**Determining Trade Union Status – A Final Project**

**The Goal of the Project**

This project entails building a model to predict whether a person will remain in a hypothetical trade union. The target variable is categorical with values Yes and No indicating whether a person left the union. Below are the independent variables and their datatypes:

gender object

Management int64

USAcitizen object

Married object

MonthsInUnion int64

ContinuingEd object

FeatureA object

Connectivity object

FeatureC object

FeatureD object

FeatureE object

FeatureF object

FeatureG object

FeatureB object

DuesFrequency object

PaperlessBilling object

PaymentMethod object

MonthlyDues int64

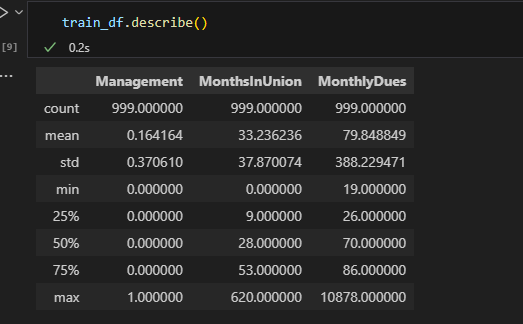
TotalDues object

**Tools and Programming Language**

Python and Jupyter Lab are the two main tools used in the project. I have chosen Python because of its simplicity and ease of use.

**Descriptive Statistics**

I have used summary statistics to summarize the numerical variables of the data set as shown in the figure below. The descriptive statistics show the mean, standard deviation, min,1Q, median, 3Q, and max, values of the variables.

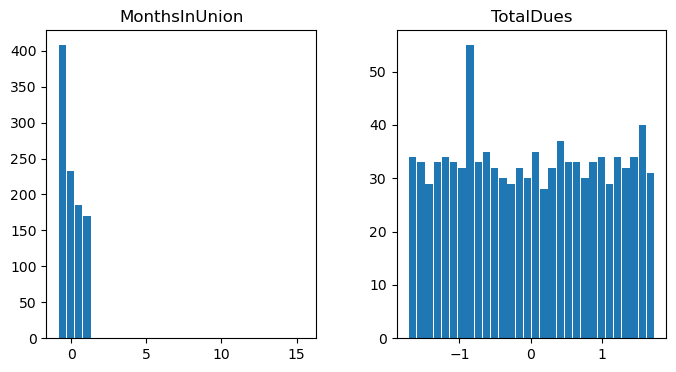


**Data Cleaning and Preprocessing**

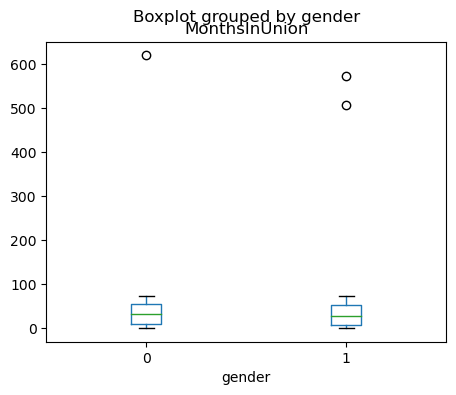
There were no missing values in both the train and test datasets. Since the dataset contained categorical variables, these variables were encoded them scaled using StandardScaler function.

**Distribution of Variables**

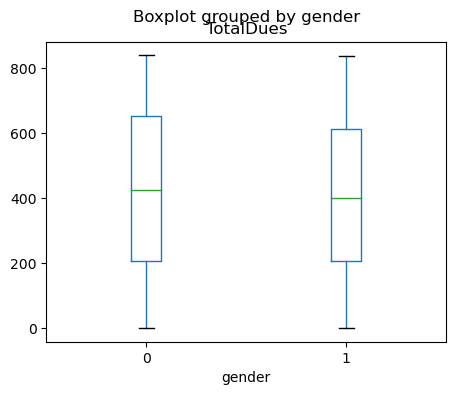
1. **Histogram**



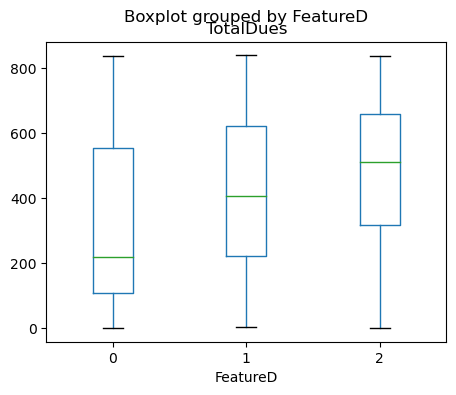
1. **Box Plots**
2. **Months in Union Grouped by Gender**



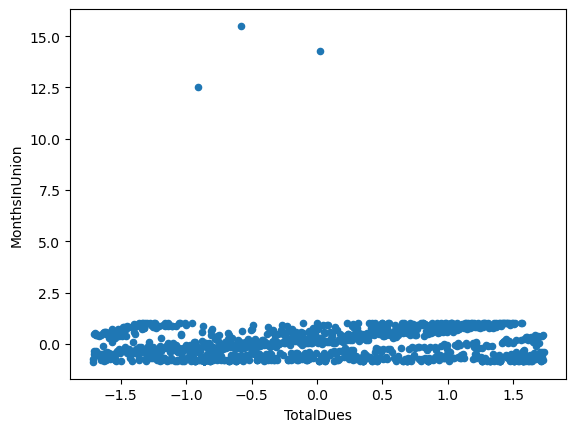
1. **Total Dues by Gender**



1. **Total Dues by Feature D**



1. **Scatter Plot**

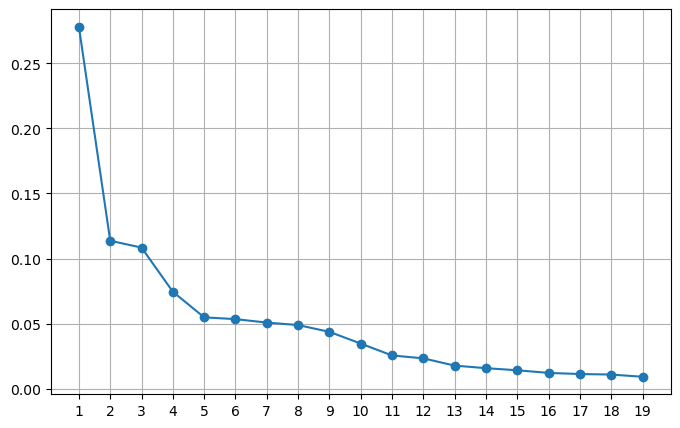


**Justification of the Methods Chosen to Visually Present Data**

The choice of visualization depends on the nature of the data and the specific insights you want to extract or communicate. In this case, histograms are chosen to understand data distribution and central tendency, box plots are used for summarizing distributions and identifying outliers, and scatter plots help visualize relationships and patterns between two continuous variables.

**Approach to Dimensionality Reduction**

Principal component analysis was used to determine the essential number of components/features. A scree plot of the components indicated that only 5 components were essential. The data was then reduced to 5 components for further analysis.



**Explanation of the Models Chosen**

Logistic Regression is a statistical model used for binary classification tasks. It estimates the probability that an instance belongs to a particular class using a logistic function. It models the relationship between the features and the probability of the positive class. When the relationships between features and the outcome are relatively linear, Logistic Regression can work well and is less prone to overfitting with smaller datasets.

Random Forest is an ensemble learning method that builds multiple decision trees during training and combines their predictions to make a final decision. Each tree is trained on a random subset of data and features, reducing overfitting. Random Forest often generalizes well to new data, making it a reliable choice for predictive modeling.

A Decision Tree is a tree-like model that recursively splits the feature space into regions based on feature values. Each leaf node represents a class or outcome. Decision Trees can be used for both classification and regression tasks. Decision Trees naturally select important features by positioning them higher up in the tree, aiding in feature selection and simplifying the model.