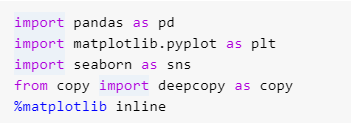
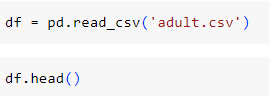
**Prediction of adult income based on census data**

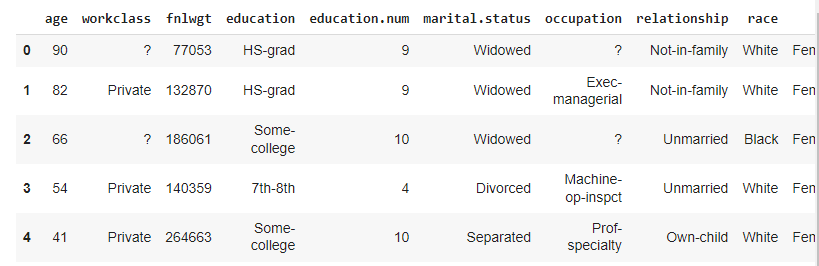
This project focuses on predicting whether an individual's income exceeds 50K dollar using machine learning classification algorithms and uncovering patterns in the dataset using Association rules. By leveraging comprehensive census data encompassing demographic attributes and income information, we aim to provide valuable insights across various domains. The predictive models generated through machine learning algorithms, including Logistic Regression, Decision Trees, and Ensemble Methods, will enable us to assess income levels and demographic trends. Additionally, association rule mining techniques such as the Apriori algorithm will help uncover relationships between demographic attributes and income thresholds. Through this analysis, we seek to inform decision-making in areas such as business strategy, financial services, real estate, and education planning. Ultimately, our goal is to contribute to efforts aimed at addressing income inequality and promoting socio-economic development.

**Exploratory Data Analysis (EDA):**

****

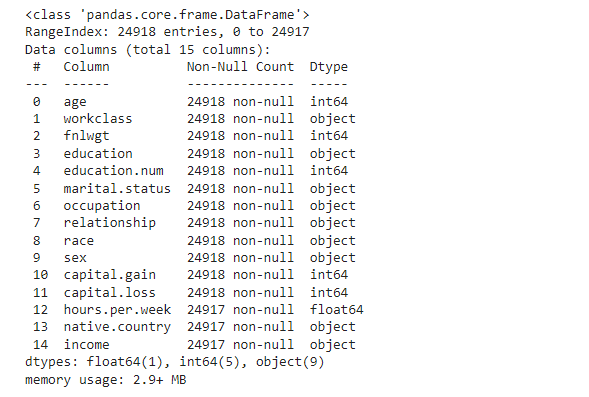
**Reading the Data**

****

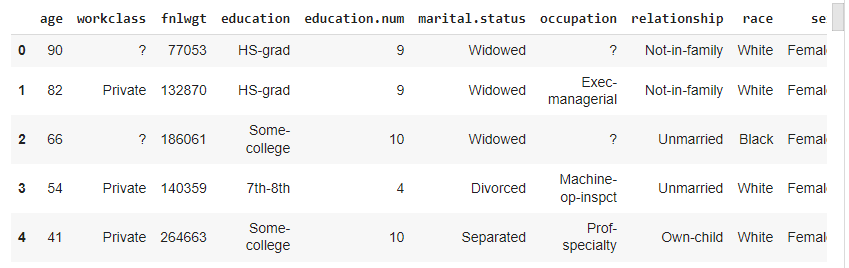
****

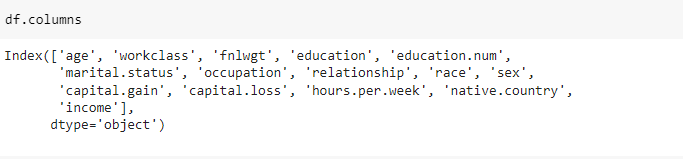
**View summary of dataframe:**

****

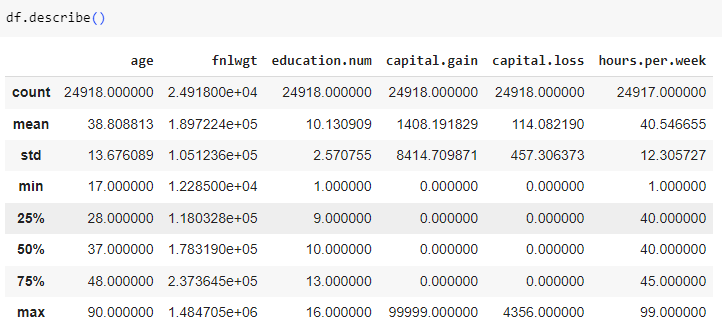
****

****

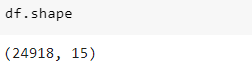
****

****

**Data description:**

****

**Data Shape:**

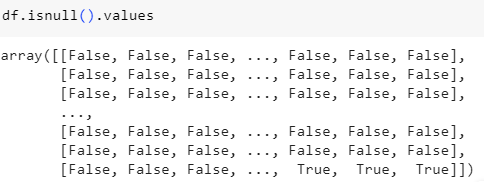
****

Number of rows :24918

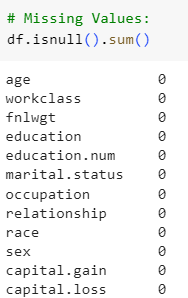
Number of columns:15

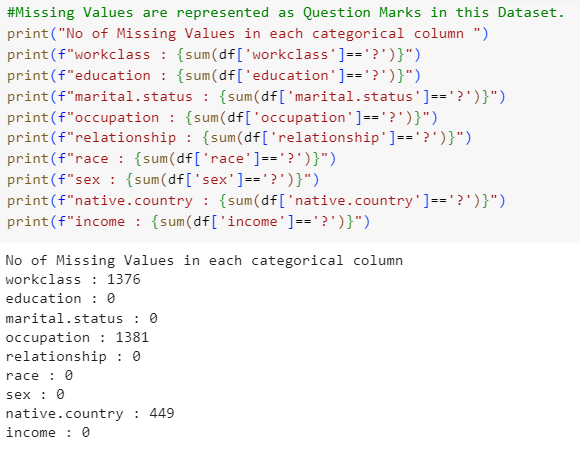
**Preprocessing of the data:**

**Checking Null Values :**

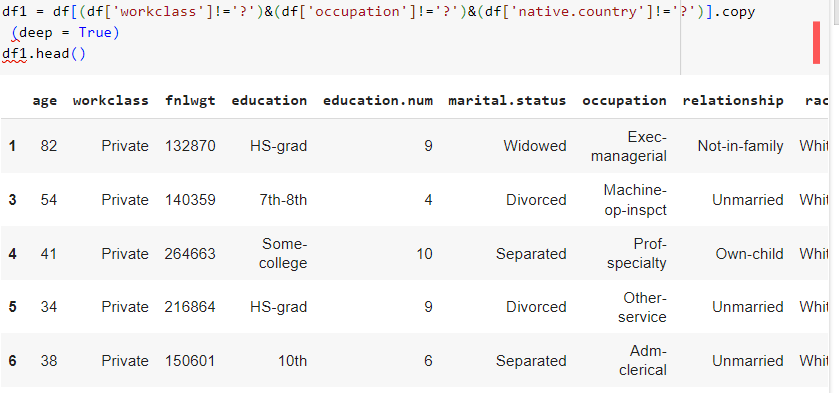
****

**Checking Missing Values :**

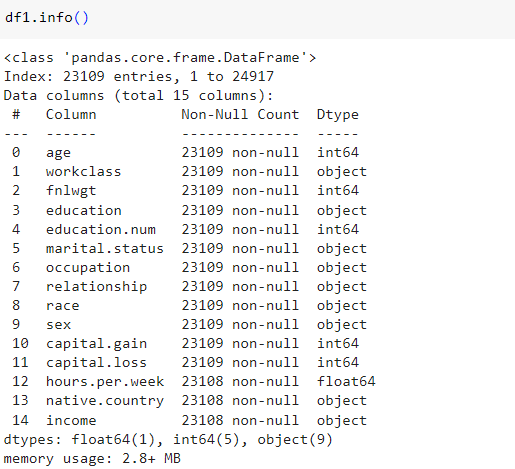
****

****

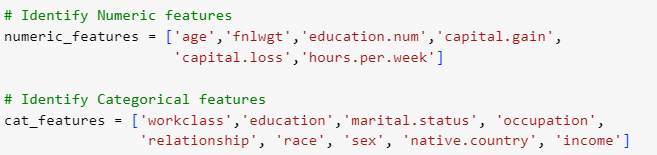
**Removing Missing Values:**

****

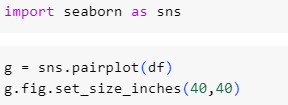
**Data Information :**

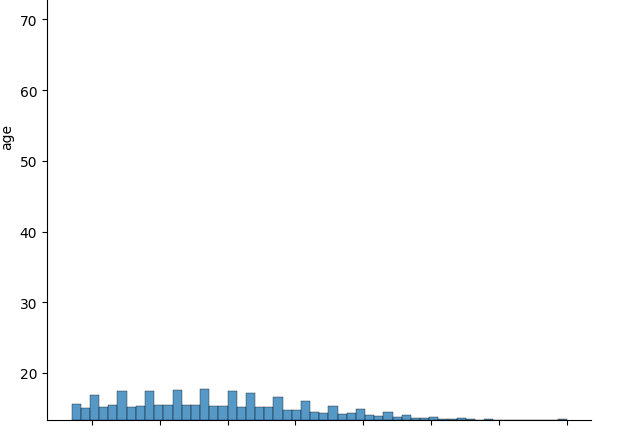
****

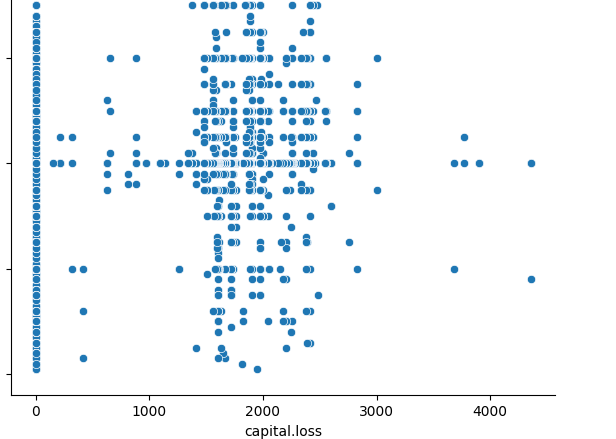
**Identify dataset features:**

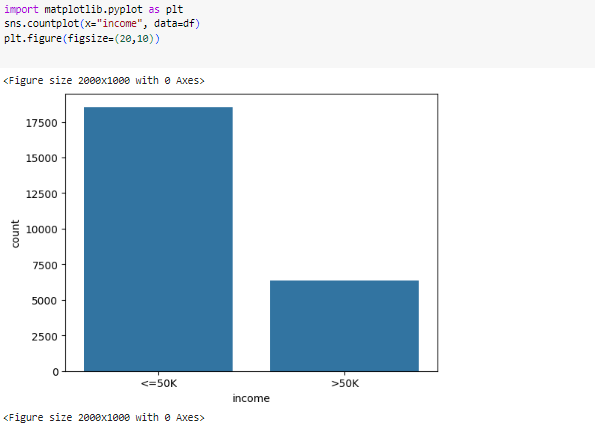
****

**Data Visualization:**

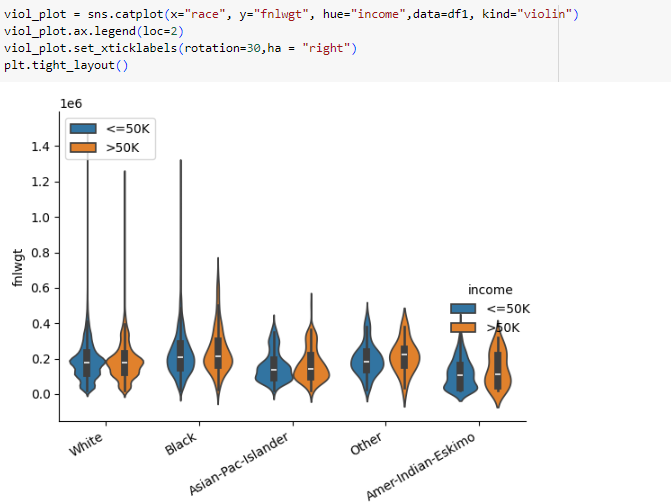
****

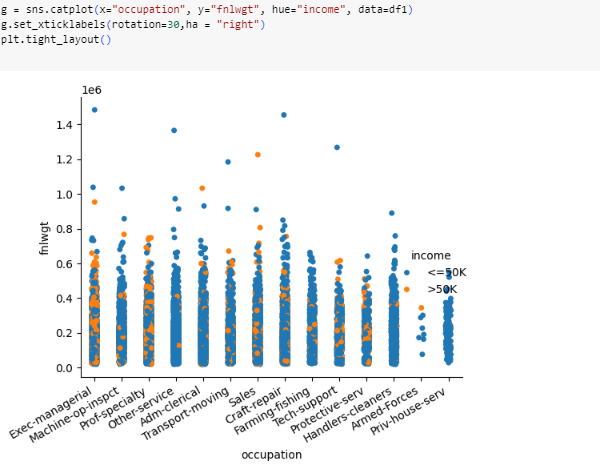
****

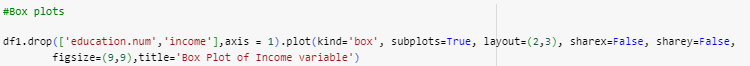
****

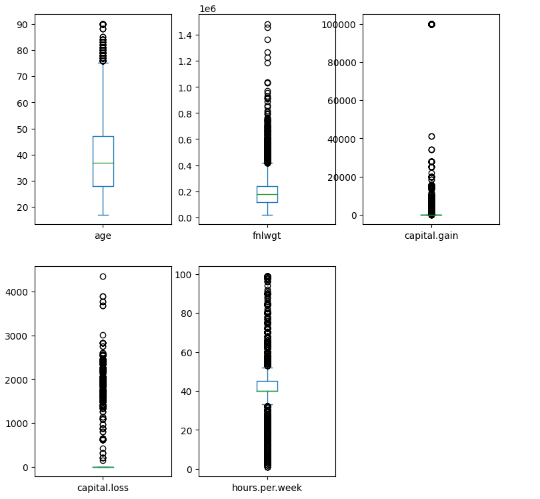
****

The proportion of records having income less than 50K dollars is more than the proportion of records having more than 50k dollars.

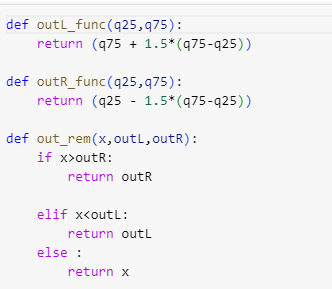
****

****

****

****

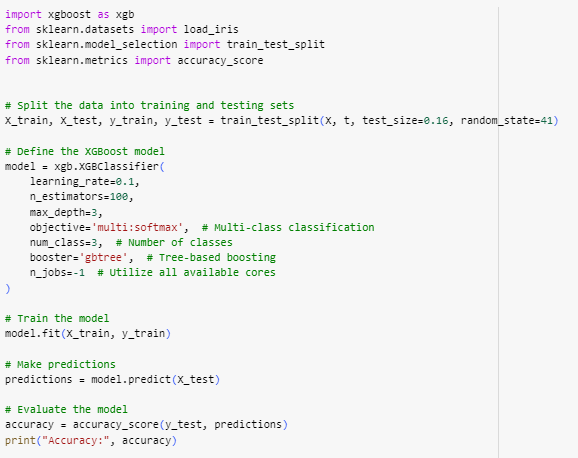
**Fucntion to Remove the Outliers:**

****

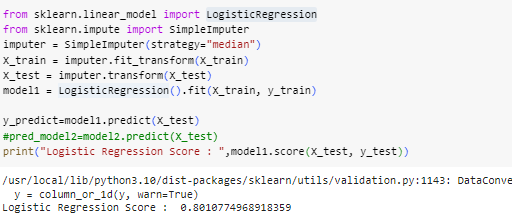
**Working on categorical variables:**

****

**XGBoost Model:**

****

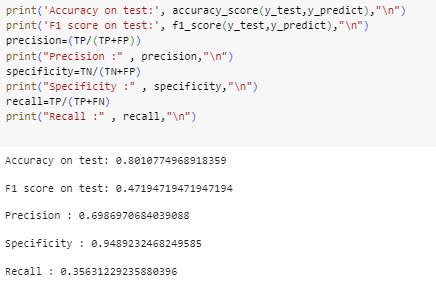
**LogisticRegression Model:**

****

**LogisticRegression Confusion Matrix:**

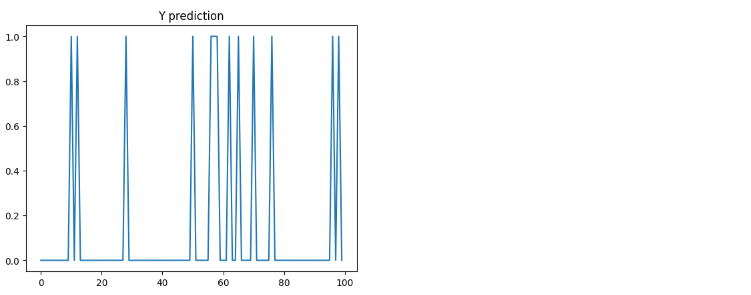
****

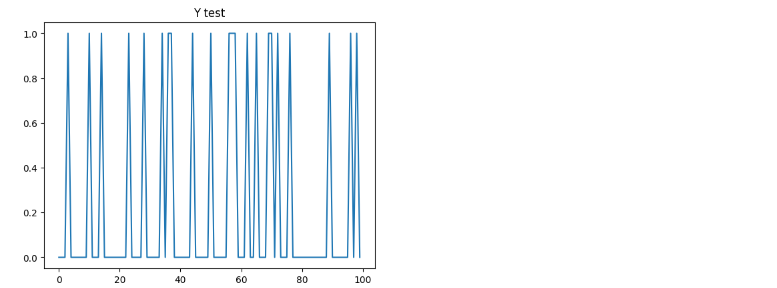
**Accuracy Testing :**

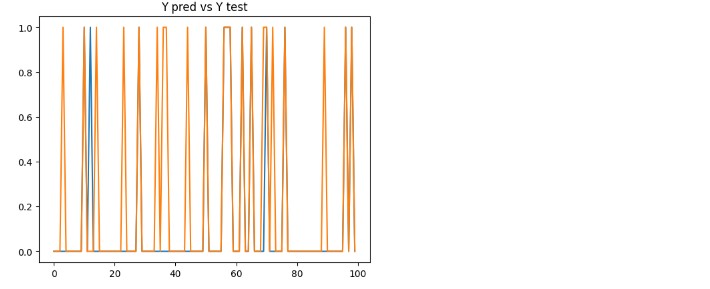
****

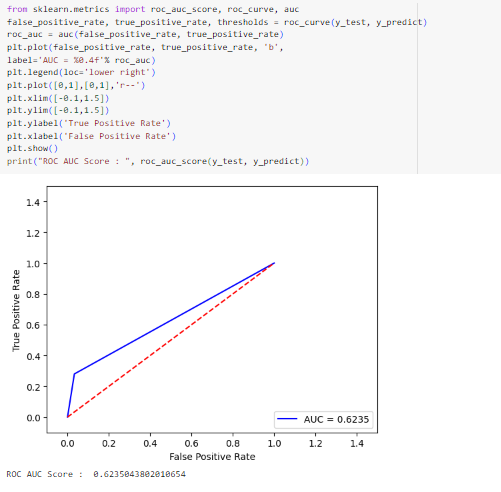
**Prediction Vs Testing :**

****

****

****

****

****

**Preliminary Observations :**

* The no of records having income less than 50k dollars is more than the no of records having income more than 50k dollars income. The dataset neeeds to be balanced with the target values so that the models do not overfit the data.
* The capital.gain and capital.loss values contain zeroes ,so these columns can be dropped.
* Scatter plots and bar plots are plotted to find the distribution of various values of categorical values.
* Hours.per.week has a value of 40 in most of the records, so this field can be dropped.
* The fnlwgt values are moslty in the range of 0-40,000 and are of age 20 to 40.
* The outliers are present in some of the contiuous variables which need to the handled properly.