Research Project Proposal

How accurately can a linear regression model predict STEM degree completion rates at public universities in the United States based on demographic and institutional characteristics, and what are the key factors influencing these rates?

Files to Use:

1. **IPEDS Completions (C)**: Contains data on degree completions by field of study.
2. **IPEDS Institutional Characteristics (IC)**: Provides information about the institutions.
3. **IPEDS Demographic Characteristics (EFFY)**: Contains enrollment data by demographic characteristics.

What is your target variable? Provide some descriptive statistics about your target variable.

Steps

1. **Data Loading:** Load the IPEDS datasets and filter for public universities and STEM fields.
2. **Feature Engineering:** Select relevant features and target variables. Handle categorical variables using one-hot encoding and scale numerical features.
3. **Model Training:** Split the data into training and test sets. Using a pipeline, define and train a RandomForestRegressor model.
4. **Evaluation:** Evaluate the model using Mean Squared Error (MSE) and R-squared metrics. Plot the actual vs predicted completions