Supervised and Unsupervised Learning

Learning Graded Project

Week 4

Table of Contents

Problem 1

You are hired by one of the leading news channels CNBE who wants to analyze recent elections. This survey was conducted on 1525 voters with 9 variables. You 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 covered by a particular party.

A Data Ingestion

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

First 5 rows of the data

	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
1	Labour	43	3	3	4	1	2	2	female
2	Labour	36	4	4	4	4	5	2	male
3	Labour	35	4	4	5	2	3	2	male
4	Labour	24	4	2	2	1	4	0	female
5	Labour	41	2	2	1	1	6	2	male

Information of the data and datatypes of each respective attribute:

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1525 entries, 1 to 1525
Data columns (total 9 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   vote                    1525 non-null   object
 1   age                     1525 non-null   int64
 2   economic.cond.national  1525 non-null   int64
 3   economic.cond.household 1525 non-null   int64
 4   Blair                   1525 non-null   int64
 5   Hague                   1525 non-null   int64
 6   Europe                  1525 non-null   int64
 7   political.knowledge     1525 non-null   int64
 8   gender                  1525 non-null   object
dtypes: int64(7), object(2)
memory usage: 119.1+ KB
```

EDA Descriptive Statistics:

	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge
count	1525.000000	1525.000000	1525.000000	1525.000000	1525.000000	1525.000000	1525.000000
mean	54.182295	3.245902	3.140328	3.334426	2.746885	6.728525	1.542295
std	15.711209	0.880969	0.929951	1.174824	1.230703	3.297538	1.083315
min	24.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.000000
25%	41.000000	3.000000	3.000000	2.000000	2.000000	4.000000	0.000000
50%	53.000000	3.000000	3.000000	4.000000	2.000000	6.000000	2.000000
75%	67.000000	4.000000	4.000000	4.000000	4.000000	10.000000	2.000000
max	93.000000	5.000000	5.000000	5.000000	5.000000	11.000000	3.000000

Check for Null Values:

There is no null values.

```
vote                        0
age                         0
economic.cond.national      0
economic.cond.household     0
Blair                       0
Hague                       0
Europe                      0
political.knowledge         0
gender                      0
dtype: int64
```

Check for dublicated values:

There was found 8 dublicated rows

Dropping these values

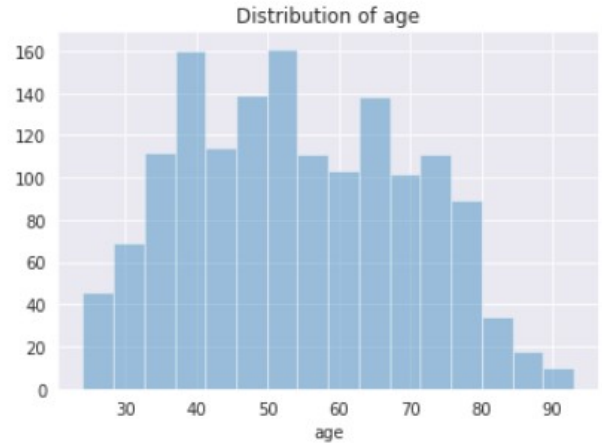
```
Before:  (1525, 9)
After:  (1517, 9)
```

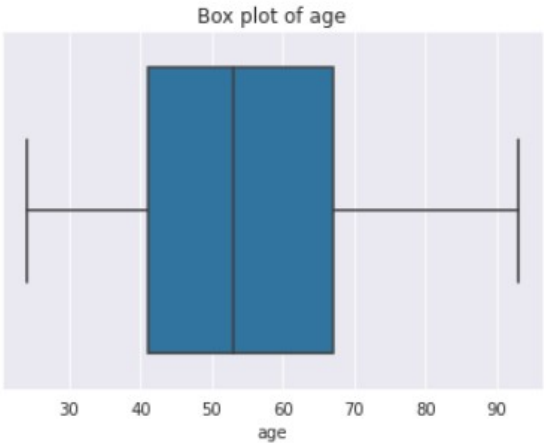
Inferences:

- On performing the descriptive analysis, we can see that there are a few columns having categorical values but are not having the data type “object”
- The Election dataset have 1525 rows and 9 columns. All the variables except vote and gender are int64 datatypes.
- ‘vote’ have two unique values Labour and Conservative, which is also a dependent variable.
- ‘gender’ has two unique values male and female.
- There are no null values in the data set.
- There are 8 duplicate rows. Even though they could represent different person with exact same profile and political outlook, we drop these rows as they are few in number and add no value to the data set.

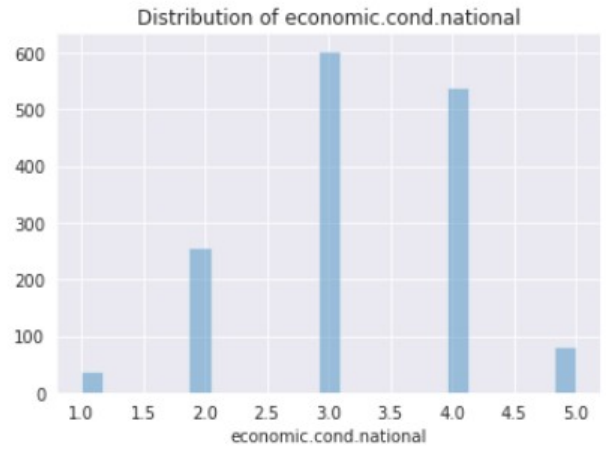
1.2 Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers.

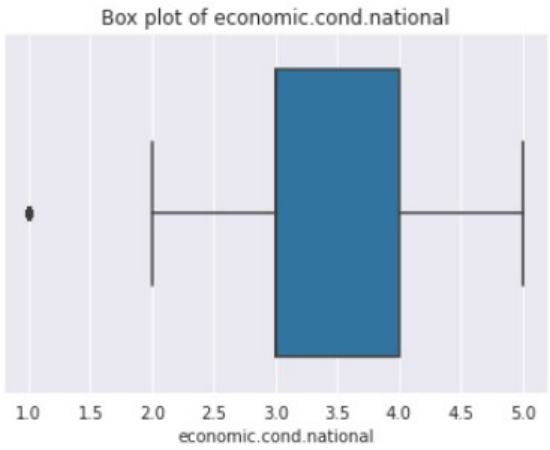
Univatiate Analysis



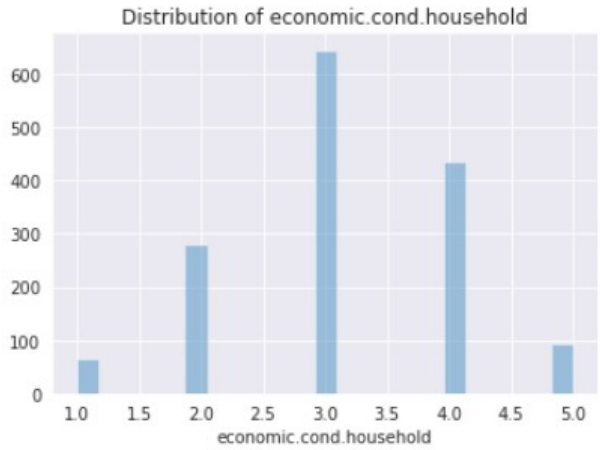


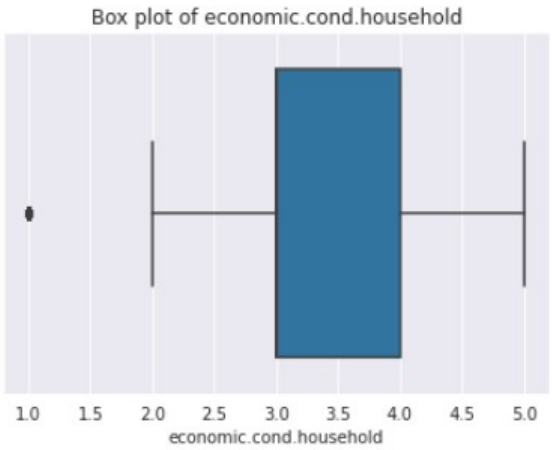
Distribution of ‘age’ resembles normal distribution and is slightly right skewed. Most of the respodent in the data is in age between 40 and 65.



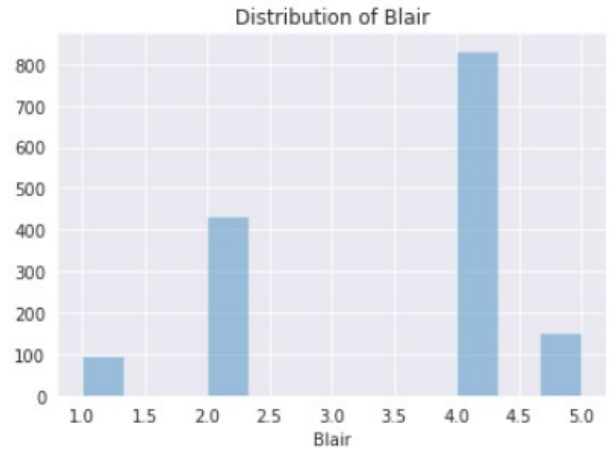


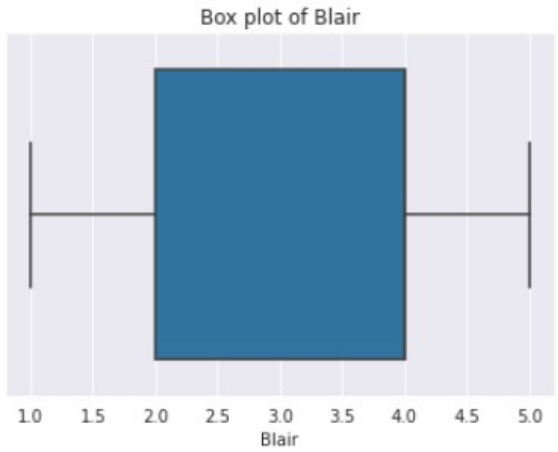
Distribution of ‘economic.cond.national’ isn’t normal distribution and is slightly left skewed. Out of the 1525 participants around 600 participants rated the national economic condition as more than average



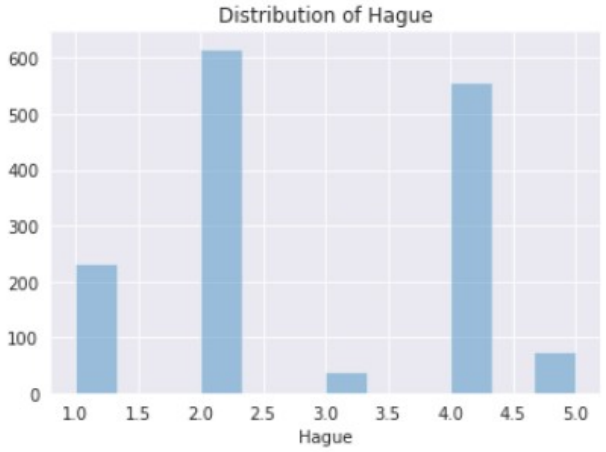


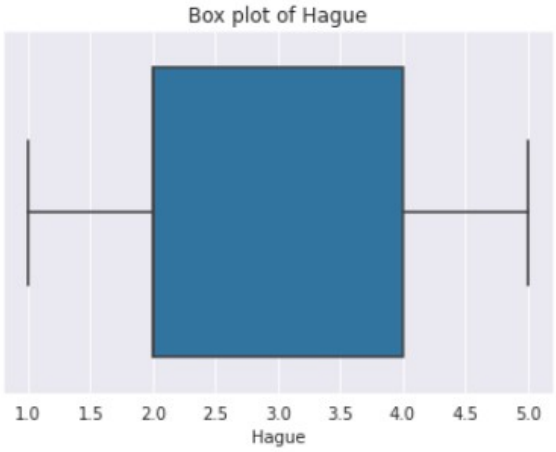
Distribution of ‘economic.cond.househols’ isn’t normal distribution and is slightly left skewed.



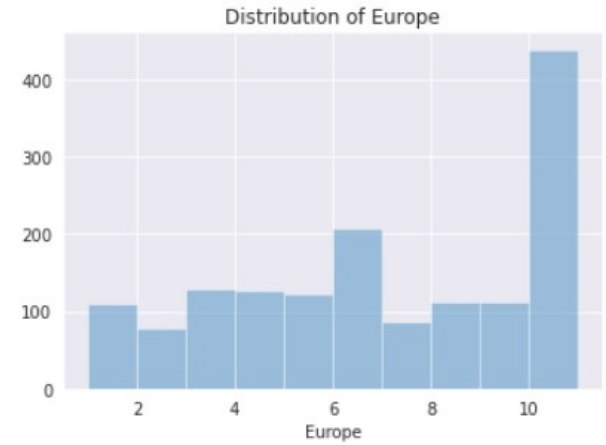


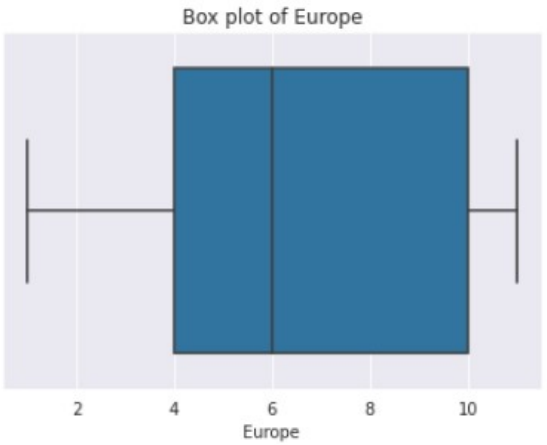
Distribution of ‘Blair’ isn’t normal distribution and is slightly left skewed, because mode is greater than mean.



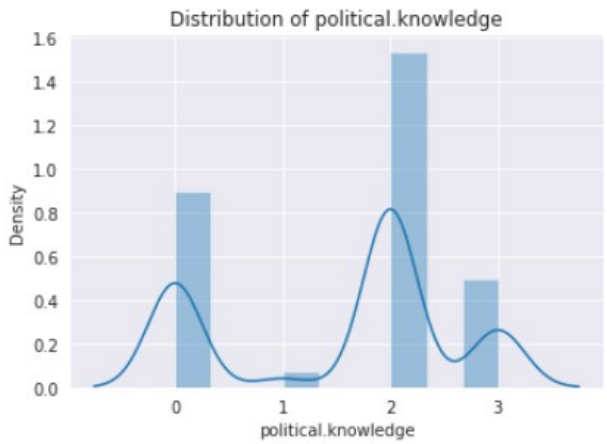


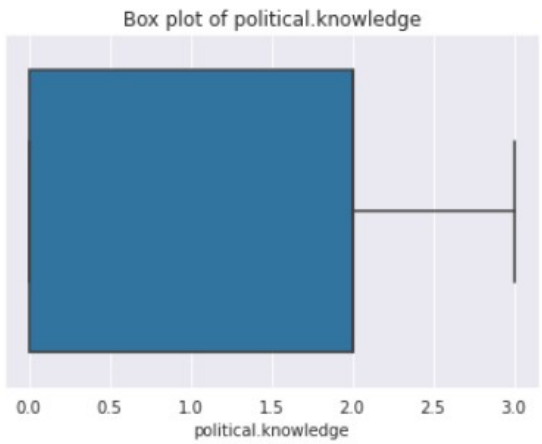
Distribution of “Hague” is not normal and it is slightly right skewed. This variable is not normally distributed and skewed.



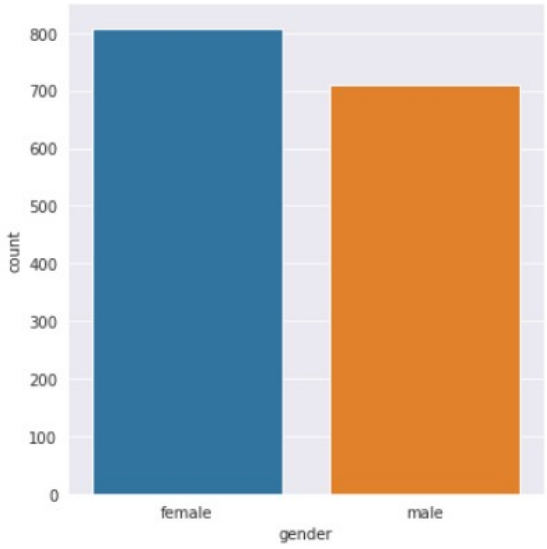


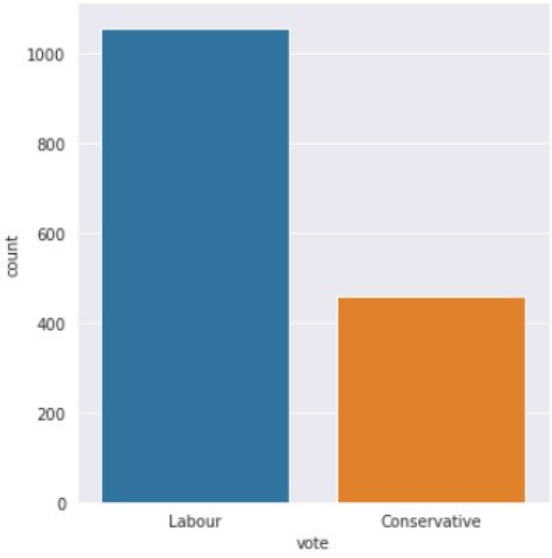
Distribution of “Europe” is somewhat normal and it is left skewed. Mode is higher than Mean. More than a half of participants have given rating of more than 6 in the scale, which shows that majority of the participants are much sceptical about European integration.





Distribution of “political.knowledge” is not normal and it is slightly right skewed.

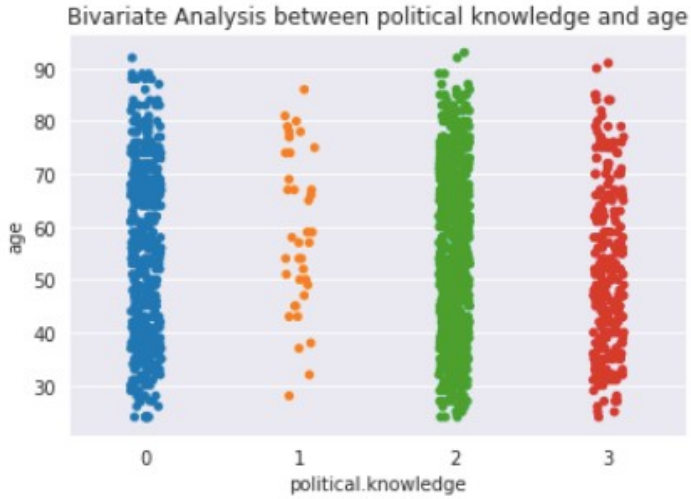




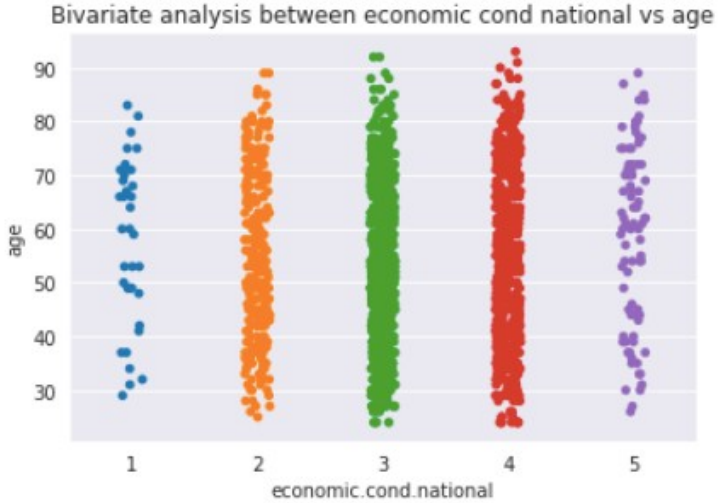
Most of the respodents in the data are women.

About 70 % of participants have chosen Labour party

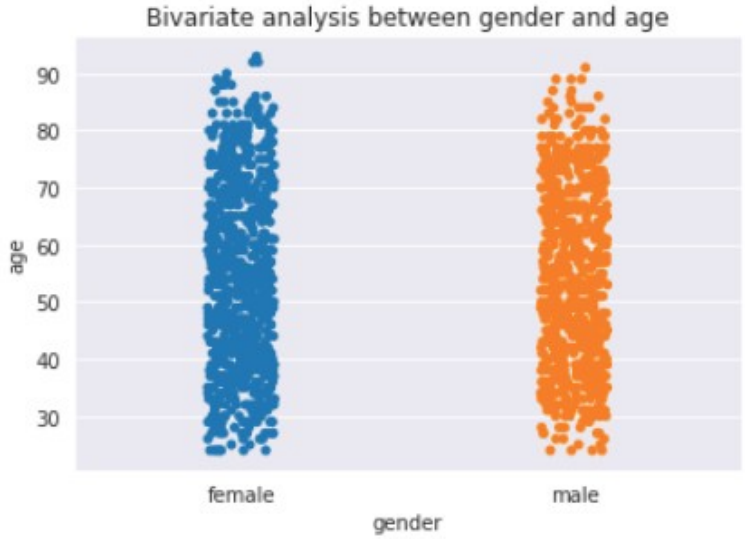
Bivatiate Analysis



It seems that respodents have a moderate understanding of the political situation.

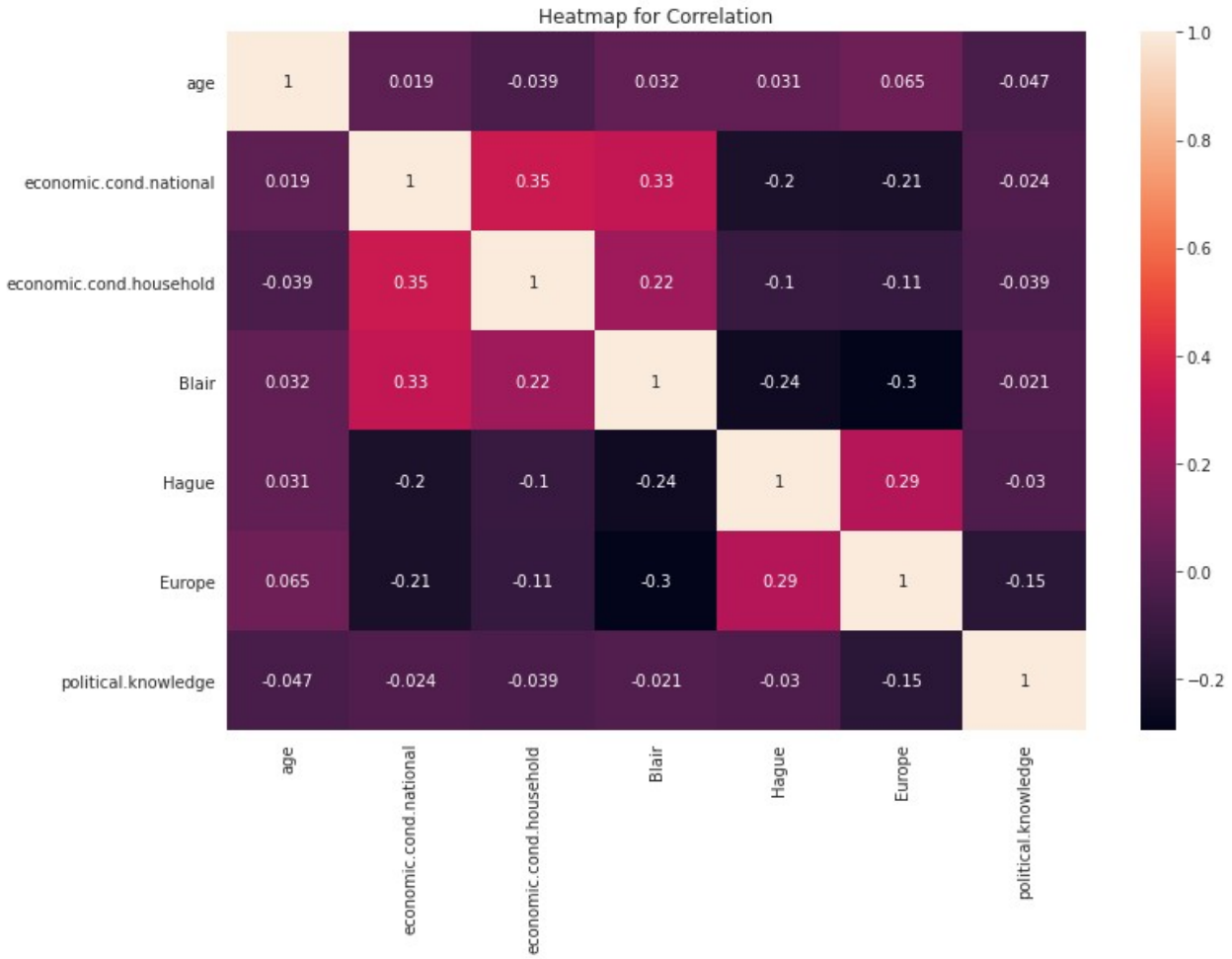


Majority of respondent rated the national economic condition as more than average



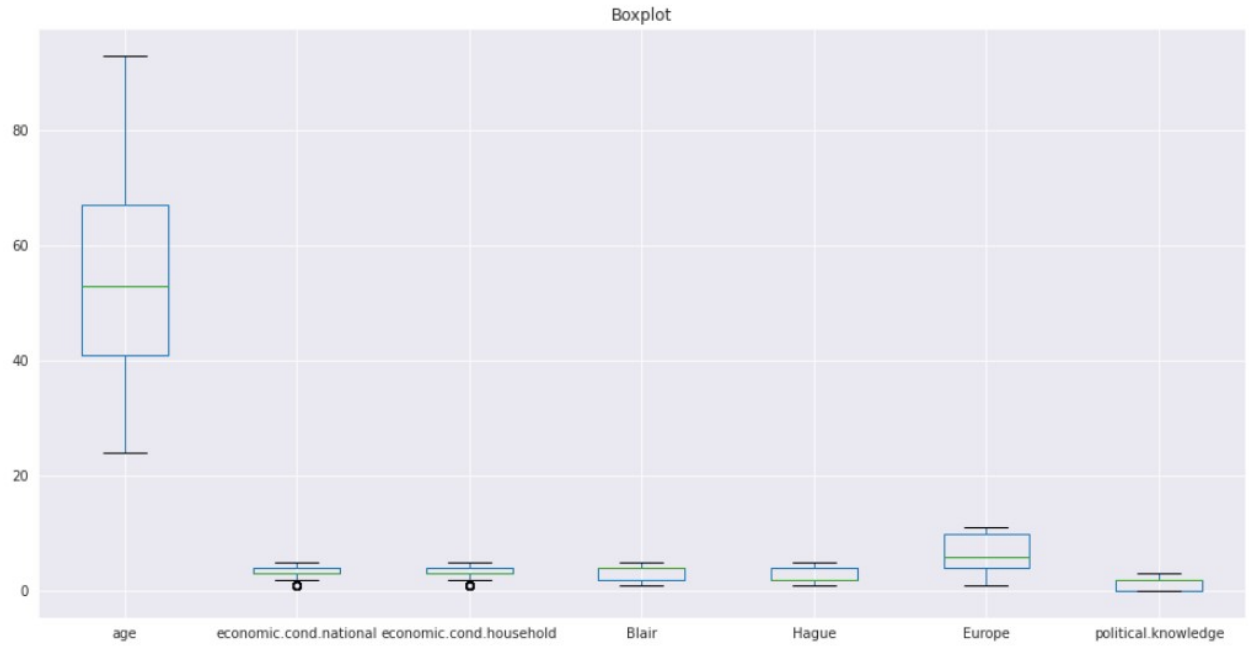
The population of both middle-aged male and female is more than the other ages.

Heatmap for Correlation



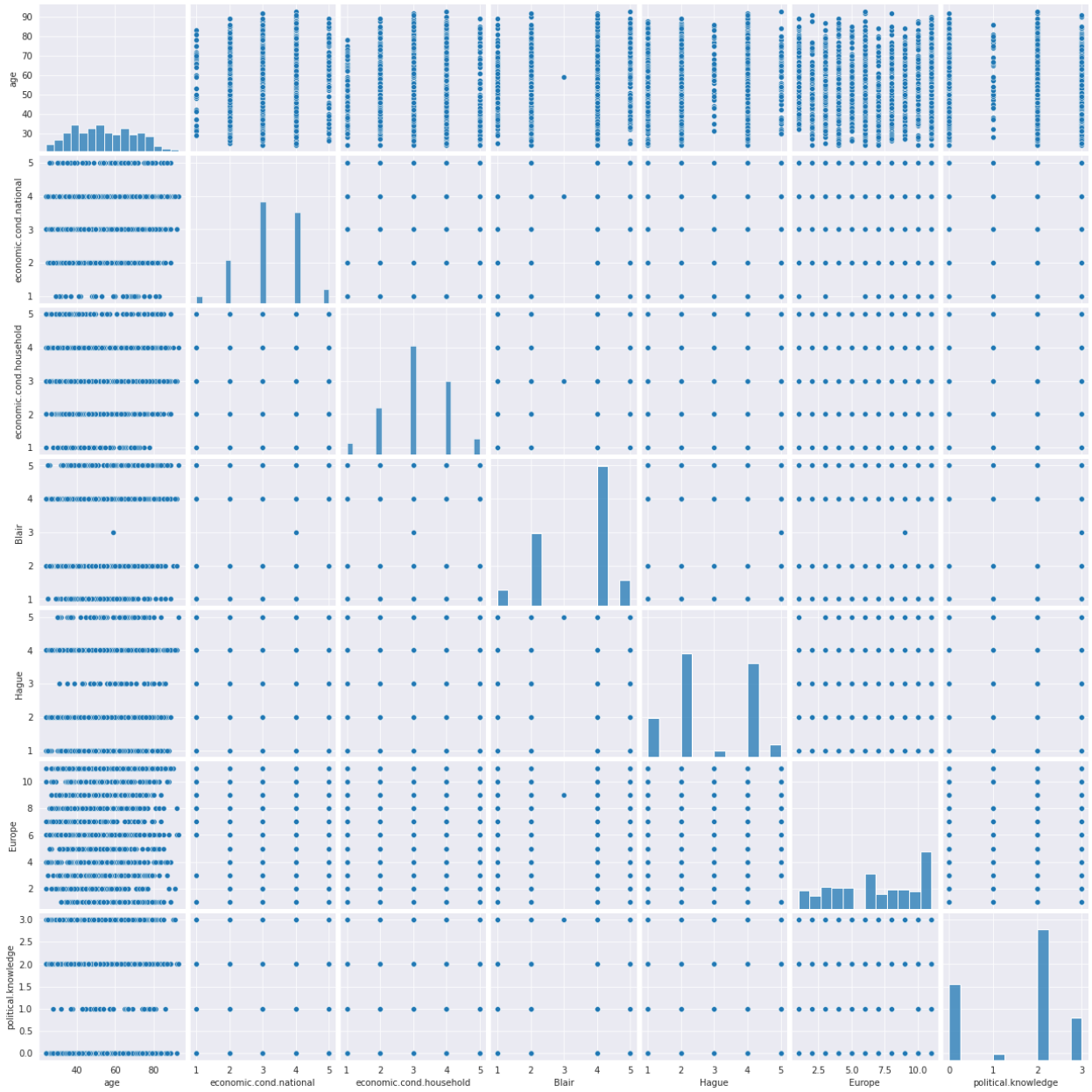
There are no highly correlated variables

Check ouliers



On performing the outliers analysis, we can see that there are a few outliets,

Pairplot for checking the correlation:



Inferences:

- None of variables are highly correlated with each other
- Ratings of 0, 2 & 3 on Political Knowledge of parties' positions on European integration has not been influenced by different age groups.
- About 70 % of participants have chosen Labour party

B Data Preparation

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

Linear regression model does not take categorical values so that we have encoded categorical values to integer for better result.

Vote and Gender variables are encoded using replace method. Vote, are label encoded :Labour : 0

Conservative : 1. Gender: Female:0, Male:1

We split the data into train and test set in a 70:30 ratio to perform further analysis and building our machine learning models.

	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
0	0	43	3	3	4	1	2	2	0
1	0	36	4	4	4	4	5	2	1
2	0	35	4	4	5	2	3	2	1
3	0	24	4	2	2	1	4	0	0
4	0	41	2	2	1	1	6	2	1

Scaling is required as continuous variables are of different scales and need to normalize the data using Standard Scaler.

C Modeling

1.4 Apply Logistic Regression

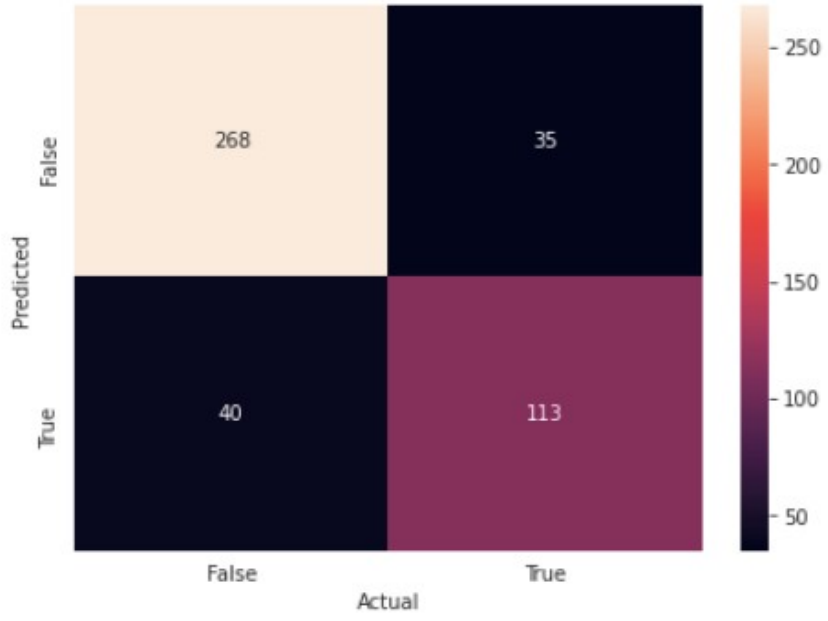
```
Score for training set:  0.8312912346842601
Score for test set:  0.8355263157894737


Coefficients of Logistic Regression

0.23   age
-0.54   economic.cond.national
-0.06   economic.cond.household
-0.7   Blair
1.01   Hague
0.68   Europe
0.34   political.knowledge
-0.1   gender

Intercept:  -1.450169249163272
```

	precision	recall	f1-score	support
0	0.87	0.88	0.88	303
1	0.76	0.74	0.75	153
accuracy			0.84	456
macro avg	0.82	0.81	0.81	456
weighted avg	0.83	0.84	0.83	456



- The model score seems to be pretty good in both training and testing instance and this looks like a fairly good model.
- The Test and Train performance is within the accepted limited of +/- 10% which makes it an acceptable model as well.
- Accuracy & prediction is good. The model is able to detect properly to predict which party a voter will vote to win the election.

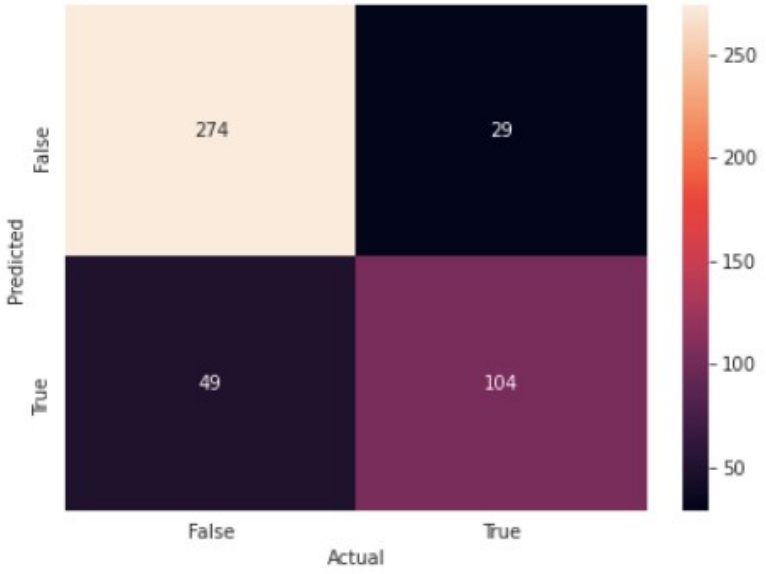
1.5 Apply KNN Model and Naïve Bayes Model. Interpret the results.

KNN

```
Score for training set:  1.0
Score for test set:  0.8289473684210527
```

- Training shows that the model is excellent with good precision and recall values.
- This KNN model have good accuracy and recall values.
- The Test and Train performance isn’t within the accepted limited of +/- 10% which makes model overtrained.

	precision	recall	f1-score	support
0	0.85	0.90	0.88	303
1	0.78	0.68	0.73	153
accuracy			0.83	456
macro avg	0.82	0.79	0.80	456
weighted avg	0.83	0.83	0.83	456



Naïve Bayes Model

```
Score for training set:  0.8350612629594723
Score for test set:  0.8223684210526315
```

	precision	recall	f1-score	support
0	0.87	0.87	0.87	303
1	0.74	0.73	0.73	153
accuracy			0.82	456
macro avg	0.80	0.80	0.80	456
weighted avg	0.82	0.82	0.82	456

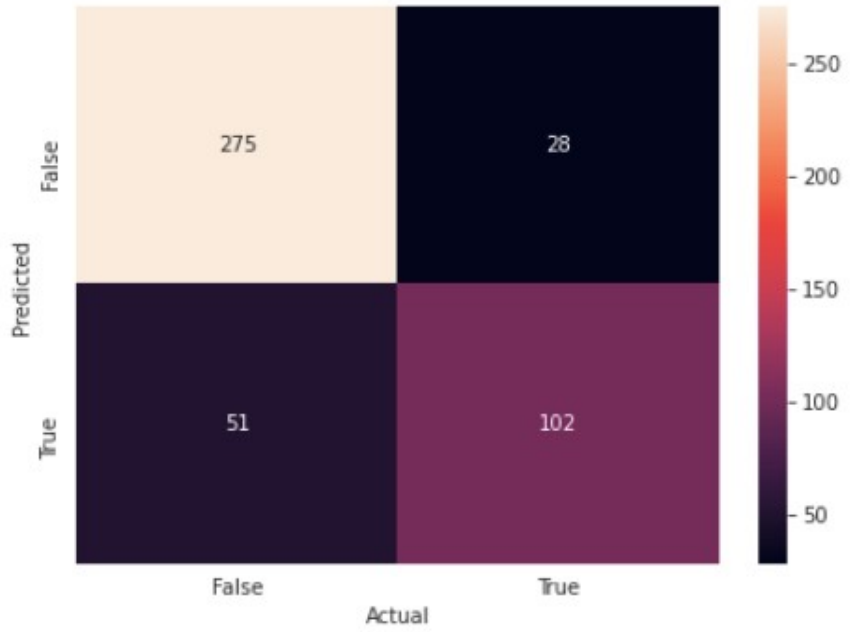
- Training and Testing results shows that the model neither overfitting nor underfitting.

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

Random Forest

```
Score for training set:  1.0
Score for test set:  0.8267543859649122
```

	precision	recall	f1-score	support
0	0.84	0.91	0.87	303
1	0.78	0.67	0.72	153
accuracy			0.83	456
macro avg	0.81	0.79	0.80	456
weighted avg	0.82	0.83	0.82	456



Bagging (Random Forest)

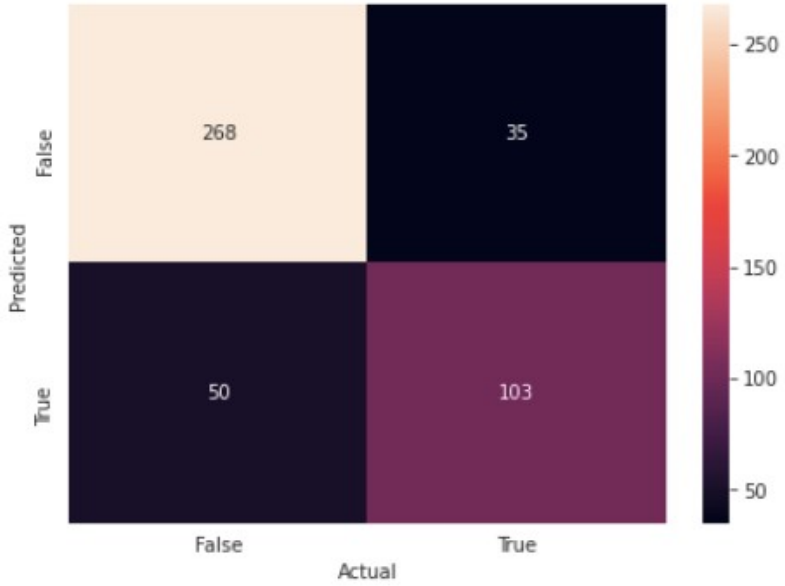
```
Score for training set:  0.9679547596606974
Score for test set:  0.8289473684210527
```

	precision	recall	f1-score	support
0	0.85	0.90	0.88	303
1	0.78	0.68	0.73	153
accuracy			0.83	456
macro avg	0.82	0.79	0.80	456
weighted avg	0.83	0.83	0.83	456

AdaBoost Model

```
Score for training set:  0.8501413760603205
Score for test set:  0.8135964912280702
```

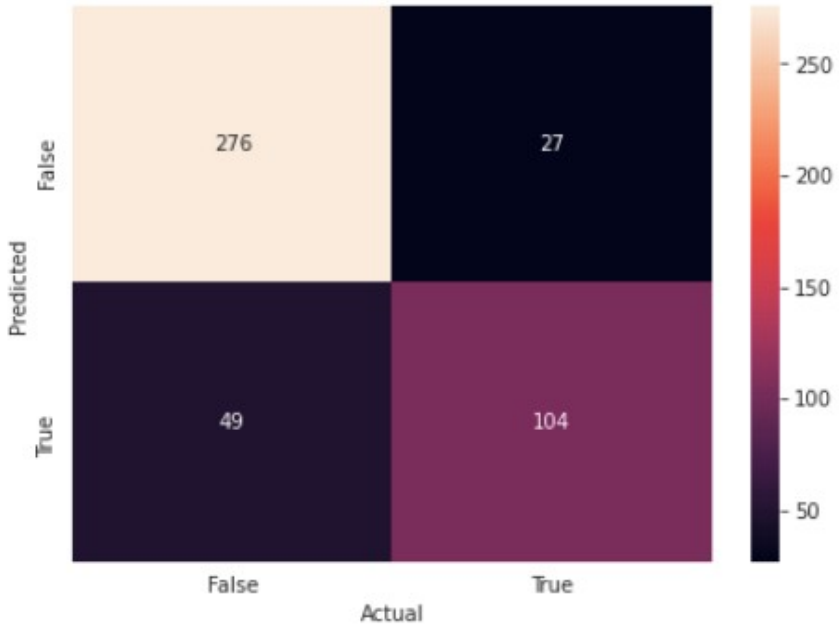
	precision	recall	f1-score	support
0	0.84	0.88	0.86	303
1	0.75	0.67	0.71	153
accuracy			0.81	456
macro avg	0.79	0.78	0.79	456
weighted avg	0.81	0.81	0.81	456



Gradient Boosting

```
Score for training set:  0.8925541941564562
Score for test set:  0.8333333333333334
```

	precision	recall	f1-score	support
0	0.85	0.91	0.88	303
1	0.79	0.68	0.73	153
accuracy			0.83	456
macro avg	0.82	0.80	0.81	456
weighted avg	0.83	0.83	0.83	456



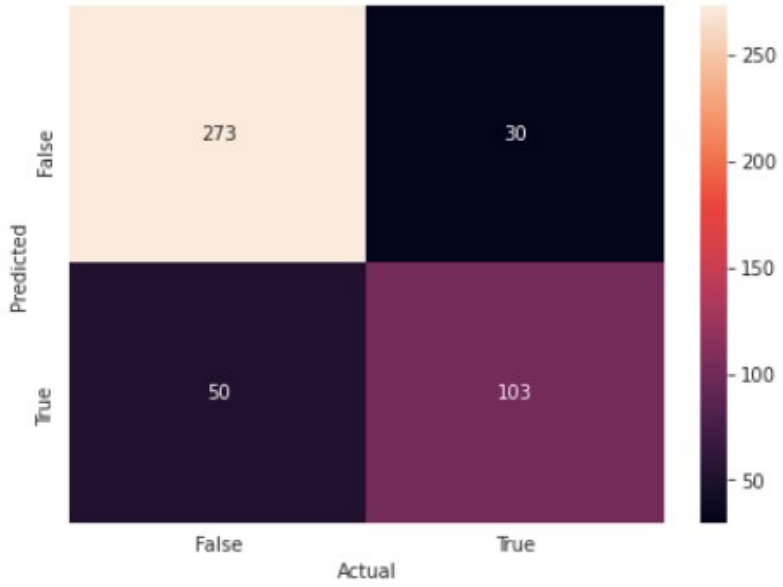
Model Tuning (Grid Search Random Forest)

```
GridSearchCV(cv=5, estimator=RandomForestClassifier(),
             param_grid={'max_depth': [7, 8, 9, 10, 11, 12],
                         'max_features': [3, 5, 8, 10],
                         'min_samples_leaf': [5, 10, 15],
                         'min_samples_split': [50, 100],
                         'n_estimators': [100, 125, 150]})
```

Best params : `{'max_depth': 12, 'max_features': 3, 'min_samples_leaf': 10, 'min_samples_split': 50, 'n_estimators': 125}`

```
Score for training set:  0.8567389255419415
Score for test set:  0.8245614035087719
```

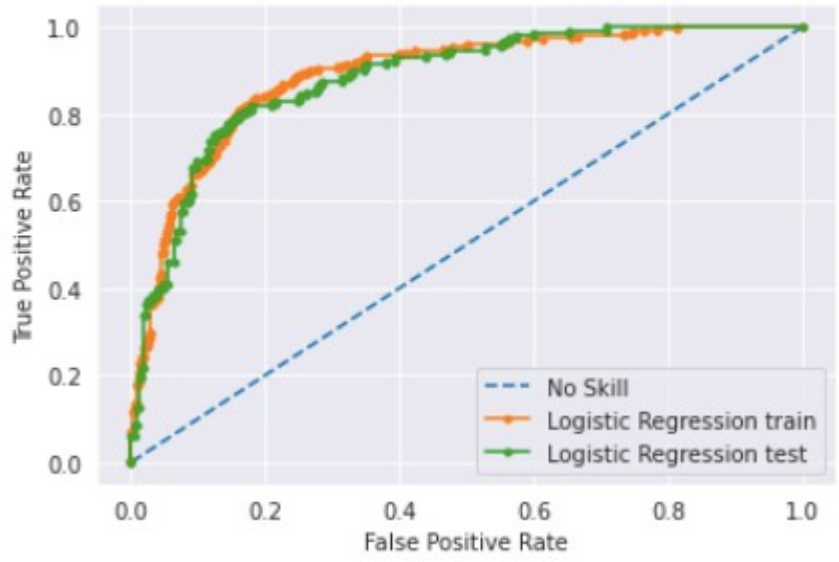
	precision	recall	f1-score	support
0	0.85	0.90	0.87	303
1	0.77	0.67	0.72	153
accuracy			0.82	456
macro avg	0.81	0.79	0.80	456
weighted avg	0.82	0.82	0.82	456



1.7 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix

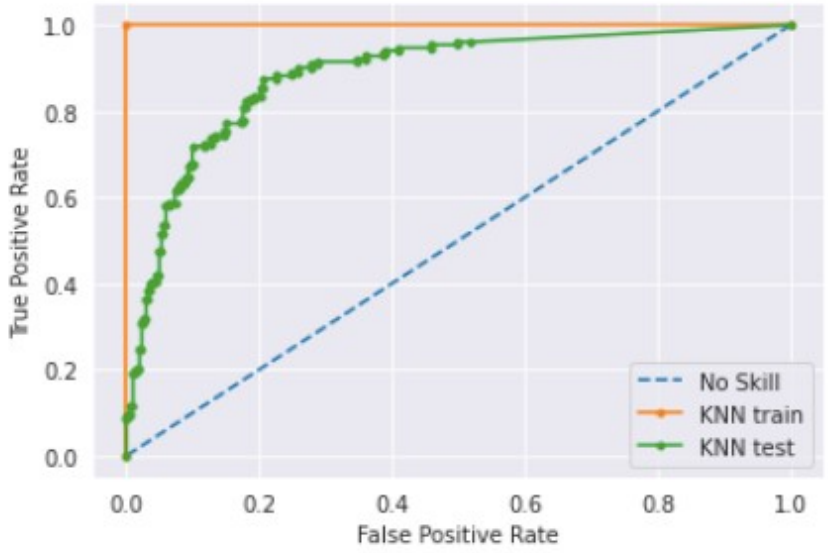
Logistic Regression

```
No Skill: ROC AUC=0.500
Logistic Regression train: ROC AUC=0.890
Logistic Regression test: ROC AUC=0.883
```



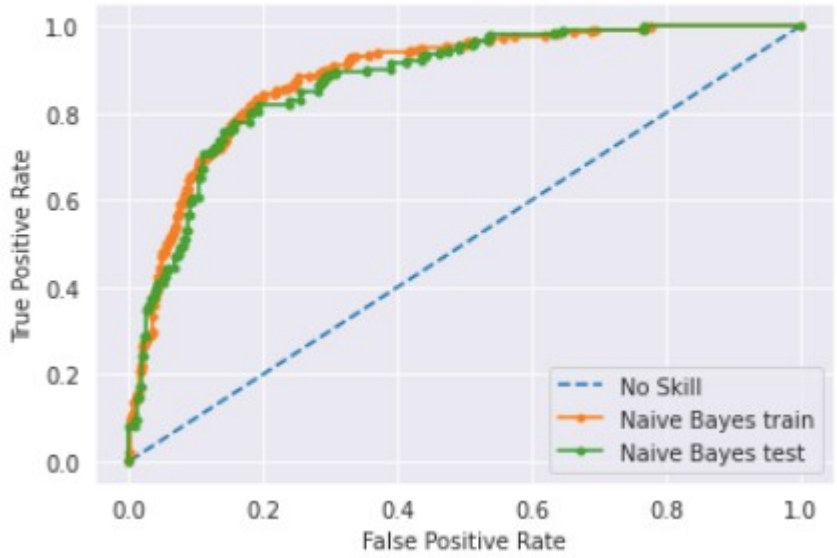
KNN

```
No Skill: ROC AUC=0.500
KNN train: ROC AUC=1.000
KNN test: ROC AUC=0.887
```



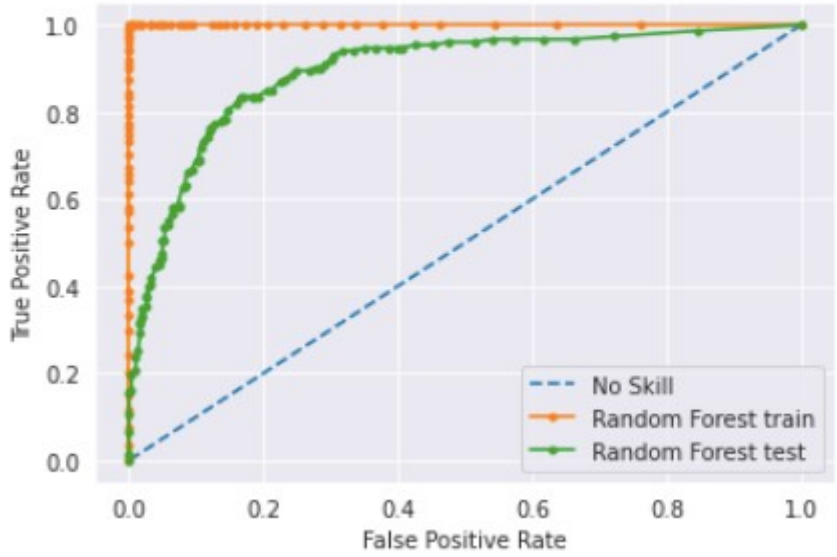
Naïve Bayes Model

```
No Skill: ROC AUC=0.500
Naive Bayes train: ROC AUC=0.888
Naive Bayes test: ROC AUC=0.876
```



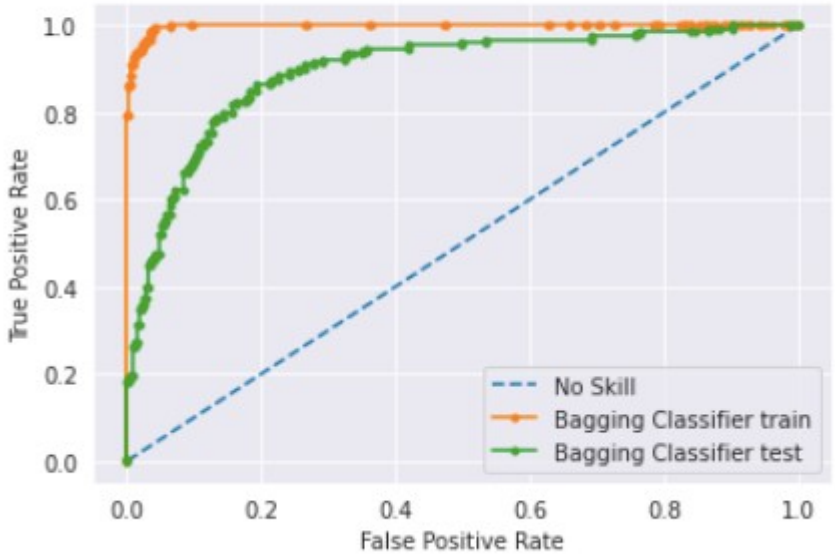
Random Forest

```
No Skill: ROC AUC=0.500
Random Forest train: ROC AUC=1.000
Random Forest test: ROC AUC=0.895
```



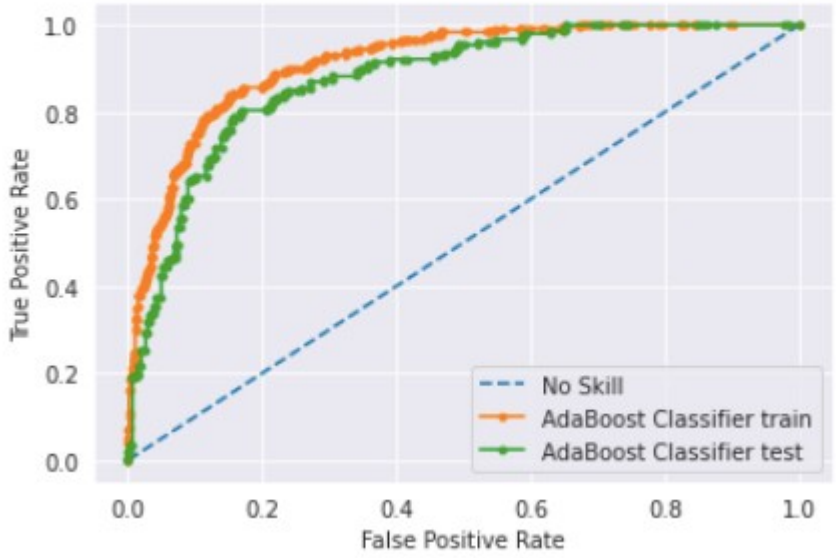
Bagging (Random Forest)

```
No Skill: ROC AUC=0.500
Bagging Classifier train: ROC AUC=0.997
Bagging Classifier test: ROC AUC=0.897
```



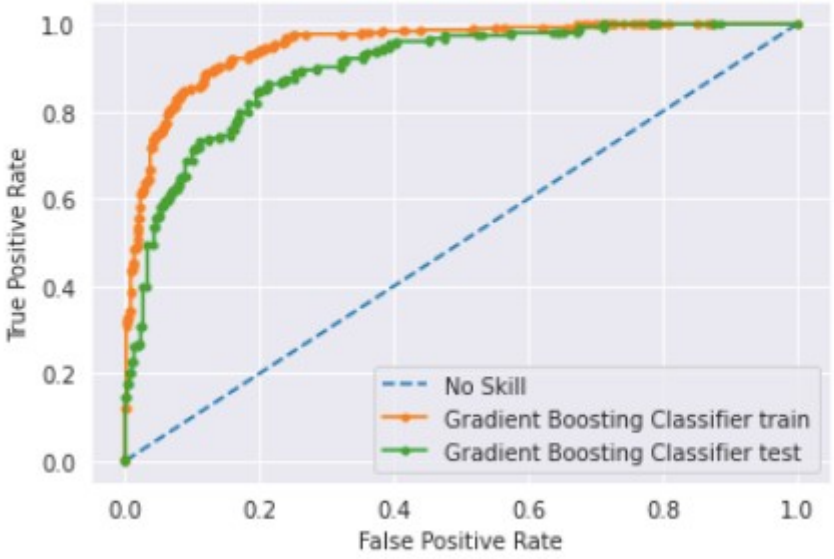
AdaBoost Model

```
No Skill: ROC AUC=0.500
AdaBoost Classifier train: ROC AUC=0.915
AdaBoost Classifier test: ROC AUC=0.877
```



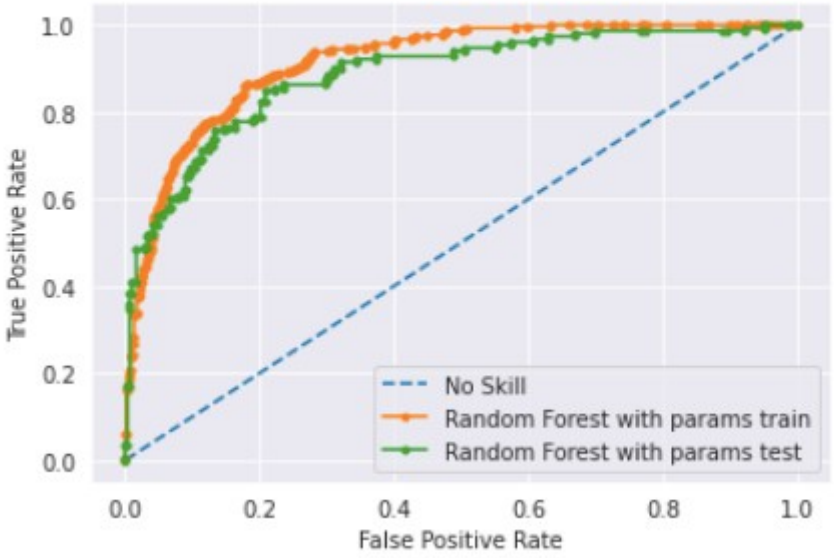
Gradient Boosting

```
No Skill: ROC AUC=0.500
Gradient Boosting Classifier train: ROC AUC=0.951
Gradient Boosting Classifier test: ROC AUC=0.899
```



Model Tuning (Grid Search Random Forest)

```
No Skill: ROC AUC=0.500
Random Forest with params train: ROC AUC=0.916
Random Forest with params test: ROC AUC=0.887
```



D Inferences

1.8 Based on these predictions, what are the insights?

	Model	Train Score	Test Score
0	Logistic Regression	0.831291	0.835526
6	Gradient Boosting	0.892554	0.833333
1	KNN	1.000000	0.828947
4	Bagging	0.967955	0.828947
3	Random Forest	1.000000	0.826754
2	Naïve Bayes	0.835061	0.822368
5	AdaBoost	0.850141	0.813596
7	Random Forest grid params	0.853911	0.811404

Inferences:

- Logistic Regression can be used to make predictions on the exit poll as to whether a particular voter would vote the Conservative or the Labour party based on the information provided.
- Majority of the population is between the ages 35-60 with considerable political knowledge and would vote mostly for Labour party.
- Bagging model is not very suitable and has overfitting issues
- Remaining models are relatively suitable for making predictions. However, the predictions are more reliable for Labour party
- Except Bagging, KNNand Random Forest other models have performed well enough in both training and test data set
- Accuracy is almost similar – in between 81 – 83%
- Accuracy & prediction is good. Hence both the model's performance is approx. equal.
- The model is able to detect properly to predict which party a voter will vote to win the election.
- The accuracy of KNN model is very good compared to other models on train set but the accuracy decreases massively on testing data
- These Models should perform even better
- Even after applying Model Tuning, bagging & boasting did not improve performance of models.
- None of the model is performed properly.

SMOTE

About 70 % of participants have chosen Labour party. To solve this problem of not equaly distributed data, we will use SMOTE to generate more additional data for Conservative class (0)

	Model	Train Score	Test Score
3	Random Forest	0.999324	0.872441
4	Bagging	0.973631	0.866142
1	KNN	0.999324	0.845669
7	Random Forest grid params	0.858688	0.845669
6	Gradient Boosting	0.894523	0.842520
5	AdaBoost	0.856660	0.840945
0	Logistic Regression	0.835700	0.822047
2	Naïve Bayes	0.822177	0.822047

- Accuracy is have increased, and now it is between 82 – 87%
- According to the results: top 3 models with the highest score are slightly overtrained
- Bagging and Random Forest are showing better results for 5%

Problem 2

Text Analysis

In this particular 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:

1. President Franklin D. Roosevelt in 1941
2. President John F. Kennedy in 1961
3. President Richard Nixon in 1973

- Find the number of characters, words and sentences for the mentioned documents.
- Remove all the stopwords from all the three speeches.
- Which word occurs the most number of times in his inaugural address for each
- president? Mention the top three words. (after removing the stopwords)
- Plot the word c

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

```
------------Roosevelt------------
Number of characters : 7571
Number of words : 1360
Number of sentences : 67

------------Kennedy------------
Number of characters : 7618
Number of words : 1390
Number of sentences : 50

------------Nixon------------
Number of characters : 9991
Number of words : 1819
Number of sentences : 64
```

2.2 Remove all the stopwords from all three speeches

We need these to remove all the English predefined words from each text file separately and with the help of tokenize we would separate each word and remove all the words from the text file.

After removing stop words and punctuation marks, quatity of words has decreased by a factor of two.

```
Roosevelt
before removing stopwords : 1360
after  : 632

Kenedy
before removing stopwords : 1390
after  : 696

Nixon
before removing stopwords : 1819
after  : 848
```

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

```
--------------TOP 3 WORDS--------------

Roosevelt
[('nation', 12), ('know', 10), ('spirit', 9)]

Kennedy
[('let', 16), ('us', 12), ('world', 8)]

Nixon
[('us', 26), ('let', 22), ('america', 21)]
```

Roosevelt's most common top three words, post removing stopwords are: **Nation**, **Know** and **Spirit**.

Kennedy's most common top three words, post removing stopwords are: **Let**, **Us** and **World**.

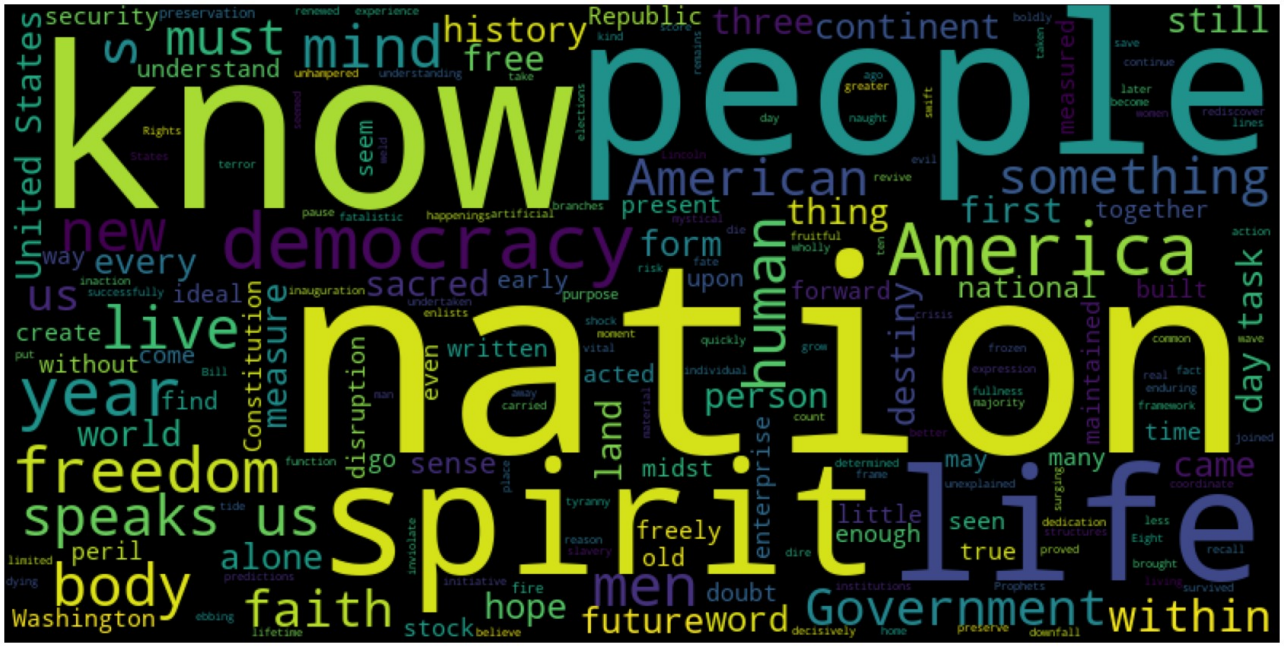
Nixon's most common top three words, post removing stopwords are: **Us, Let** and **America**.

2.4 Plot the word cloud of each of the speeches of the variable. (after removing the stopwords)

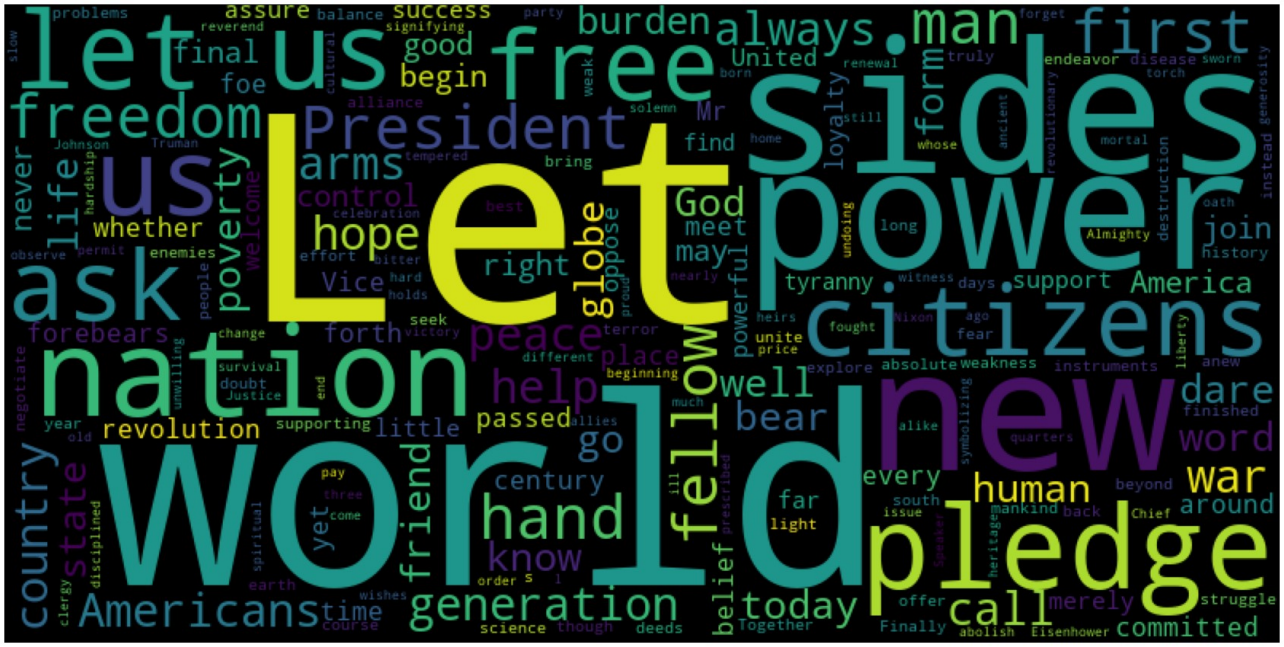
Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud. Word clouds are widely used for analysing data from social network websites.

Here we are creating the wordcloud for Roosevelt speech and we have imported the wordcloud by importing libraries.

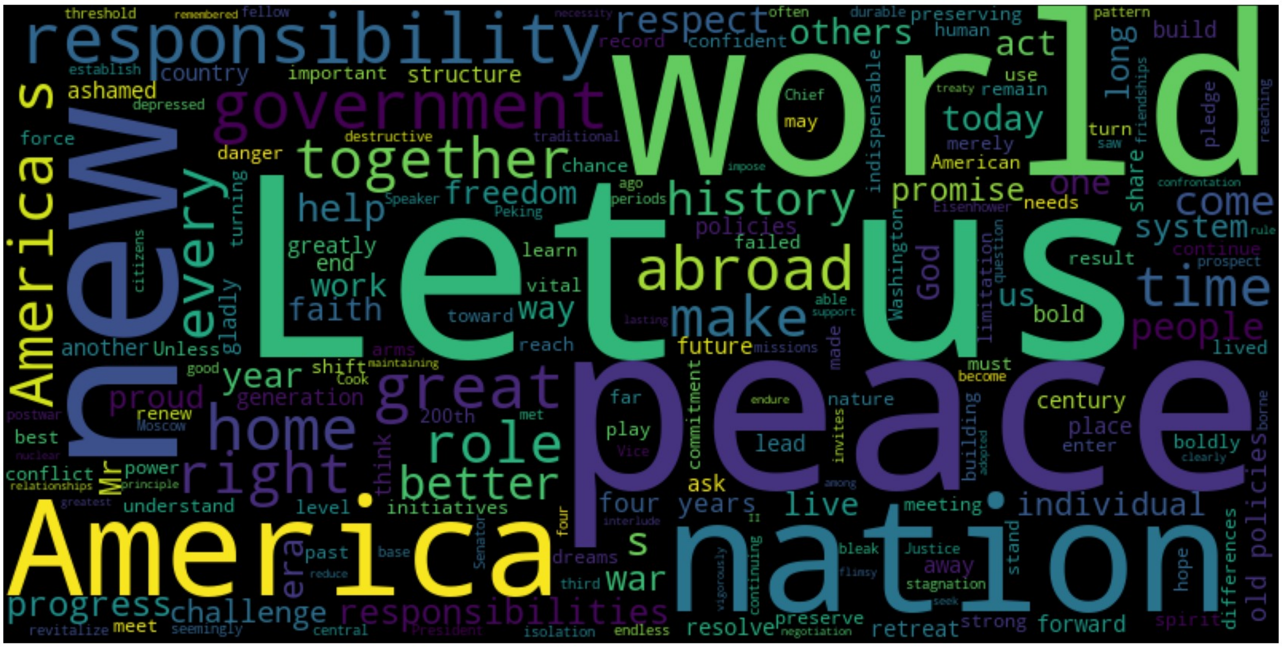
Roosevelt



Kennedy



Nixon



world
Let
us
peace
new
America
nation
responsibility
government
together
history
abroad
make
great
home
role
right
better
America s
every
time
people
responsibilities
progress
challenge
freedom
help
faith
work
promise
today
act
others
respect
individual
live
years
four
war
come
system
one
long
century
place
lead
play
structure
important
country
ashamed
force
danger
year
proud
generation
best
conflict
power
renew
understand
level
initiatives
past
era
meet
resolve
preserve
retreat
strong
forward
spirit
hope
old policies
differences
stand
stagnation
Justice
bleak
meeting
boldly
enter
nature
future
ask
far
commitment
building
must
bold
result
needs
American
turn
merely
remain
use
build
pledge
share
today
chance
way
vital
learn
failed
policies
end
greatly
gladly
turning
citizens
another
Unless
shift
200th
think
Senator
isolation
endless
central
seemingly
revitalize
threshold
remembered
fellow
preserving
confident
human
pattern
often
durable
record
establish
depressed
destructive
traditional
Speaker
Peking
Washington
indispensable
God
limitation
question
Eisenhower
continue
prospect
lived
reaching
friendships
confrontation
treaty
Chief
may
saw
lasting
support
able
missions
made
invites
endure
become
maintaining
Cook
good
arms
reach
toward
Mr
Moscow
postwar
nuclear
relationships
principle
greatest
interlude
base
reduce
third
dreams
seek
flimsy
negotiation
vigorously
continuing
among
adopted
clearly
borne
rule
President
Vice
met
II
ago
periods
impose