Data Preprocessing and Feature Engineering in Machine Learning

**Objective:**

This assignment aims to equip you with practical skills in data preprocessing, feature engineering, and feature selection techniques, which are crucial for building efficient machine learning models. You will work with a provided dataset to apply various techniques such as scaling, encoding, and feature selection methods including isolation forest and PPS score analysis.

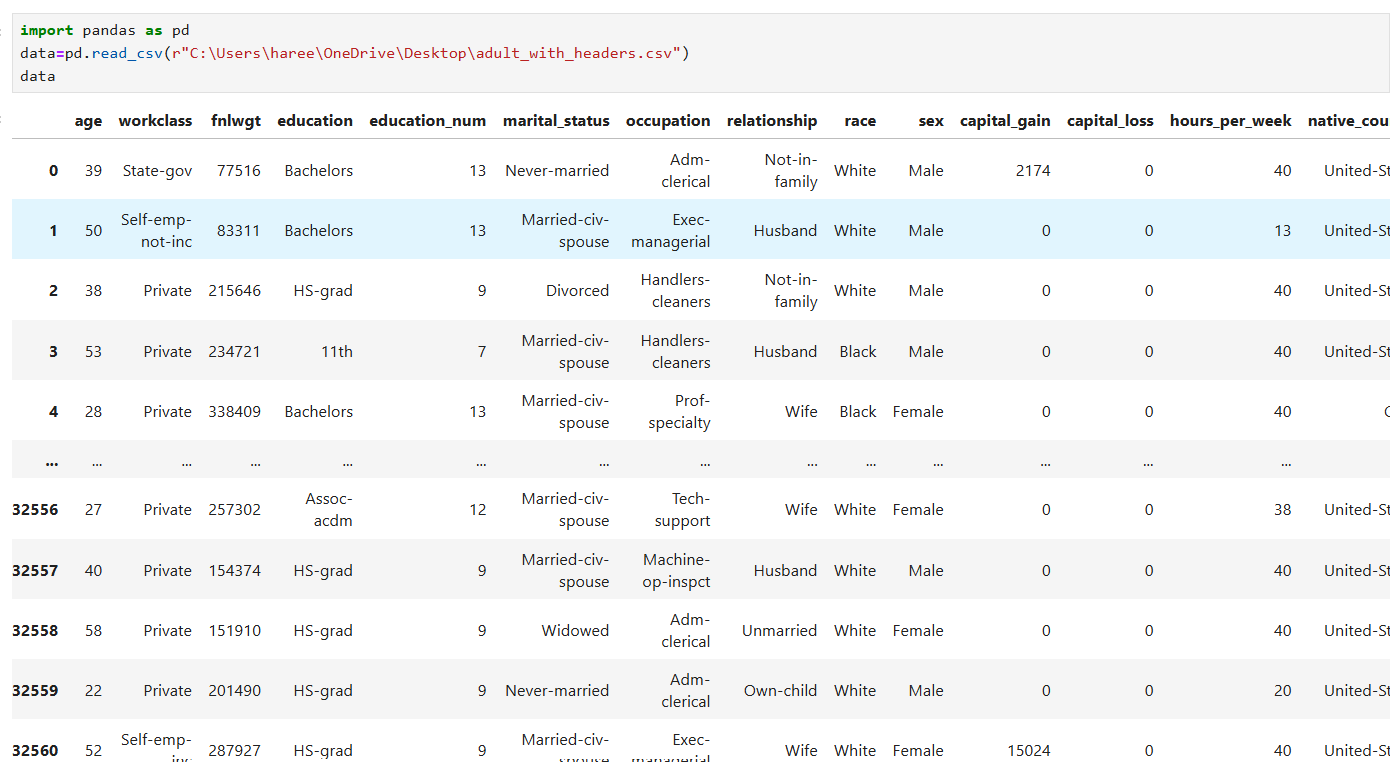
Dataset:

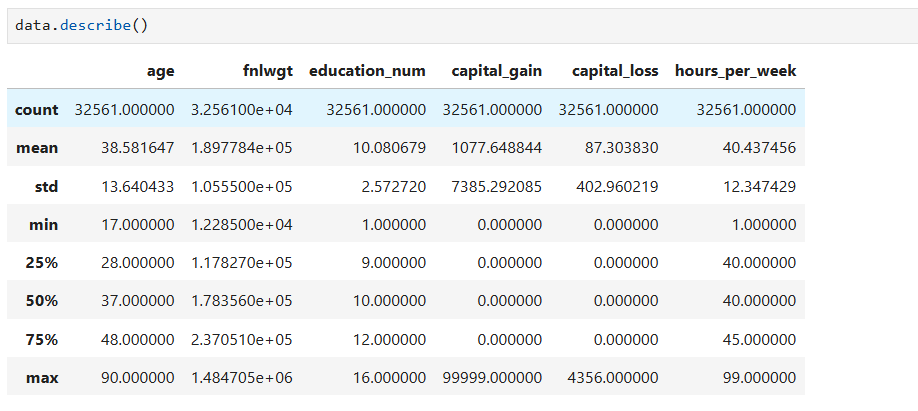
Given "Adult" dataset, which predicts whether income exceeds $50K/yr based on census data.

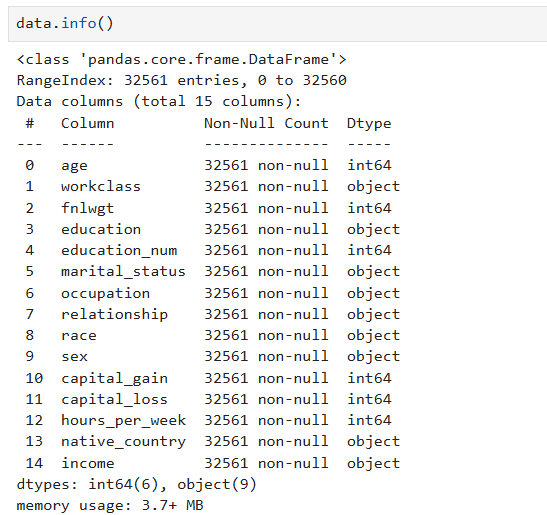
Tasks:

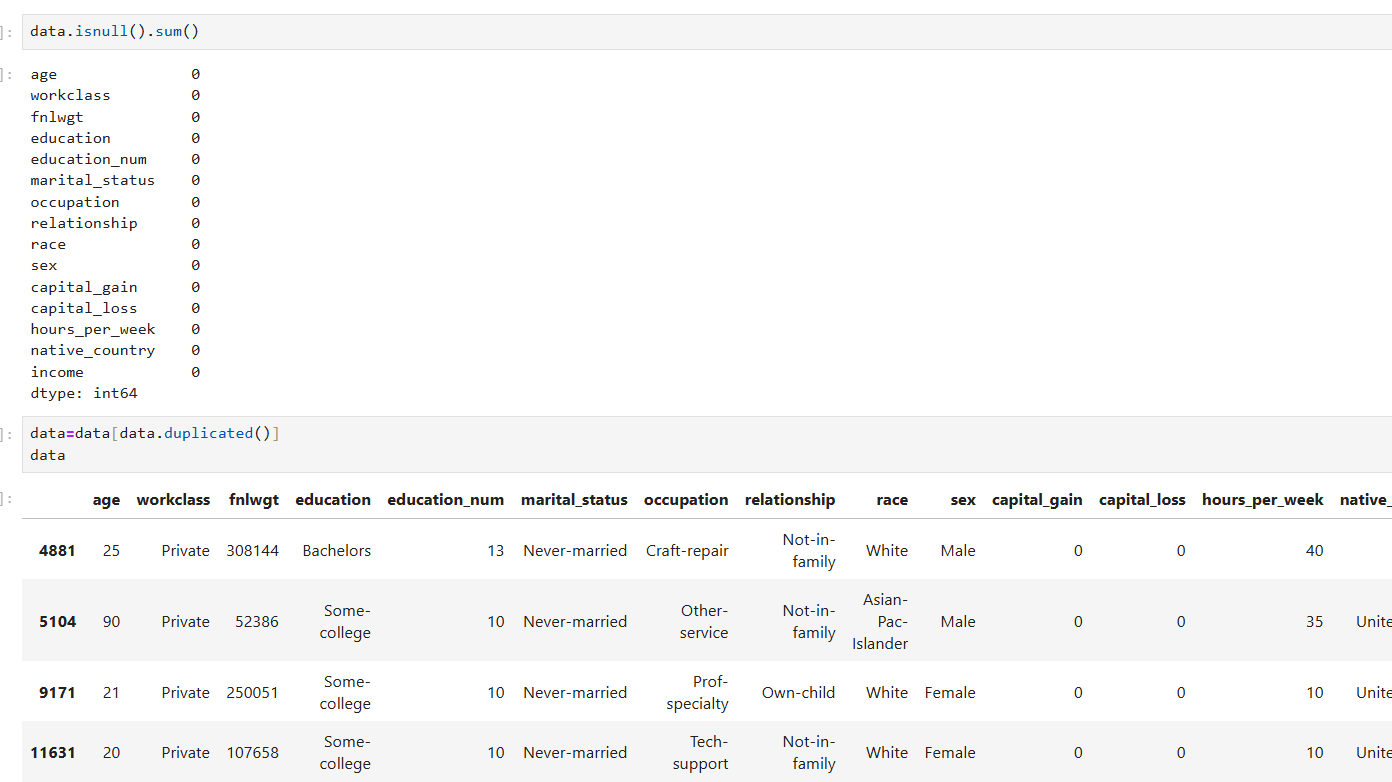
**1. Data Exploration and Preprocessing:**

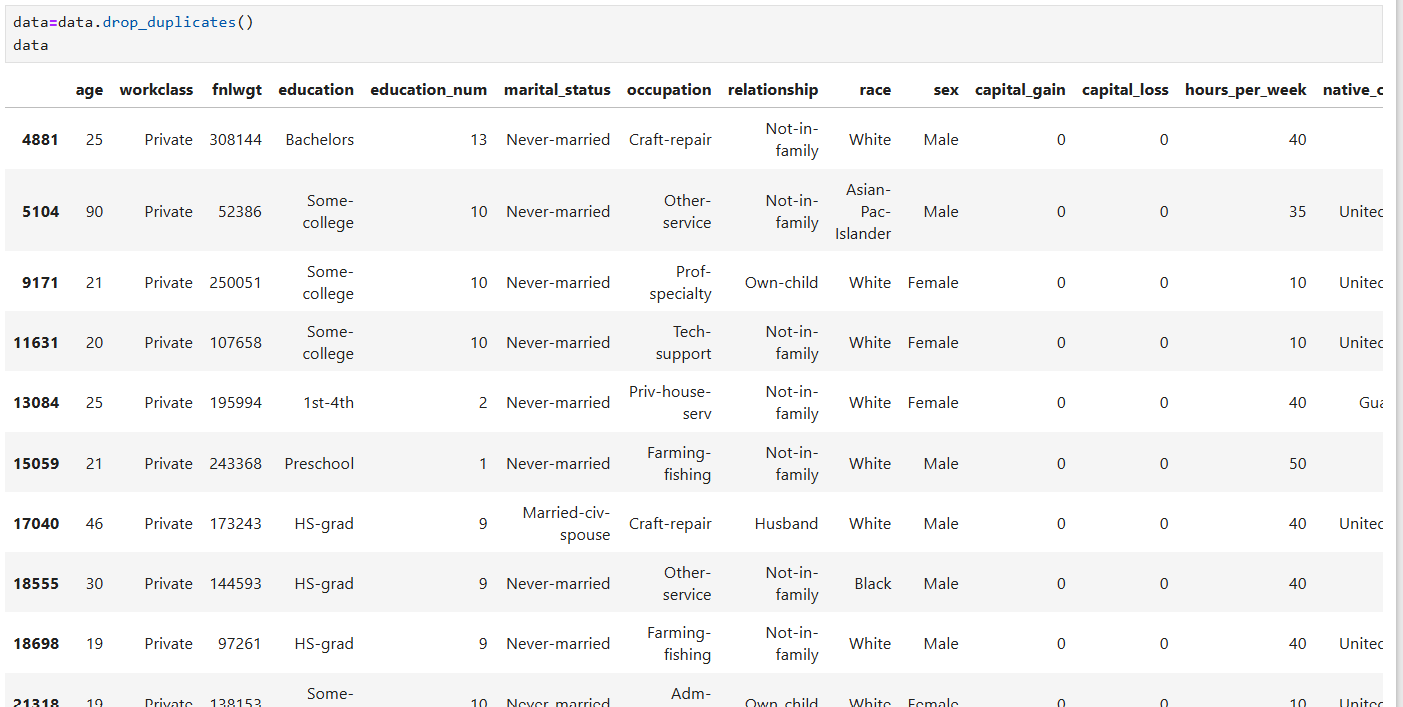
* Load the dataset and conduct basic data exploration (summary statistics, missing values, data types).
* Handle missing values as per the best practices (imputation, removal, etc.).
* Apply scaling techniques to numerical features:
  + Standard Scaling
  + Min-Max Scaling
* Discuss the scenarios where each scaling technique is preferred and why.

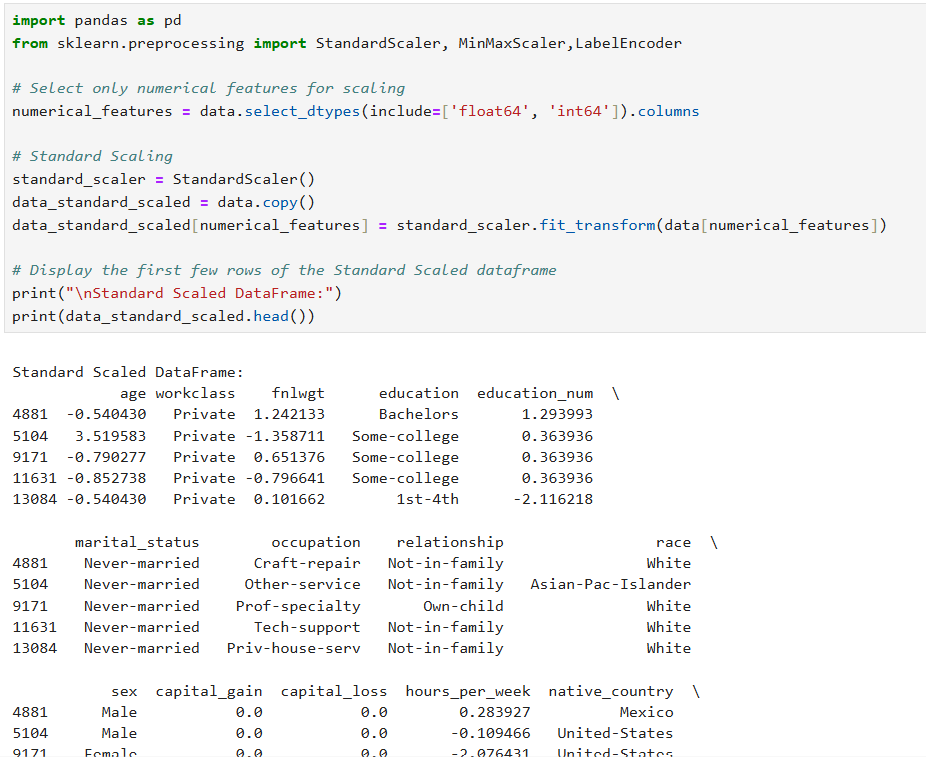


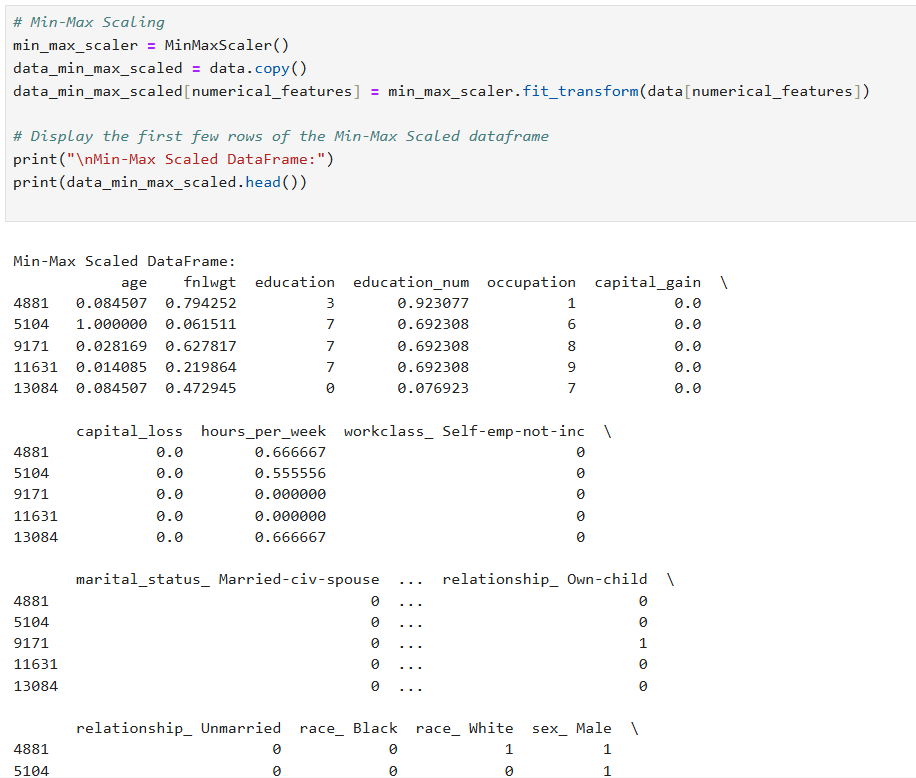






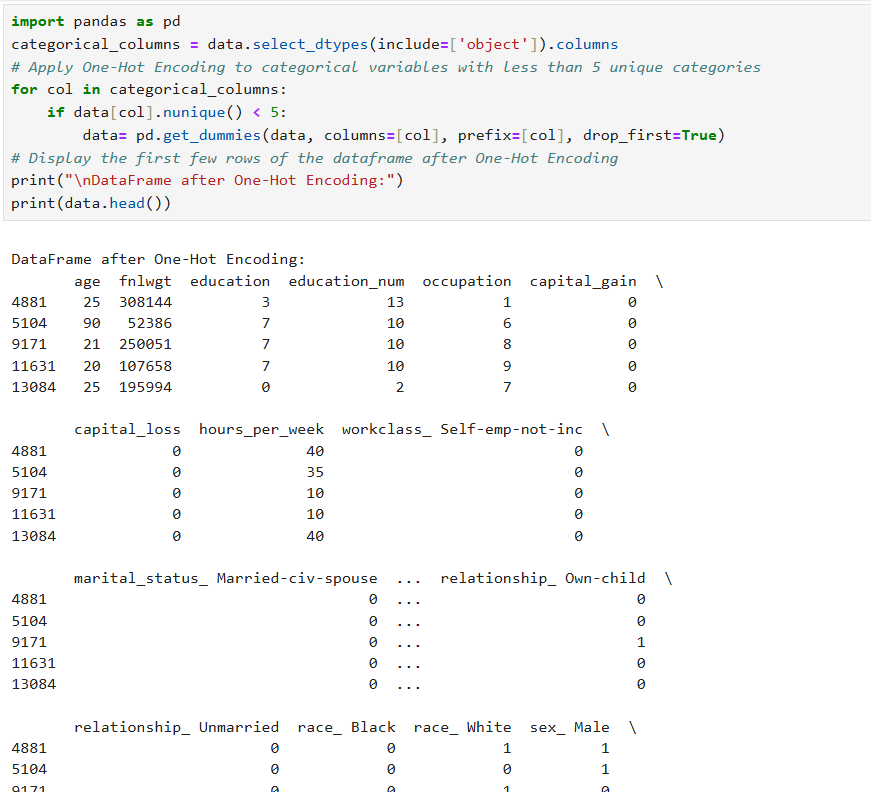






**2. Encoding Techniques:**

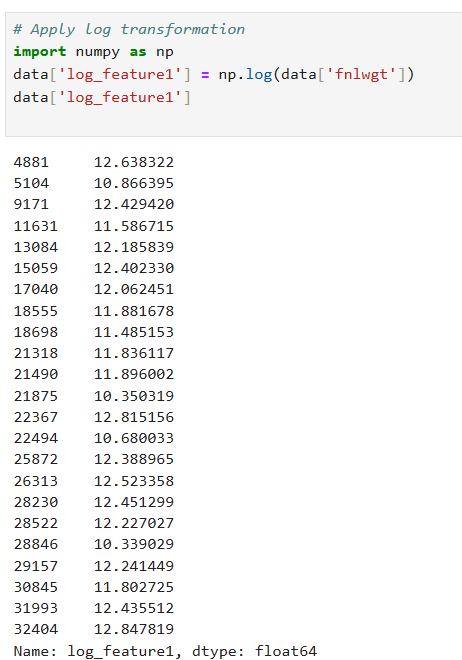
* Apply One-Hot Encoding to categorical variables with less than 5 categories.
* Use Label Encoding for categorical variables with more than 5 categories.
* Discuss the pros and cons of One-Hot Encoding and Label Encoding.

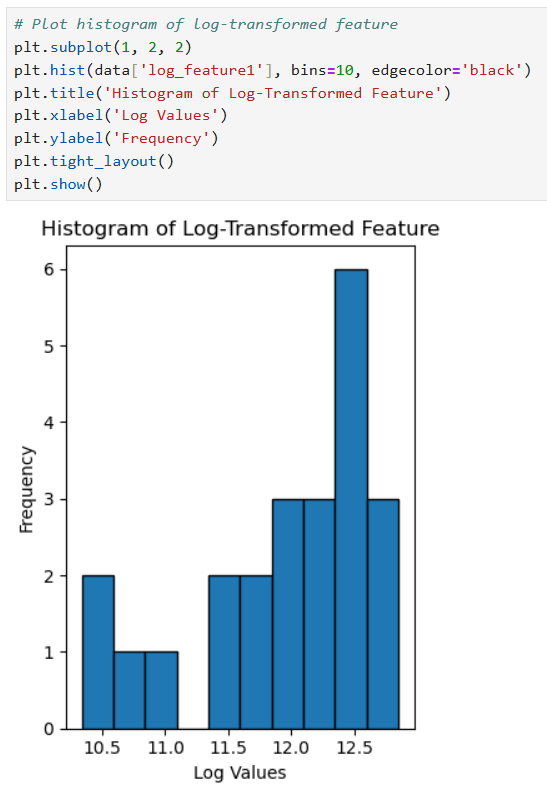




**3. Feature Engineering:**

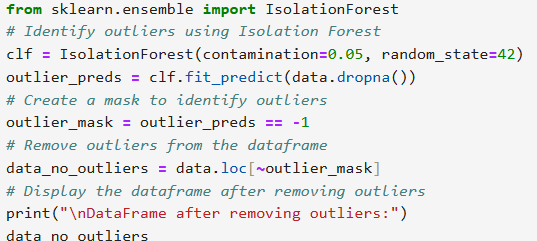
* Create at least 2 new features that could be beneficial for the model. Explain the rationale behind your choices.
* Apply a transformation (e.g., log transformation) to at least one skewed numerical feature and justify your choice.

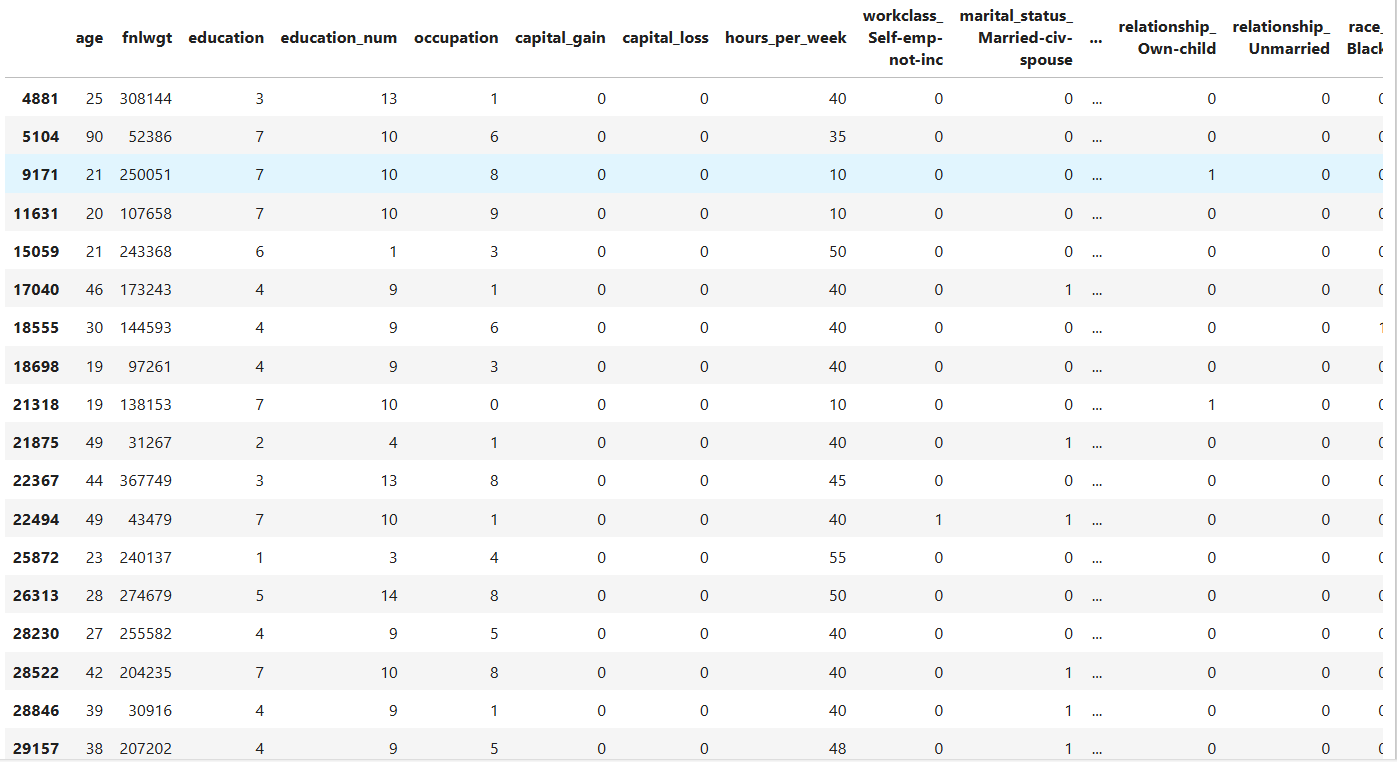




**4. Feature Selection:**

* Use the Isolation Forest algorithm to identify and remove outliers. Discuss how outliers can affect model performance.





* Apply the PPS (Predictive Power Score) to find and discuss the relationships between features. Compare its findings with the correlation matrix.

