Google Play Store Data Analysis — Detailed Summary Report

K. Kiran Sai Karthik

Contents

1 Introduction

This report presents a comprehensive analysis of the Google Play Store dataset, comprising 10,841 apps with 13 attributes. The analysis involves data cleaning, exploratory data analysis, statistical testing, visualizations, and predictive modeling to uncover insights about app ratings, categories, pricing strategies, and user engagement. The workflow follows a structured approach, from importing libraries to building enhanced regression models with engineered features.

2 Importing Required Libraries

To initiate the analysis, the following Python libraries were imported:

- **Pandas and NumPy**: For efficient data manipulation and numerical computations.
- **Matplotlib and Seaborn**: For creating high-quality visualizations to identify patterns.
- **Scipy**: For conducting statistical tests such as T-tests and Z-tests.
- **Scikit-learn**: For building regression models, data splitting, and model evaluation.
- **Warnings**: Suppressed to ensure cleaner outputs during the workflow.

These libraries provide the foundation for a robust end-to-end data science pipeline.

3 Data Cleaning and Preprocessing

This step ensures the dataset is consistent and ready for analysis:

- **Dataset Loaded**: Initially consisted of 10,841 rows and 13 columns.
- **Missing Values**:
  - *Rating*: Filled with the median value.
  - *Android Ver* and *Current Ver*: Filled with the most frequent value (mode).
  - Remaining null values were dropped to maintain data quality.
- **Data Type Conversions**:
  - *Reviews*, *Installs*, and *Price*: Converted to numeric formats.
  - *Size*: Converted from strings (e.g., ”19M”, ”14k”) to float (MB).
  - *Last Updated*: Parsed as datetime for temporal analysis.
- **Duplicates**: Removed 483 duplicate records.
- **Final Dataset**: 10,356 rows with consistent, numeric, and well-structured data.

4 Basic Data Insights

Exploratory insights were derived to understand the dataset's structure:

- Summary statistics provided distributions and ranges for key variables.
- Unique value counts:
  - 33 unique app categories.
  - 6 unique content ratings.
  - 2 app types (Free, Paid).

These statistics highlight the diversity and scope of the dataset.

5 Visualization 1 (Distribution of App Ratings)

A histogram with a Kernel Density Estimate (KDE) plot was used to visualize the distribution of app ratings:

- Most ratings clustered between 4.0 and 4.5.
- Few apps were rated below 3.0 or above 4.8.
- This confirms that the majority of apps receive positive ratings.

6 Visualization 2 (Ratings by Content Rating)

A boxplot displayed how app ratings vary by content rating:

- Categories like *Everyone*, *Teen*, and *Mature 17+* showed distinct rating distributions.
- This analysis assessed the impact of age-based content restrictions on user ratings.

7 Visualization 3 (Correlation Heatmap)

A heatmap visualized pairwise correlations between *Rating*, *Reviews*, *Size*, *Installs*, and *Price*:

- **Positive correlation**: Between *Reviews* and *Installs*.
- **Weak correlation**: Between *Price* and *Ratings*.
- This guided feature selection for regression modeling.

8 Visualization 4 (Top 10 App Categories by Frequency)

A bar chart displayed the top 10 app categories by frequency:

- Dominant categories: *FAMILY*, *GAME*, and *TOOLS*.
- Provided insight into the most prevalent app types in the Play Store.

9 Central Tendency and Spread

Key statistical metrics were computed for numeric columns (*Rating*, *Reviews*, *Size*, *Installs*, *Price*):

- Metrics: Mean, Median, Mode, and Standard Deviation.
- These measures summarized the distribution and variability of key variables.

10 Free vs Paid App Analysis

A pie chart showed the distribution of free vs paid apps:

- Approximately 93% of apps were free, indicating a freemium-dominated ecosystem.
- This set the stage for comparing ratings between free and paid apps.

11 Top 10 Apps by Reviews

The top 10 apps with the highest number of reviews were identified:

- Apps included: *Facebook*, *WhatsApp*, *Instagram*, and *Clash of Clans*.
- Each app's review count and rating were listed, highlighting the most popular apps.

12 Average Rating by Category

The dataset was grouped by category to compute average ratings:

- Top-rated categories:
  - *EVENTS*: 4.39
  - *EDUCATION*: 4.37
  - *ART_AND_DESIGN*: 4.35
- This revealed categories with higher user satisfaction.

13 Price Distribution of Paid Apps

A histogram visualized the price range of paid apps:

- Most apps were priced below $10, with outliers up to $400.
- This highlighted monetization trends and pricing strategies.

14 T-Test (Free vs Paid App Ratings)

A T-test compared ratings of free and paid apps:

- **Null Hypothesis**: No difference in mean ratings.
- **Result**: Significant difference ($p \approx 0.00017$).
- **Conclusion**: Paid apps have slightly different ratings than free apps.

15 Z-Test (Mean Rating vs 4.0)

A Z-test assessed whether the average rating differed from an industry baseline of 4.0:

- **Result**: Statistically significant difference ($p < 0.05$).
- **Conclusion**: App ratings are slightly higher than the expected average.

16 Regression Model (Base Features)

A linear regression model was built using *Reviews*, *Size*, *Installs*, and *Price* to predict *Rating*:

- Evaluated using:
  - R^2 Score
  - Mean Squared Error
  - Residual Plot
  - Cross-Validation
- Provided a baseline for predicting app quality.

17 Feature Engineering

New features were created to enhance model accuracy:

- *App Age* (in days)
- *Log_Reviews*, *Log_Installs*
- *High Price* binary flag
- *Size Bins* (Small, Medium, Large, Very Large)
- *Reviews* $\times$ *Rating* interaction term
- One-hot encoding applied for categorical bins.

These features added business logic and granularity.

18 Regression with Feature Engineering and Categories

The final model incorporated engineered features and one-hot encoded *Category*:

- Retrained and tested on the enhanced dataset.
- Improved R^2 Score confirmed the effectiveness of additional features.
- Captured complex relationships in the dataset.

19 Conclusion

This analysis provided actionable insights into the Google Play Store ecosystem, highlighting trends in app ratings, categories, pricing, and user engagement. Statistical tests and visualizations uncovered significant patterns, while regression models demonstrated the predictive power of engineered features. The results can inform app developers and marketers about user preferences and market dynamics.