Research Project Proposal

**What factors affect STEM degree completion rates at public universities in the United States?**

Files to Use: all data will be for the **2022** Year

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?

Target variable target is AWLEVELC which is Award Level code going to filter all Steam Degrees

Provide some descriptive statistics about your target variable.

A computer screen shot of a black screen

Description automatically generated

A blue and white graph

Description automatically generated

Steps

1. **Data Loading:** Load the datasets from the provided files.
2. **Filter and Preprocess Data:** Filter the institutions to include only public universities. Identify the CIP codes corresponding to STEM fields and filter the completions dataset.
3. **Feature Engineering:** Select relevant features and target variables. Handle categorical variables using one-hot encoding and scale numerical features.
4. **Model Training and Evaluation:** Split the data into training and test sets. Define and train a LinearRegression model using a pipeline. Evaluate the model using Mean Squared Error (MSE) and R-squared metrics. Plot the actual vs predicted completions.
5. **Feature Importance:** Display the coefficients of the linear regression model to identify the impact of different features.

Member of the Group

Biniam Abebe