Problem 1: Vote Prediction

You are hired by one of the leading news channels CNBE who wants to analyse recent elections. This survey was conducted on 1525 voters with 9 variables. You must build a model, to predict which party a voter will vote for based on the given information, to create an exit poll that will help in predicting overall win and seats covered by a particular party.

**Data Inferences**

1. Features Vote and Gender are of Object Type and Categorical in nature.
2. Age feature provides information about the voter’s age.
3. National and Household economy are provided in the features “economic.cond.national” and “economic.cond.household”.
4. Blair feature will give information about the Voter’s assessment of the Labour Leader, rating will be between 1 and 5, latter being better.
5. Hague feature will give information about the Voter’s assessment of the Conservative Leader, rating will be between 1 and 5, latter being better.
6. Europe feature will explain the sentiment of Eurosceptic among the voters, range beings at 1 and ends at 11, where in the latter is maximum Eurosceptic.
7. Feature “political.knowledge” gives information about the knowledge of Parties on the integration of US policies with Europe, range starts from 0 to 3, where in the latter is of the best knowledge.

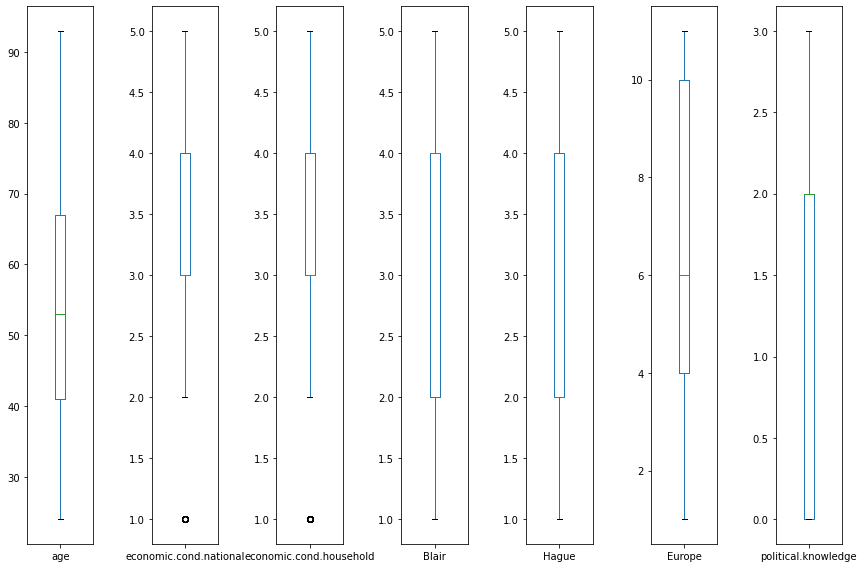
**Data Shape**

1. Data set contains 9 Features and 1525 Voter details.

**Data Validation**

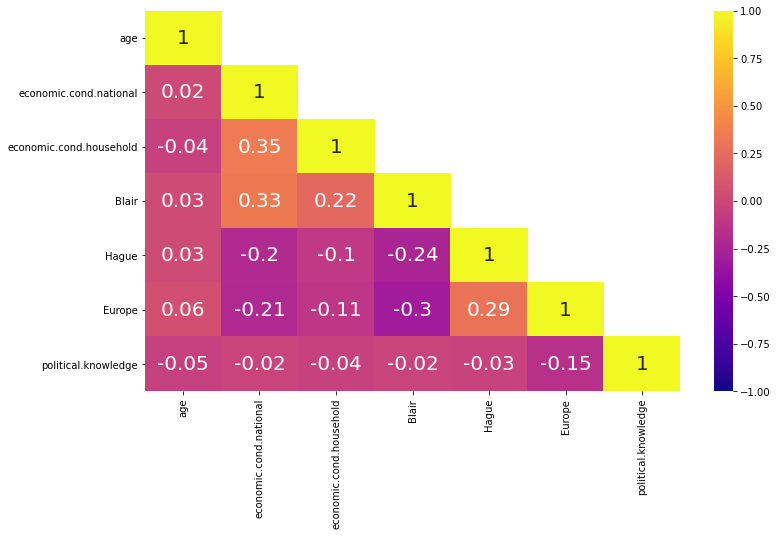
1. No nulls are seen in the data set.
2. Total duplicates found are 8 in the data set, will be dropping them as their count is insignificant, considering data set size.
3. Vote feature has two Categories “Labour” and “Conservative” and has no anomalies.
4. Gender feature has two Categories “Female” and “Male” and has no anomalies.

**Outliers**



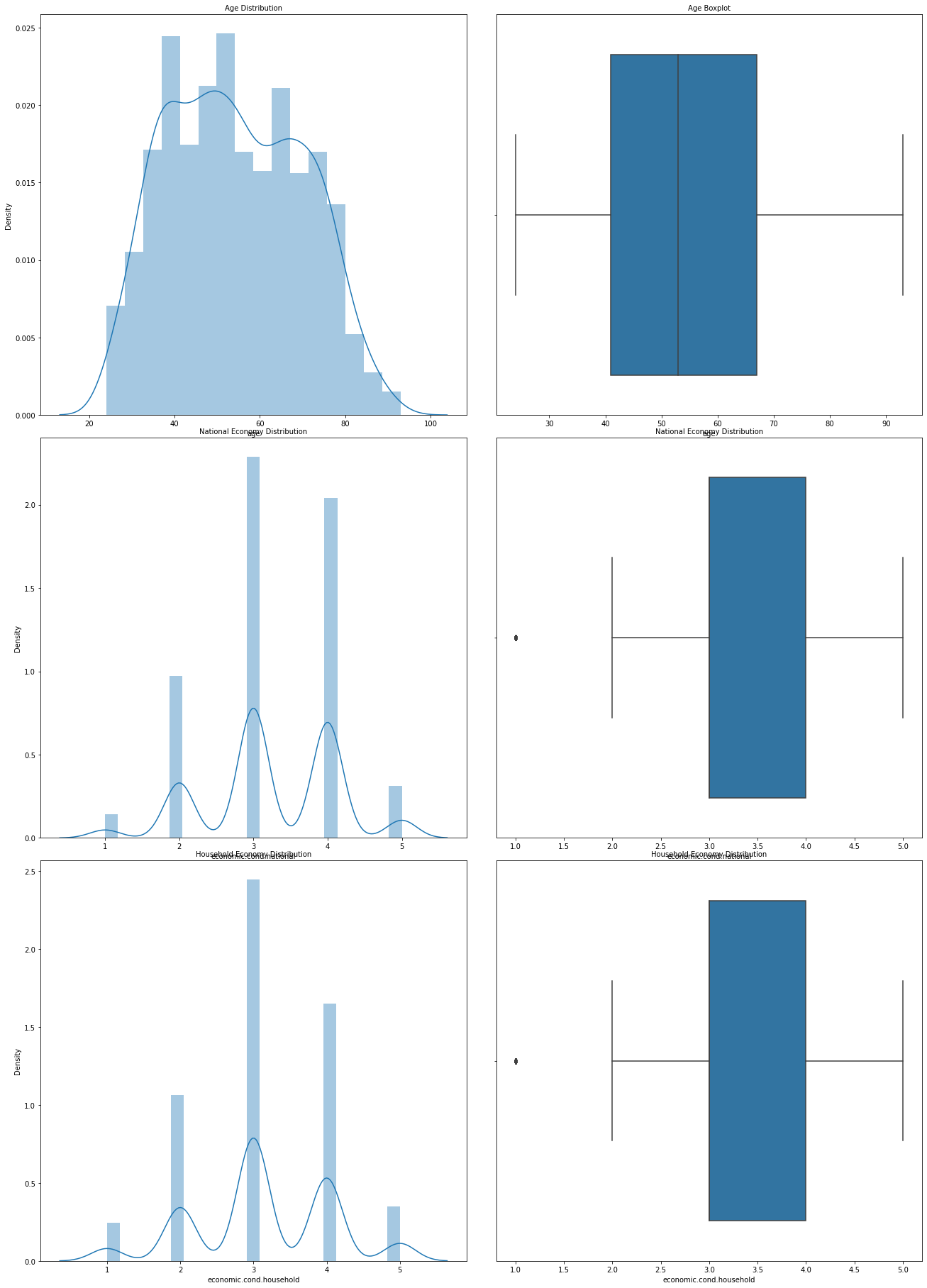
1. National and Household economy features have very insignificant Q1 outliers.
2. Other columns have no outliers.

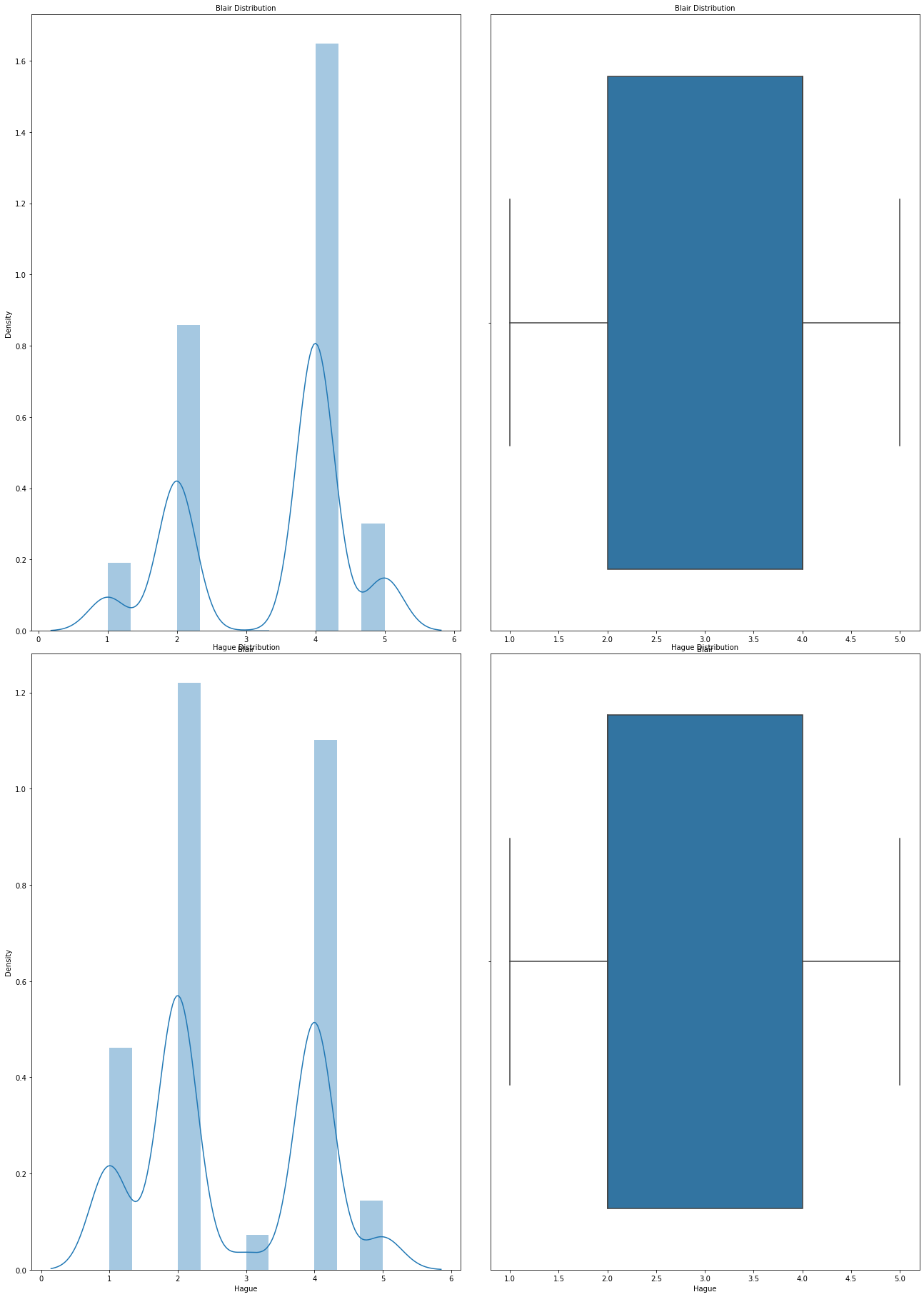
**Correlation Analysis**

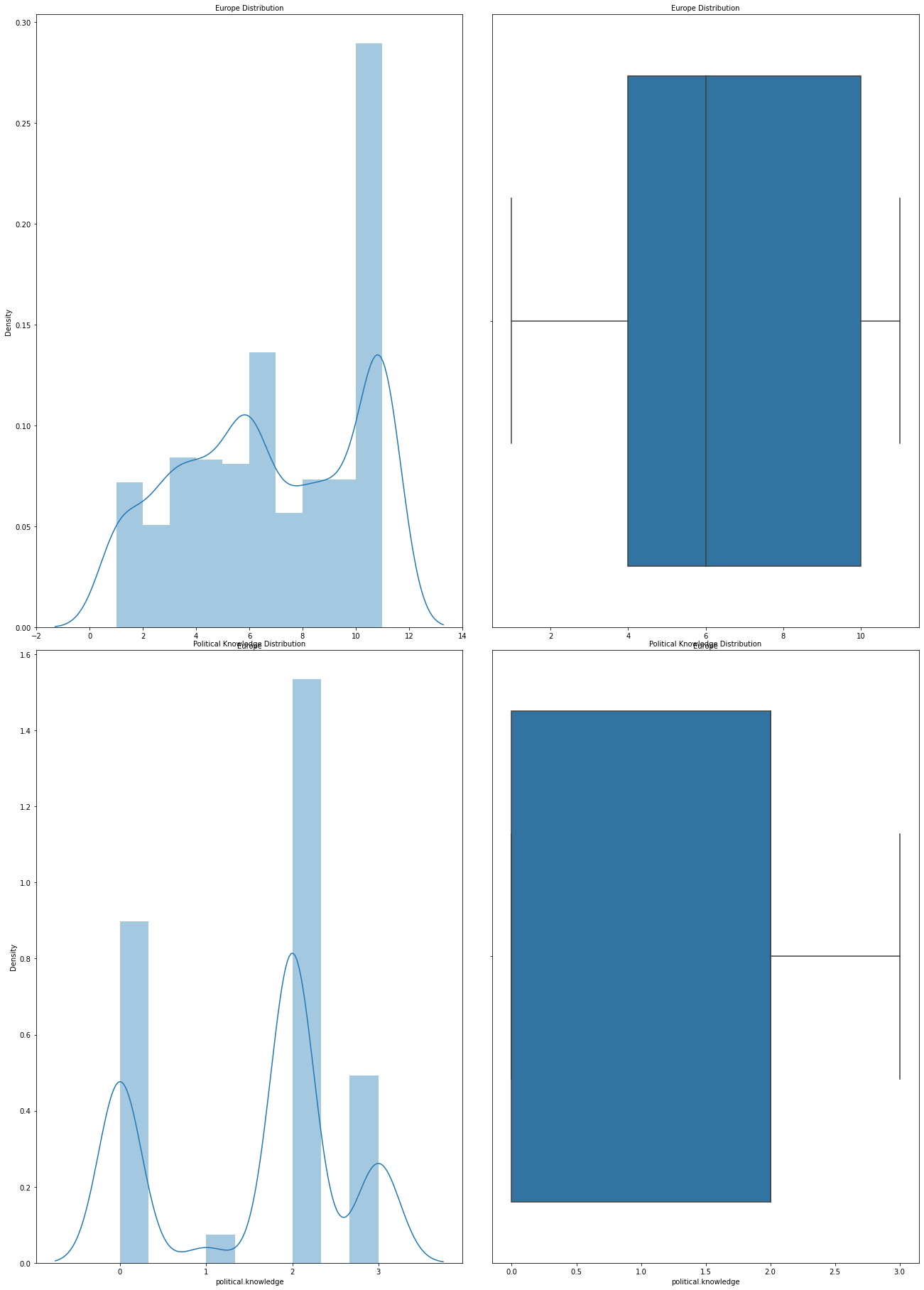


1. Age feature has no correlation with any other features of the dataset, it means, age has no effect in the voter’s understanding on the other features of the dataset.
2. National Economy has a weak-positive-correlation with household economy and Blair-Labour Leader’s rating.
3. National Economy has a weak-negative-correlation with Eurosceptic sentiment. Other features will not have any affect.
4. Labour Leader’s rating-Blair has a negative correlation with Conservative Leader’s rating-Hague and vice versa.
5. Eurosceptic rating has a negative correlation with the political knowledge of the voter.

**Uni-Variate Analysis**







**Vote vs Other Features**

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

**Categorical column analysis**

Chart, bar chart

Description automatically generated

**Pair plot**

**A picture containing shoji, crossword puzzle, window

Description automatically generated**

**Data Encoding**

Table

Description automatically generated

**Data Scaling**

Table

Description automatically generated

**Train and Test split**

 **and** 

**Logistic Regression**

**LogisticRegression(max\_iter=10000, n\_jobs=2, penalty='none', random\_state=123,solver='newton-cg', verbose=True)**

**Test Prediction probabilities**

Graphical user interface, text, application

Description automatically generated

**Accuracy Score – Training data**

**0.83**

**ROC Curve – Training data**

**A picture containing graphical user interface

Description automatically generated**

**Accuracy Score – Testing Data**

**0.85**

**ROC Curve – Testing Data**

**A picture containing graphical user interface

Description automatically generated**

**Confusion Matrix and Classification Report – Training data**

**Chart, treemap chart

Description automatically generated**

Table

Description automatically generated

**Confusion Matrix and Classification Report – Testing data**

**Chart, treemap chart

Description automatically generated**

Table

Description automatically generated

**GridSearchCV for Logistic Regression**

Text, letter

Description automatically generated

**Best Model**



**Test probabilities for the best model**

Graphical user interface, text, application, chat or text message

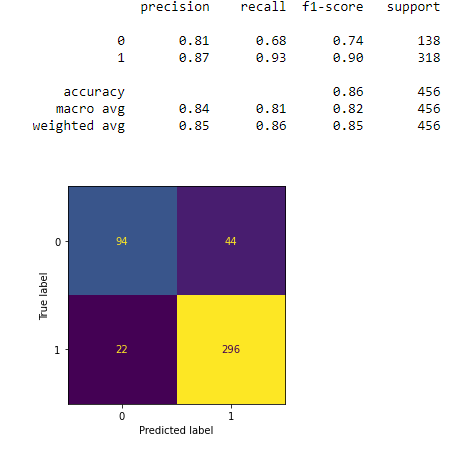
Description automatically generated

**Confusion matrix and Classification report – Training data**

Chart, treemap chart

Description automatically generated

**Confusion matrix and Classification report – Testing data**



Chart, treemap chart

Description automatically generated

**Important Features**

**Chart

Description automatically generated**

**Linear Discriminant Analysis**

Confusion matrix and Classification report – Training data

Chart, treemap chart

Description automatically generated

Confusion matrix and Classification report – Testing data

Chart, treemap chart

Description automatically generated

AUC and ROC curve – Training and Testing data

A picture containing graphical user interface

Description automatically generated

Naïve Bayes

Training data

Table

Description automatically generated with low confidence

Testing data

Table

Description automatically generated

KNN

KNN-5 Training data

Table

Description automatically generated

KNN-5 Testing data

Table

Description automatically generated

KNN-7 Training data

Table

Description automatically generated

KNN-7 Testing data

Table

Description automatically generated

Bagging

Metrics on Training data

Chart, treemap chart

Description automatically generated

Metrics on Test data

Chart, treemap chart

Description automatically generated

Ada Boost

Training data

Chart, treemap chart

Description automatically generated

Testing data

Chart, treemap chart

Description automatically generated

Gradient Boost

Training data

Chart, treemap chart

Description automatically generated

Testing data

Chart, treemap chart

Description automatically generated

Summary

Table

Description automatically generated

# 2  Problem Statement

**In this project, we are going to work on the inaugural corpora from the nltk in Python. We will be looking at the following speeches of the Presidents of the United States of America:**

**President Franklin D. Roosevelt in 1941 President John F. Kennedy in 1961 President Richard Nixon in 1973**

Total characters in President Roosvelt speech is: 7571

Total sentences in President Roosvelt speech is: 68

Total words in President Roosvelt speech is: 1526

A picture containing text, newspaper

Description automatically generated

Total characters in President Kennedy speech is: 7618

Total sentences in President Kennedy speech is: 52

Total words in President Kennedy speech is: 1543

Text

Description automatically generated

Total characters in President Nixon speech is: 9991

Total sentences in President Nixon speech is: 68

Total words in President Nixon speech is: 2006

A picture containing text, newspaper

Description automatically generated