Data Cleaning, EDA, Feature Engineering, and Modeling Report

1. Dataset Loading and Inspection:

- The dataset was loaded from a CSV file containing overseas trade indexes.
- The dataset structure was inspected using data info and sample rows.

2. Data Cleaning:

a. Handling Missing Values:

- Rows with missing critical values were removed.
- Numerical missing values were imputed with the column mean.

b. Removing Duplicate Records:

- Duplicate rows were identified and removed.

c. Outlier Detection and Treatment:

- Outliers were detected using the IQR method and removed.

d. Data Standardization:

- Numerical features were standardized using z-score normalization.
- Categorical features were standardized by converting text to lowercase and trimming spaces.

3. Exploratory Data Analysis (EDA):

- Histograms were plotted to show the distribution of numerical features.
- Boxplots were used to visualize data spread and detect outliers.
- Scatter plots were created to examine pairwise relationships.
- A correlation matrix was computed to identify relationships between numerical variables.

4. Feature Engineering:

- Categorical variables were manually encoded into numerical values.

- The dataset was split into training (80%) and testing (20%) sets, with the first numerical column as the target.

5. Modeling: Linear Regression:

- A simple linear regression model was built manually using a pseudoinverse approach.
- The model performance was evaluated using Mean Squared Error (MSE) and R-Squared (R^2).

6. Final Output:

- The cleaned and processed dataset was saved as a new CSV file for further use.

Conclusion:

This report outlines a systematic approach to prepare and analyze the dataset. The steps taken ensure that the data is clean, well-explored, and properly modeled, providing a strong foundation for subsequent predictive analysis.