Adv EDA + Regression (problem statement)

Objective:

To conduct a comprehensive Exploratory Data Analysis (EDA) and build a Linear Regression model to predict app ratings on the Google Play Store using two datasets: **googleplaystore.csv** and **googleplaystore_user_reviews.csv**.

Datasets:
googleplaystore.csv
googleplaystore_user_reviews.csv.zip

Description:

This project aims to give students hands-on experience with real-world data analysis and predictive modeling. Students will explore, visualize, clean, and preprocess the dataset, then apply Linear Regression to predict app ratings.

Data Description:

1. **googleplaystore.csv:** Contains details of applications on Google Play Store. It includes 13 features that describe a given app.
2. **googleplaystore_user_reviews.csv:** Contains the first 'most relevant' 100 reviews for each app, with each review text/comment pre-processed and attributed with three new features - Sentiment, Sentiment Polarity, and Sentiment Subjectivity.

Assignment Tasks:

Part 1: Exploratory Data Analysis (EDA)

1. **Data Loading:** Import the datasets into a suitable Python environment.
2. **Data Inspection:** Examine the datasets for dimensions, data types, and summary statistics.
3. **Data Cleaning:** Handle missing values, incorrect data types, and outliers.
4. **Data Visualization:** Create visualizations to understand distributions, relationships, and patterns in the data. Suggested plots include histograms, scatter plots, box plots, and heatmaps.
5. **Feature Engineering:** Generate new features if necessary, based on the insights gained from EDA.

Part 2: Data Preprocessing

1. **Merge Datasets:** If relevant, merge the two datasets on a common key.
2. **Handling Categorical Data:** Use techniques like one-hot encoding or label encoding for categorical variables.
3. **Data Splitting:** Split the dataset into training and test sets.

Part 3: Linear Regression Model

1. **Model Building:** Build a Linear Regression model to predict the rating of apps.
2. **Model Evaluation:** Evaluate the model using appropriate metrics such as R-squared, Mean Squared Error (MSE), or Mean Absolute Error (MAE).
3. **Interpretation:** Interpret the model coefficients to understand the impact of different features on app ratings.

Part 4: Conclusion and Recommendations

1. **Insights:** Summarize key insights from the EDA and the predictive model.
2. **Recommendations:** Provide recommendations to app developers based on your findings.