Write Up on App Rating Prediction

In this project, I built a model to predict app ratings using the Google Play Store dataset. The goal was to understand which features influence app ratings and help identify apps that may perform well.

First, the dataset was loaded using pandas. I checked for missing values and removed rows with null data to ensure accuracy. Then I cleaned the dataset. The Size column had values in MB and KB, so I converted them into numeric format. The Reviews column was converted from text to numbers. In the Installs column, commas and "+" symbols were removed before converting the values into integers. The dollar symbol was removed from the Price column and converted into numeric format.

After cleaning, I applied sanity checks. I kept ratings only between 1 and 5, removed records where reviews were greater than installs, and filtered free apps with price greater than zero.

Next, I performed univariate and bivariate analysis using boxplots and histograms. Some extreme outliers were found, especially in Price, Reviews, and Installs. These were removed to improve model performance.

In the preprocessing step, I applied log transformation to Reviews and Installs to reduce skewness. Unnecessary columns were removed, and categorical columns were converted into numeric values using dummy encoding.

The data was split into 70% training and 30% testing sets. A Linear Regression model was trained. The model achieved a Train R^2 score of about 0.16 and a Test R^2 score of about 0.11. This shows moderate predictive ability.

Overall, the project successfully demonstrates data cleaning, analysis, preprocessing, and model building.