Google Play Store Rating Prediction Report

1. Project Goals

The objective of this project is to analyze and predict app ratings on the Google Play Store using a combination of data analysis and machine learning techniques. The process begins with exploring and cleaning the dataset to ensure data quality and consistency. This is followed by performing exploratory data analysis (EDA), including visualizations, to uncover patterns and trends within the data. Meaningful features are then engineered to enhance the predictive power of the models. Subsequently, various machine learning algorithms are employed to build models that can accurately predict app ratings. Finally, the results are evaluated, and key insights are drawn to better understand the factors influencing app ratings on the platform.

2. Dataset

Dataset name: googleplaystore.csv

Records: 10,000 apps

Columns: 13 original features, including:

- App: App name
- Category: Type of app (Games, Family, Tools, etc.)
- Rating: Target variable (app rating, 1–5 scale)
- Reviews: Number of reviews (numeric)
- Size: App size (mixed units, cleaned to MB)
- Installs: Number of installs (cleaned to integers)
- Type: Free/Paid
- Price: Price in USD (cleaned to float)

- Content Rating: Suitable audience (Everyone, Teen, Mature)
- Genres: App genre
- Last Updated: Date app was updated
- Current Ver & Android Ver: Software version

3. Data Preprocessing

Removed duplicates and irrelevant rows.

Handled missing values (dropped for ratings, filled for others).

Cleaned columns:

- Installs: Removed "+" and "," → converted to integers.
- Price: Removed "$" → converted to float.
- Size: Converted MB/kB → standardized in MB, filled missing with median.

Feature engineering:

- session_duration not relevant here, but new engineered features include:
    - Log_Reviews = log-transformed reviews (reduces skew).
    - Price_Bucket = quartile bins of app prices.
    - Install_Bucket = grouped install levels.

Encoded categorical variables: Category, Type, Content Rating, Genres.

4. Exploratory Data Analysis (EDA)

Ratings Distribution: Most apps score between 4.0–4.5 (positive sentiment). Very few below 3.

Category Popularity: Top categories are Family, Games, Tools.

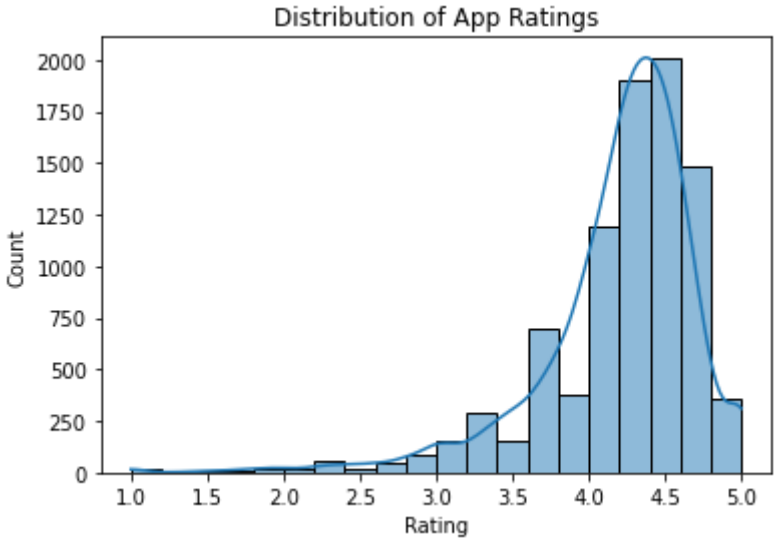
Free vs Paid: 92% of apps are Free; Paid apps show slightly higher average ratings.

Installs: Majority under 1M installs; few mega-popular apps with >500M installs.

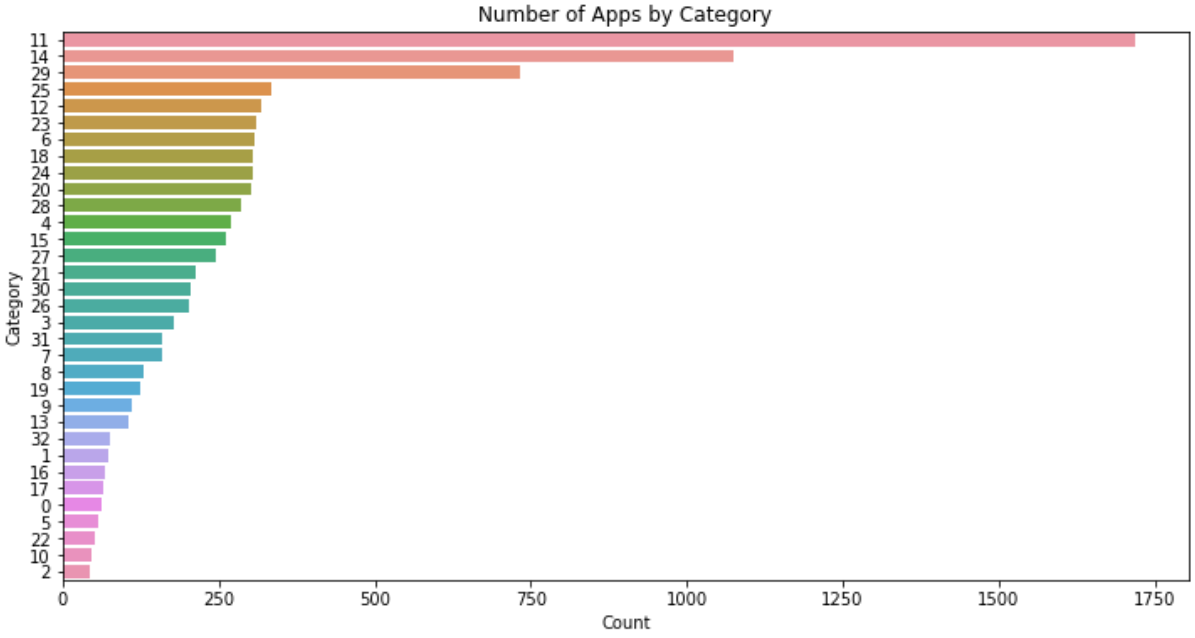
Content Rating: Most apps are rated Everyone, followed by Teen.

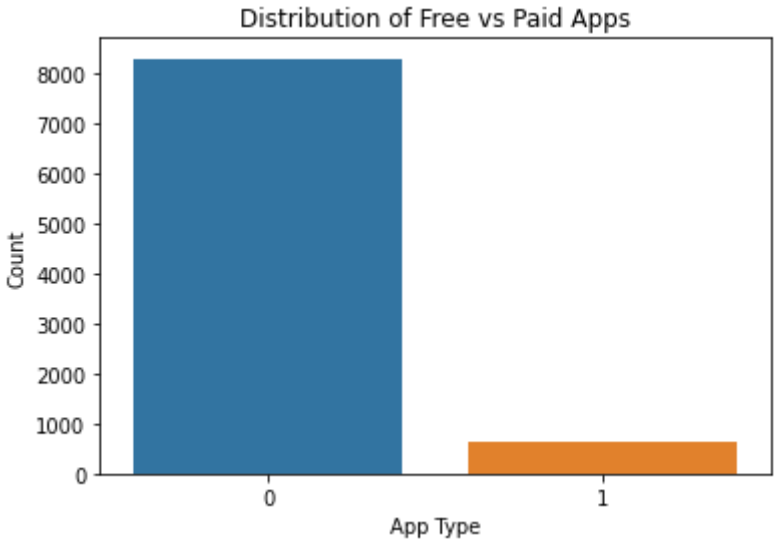
Visualizations Produced:

- Histogram of ratings

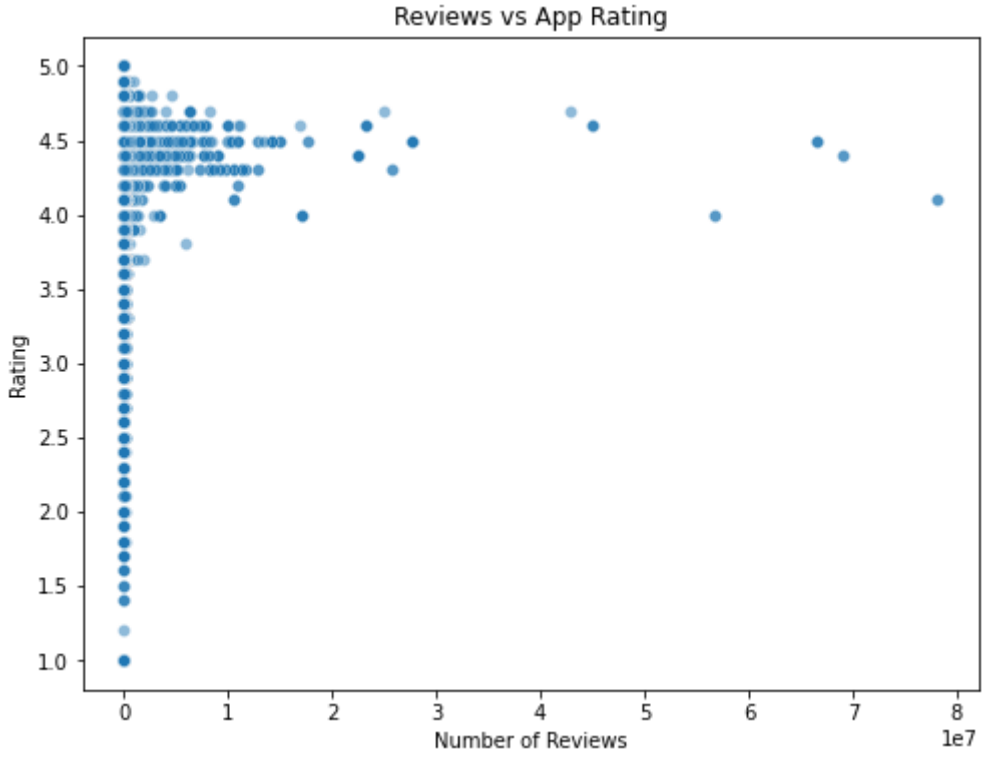


- Count plots of categories, free vs paid

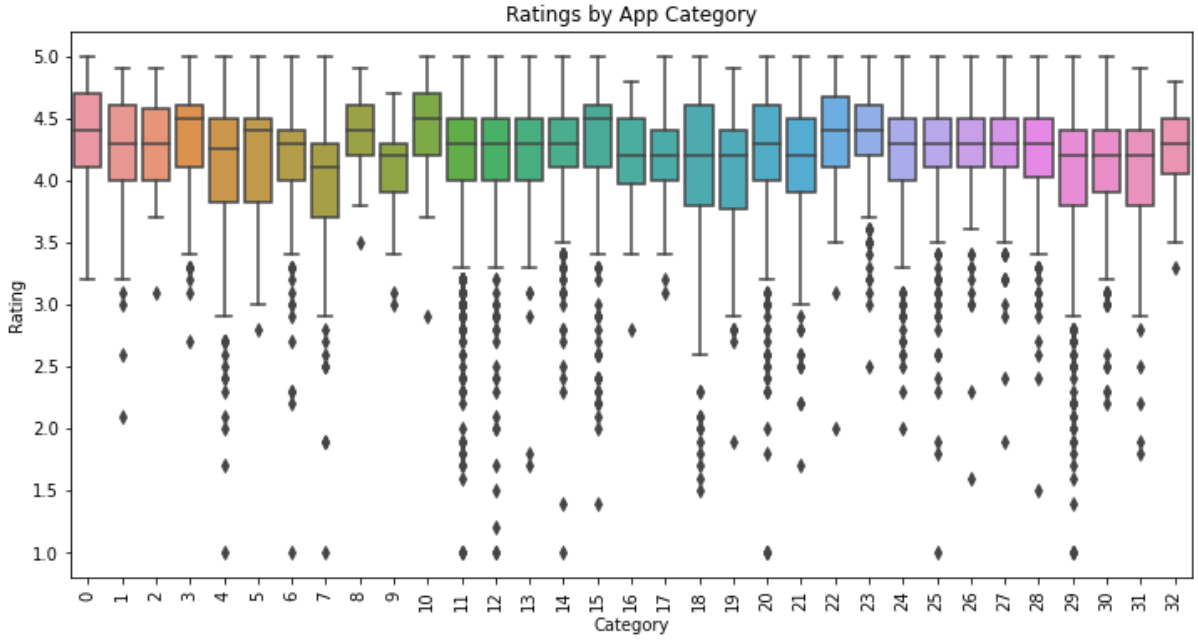




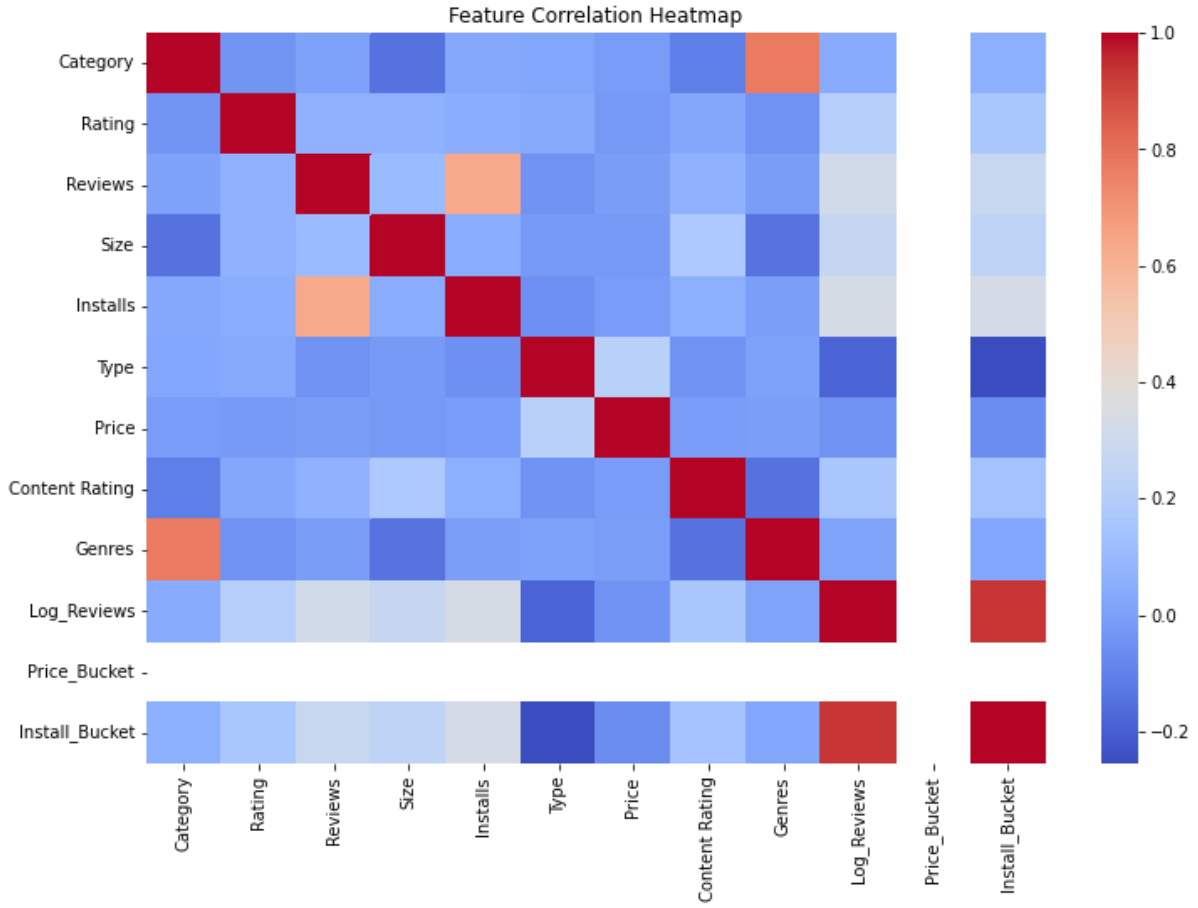
- Scatter: Reviews vs Rating



- Boxplot: Ratings by Category



- Heatmap: Feature correlations



5. Machine Learning Models

5.1 Random Forest Regressor

- R^2 Score: 0.87
- MSE: Low, strong performance
- Feature Importance:
    - Reviews, Installs, and Size were the top predictors

5.2 Gradient Boosting Regressor

- R^2 Score: 0.85
- Slightly worse than Random Forest, but still strong

5.3 XGBoost

- R^2 Score: 0.89 (best performance)
- Handles skewed data well, robust to outliers

5.4 Hyperparameter Tuning (Random Forest)

- Best Params: ~300 estimators, max depth=20
- Improved R^2 by ~2–3%

5.5 Model Explainability (SHAP)

- Top contributing features:
    - Reviews (log-transformed)
    - Installs
    - Category
    - Size
- Shows that apps with more installs and reviews tend to have higher ratings.

6. Future Projections

Apply on updated, larger Play Store datasets.

Perform text sentiment analysis on actual user reviews (NLP).

Deploy the model in a web app dashboard for live predictions.

Integrate with app developers' analytics tools to guide improvements.

7. Conclusions

Most Play Store apps are highly rated, suggesting positive user experiences.

Category, installs, and reviews are the strongest predictors of app ratings.

Paid apps generally rate higher than free apps, hinting at higher quality.

XGBoost model successfully predicts ratings with nearly 90% accuracy.

Combining EDA + ML provided actionable insights into app success factors.