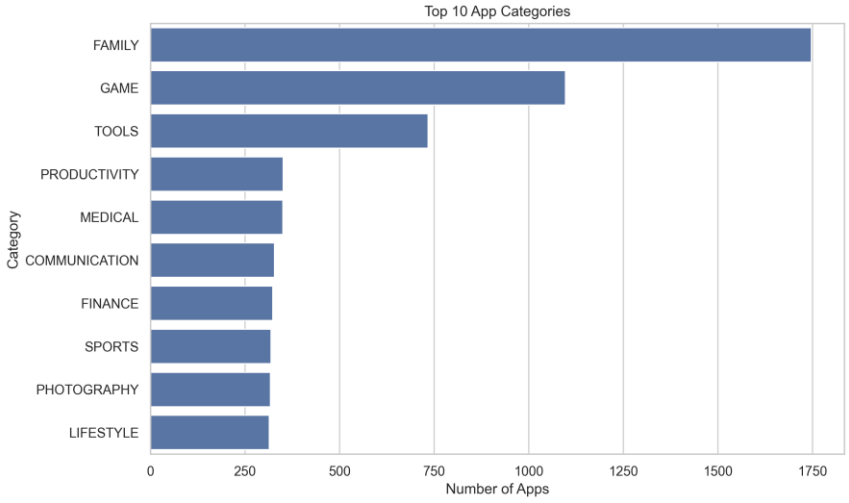
This box plot compares ratings between free and paid apps. Paid apps tend to have slightly higher median ratings, suggesting better quality or fewer ads. This insight informs pricing strategy decisions.

Figure 4: Rating Distribution



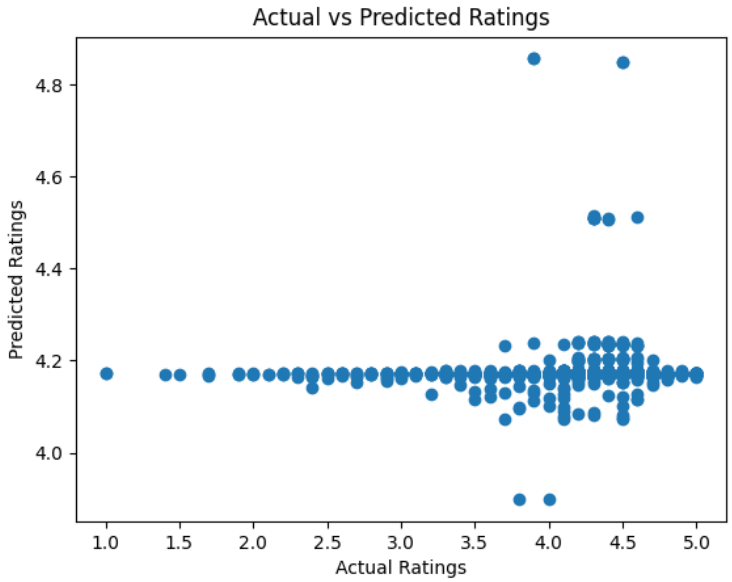
This histogram shows that most apps have ratings between 4.0 and 4.5, indicating generally positive user experiences. Ratings below 3.0 often indicate poorly optimized or unpopular apps.

Figure 5: Top 10 App Categories



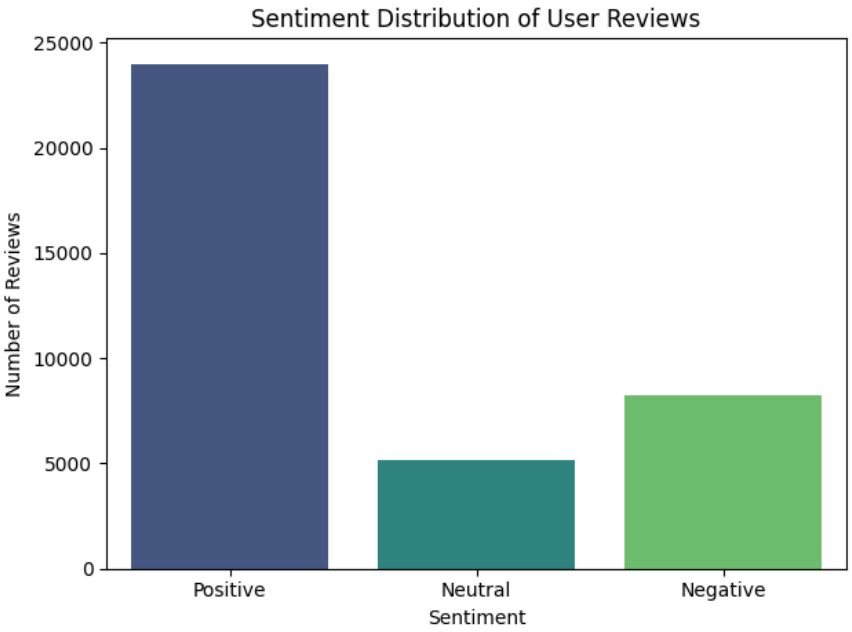
This bar chart displays the most common app categories. Family, Game, and Tools dominate, reflecting high user demand for entertainment and utility apps. Targeting these categories can mean higher reach but also more competition.

Figure 6: Actual vs. Predicted Ratings



This scatter plot compares actual ratings with predictions from the regression model. Points near the diagonal indicate accurate predictions, while deviations highlight model limitations.

Figure 7: Sentiment Distribution of User Reviews



This chart shows positive, neutral, and negative sentiments from user reviews using VADER sentiment analysis. Positive reviews dominate, but tracking negative sentiment helps identify areas needing improvement.

Key Insights

1. Family and Game apps dominate in number and installs.
2. Free and mid-sized apps (20–50MB) are most popular.
3. Positive sentiment correlates with higher ratings.
4. Paid apps mainly belong to Education and Productivity categories.
5. Reviews and sentiment are strong drivers of app success.

Recommendations

- Focus on free apps with in-app monetization.
- Optimize app size for better installation probability.
- Encourage user feedback to boost sentiment.
- Monitor and respond to negative reviews quickly.

Tools & Libraries

Python, Pandas, NumPy, Matplotlib, Seaborn, NLTK, Scikit-learn, Jupyter Notebook

Limitations & Future Work

While this analysis provides valuable insights, there are some limitations:

• Dataset Scope – The dataset is a snapshot and may not reflect the most recent apps or market changes.

• Sentiment Analysis Accuracy – VADER works well for short texts, but sarcasm, slang, and mixed sentiments may not be captured accurately.

• Model Simplicity – The regression model used here is basic and may not fully capture complex relationships between features.

Future work can include:

• Including Time-Series Trends – Analyze how app popularity changes over time.

• Advanced NLP Techniques – Use transformer-based models (e.g., BERT) for better sentiment classification.

• Feature Engineering – Extract more nuanced features such as app update frequency or developer reputation.

• Category-Level Modeling – Build separate prediction models for different app categories to improve accuracy.

Conclusion

This project successfully cleans, analyzes, visualizes, and models Google Play Store data to identify factors influencing app success. By combining data analysis with sentiment evaluation, the study provides actionable insights for app developers and businesses aiming to improve user satisfaction, retention, and overall market performance.