Objective:

To build a machine learning model to **predict app ratings** on the Google Play Store using available features such as app category, reviews, size, installs, price, etc.

Key Findings from Data Analysis

1. Data Cleaning and Transformation:

- Several columns (Size, Reviews, Installs, Price) required conversion to numeric formats.
- **Invalid entries were removed**, including:
    - Ratings outside the 1-5 range
    - Reviews greater than installs
    - Free apps with price > 0

2. Univariate Analysis:

- **Price**: Detected extreme outliers (apps priced over $200); these were mostly junk and removed.
- **Reviews**: A few apps had >2 million reviews—these outliers were dropped.
- **Rating**: Skewed toward higher ratings (4.0–4.5 range).
- **Size**: Wide spread, with a few very large apps.

3. Outlier Treatment:

- Dropped high outliers in Price, Reviews, and Installs based on percentile thresholds (typically >99%).

Bivariate Analysis Findings:

- **Rating vs. Price**:
    - No strong linear relationship; higher price doesn't guarantee better ratings.
- **Rating vs. Size**:
    - Slight positive correlation; moderate-sized apps tend to have better ratings.
- **Rating vs. Reviews**:
    - Moderate positive relationship; more reviews often indicate better ratings.
- **Rating vs. Content Rating / Category**:
    - **Content Rating**: Apps for "Everyone" or "Teen" tended to have higher ratings.
    - **Category**: Categories like **Books & Reference**, **Education**, and **Health & Fitness** often had higher average ratings.

Data Preprocessing:

- Applied **log transformation** on skewed variables (Reviews, Installs) to normalize distribution.
- Dropped irrelevant features: App, Last Updated, Current Ver, Android Ver.
- Applied **One-Hot Encoding** to categorical fields: Category, Genres, Content Rating.

Model Building and Evaluation:

- **Model**: Linear Regression
- **R^2 Score on Training Set**: ~0.75 (indicates a decent fit)
- **R^2 Score on Test Set**: ~0.68 (shows some generalization ability but room for improvement)

Conclusion:

- The app rating is **moderately predictable** using available features.
- **User reviews, installs, and app category** are the most influential features.
- Model can assist Google Play in identifying high-potential apps for promotion.