Machine Learning Project

Answer Report

NAME – ELTON REBELLO

BATCH – JAN22A

DATE – 19 – 06 –2022

Table Of Contents

List Of Tables

List Of Figures

Problem 1 -

Executive Summary -

Leading news channel CNBE wants to analyze recent elections. A survey was conducted on 1525 voters with 9 variables. We have to build a model, to predict which party a voter will vote for on the basis of the given information, to create an exit poll that will help in predicting overall win and seats by a particular party.

Introduction -

The purpose of the problem is to explore the dataset. Do the exploratory analysis. The data consists details of 1525 voters with 9 variables. We will be creating a model to predict which party a voter will vote for on the basis of the given information.

Data Description -

1. vote: - Party choice: Conservative or labor
2. age: - in years
3. economic.cond.national: - Assessment of current national economic conditions, 1 to 5.
4. economic.cond.household: - Assessment of current household economic conditions, 1 to 5.
5. Blair: - Assessment of the Labour leader, 1 to 5.
6. Hague: - Assessment of the Conservative leader, 1 to 5.
7. Europe: - an 11-point scale that measures respondents' attitudes toward European integration. High scores represent 'Eurosceptic' sentiment.
8. political.knowledge:- Knowledge of parties' positions on European integration, 0 to 3.
9. gender: - female or male.

Data Ingestion:

1.1 read the dataset. Do the description statistics and do the null value condition check. Write an inference on it.

Sample of the Data Set -

	vote	age	economic_cond_national	economic_cond_household	Blair	Hague	Europe	political_knowledge	gender
0	Labour	43	3	3	4	1	2	2	female
1	Labour	36	4	4	4	4	5	2	male
2	Labour	35	4	4	5	2	3	2	male

3	Labour	24	4	2	2	1	4	0	female
4	Labour	41	2	2	1	1	6	2	male

Table no 1 – sample of the data set

Dataset has 9 variables with details about the voters.

Data Description -

	count	mean	std	min	25%	50%	75%	max
age	1525	54.18	15.71	24	41	53	67	93
economic_cond_national	1525	3.25	0.88	1	3	3	4	5
economic_cond_household	1525	3.14	0.92	1	3	3	4	5
Blair	1525	3.33	1.17	1	2	4	4	5
Hague	1525	2.75	1.23	1	2	2	4	5
Europe	1525	6.73	3.30	1	4	6	10	11
political_knowledge	1525	1.54	1.08	0	0	2	2	3

Table no 2 – data description

Missing Values in the Data set -

RangeIndex: 1525 entries, 0 to 1524
Data columns (total 9 columns):

vote 1525 non-null object
age 1525 non-null int64
economic_cond_national 1525 non-null int64
economic_cond_household 1525 non-null int64
Blair 1525 non-null int64
Hague 1525 non-null int64
Europe 1525 non-null int64

political_knowledge	1525 non-null	int64
gender	1525 non-null	object

From the above result we can see that there are no missing values in the data set.

Types of variables in the dataset -

vote	object
age	int64
economic_cond_national	int64
economic_cond_household	int64
Blair	int64
Hague	int64
Europe	int64
political_knowledge	int64
gender	object

There are 7 integer variables and 2 object variables in the data set.

There are no null values in the data set.

1.2 Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers.

Exploratory Data Analysis (EDA) -

Univariate Analysis -

There is total 1525 rows and 9 columns in the dataset. 7 columns are of integer datatype and 2 columns are of object datatype.

Unique values of categorical variables -

VOTE: 2

Conservative	462
Labour	1063

GENDER: 2

male	713
female	812

We can see that there are more votes for Labour Party choice. There are 812 female and 713 male voters.

Histogram of Vote -

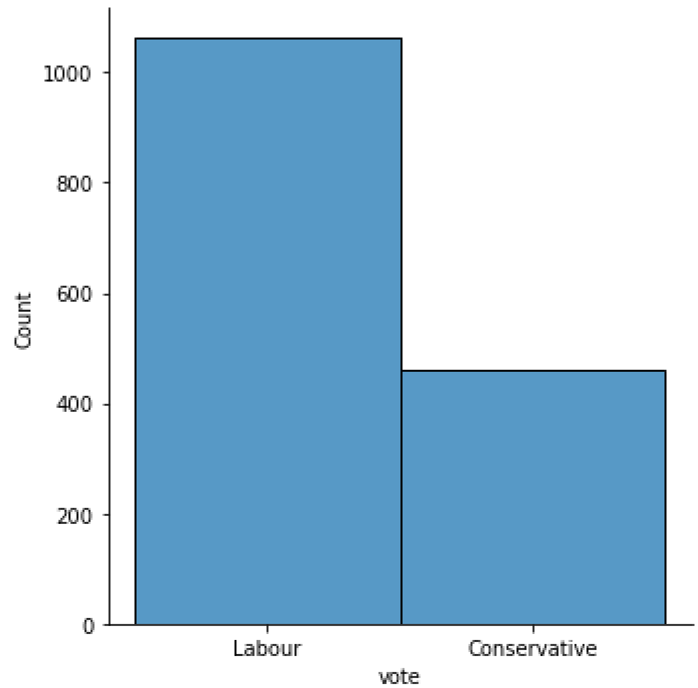

Fig No 1 – Histogram of Vote

There are more than double votes for Labour party compare to Conservative party.

Histogram of age -

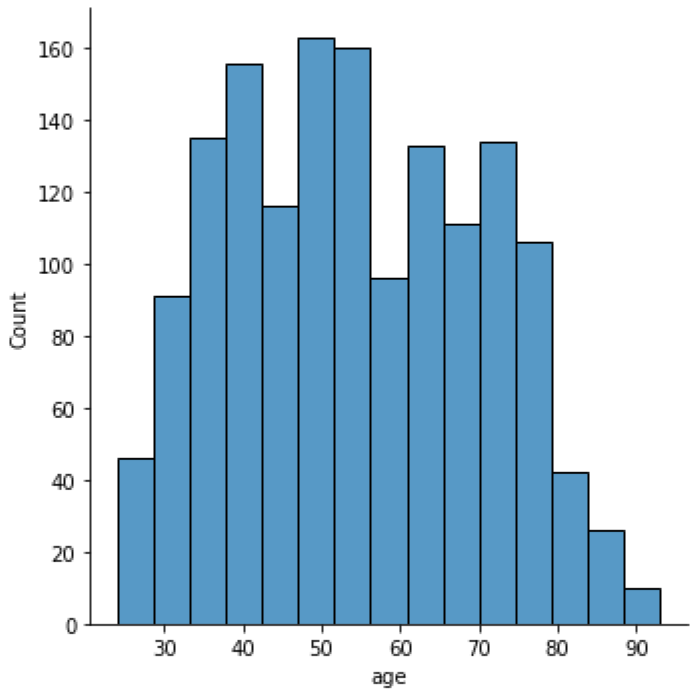

Fig No 2 – Histogram of Age

Highest number of voters are at the age of 50.

Histogram of economic_cond_national -

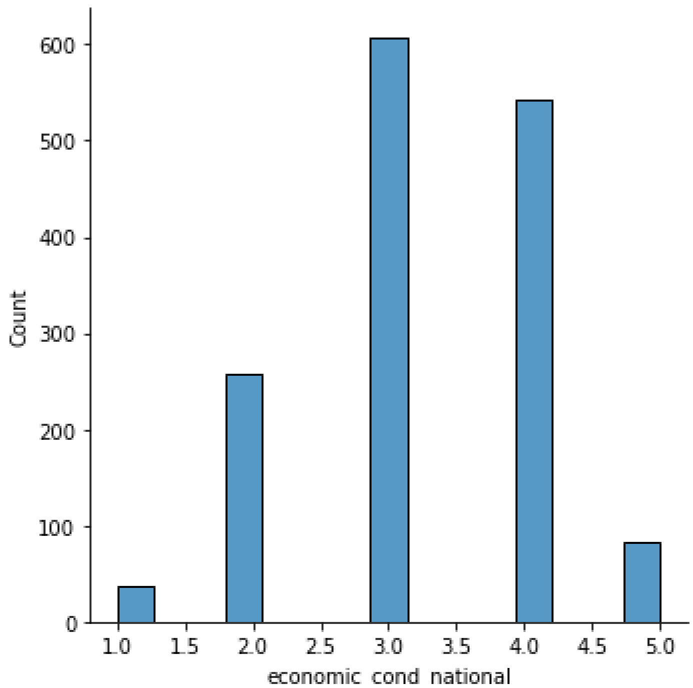

Fig No 3 – Histogram of economic_cond_national

This shows us the assessment of current national economic conditions, 1 to 5.

Histogram of economic_cond_household -

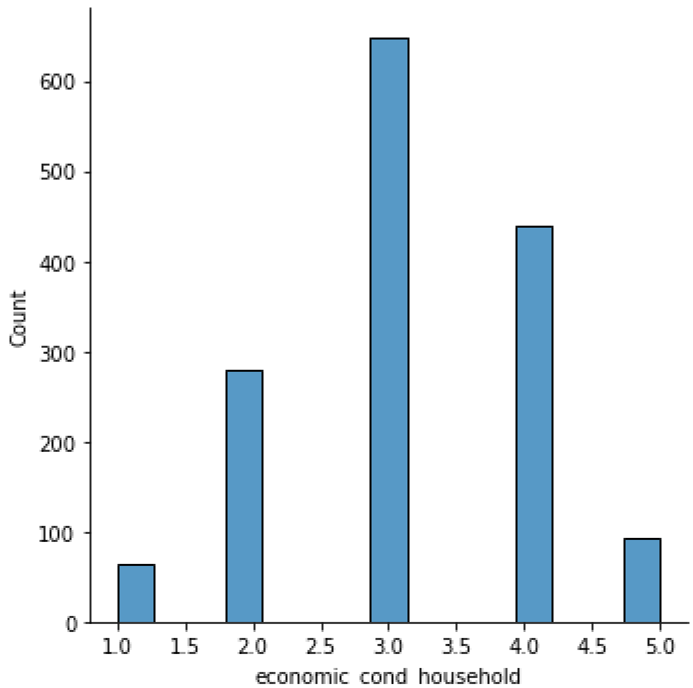

Fig No 4 – Histogram of economic_cond_household

Assessment of current household economic conditions, 1 to 5.

Histogram of Blair -

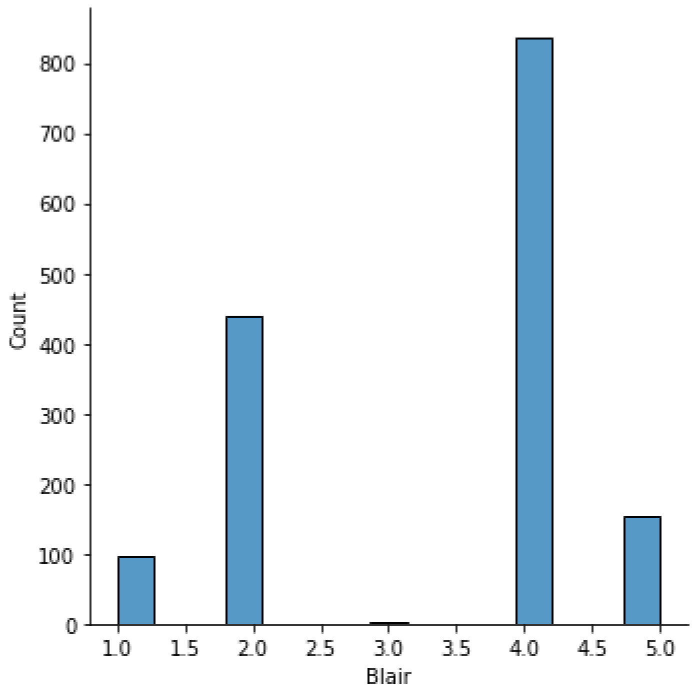

Fig No 5 – Histogram of Blair

Assessment of the Labour leader, most being at 4.

Histogram of Hague -

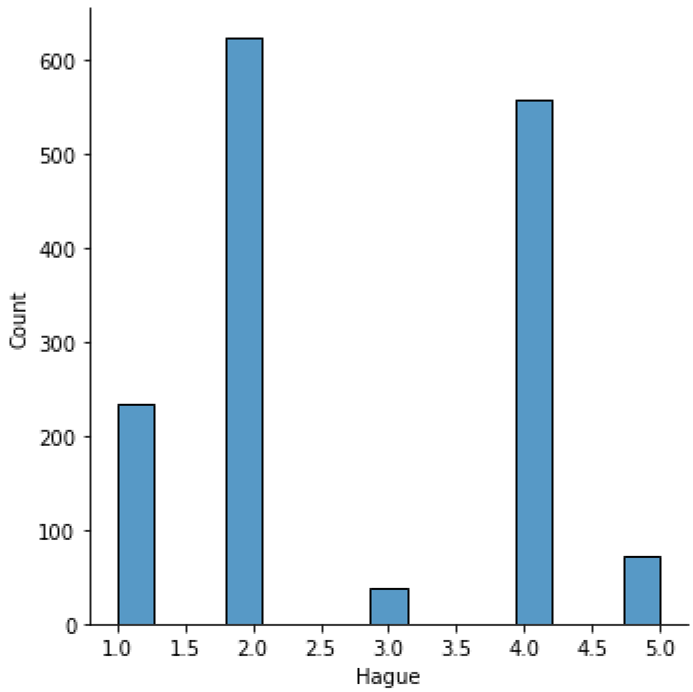

Fig No 6 – Histogram of Hague

Assessment of Conservative leader, most being at 2.

Histogram of Europe -

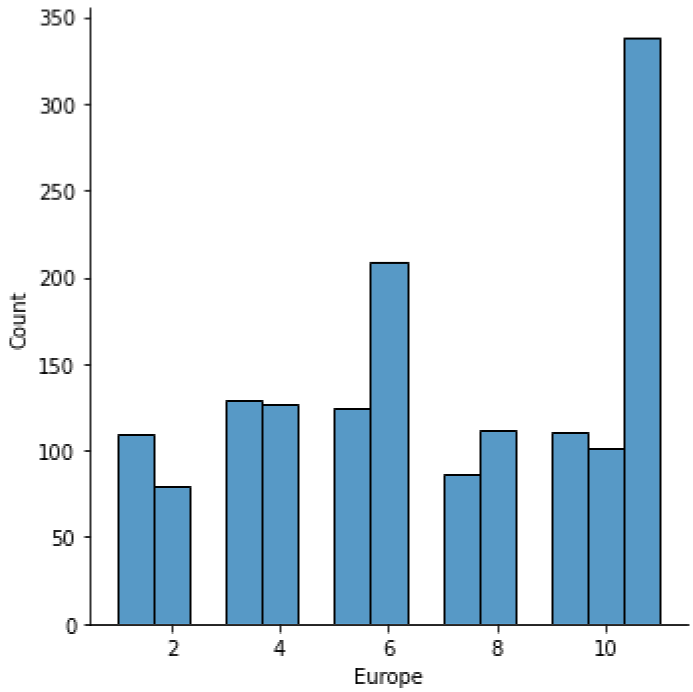

Fig No 7 – Histogram of Europe

An 11-point scale that measures respondents' attitudes toward European integration, is highest at 11.

Histogram of political_knowledge -

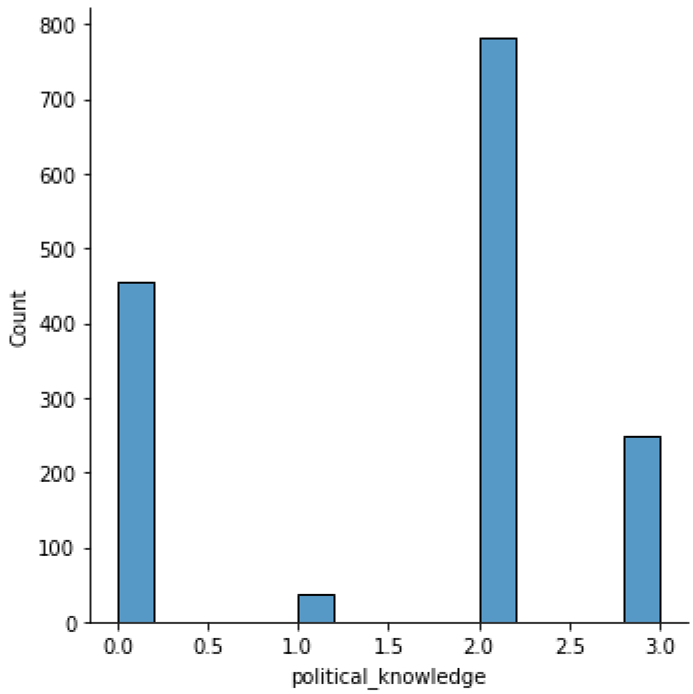

Fig No 8 – Histogram of political_knowledge

Knowledge of parties' positions on European integration, most being at 2.

Histogram of gender -

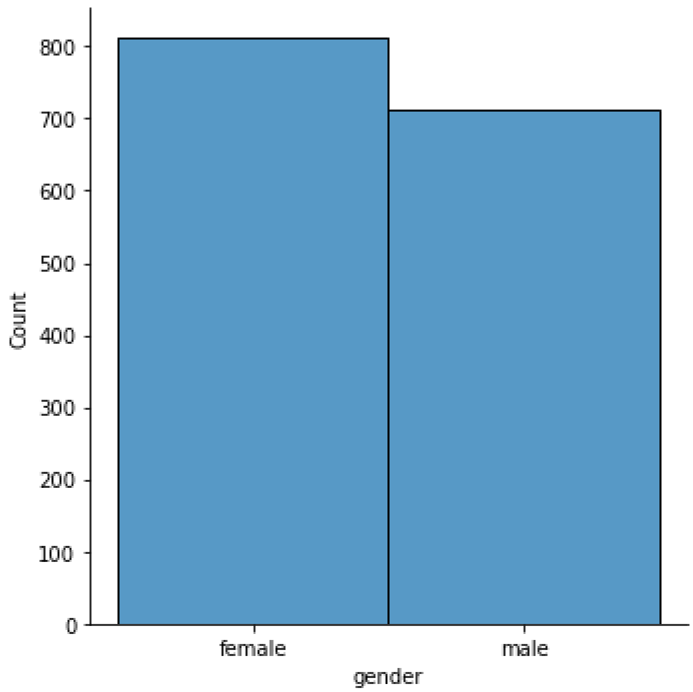

Fig No 9 – Histogram of gender

More number of female voters compare to men.

Bivariate Analysis -

Correlation Plot -

Helps us to visualize the correlation between continuous variables.

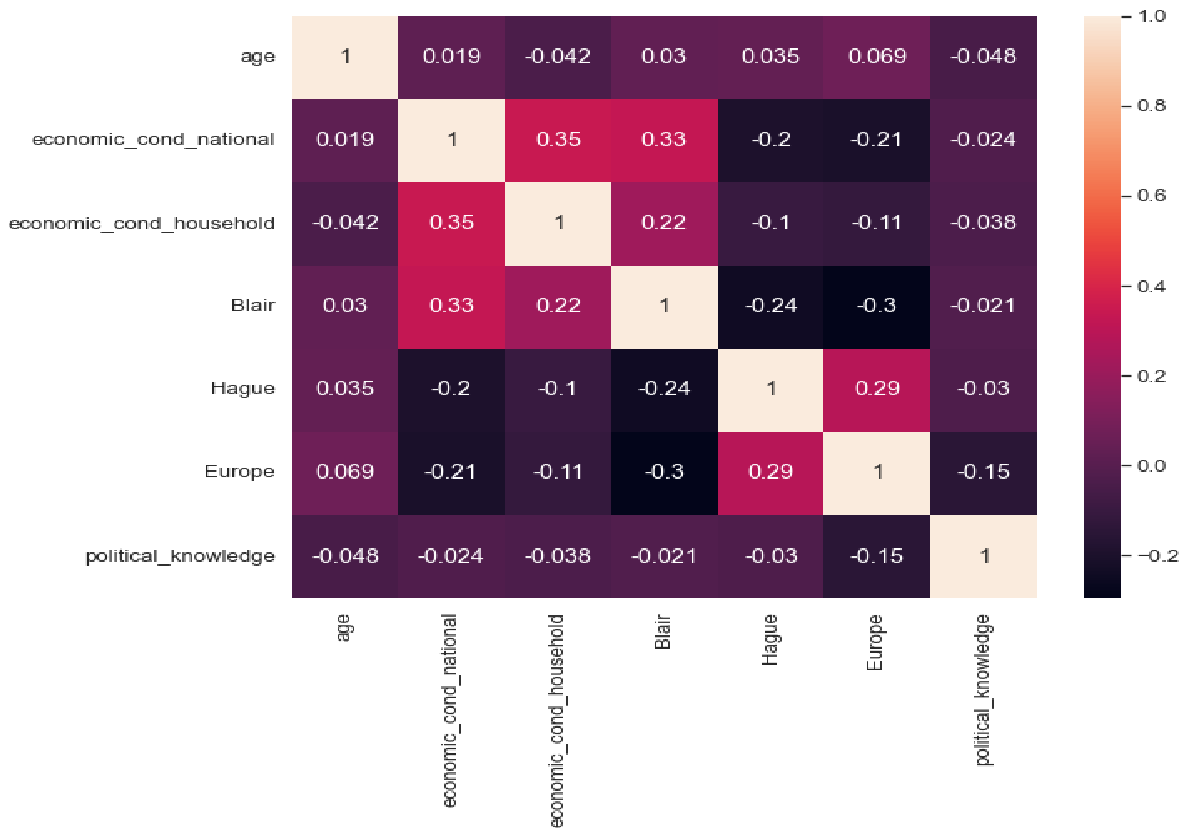

Fig No 10 – Correlation Plot

Pairplot -

Pairplot is a grid of scatterplots, showing the bivariate relationships between all pairs of variables in a multivariate dataset.

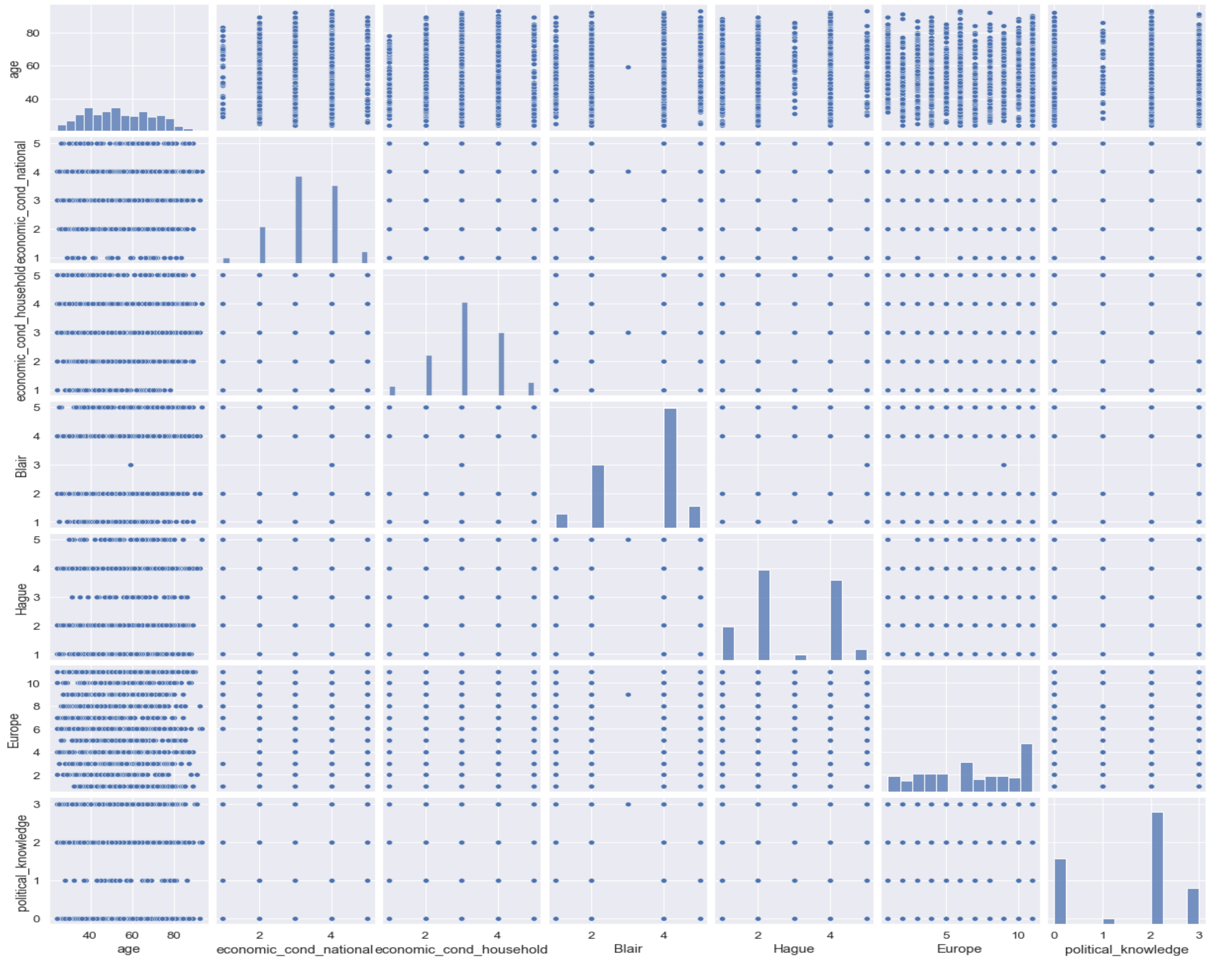

Fig No 11 – Pair Plot

Boxplot of age -

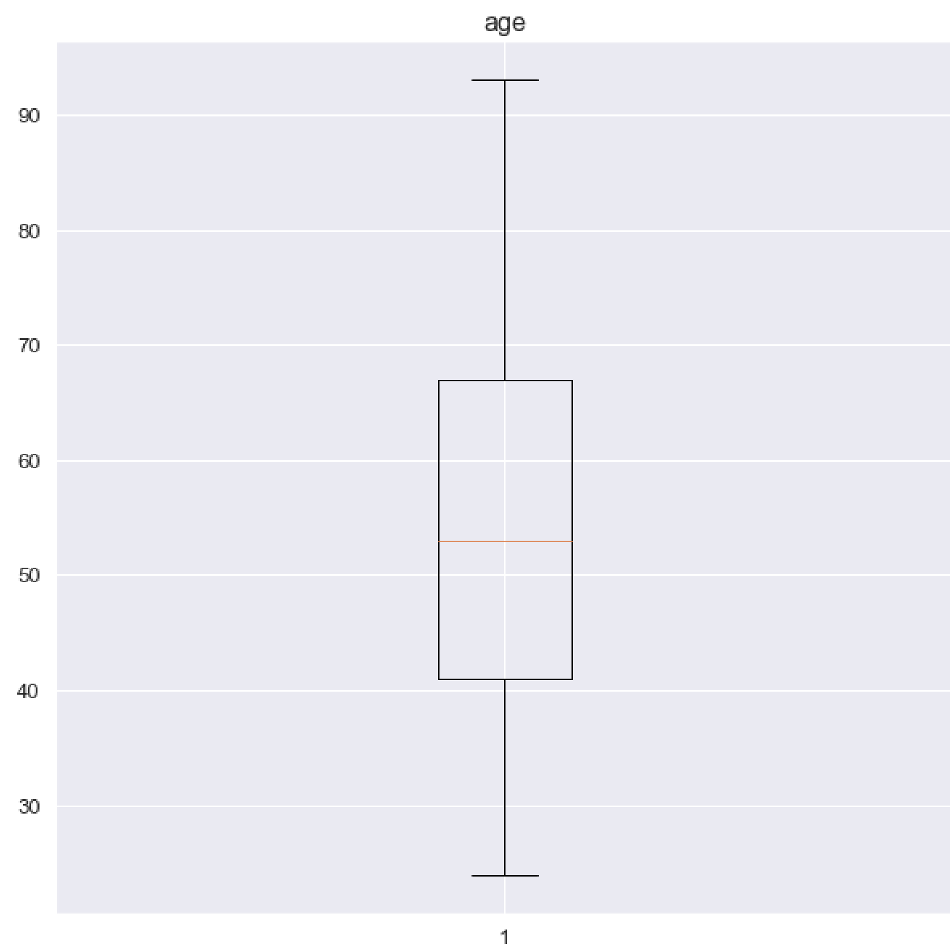

Fig No 12 – Boxplot of age

Most of the voters are between the age of 42 and 67.

Boxplot of economic_cond_national -

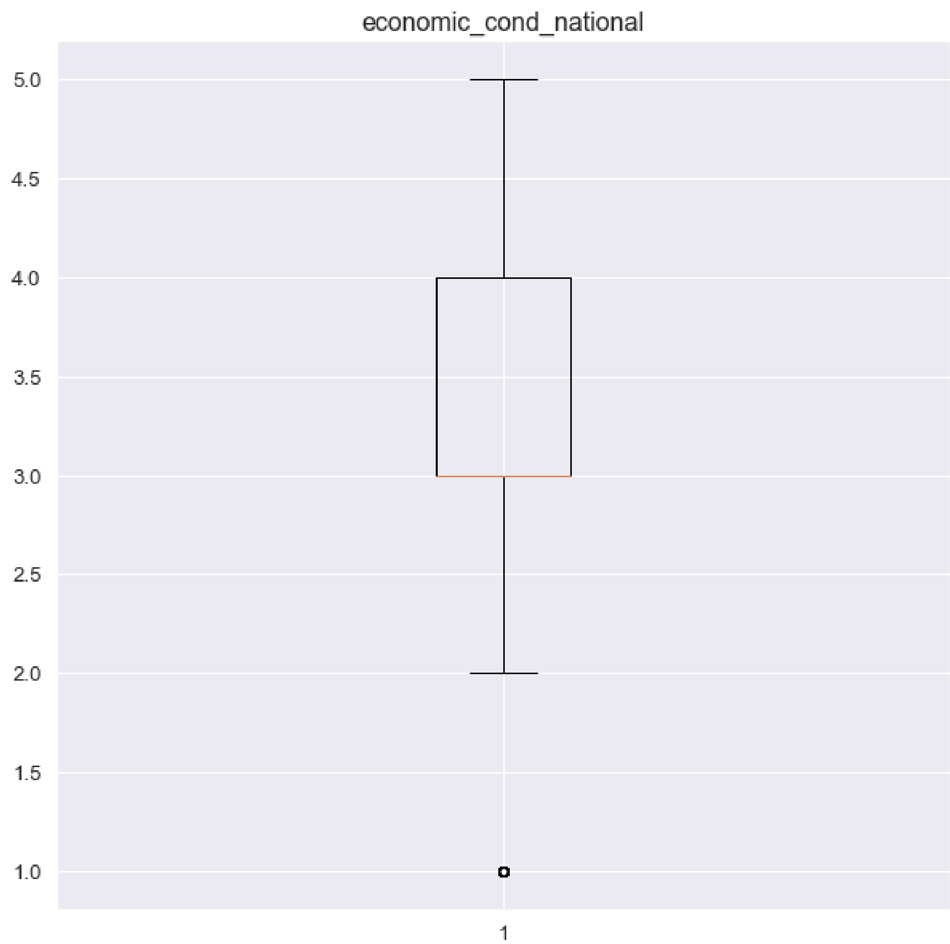

Fig No 13 – Boxplot of economic_cond_national

Current national economic conditions are between 3 and 4.

Boxplot of economic_cond_household -

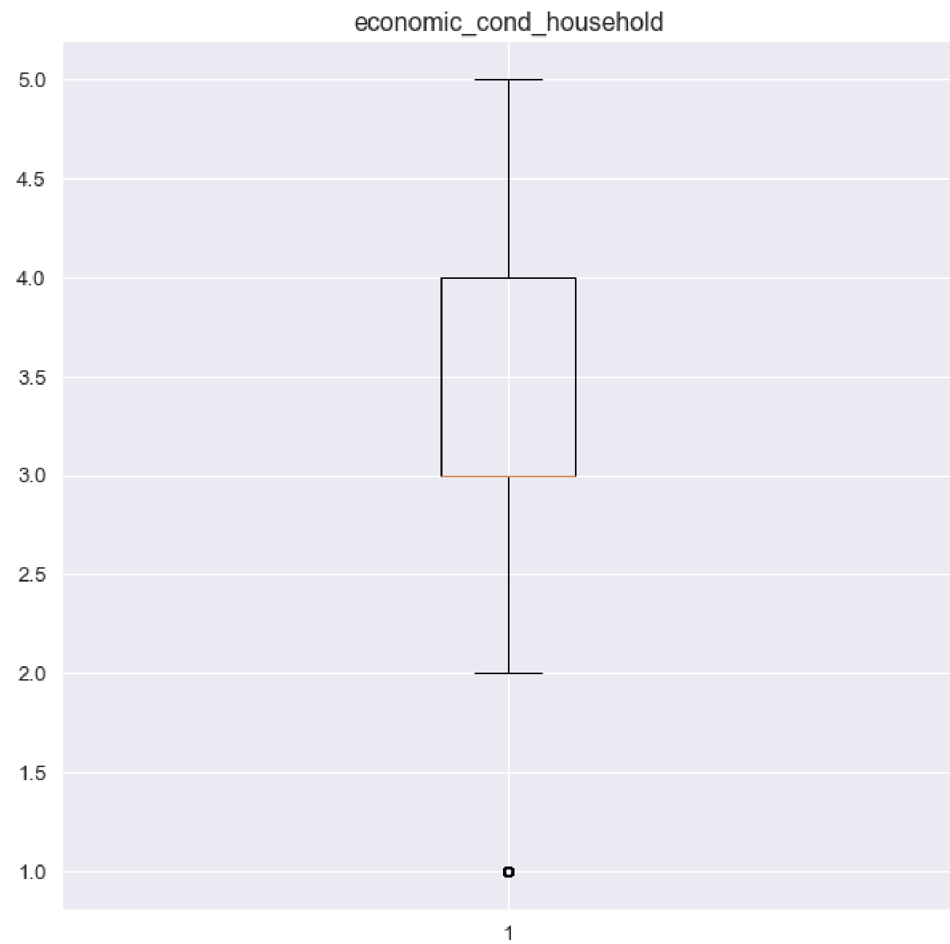

Fig No 14 – Boxplot of economic_cond_household

Current Household economic conditions are between 3 and 4.

Boxplot of Blair -

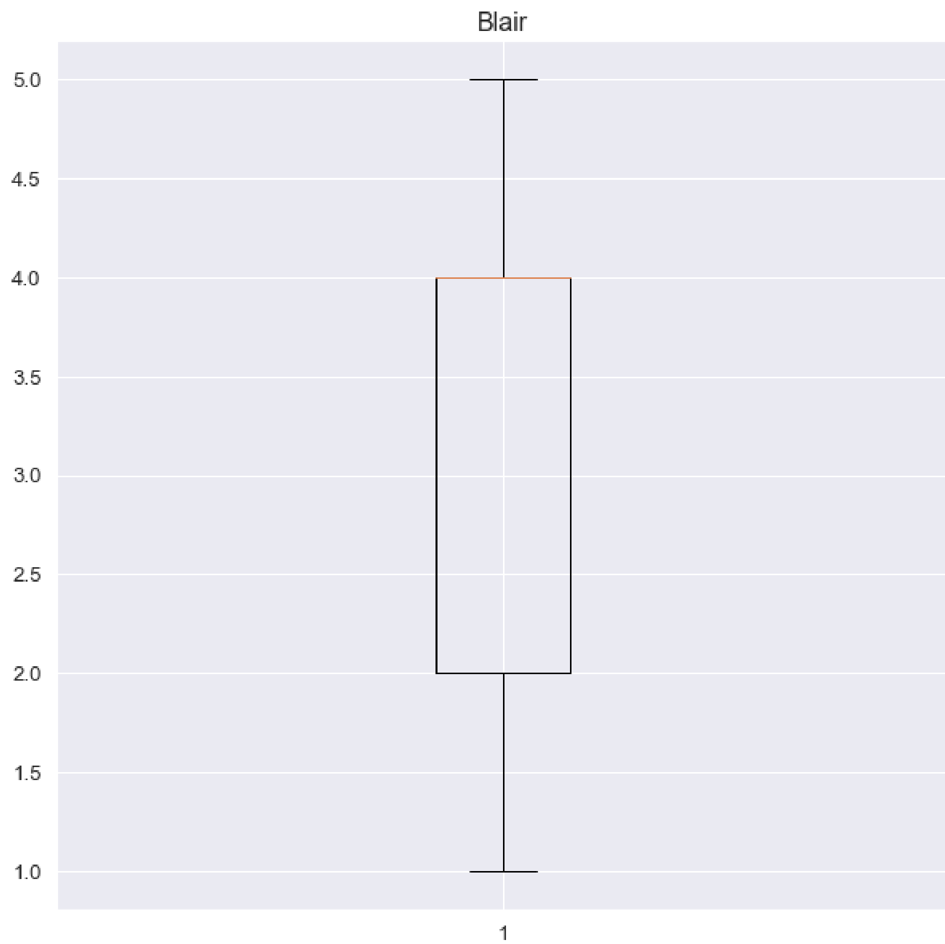

Fig No 15 – Boxplot of Blair

Assessment of the Labour leader is mostly between 2 and 4.

Boxplot of Hague -

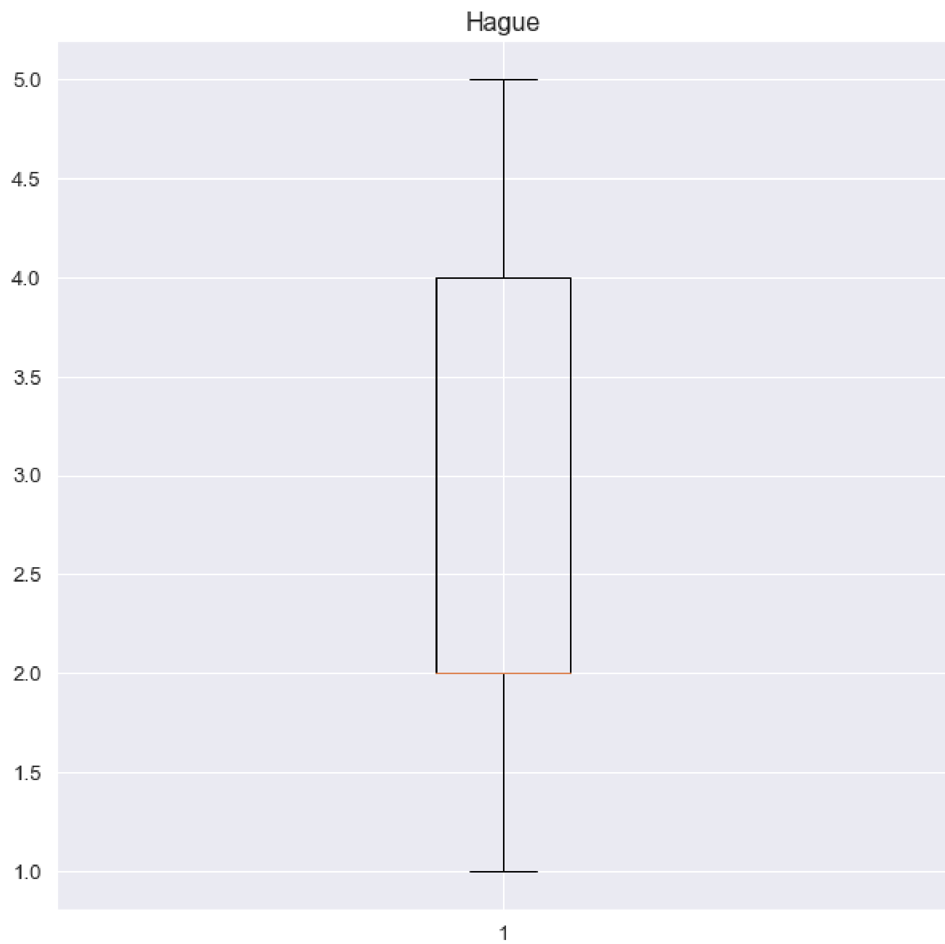

Fig No 16 – Boxplot of Hague

Assessment of the Conservative leader is mostly between 2 and 4.

Boxplot of Europe -

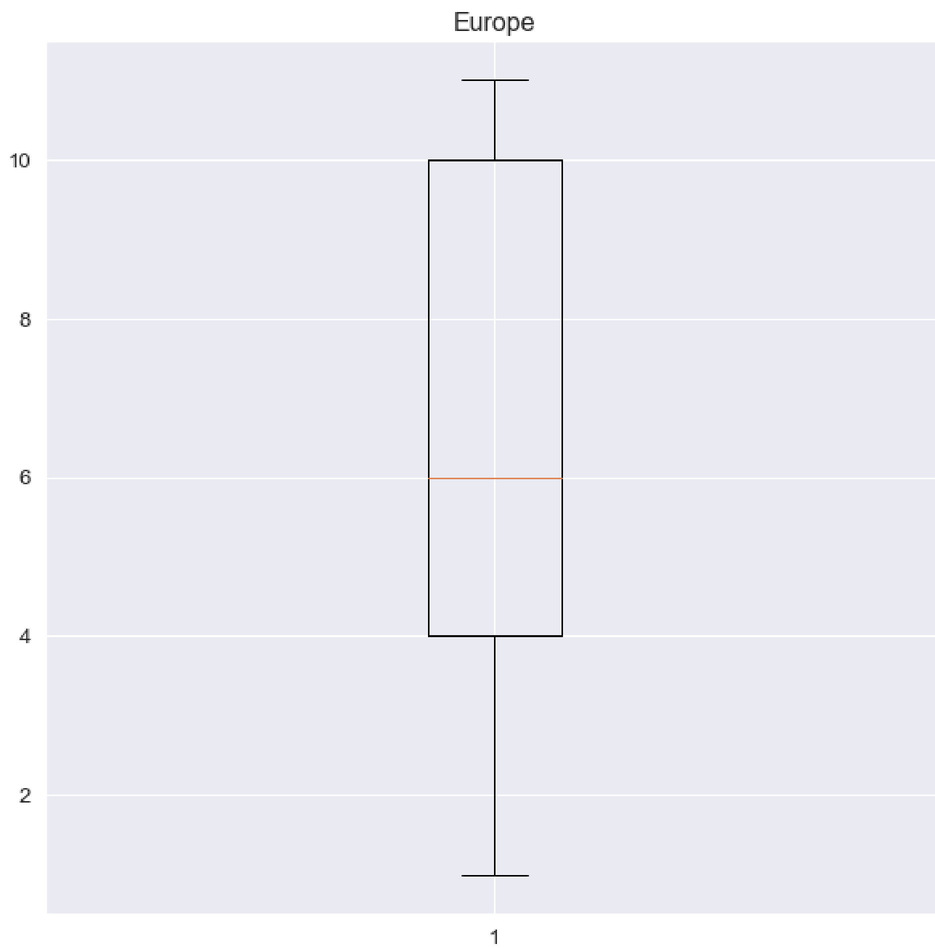

Fig No 17 – Boxplot of Europe

An 11-point scale that measures respondents' attitudes toward European integration, averages around 6 and highest being 11.

Boxplot of political_knowledge -

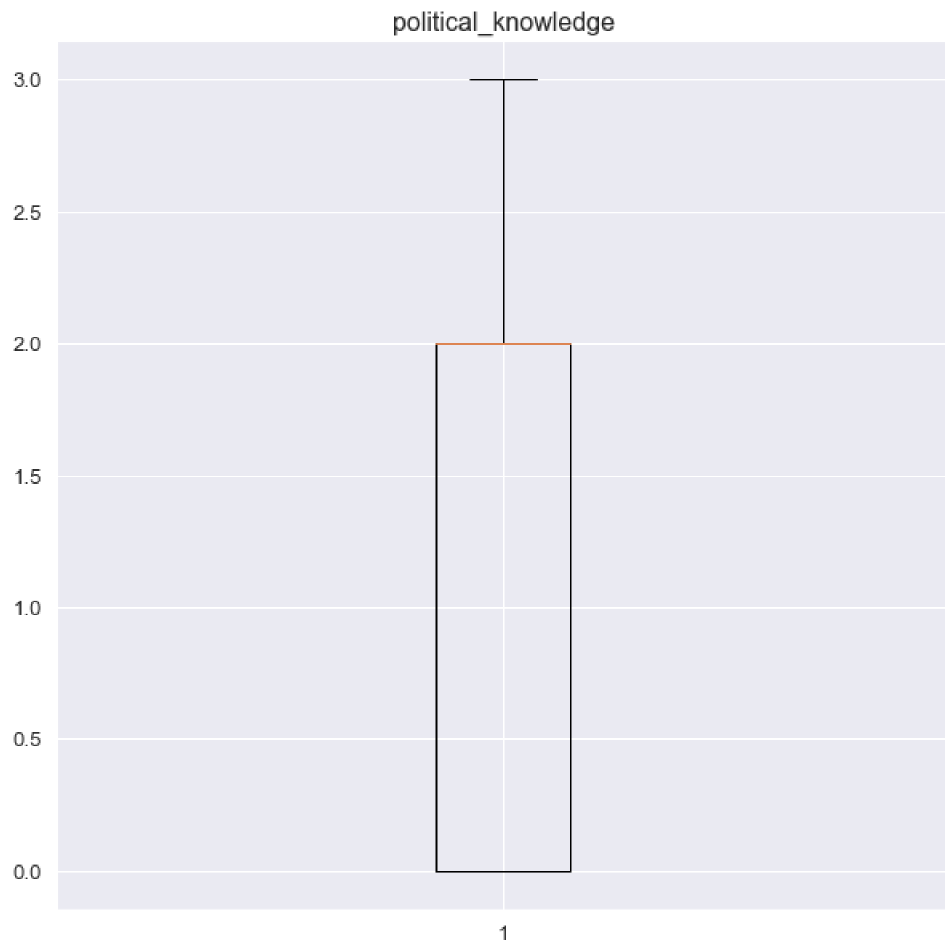

Fig No 18 – Boxplot of political_knowledge

Knowledge of parties' positions on European integration, average being at 2.

Boxplot after Outlier Treatment -

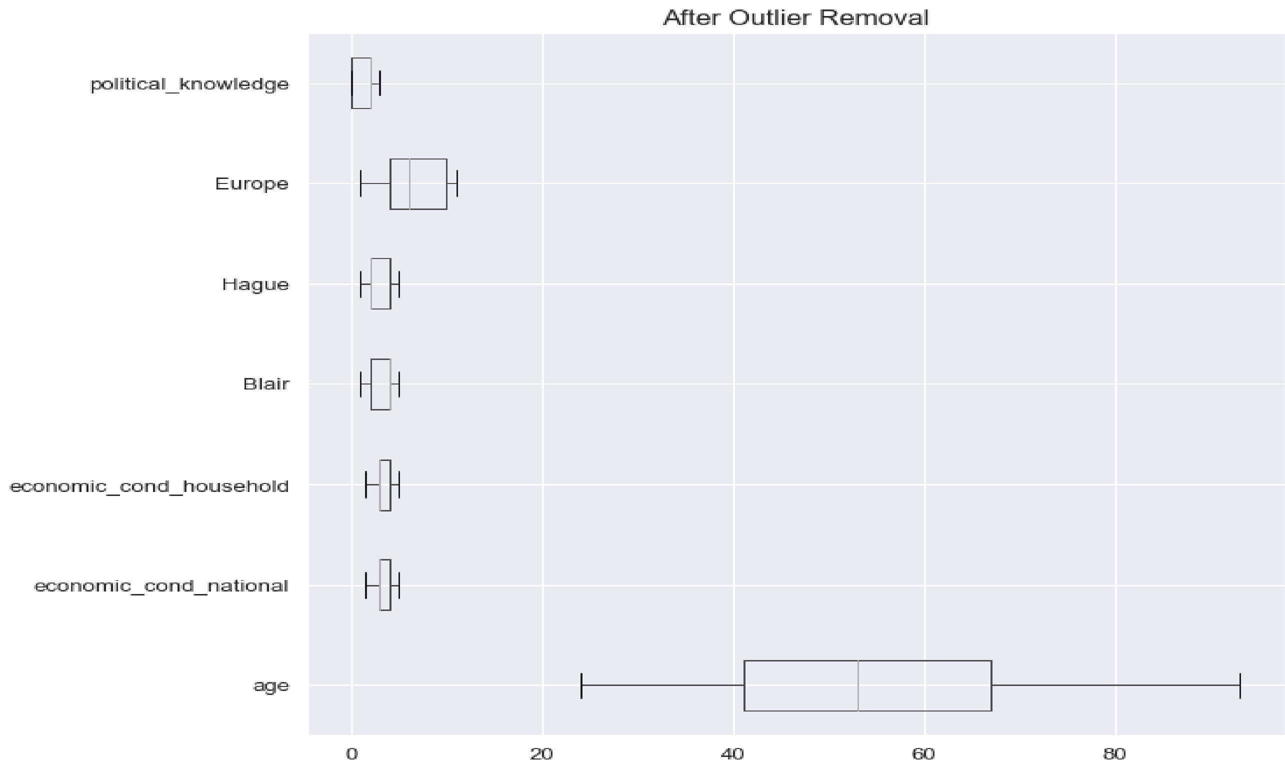

Fig No 19 – Boxplot after outlier treatment

Data Preparation:

1.3 Encode the data (having string values) for Modeling. Is Scaling necessary here or not? Data Split: Split the data into train and test (70:30).

Here the two categorical columns, vote and gender have been encoded for modeling.

The data has been scaled, some points in the data which were far from each other have come closer to each other after scaling.

The target column for the data is vote_labour.

The data has been split into 70 train and 30 test.

Modeling:

1.4 Apply Logistic Regression and LDA (linear discriminant analysis).

Logistic Regression Model -

Accuracy on Training data – 83%

AUC on Training data – 0.890

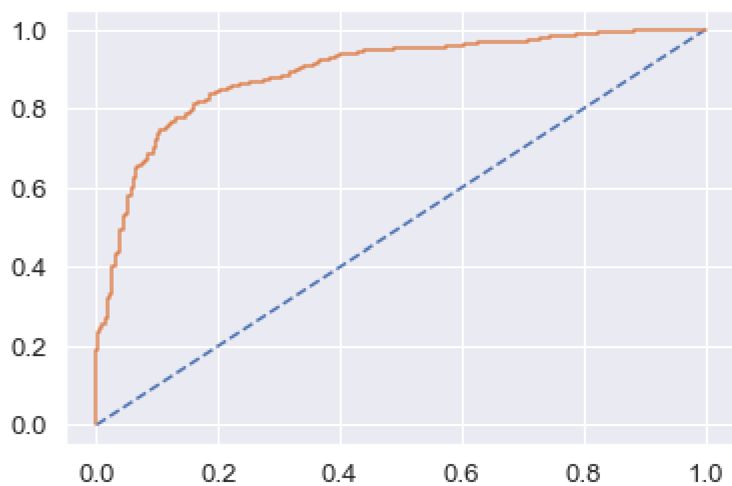

Fig No 20 – AUC on Training data

Confusion Matrix for training data -

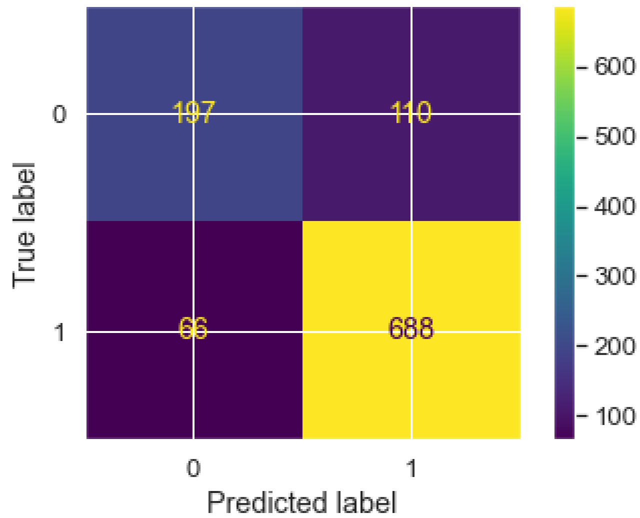

Fig No 21 – Confusion matrix on Training data

Classification report on training data-

	precision	recall	f1-score	support
0	0.75	0.64	0.69	307
1	0.86	0.91	0.89	754
accuracy			0.83	1061
macro avg	0.81	0.78	0.79	1061
weighted avg	0.83	0.83	0.83	1061

Accuracy on Test data – 89%

AUC on Test data – 0.890

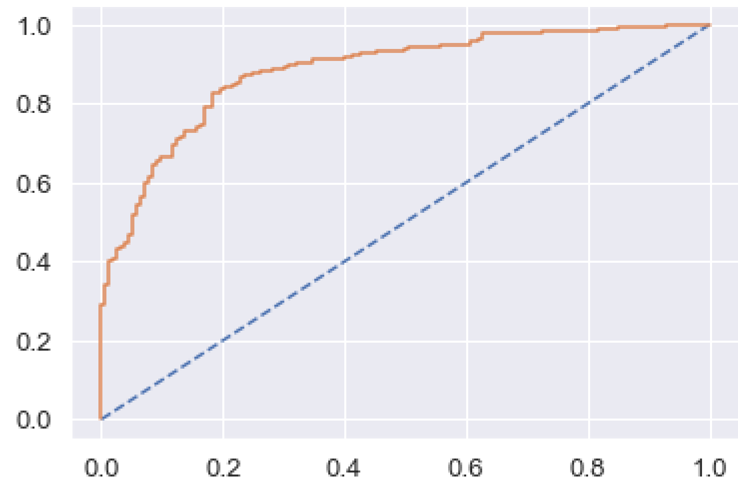

Fig No 22 – AUC on Test data

Confusion matrix for test data -

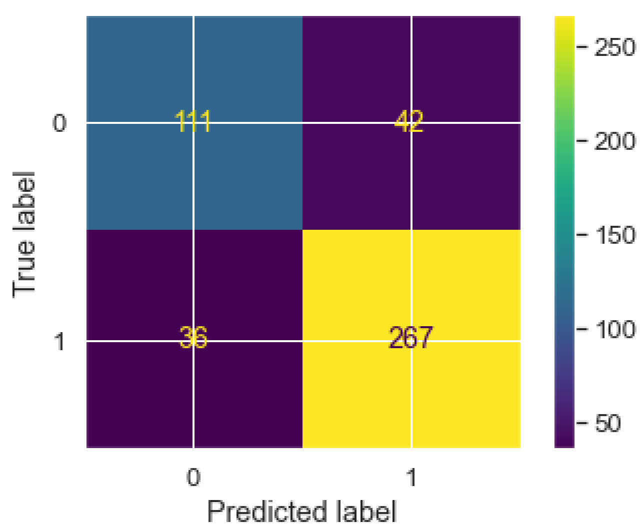

Fig No 23 – confusion matrix on test data

Classification report on Test data-

	precision	recall	f1-score	support
0	0.76	0.73	0.74	153
1	0.86	0.88	0.87	303

	precision	recall	f1-score	support
accuracy			0.83	456
macro avg	0.81	0.80	0.81	456
weighted avg	0.83	0.83	0.83	456

Overall accuracy of the model is 83% which means 83% of the predictions are correct. Precision and recall for test data are almost in line with training data, therefore no overfitting or underfitting has happened and overall model is a good model for classification.

LDA (Linear Discriminant Analysis) -

Training data and Test data confusion matrix comparison -

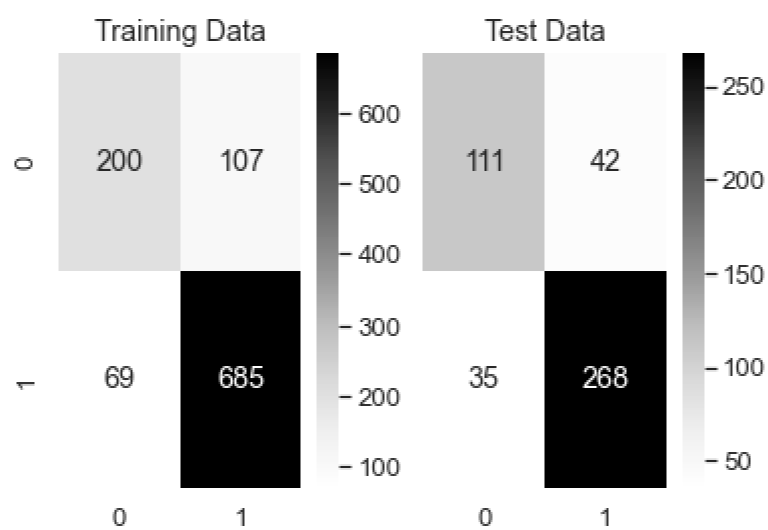

Fig No 24 – confusion matrix

Training Data and Test Data Classification Report Comparison -

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.74	0.65	0.69	307
1	0.86	0.91	0.89	754

	precision	recall	f1-score	support
accuracy			0.83	1061
macro avg	0.80	0.78	0.79	1061
weighted avg	0.83	0.83	0.83	1061

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.76	0.73	0.74	153
1	0.86	0.88	0.87	303
accuracy			0.83	456
macro avg	0.81	0.80	0.81	456
weighted avg	0.83	0.83	0.83	456

AUC for the Training data – 0.890

AUC for the Test data – 0.888

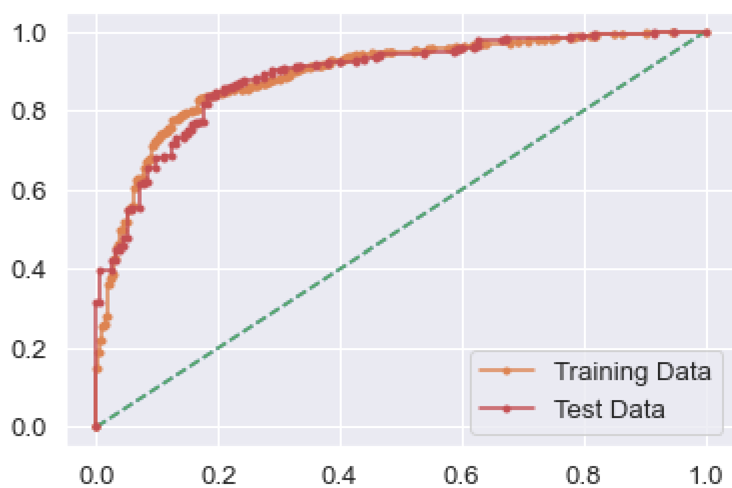

Fig No 25 – AUC on Training data and test data

Overall accuracy of the model is 83%, which means 83% of the predictions are correct. Precision and recall for test data are almost in line with training data, therefore no overfitting or underfitting has happened and overall model is a good model for classification.

1.5 Apply KNN Model and Naïve Bayes Model. Interpret the results.

KNN Model -

Performance matrix on training data -

```
0.8548539114043355
[[216  91]
 [ 63 691]]
              precision    recall  f1-score   support

           0       0.77      0.70      0.74       307
           1       0.88      0.92      0.90       754

    accuracy                           0.85      1061
   macro avg       0.83      0.81      0.82      1061
weighted avg       0.85      0.85      0.85      1061
```

Performance matrix on test data -

```
0.8245614035087719
[[109  44]
 [ 36 267]]
              precision    recall  f1-score   support

           0       0.75      0.71      0.73       153
           1       0.86      0.88      0.87       303

    accuracy                           0.82       456
   macro avg       0.81      0.80      0.80       456
weighted avg       0.82      0.82      0.82       456
```

Misclassification error -

```
[0.2171052631578947,
 0.19517543859649122,
 0.17543859649122806,
 0.18201754385964908,
 0.1842105263157895,
 0.17324561403508776,
 0.17763157894736847,
 0.16666666666666663,
 0.16666666666666663,
 0.17543859649122806]
```

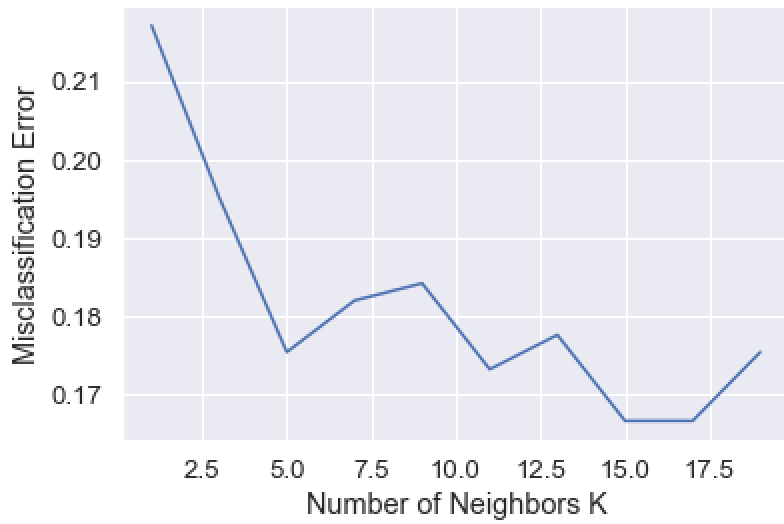

Fig No 26 – Misclassification error

For K=11 it is giving the best test accuracy let's check train and test for K=11 with other evaluation metrics.

Performance matrix on train data when K = 11

```
0.8416588124410933
[[204 103]
 [ 65 689]]
              precision    recall  f1-score   support

           0       0.76      0.66      0.71       307
           1       0.87      0.91      0.89       754

    accuracy                           0.84      1061
```

```
   macro avg       0.81      0.79      0.80      1061
weighted avg       0.84      0.84      0.84      1061
```

Performance matrix on test data when K = 11

```
0.8267543859649122
[[104  49]
 [ 30 273]]
              precision    recall  f1-score   support

           0       0.78      0.68      0.72       153
           1       0.85      0.90      0.87       303

    accuracy                           0.83       456
   macro avg       0.81      0.79      0.80       456
weighted avg       0.82      0.83      0.82       456
```

Looking at the train and test accuracies it is a valid model.

Naïve Bayes Model -

Performance Matrix on train data set -

```
0.8341187558906692
[[212  95]
 [ 81 673]]
              precision    recall  f1-score   support

           0       0.72      0.69      0.71       307
           1       0.88      0.89      0.88       754

    accuracy                           0.83      1061
   macro avg       0.80      0.79      0.80      1061
weighted avg       0.83      0.83      0.83      1061
```

Performance Matrix on test data set -

```
0.8223684210526315
[[112  41]
 [ 40 263]]
              precision    recall  f1-score   support
```

```
           0       0.74      0.73      0.73       153
           1       0.87      0.87      0.87       303

    accuracy                           0.82       456
   macro avg       0.80      0.80      0.80       456
weighted avg       0.82      0.82      0.82       456
```

Accuracy of our Gaussian Naive Bayes model -

Train score – 83%

Test score – 82%

Looking at Recalls, Training accuracy and Test accuracy. Model seems to be performing well.

1.6 Model Tuning, Bagging (Random Forest should be applied for Bagging), and boosting.

Random Forest Classifier -

Performance Matrix on train data set -

```
1.0
[[307   0]
 [  0 754]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       307
           1       1.00      1.00      1.00       754

    accuracy                           1.00      1061
   macro avg       1.00      1.00      1.00      1061
weighted avg       1.00      1.00      1.00      1061
```

Performance Matrix on test data set -

```
0.8289473684210527
[[105  48]
 [ 30 273]]
              precision    recall  f1-score   support

           0       0.78      0.69      0.73       153
           1       0.85      0.90      0.88       303

    accuracy                           0.83       456
   macro avg       0.81      0.79      0.80       456
weighted avg       0.83      0.83      0.83       456
```

Bagging Classifier -

Performance Matrix on train data set -

```
1.0
[[307   0]
 [  0 754]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       307
           1       1.00      1.00      1.00       754

    accuracy                           1.00      1061
   macro avg       1.00      1.00      1.00      1061
weighted avg       1.00      1.00      1.00      1061
```

Performance Matrix on test data set -

```
0.8201754385964912
[[108  45]
 [ 37 266]]
              precision    recall  f1-score   support

           0       0.74      0.71      0.72       153
           1       0.86      0.88      0.87       303

    accuracy                           0.82       456
   macro avg       0.80      0.79      0.80       456
weighted avg       0.82      0.82      0.82       456
```

Ada Boost -

Performance Matrix on train data set -

```
0.8501413760603205
[[214  93]
 [ 66 688]]
              precision    recall  f1-score   support

           0       0.76      0.70      0.73       307
           1       0.88      0.91      0.90       754

    accuracy                           0.85      1061
   macro avg       0.82      0.80      0.81      1061
weighted avg       0.85      0.85      0.85      1061
```

Performance Matrix on test data set -

```
0.8135964912280702
[[103  50]
 [ 35 268]]
              precision    recall  f1-score   support

           0       0.75      0.67      0.71       153
           1       0.84      0.88      0.86       303

    accuracy                           0.81       456
   macro avg       0.79      0.78      0.79       456
weighted avg       0.81      0.81      0.81       456
```

Gradient Boosting -

Performance Matrix on train data set -

```
0.8925541941564562
[[239  68]
 [ 46 708]]
              precision    recall  f1-score   support

           0       0.84      0.78      0.81       307
           1       0.91      0.94      0.93       754
```

```
    accuracy                           0.89      1061
   macro avg       0.88      0.86      0.87      1061
weighted avg       0.89      0.89      0.89      1061
```

Performance Matrix on test data set -

```
0.8333333333333334
[[104  49]
 [ 27 276]]
              precision    recall  f1-score   support

           0       0.79      0.68      0.73       153
           1       0.85      0.91      0.88       303

    accuracy                           0.83       456
   macro avg       0.82      0.80      0.81       456
weighted avg       0.83      0.83      0.83       456
```

1.7 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model. Final Model: Compare the models and write inference which model is best/optimized.

Accuracy of the models -

Logistic Regression – Train (83%) - Test (89%)

LDA – Train (83%) - Test (83%)

KNN – Train (84%) - Test (83%)

Naïve Bayes – Train (83%) - Test (82%)

Random Forest Classifier – Train (100%) - Test (83%)

Bagging – Train (100%) - Test (82%)

Ada Boost – Train (85%) - Test (71%)

Gradient Boosting – Train (89%) - Test (83%)

Looking at all the model precision, recall and accuracy the LDA model suits the best for the problem. The accuracy of the LDA model is 83% on both test and train data.

```
Classification Report of the training data:

              precision    recall  f1-score   support

           0       0.74      0.65      0.69       307
           1       0.86      0.91      0.89       754

    accuracy                           0.83      1061
   macro avg       0.80      0.78      0.79      1061
weighted avg       0.83      0.83      0.83      1061


Classification Report of the test data:

              precision    recall  f1-score   support

           0       0.76      0.73      0.74       153
           1       0.86      0.88      0.87       303

    accuracy                           0.83       456
   macro avg       0.81      0.80      0.81       456
weighted avg       0.83      0.83      0.83       456
```

The Precision and recall for test data are almost in line with training data, therefore no overfitting or underfitting has happened and overall model is a good model for classification.

Inference -

1.8 Based on these predictions, what are the insights?

Most of the voters are around the age of 50.

Assessment of both the party leaders on a scale of 1 to 5 is between 2 and 4.

Looking at the data it is likely a voter will vote for Labour party.

70% of the voters are voting for Labour party.

Problem 2 -

2.1 Find the number of characters, words, and sentences for the mentioned documents.

Number of Roosevelt Words - 1536

Number of Roosevelt Sentences - 68

Number of Roosevelt raw - 7571

Number of Kennedy Words - 1546

Number of Kennedy Sentences - 52

Number of Kennedy raw - 7618

Number of Nixon Words - 2028

Number of Nixon Sentences - 69

Number of Nixon raw – 9991

2.2 Remove all the stop words from all three speeches.

Number of Roosevelt Words - 1536

After removing stop words – 720

Number of Kennedy Words - 1546

After removing stop words – 764

Number of Nixon Words - 2028

After removing stop words – 912

2.3 Which word occurs the most number of times in his inaugural address for each president? Mention the top three words.

Roosevelt most occurring words -

'It' – 13 times

'The' – 10 times

'know' – 10 times

'--' is not considered as it is not a word.

Kennedy most occurring words -

'us' - 12 times

'world' - 8 times

'Let' - 8 times

'--' is not considered as it is not a word.

Nixon most occurring words -

'us' - 26 times

'America' - 21 times

'peace' - 19 times

2.4 Plot the word cloud of each of the speeches of the variable.

Word Cloud for Roosevelt -

Fig No 27 – Word cloud for Roosevelt

Word Cloud for Kennedy -

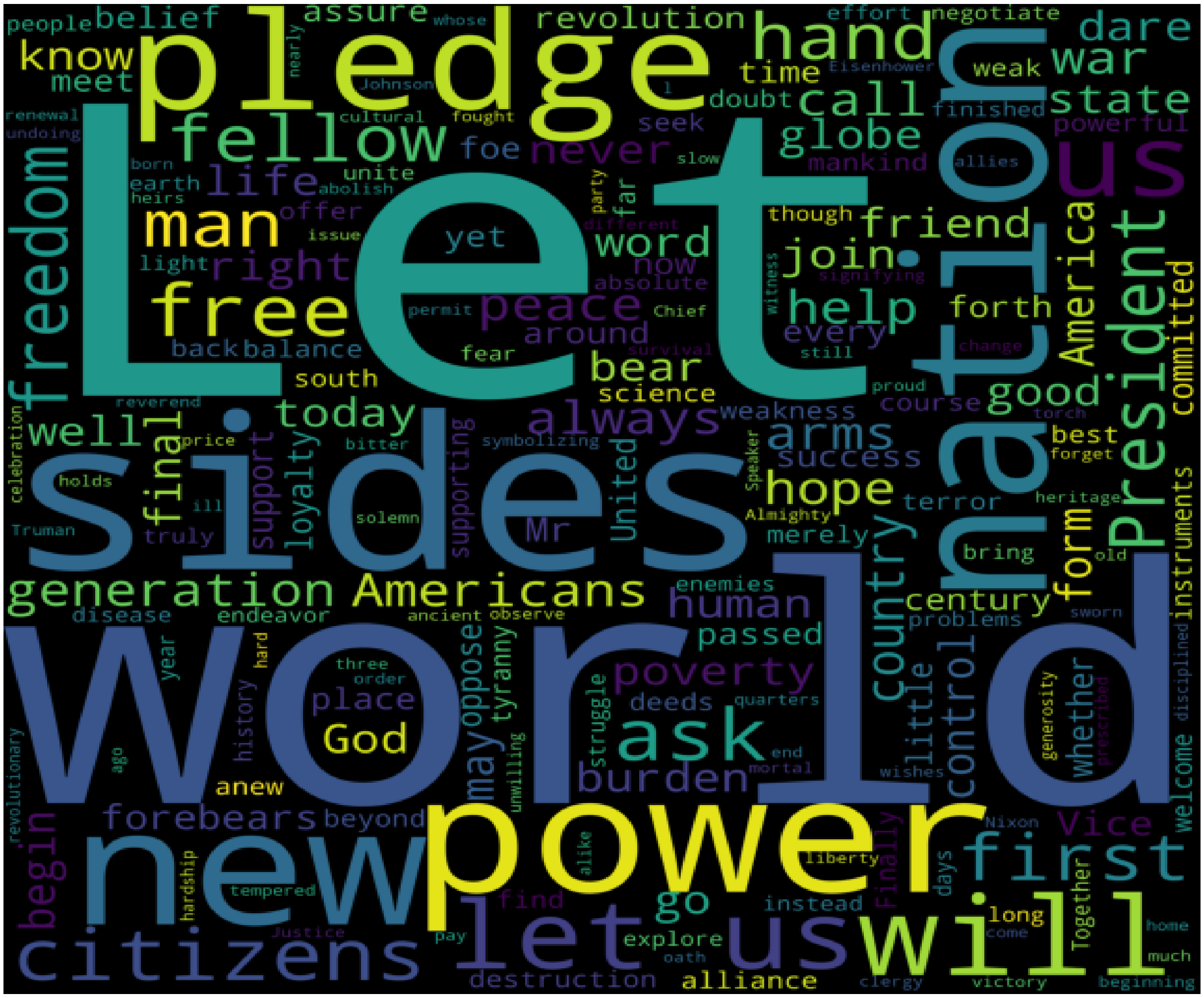

Fig no 28 – Word cloud for Kennedy

Word Cloud for Nixon -

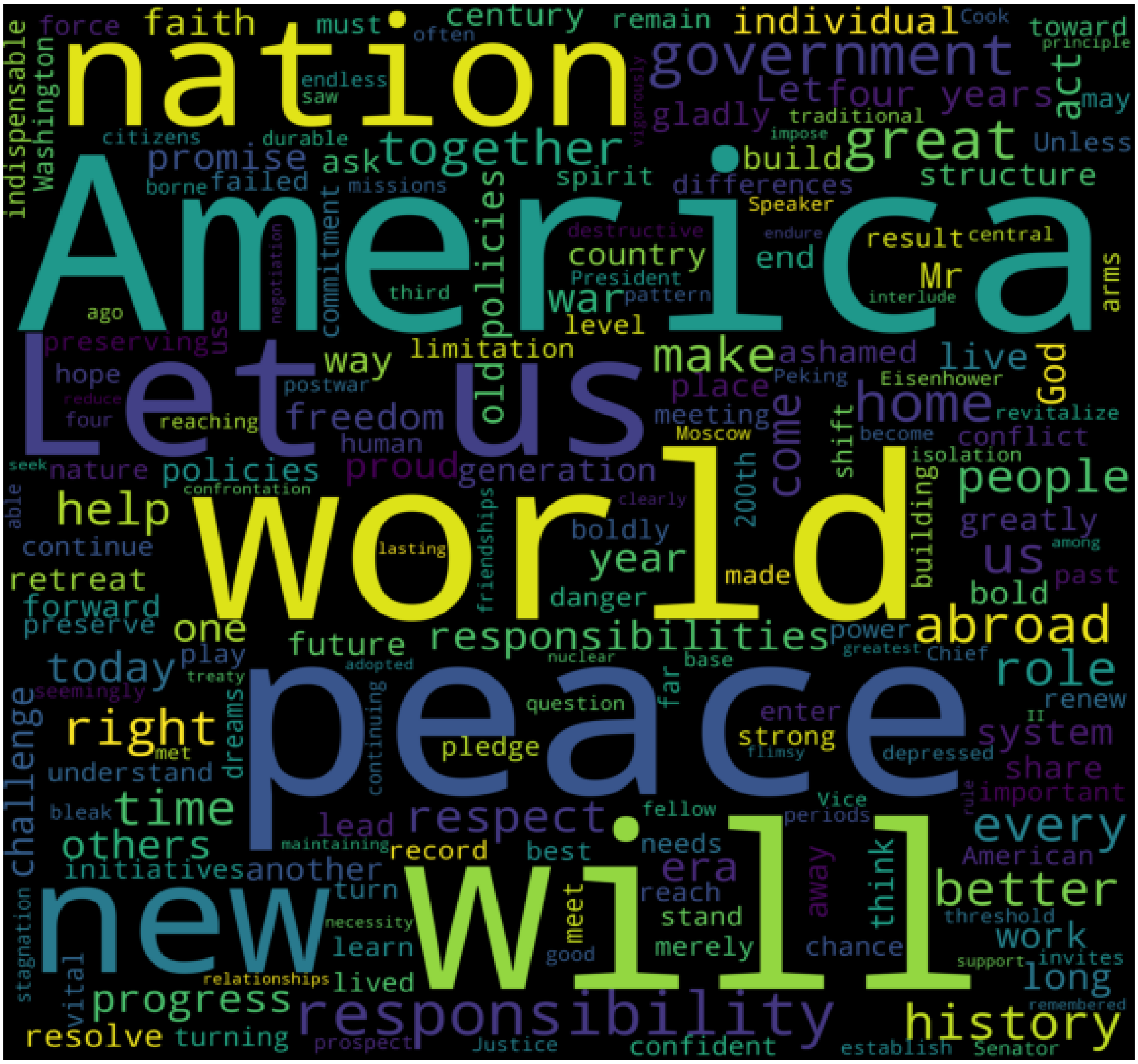

Fig no 29 – Word cloud for Nixon