App Rating Prediction

Univariate Analysis:

- **Price:** From the boxplot, it is clear that price > 50 can be considered outliers
- **Reviews:** Records with reviews > 2 Million can be considered as outliers
- **Size:** The distribution of Size variable is right skewed. The data has high no of small sized apps
- **Install:** 95th percentile i.e., installs > 50000000 can be considered outliers and can be removed from the analysis
- **Rating:** The distribution of 'Ratings' is left skewed. The no of high rated apps present in the data is more than the low rated apps

Bivariate Analysis:

- **Rating vs Price:** There is no clear pattern. High priced apps have better ratings(though it is not a linear relationship). But no of high priced apps present in the data is very low
- **Rating vs Size:** No clear pattern is observed. Heavier apps have better ratings(again the relation is not linear)
- **Rating vs Reviews:** Highly reviewed apps have better ratings. However, there is no clear pattern present
- **Rating vs Content Rating:**
  - There is significant overlap in the ratings among all the categories of 'Content Rating'.
  - 'Adults only 18+' has the highest median value of ratings. But the no of apps present in this category is very less
- **Rating vs Category:** Most of the categories have similar ratings. Apps in the 'Books and reference', 'Events' categories have the highest medians

- Linear Regression technique was used to build the model

Results:

- R^2 of train data = 0.14892699986877223
- R^2 of test data = 0.15016339300746961
- Root Mean Squared Error= 0.521288642802937