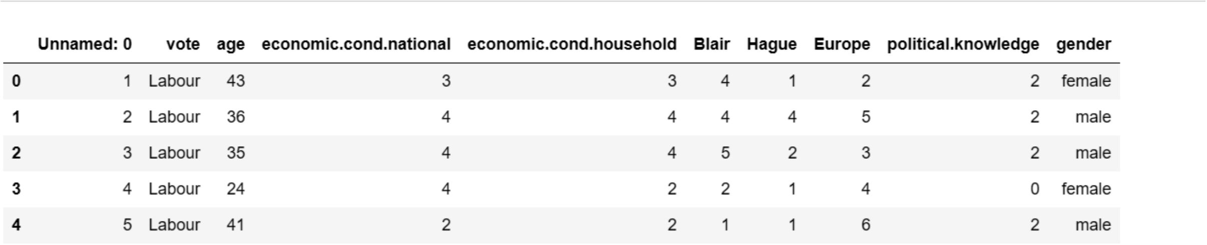
Problem 1: You are hired by one of the leading news channels CNBE who wants to analyse recent elections. This survey was conducted on 1525 voters with 9 variables. You have to build a model to predict which party a voter will vote for on the basis of the given information and create an exit poll that will help predict overall wins and seats covered by a particular party.

Dataset for Problem: Election_Data.xlsx

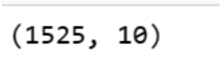
Data Ingestion:

1. Read the dataset. Do the descriptive statistics and do the null value condition check. Write an inference on it. Exploratory Data Analysis



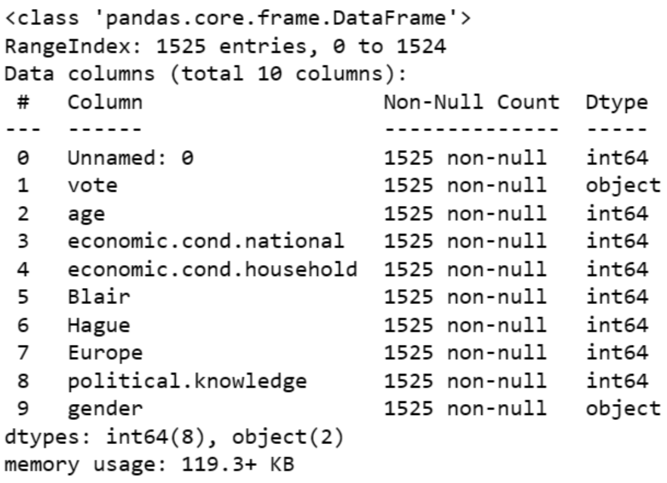
	Unnamed: 0	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
0	1	Labour	43	3	3	4	1	2	2	female
1	2	Labour	36	4	4	4	4	5	2	male
2	3	Labour	35	4	4	5	2	3	2	male
3	4	Labour	24	4	2	2	1	4	0	female
4	5	Labour	41	2	2	1	1	6	2	male

HEAD OF THE DATASET



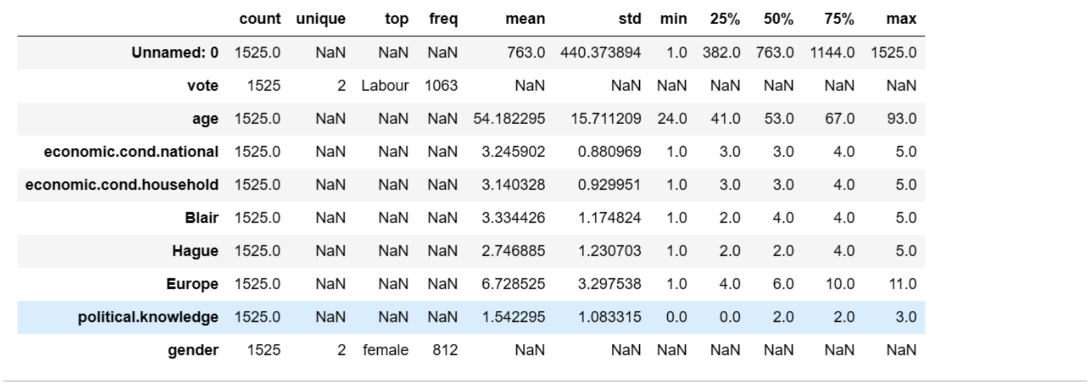
```
(1525, 10)
```

SHAPE OF THE DATASET



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1525 entries, 0 to 1524
Data columns (total 10 columns):
 #   Column                   Non-Null Count  Dtype 
---  ------                   --------------  ----- 
 0   Unnamed: 0               1525 non-null   int64 
 1   vote                     1525 non-null   object
 2   age                      1525 non-null   int64 
 3   economic.cond.national   1525 non-null   int64 
 4   economic.cond.household  1525 non-null   int64 
 5   Blair                    1525 non-null   int64 
 6   Hague                    1525 non-null   int64 
 7   Europe                   1525 non-null   int64 
 8   political.knowledge      1525 non-null   int64 
 9   gender                   1525 non-null   object
dtypes: int64(8), object(2)
memory usage: 119.3+ KB
```

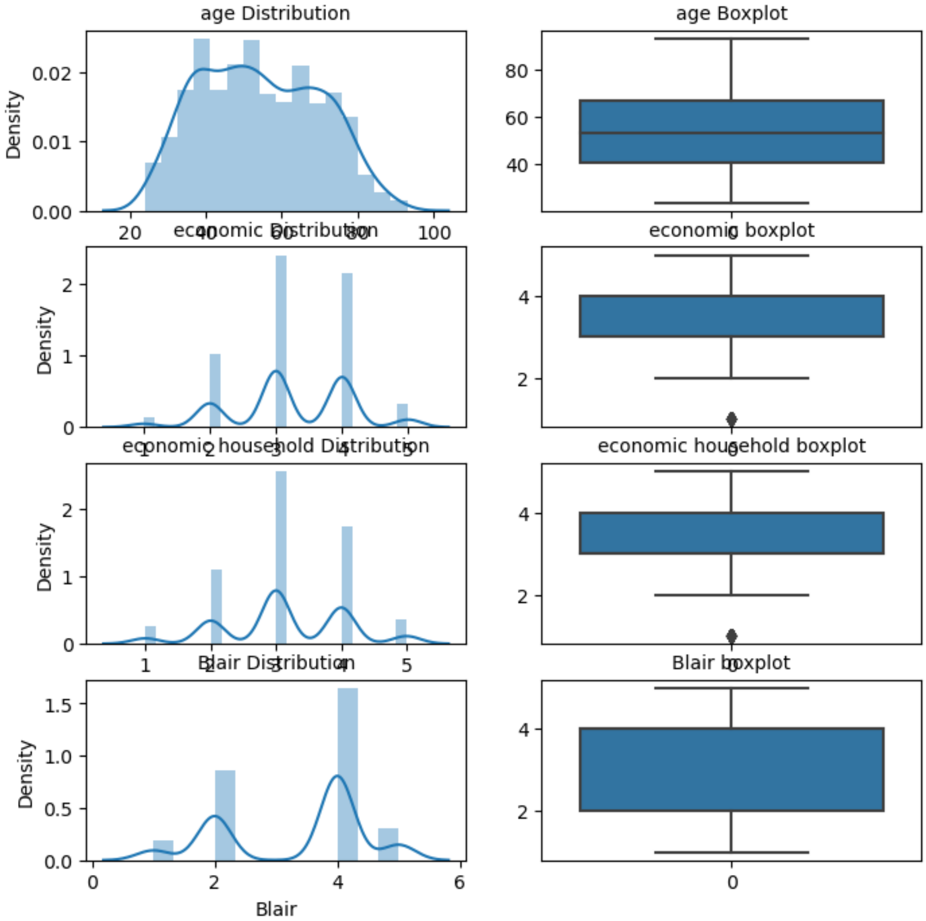
BASIC INFORMATION ABOUT THE DATASET

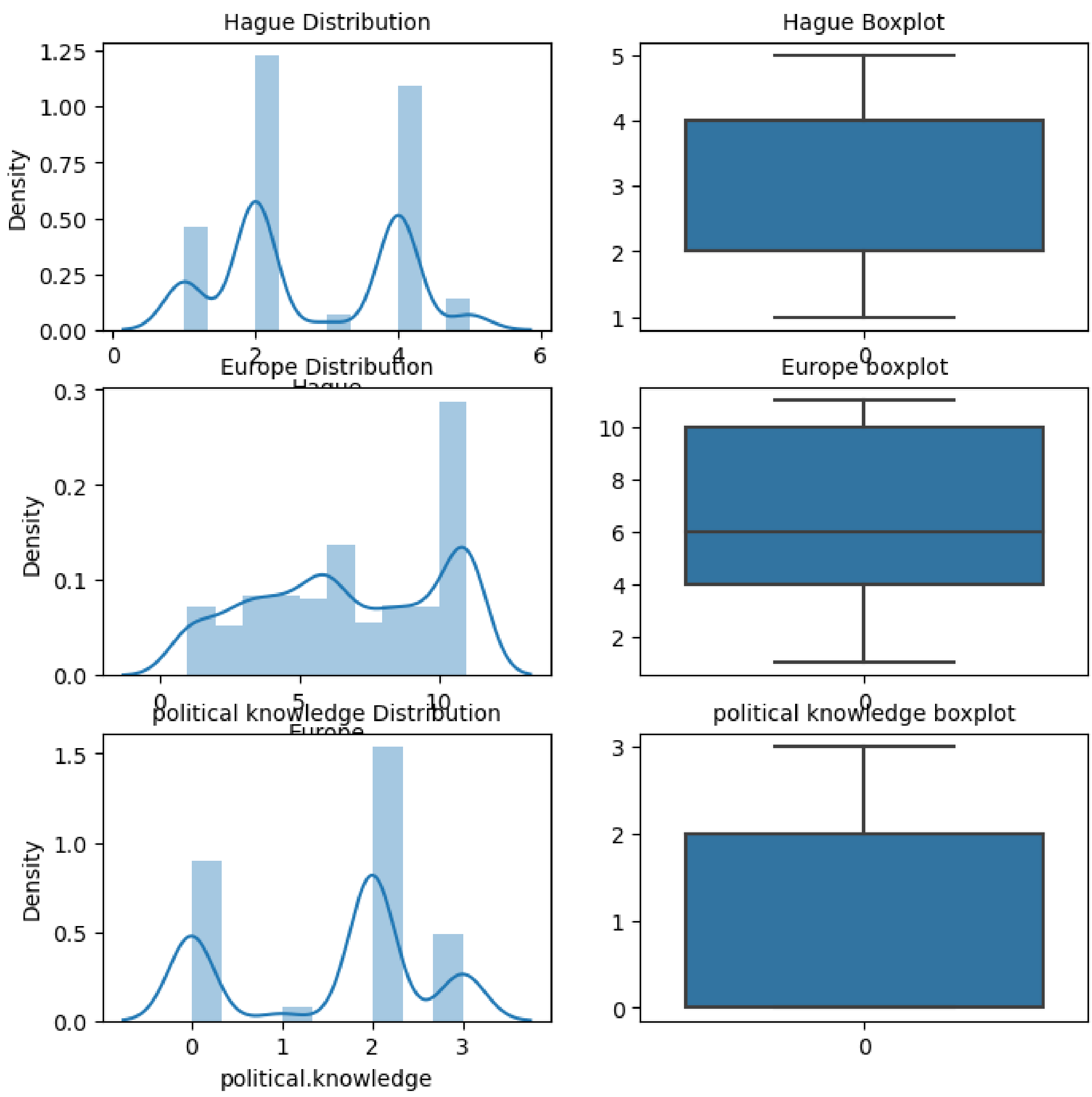


	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Unnamed: 0	1525.0	NaN	NaN	NaN	763.0	440.373894	1.0	382.0	763.0	1144.0	1525.0
vote	1525	2	Labour	1063	NaN	NaN	NaN	NaN	NaN	NaN	NaN
age	1525.0	NaN	NaN	NaN	54.182295	15.711209	24.0	41.0	53.0	67.0	93.0
economic.cond.national	1525.0	NaN	NaN	NaN	3.245902	0.880969	1.0	3.0	3.0	4.0	5.0
economic.cond.household	1525.0	NaN	NaN	NaN	3.140328	0.929951	1.0	3.0	3.0	4.0	5.0
Blair	1525.0	NaN	NaN	NaN	3.334426	1.174824	1.0	2.0	4.0	4.0	5.0
Hague	1525.0	NaN	NaN	NaN	2.746885	1.230703	1.0	2.0	2.0	4.0	5.0
Europe	1525.0	NaN	NaN	NaN	6.728525	3.297538	1.0	4.0	6.0	10.0	11.0
political.knowledge	1525.0	NaN	NaN	NaN	1.542295	1.083315	0.0	0.0	2.0	2.0	3.0
gender	1525	2	female	812	NaN	NaN	NaN	NaN	NaN	NaN	NaN

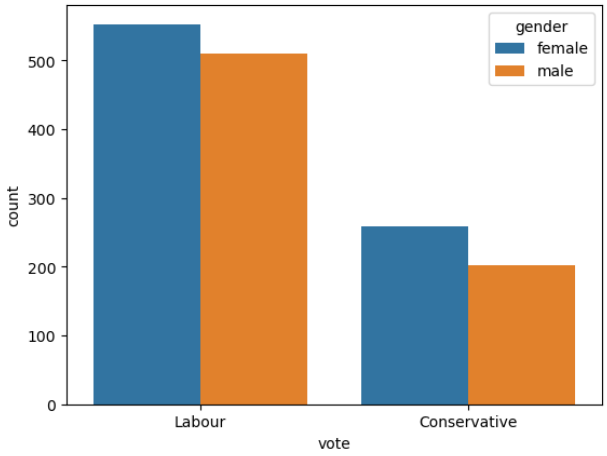
DESCRIPTIVE STATISTICS FOR DATASET

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

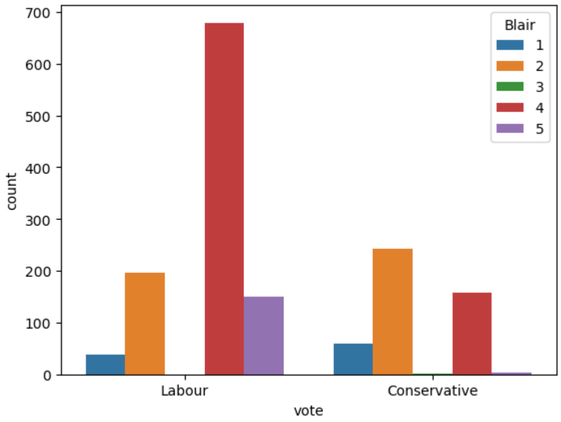


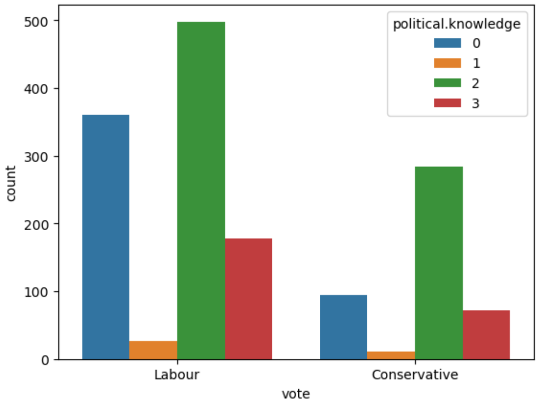


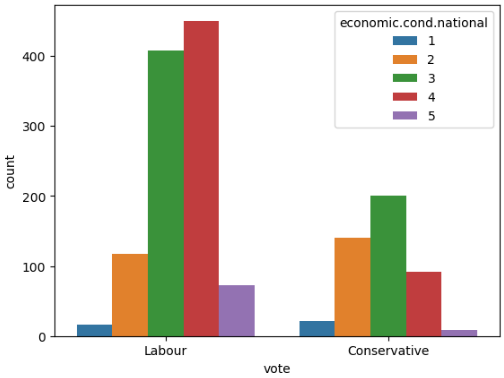
BIVARIENT ANALYSIS:

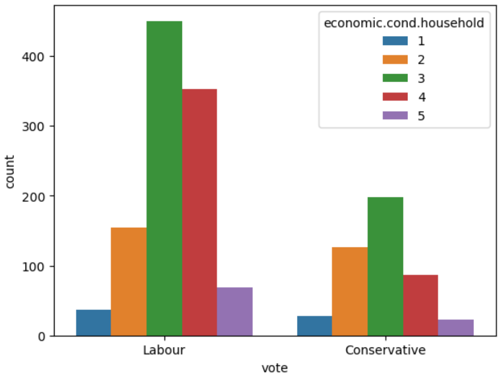


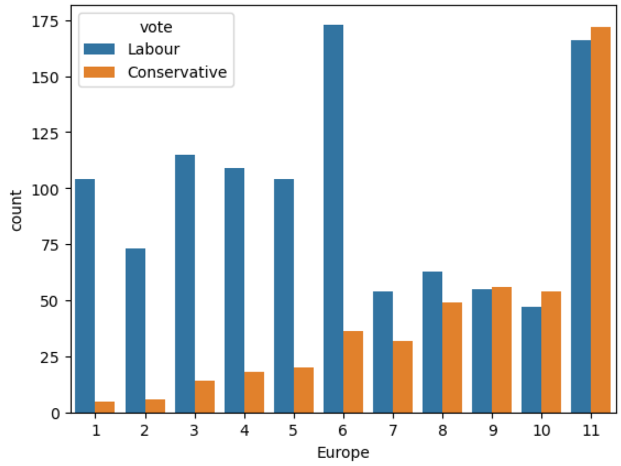
NUMBER OF VOTES BASED ON GENDER


Blair
1
2
3
4
5
count
0
100
200
300
400
500
600
700
Labour
Conservative
vote


political.knowledge
0
1
2
3
count
0
100
200
300
400
500
Labour
Conservative
vote


economic.cond.national
1
2
3
4
5
count
0
100
200
300
400
Labour
Conservative
vote

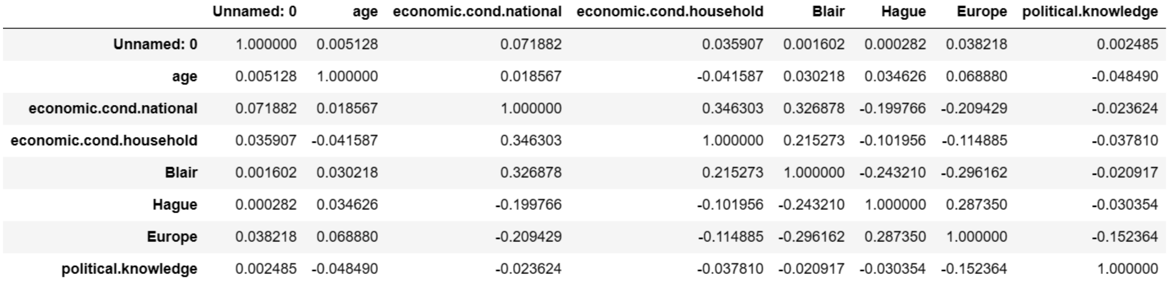

economic.cond.household
1
2
3
4
5
count
0
100
200
300
400
Labour
Conservative
vote



VOTES BASED ON *ECONOMIC CONDITIONS, POLITICAL KNOWLEDGE, AND BLAIR*

- Labour gets the highest voting from both female and male voters. Almost in all the categories
- Labour is getting the maximum votes.
- Conservative gets a little bit higher votes from Europe ‘11’.

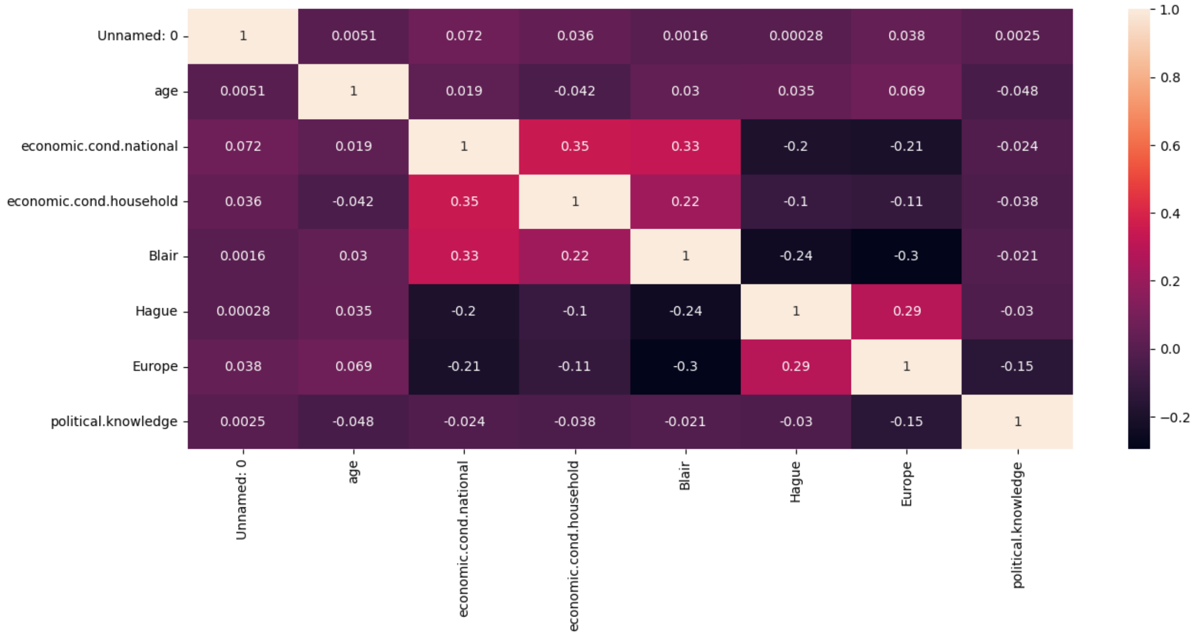
MULTIVARIATE ANALYSIS:



	Unnamed: 0	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge
Unnamed: 0	1.000000	0.005128	0.071882	0.035907	0.001602	0.000282	0.038218	0.002485
age	0.005128	1.000000	0.018567	-0.041587	0.030218	0.034626	0.068880	-0.048490
economic.cond.national	0.071882	0.018567	1.000000	0.346303	0.326878	-0.199766	-0.209429	-0.023624
economic.cond.household	0.035907	-0.041587	0.346303	1.000000	0.215273	-0.101956	-0.114885	-0.037810
Blair	0.001602	0.030218	0.326878	0.215273	1.000000	-0.243210	-0.296162	-0.020917
Hague	0.000282	0.034626	-0.199766	-0.101956	-0.243210	1.000000	0.287350	-0.030354
Europe	0.038218	0.068880	-0.209429	-0.114885	-0.296162	0.287350	1.000000	-0.152364
political.knowledge	0.002485	-0.048490	-0.023624	-0.037810	-0.020917	-0.030354	-0.152364	1.000000

CHECKING THE CORRELATIONS IN THE DATASET

- We can use the correlation matrix to view them more clearly. The correlation matrix is a table which shows the correlation coefficient between variables.
- Correlation values range from -1 to +1.
- For values closer to zero, it means that there is no linear trend between two variables.
- Values close to 1 mean that the correlation is positive.



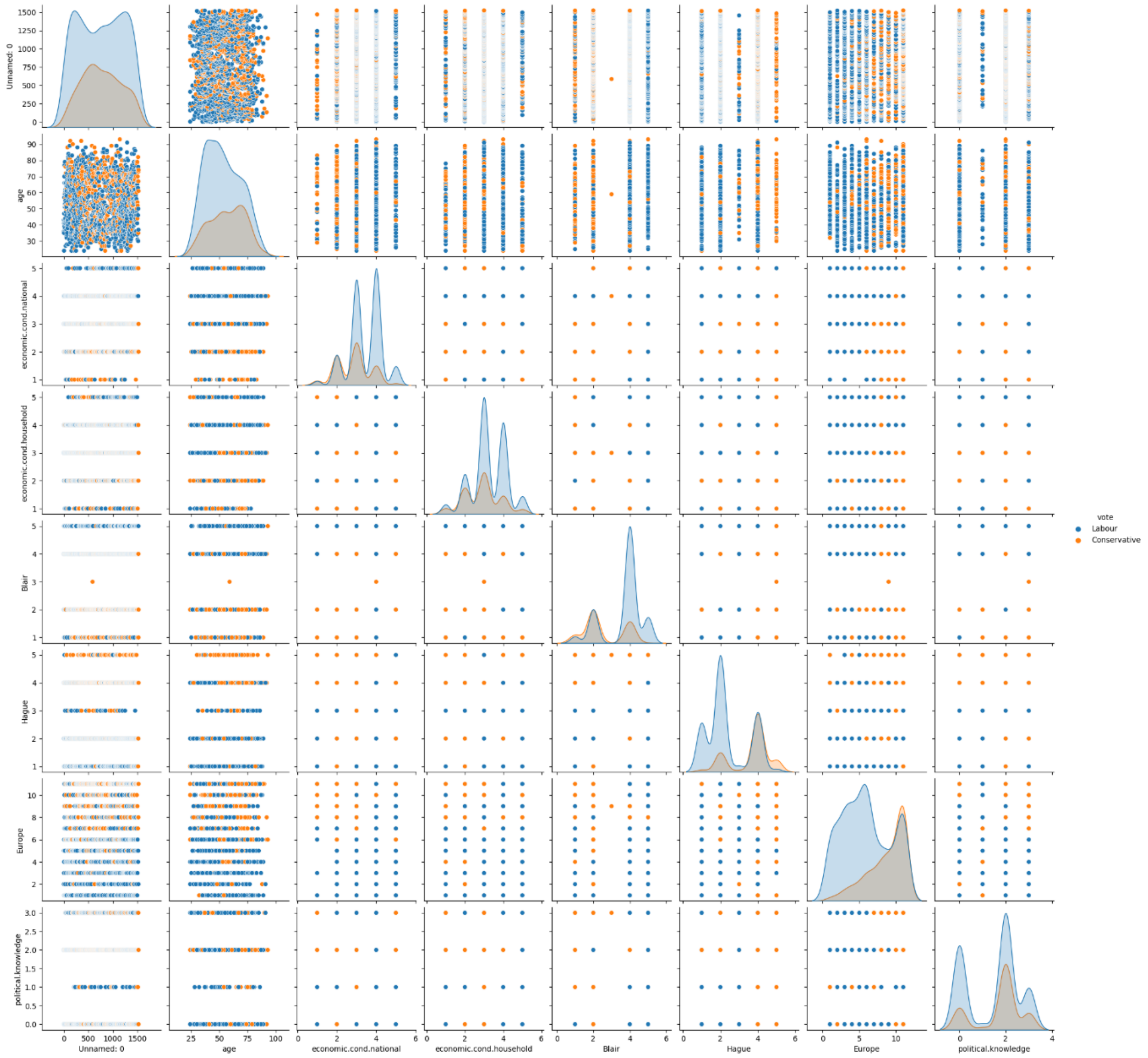
	Unnamed: 0	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge
Unnamed: 0	1	0.0051	0.072	0.036	0.0016	0.00028	0.038	0.0025
age	0.0051	1	0.019	-0.042	0.03	0.035	0.069	-0.048
economic.cond.national	0.072	0.019	1	0.35	0.33	-0.2	-0.21	-0.024
economic.cond.household	0.036	-0.042	0.35	1	0.22	-0.1	-0.11	-0.038
Blair	0.0016	0.03	0.33	0.22	1	-0.24	-0.3	-0.021
Hague	0.00028	0.035	-0.2	-0.1	-0.24	1	0.29	-0.03
Europe	0.038	0.069	-0.21	-0.11	-0.3	0.29	1	-0.15
political.knowledge	0.0025	-0.048	-0.024	-0.038	-0.021	-0.03	-0.15	1

We can see that, mostly there is no correlation in the dataset through this matrix.

Some variables are moderately positively correlated and some are slightly negatively correlated.

- ‘economic.cond.national’ with ‘economic.cond.household’ have moderate positive correlation
- ‘Blair' with 'economic.cond.national' and ‘economic.cond.household’ have moderate positive correlation.
- ‘Europe’ with ‘Hague’ have moderate positive correlation.
- 'Hague' with 'economic.cond.national' and 'Blair' have a moderate negative correlation.
- 'Europe' with 'economic.cond.national' and 'Blair' have a moderate negative correlation.

PAIRPLOT:

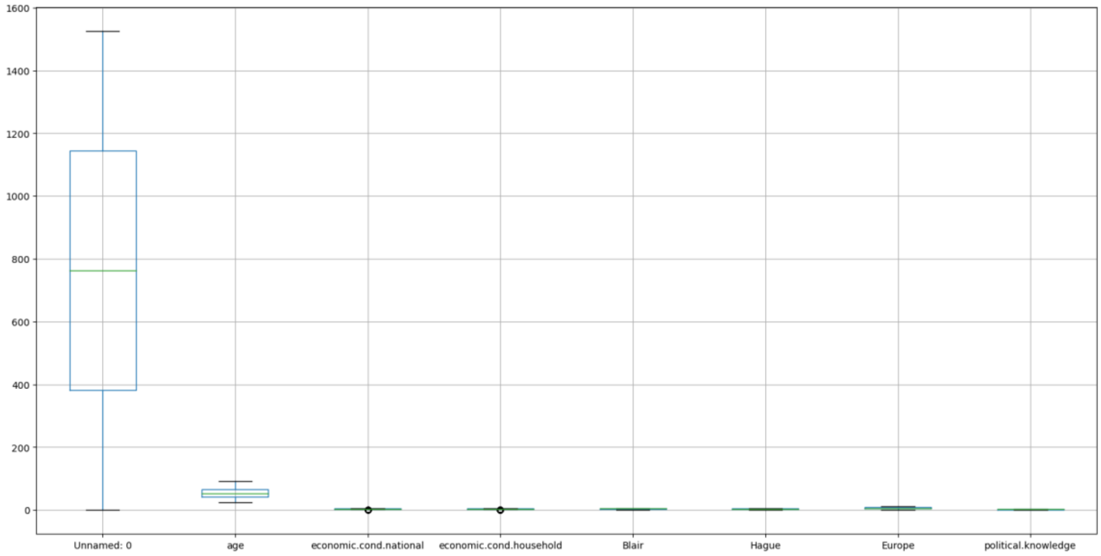


Multivariate Analysis for Election dataset based on votes

- A pair plot is a combination of histograms and scatter plots.
- From the histogram, we can see that, the 'Blair',' Europe' and 'political. knowledge variables are slightly left skewed.
- All other variables seem to be normally distributed.

- From the scatter plots, we can see that there is mostly no correlation between the variable

CHECKING OUTLIERS



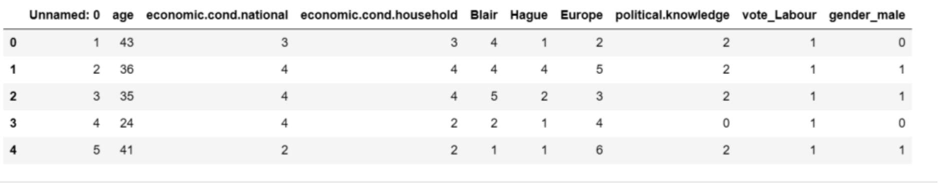
As we can see there are no outliers and the data is **clean** so there is no need for outliers

1.3 Data Preparation: 1. 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).

Encoding the dataset,

The variables 'vote' and 'gender' have string values. Converting them into numeric values for modelling,



	Unnamed: 0	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	vote_Labour	gender_male
0	1	43	3	3	4	1	2	2	1	0
1	2	36	4	4	4	4	5	2	1	1
2	3	35	4	4	5	2	3	2	1	1
3	4	24	4	2	2	1	4	0	1	0
4	5	41	2	2	1	1	6	2	1	1

Scaling :

We are not going to scale the data for Logistic regression, LDA and Naive Baye's models as it is not necessary.

But in the case of KNN, it is necessary to scale the data, as it is a distance-based algorithm (typically based on Euclidean distance).

Scaling the data gives similar weightage to all the variables.

Train-test-split:

Our model will use all the variables and 'vote_Labour' is the target variable.

The train-test split is a technique for evaluating the performance of a machine-learning algorithm.

The procedure involves taking a dataset and dividing it into two subsets. •

Train Dataset: Used to fit the machine learning model.

Test Dataset: Used to evaluate the fit machine learning model

.The data is divided into 2 subsets, training and testing sets.

Earlier, we extracted the target variable 'vote_Labour' in a separate vector for subsets.

The random state was chosen as 1.

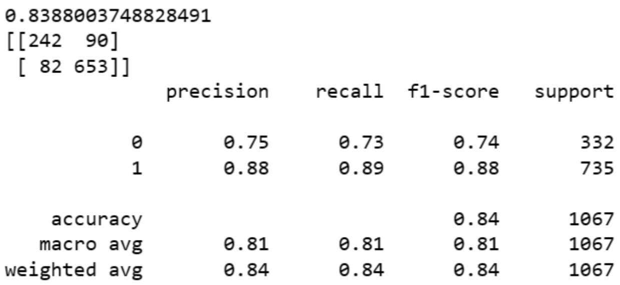
Training Set: 70 per cent of data.

Testing Set: 30 per cent of the data.

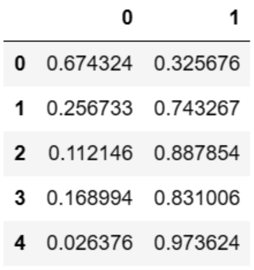
Importing GaussianNB from sklearn and applying NB model.

Fitting the training data

Train Accuracy, Confusion Matrix and matrix & classification report of GaussianNB

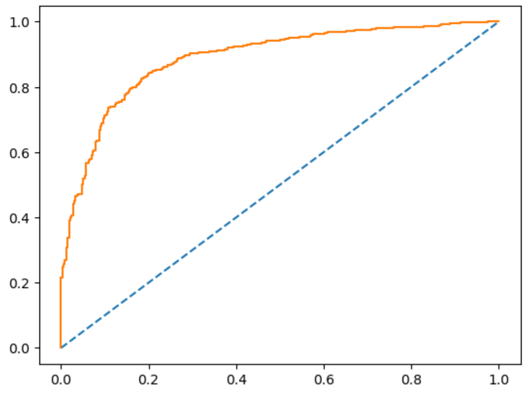


```
0.8388003748828491
[[242  90]
 [ 82 653]]
              precision    recall  f1-score   support

           0       0.75      0.73      0.74       332
           1       0.88      0.89      0.88       735

    accuracy                           0.84      1067
   macro avg       0.81      0.81      0.81      1067
weighted avg       0.84      0.84      0.84      1067
```



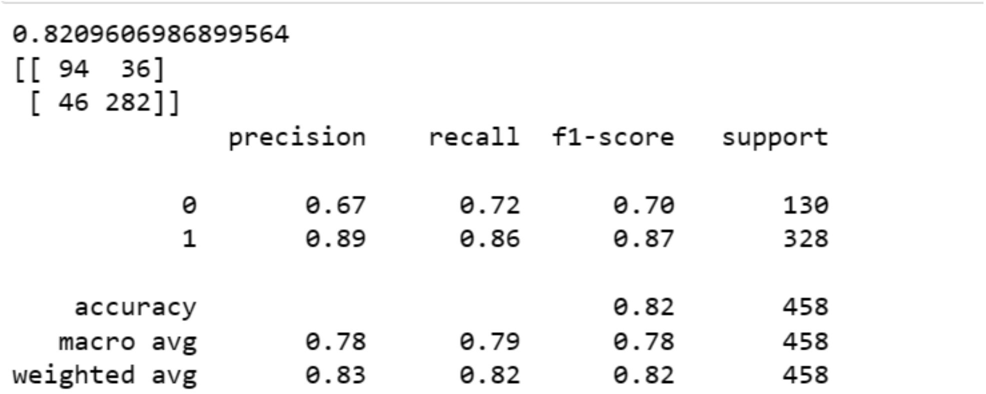
	0	1
0	0.674324	0.325676
1	0.256733	0.743267
2	0.112146	0.887854
3	0.168994	0.831006
4	0.026376	0.973624

-Probability of train data

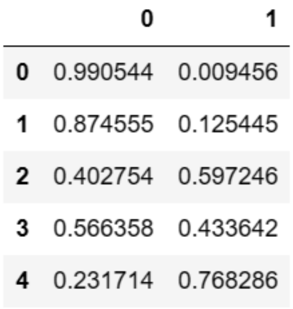


-AUC CURVE OF TRAIN DATA(the auc 0.888)

Test Accuracy, Confusion Matrix and matrix & classification report of GaussianNB:

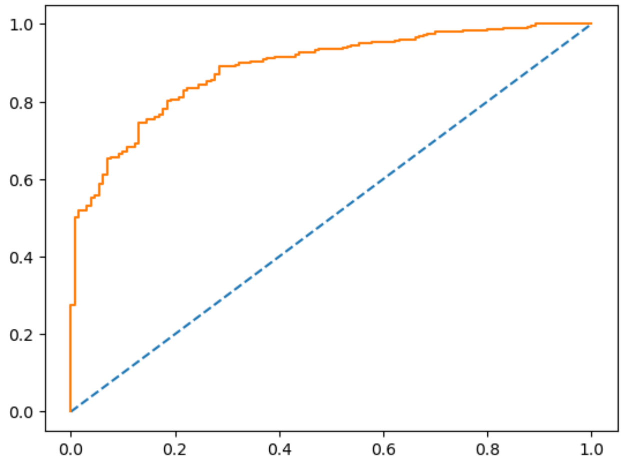


```
0.8209606986899564
[[ 94  36]
 [ 46 282]]
              precision    recall  f1-score   support

           0       0.67      0.72      0.70       130
           1       0.89      0.86      0.87       328

    accuracy                           0.82       458
   macro avg       0.78      0.79      0.78       458
weighted avg       0.83      0.82      0.82       458
```



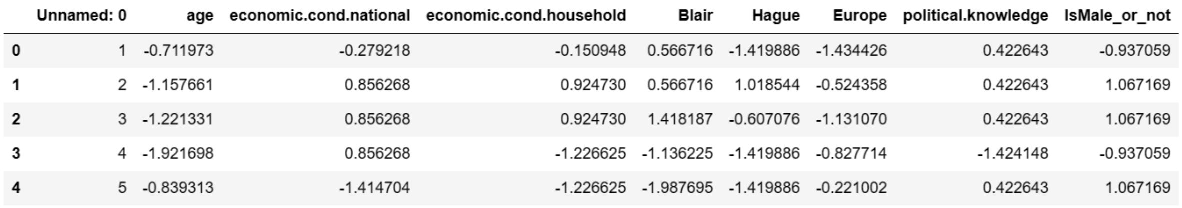
	0	1
0	0.990544	0.009456
1	0.874555	0.125445
2	0.402754	0.597246
3	0.566358	0.433642
4	0.231714	0.768286

-Probability of test data

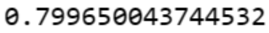


-AUC CURVE OF TEST DATA(the AUC curve 0.886)

AFTER SCALING DATASET:



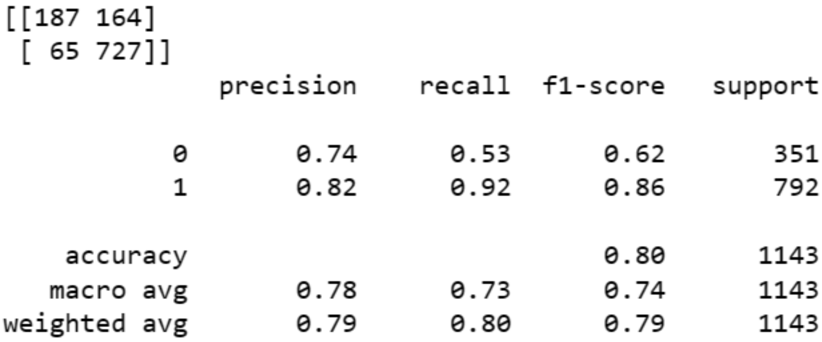
	Unnamed: 0	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	IsMale_or_not
0	1	-0.711973	-0.279218	-0.150948	0.566716	-1.419886	-1.434426	0.422643	-0.937059
1	2	-1.157661	0.856268	0.924730	0.566716	1.018544	-0.524358	0.422643	1.067169
2	3	-1.221331	0.856268	0.924730	1.418187	-0.607076	-1.131070	0.422643	1.067169
3	4	-1.921698	0.856268	-1.226625	-1.136225	-1.419886	-0.827714	-1.424148	-0.937059
4	5	-0.839313	-1.414704	-1.226625	-1.987695	-1.419886	-0.221002	0.422643	1.067169



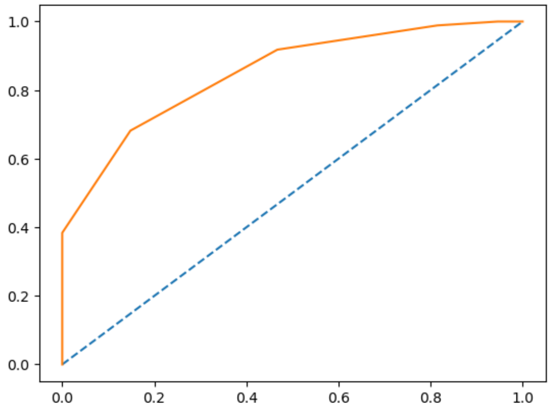
```
0.799650043744532
```

-Model score of KNN

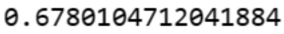
Train confusion matrix and classification of train data: KNN



```
[[187 164]
 [ 65 727]]
              precision    recall  f1-score   support

           0       0.74      0.53      0.62       351
           1       0.82      0.92      0.86       792

    accuracy                           0.80      1143
   macro avg       0.78      0.73      0.74      1143
weighted avg       0.79      0.80      0.79      1143
```

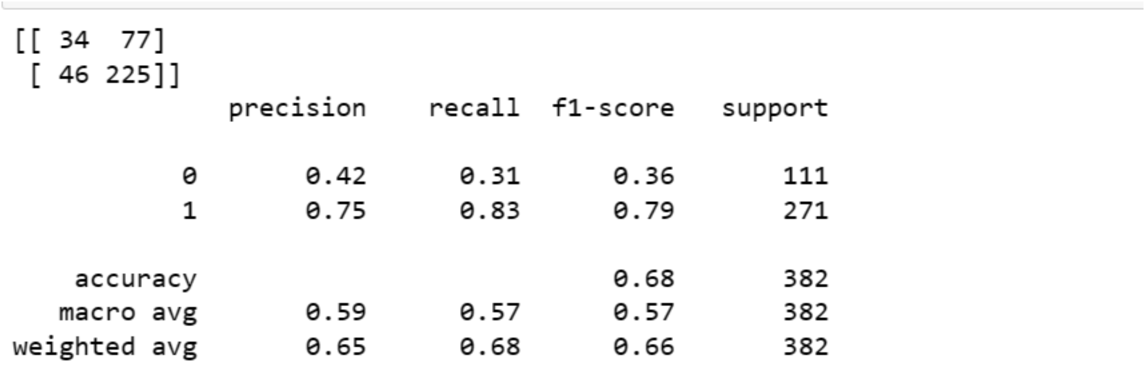


-AUC CURVE OF TRAIN DATA(0.850)



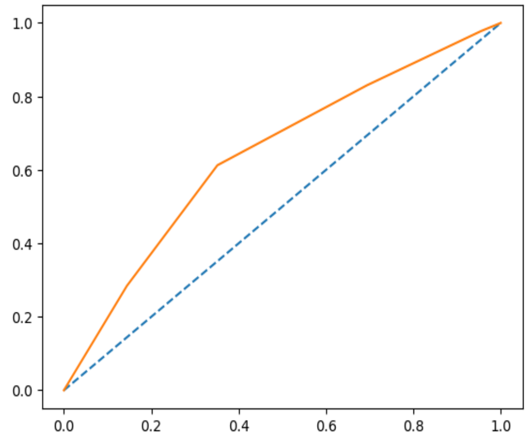
```
0.6780104712041884
```

-MODEL SOCRE OF TEST DATA OF KNN MODEL



```
[[ 34  77]
 [ 46 225]]
              precision    recall  f1-score   support

           0       0.42      0.31      0.36       111
           1       0.75      0.83      0.79       271

    accuracy                           0.68       382
   macro avg       0.59      0.57      0.57       382
weighted avg       0.65      0.68      0.66       382
```

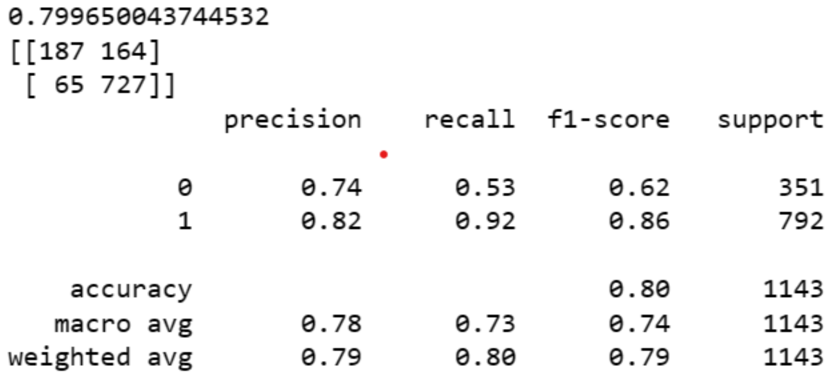
CONFUSION MATRIX AND CLASSIFICATION OF TEST DATA OF KNN MODEL



- AUC CURVE OF KNN MODEL(0.641)

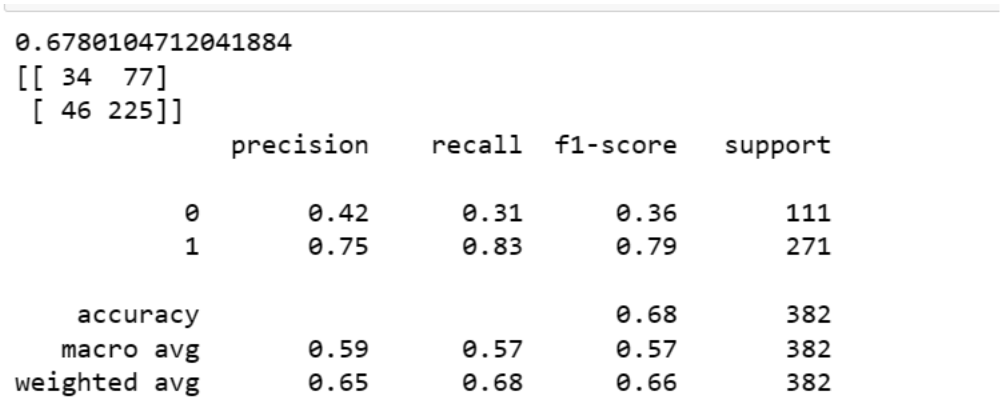
Train Accuracy, Confusion Matrix and matrix & classification report of

KNN NEIGHBOUR CLASSIFIER:



```
0.799650043744532
[[187 164]
 [ 65 727]]
              precision    recall  f1-score   support

           0       0.74      0.53      0.62       351
           1       0.82      0.92      0.86       792

    accuracy                           0.80      1143
   macro avg       0.78      0.73      0.74      1143
weighted avg       0.79      0.80      0.79      1143
```

Test Accuracy, Confusion Matrix and matrix & classification report of KNN NEIGHBOUR CLASSIFIER:



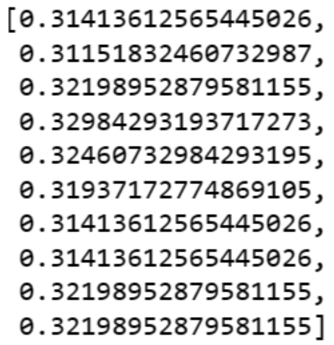
```
0.6780104712041884
[[ 34  77]
 [ 46 225]]
              precision    recall  f1-score   support

           0       0.42      0.31      0.36       111
           1       0.75      0.83      0.79       271

    accuracy                           0.68       382
   macro avg       0.59      0.57      0.57       382
weighted avg       0.65      0.68      0.66       382
```

Overall, it is a good model.

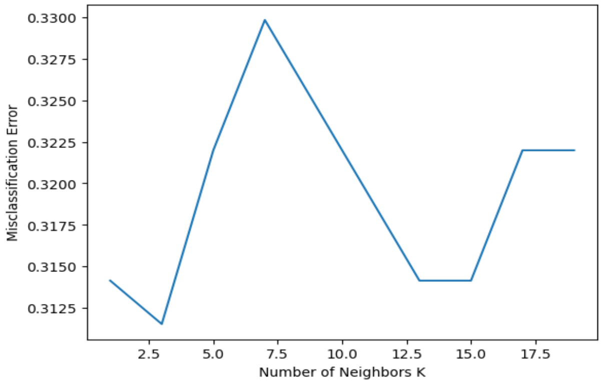
- Comparison between the regular KNN model and tuned KNN model:
- As we can see, the regular KNN model was over-fitted. But model tuning has helped the model to recover from over-fitting.
- The values are better in the tuned KNN model.

Therefore, the tuned KNN model is a better model.

ACCURACY SCORE OF TRAIN AND TEST DATA:

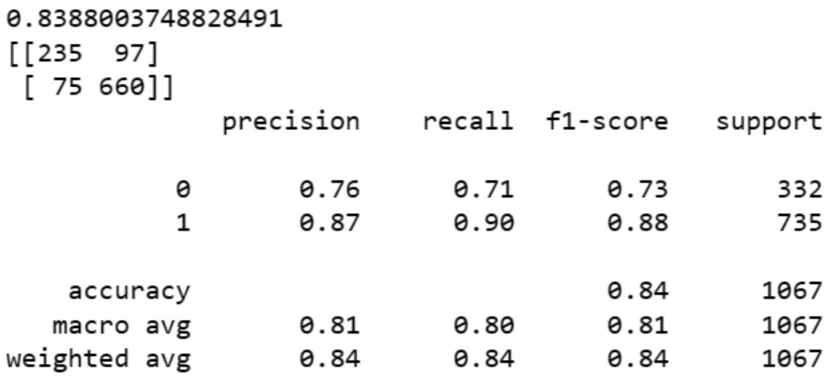


```
[0.31413612565445026,
 0.31151832460732987,
 0.32198952879581155,
 0.32984293193717273,
 0.32460732984293195,
 0.31937172774869105,
 0.31413612565445026,
 0.31413612565445026,
 0.32198952879581155,
 0.32198952879581155]
```

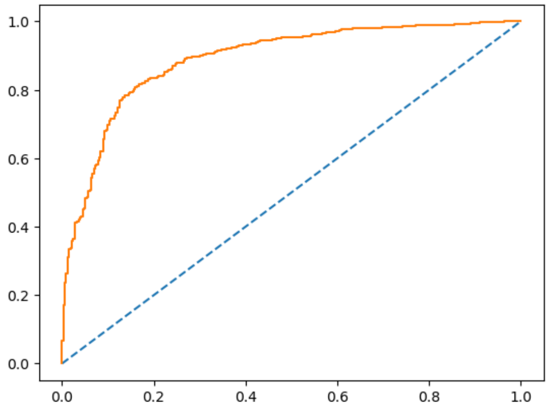


-MISCALCULATION ERROR VS K

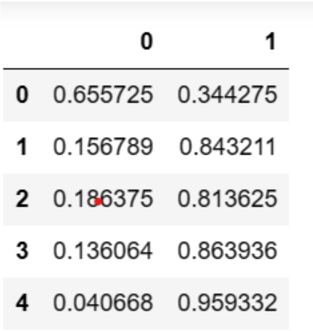
TRAIN ACCURACY ,CONFUSION MATRIX AND CLASSIFICATION OF LDA



```
0.8388003748828491
[[235  97]
 [ 75 660]]
              precision    recall  f1-score   support

           0       0.76      0.71      0.73       332
           1       0.87      0.90      0.88       735

    accuracy                           0.84      1067
   macro avg       0.81      0.80      0.81      1067
weighted avg       0.84      0.84      0.84      1067
```



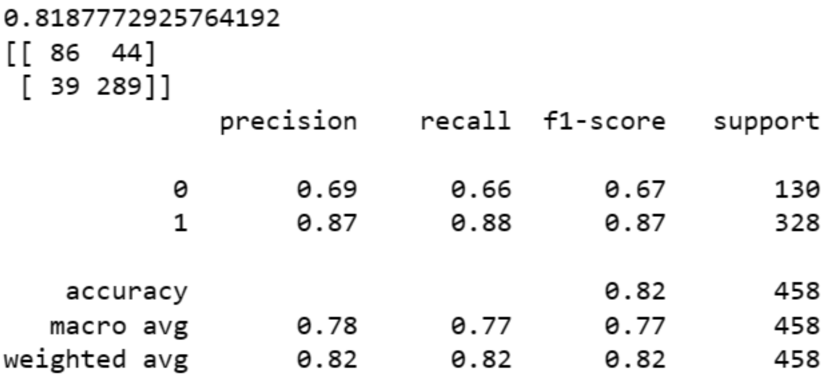
-AUC CURVE OF TRAIN DATA(0.889)



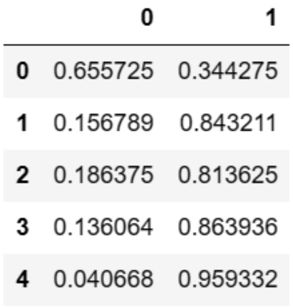
	0	1
0	0.655725	0.344275
1	0.156789	0.843211
2	0.186375	0.813625
3	0.136064	0.863936
4	0.040668	0.959332

-PROBABILITY OF TRAIN DATA OF LDA

TEST ACCURACY, CONFUSION MATRIX & CLASSIFICATION OF LDA

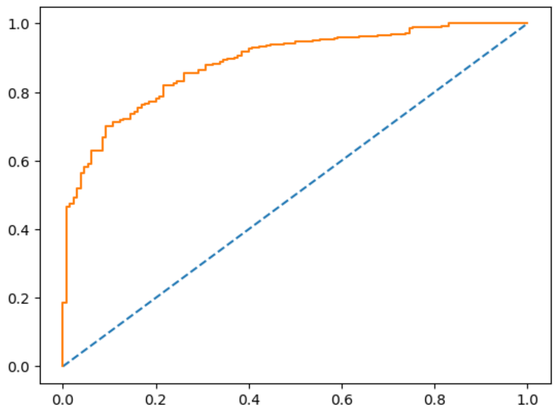


```
0.8187772925764192
[[ 86  44]
 [ 39 289]]
              precision    recall  f1-score   support

           0       0.69      0.66      0.67       130
           1       0.87      0.88      0.87       328

    accuracy                           0.82       458
   macro avg       0.78      0.77      0.77       458
weighted avg       0.82      0.82      0.82       458
```



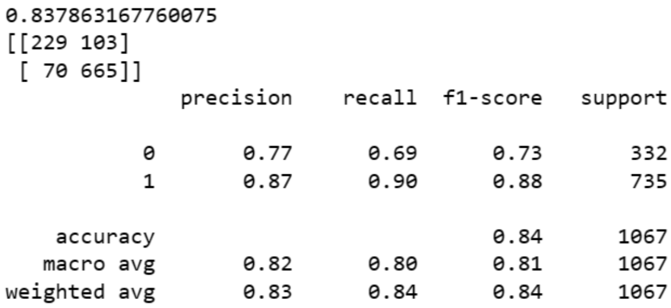
	0	1
0	0.655725	0.344275
1	0.156789	0.843211
2	0.186375	0.813625
3	0.136064	0.863936
4	0.040668	0.959332

-PROBABILITY OF TEST DATA OF LDA

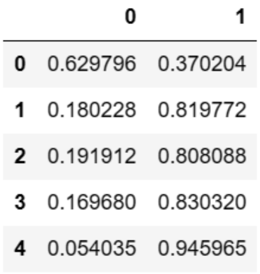


-AUC CURVE OF TEST DATA(0.884)

TRAIN ACCURACY, CONFUSION MATRIX AND CLASSIFICATION OF LOGISTIC REGRESSION:

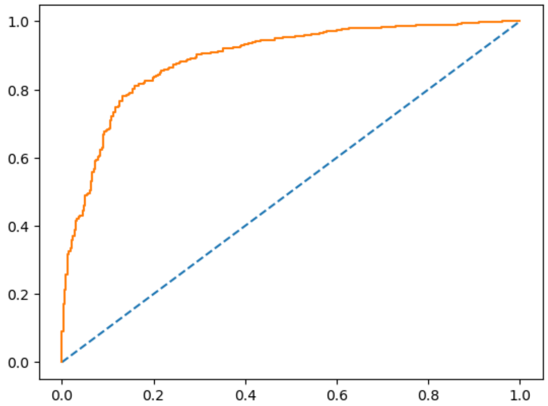


```
0.837863167760075
[[229 103]
 [ 70 665]]
              precision    recall  f1-score   support

           0       0.77      0.69      0.73       332
           1       0.87      0.90      0.88       735

    accuracy                           0.84      1067
   macro avg       0.82      0.80      0.81      1067
weighted avg       0.83      0.84      0.84      1067
```



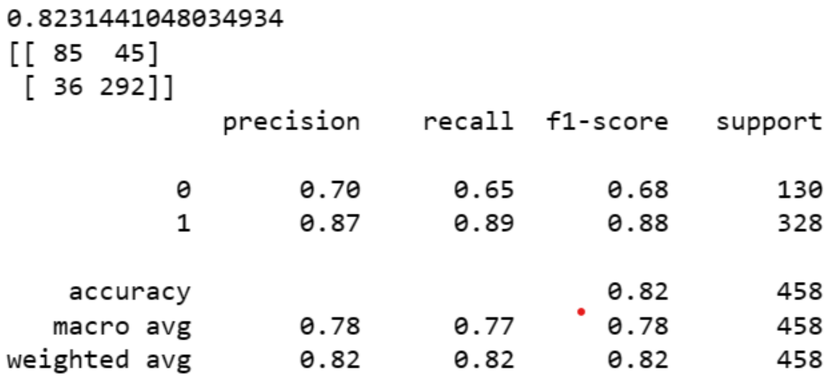
	0	1
0	0.629796	0.370204
1	0.180228	0.819772
2	0.191912	0.808088
3	0.169680	0.830320
4	0.054035	0.945965

- PROBABILITY OF TRAIN DATA OF LOGISTIC REGRESSION

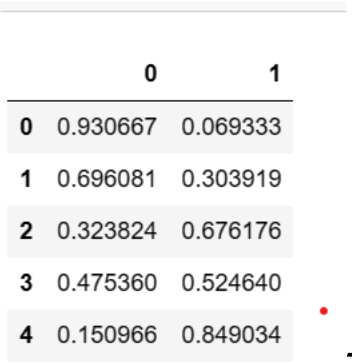


-AUC CURVE OF TRAIN DATA(0.899)

TEST ACCURACY, CONFUSION MATRIX AND CLASSIFICATION OF LOGISTIC REGRESSION:

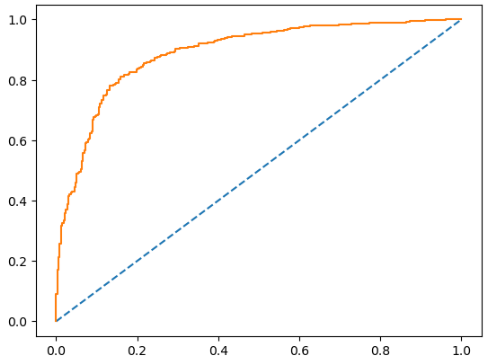


```
0.8231441048034934
[[ 85  45]
 [ 36 292]]
              precision    recall  f1-score   support

           0       0.70      0.65      0.68       130
           1       0.87      0.89      0.88       328

    accuracy                           0.82       458
   macro avg       0.78      0.77      0.78       458
weighted avg       0.82      0.82      0.82       458
```



	0	1
0	0.930667	0.069333
1	0.696081	0.303919
2	0.323824	0.676176
3	0.475360	0.524640
4	0.150966	0.849034

-PROBABILITY OF TEST DATA OF LOGISTIC REGRESSION

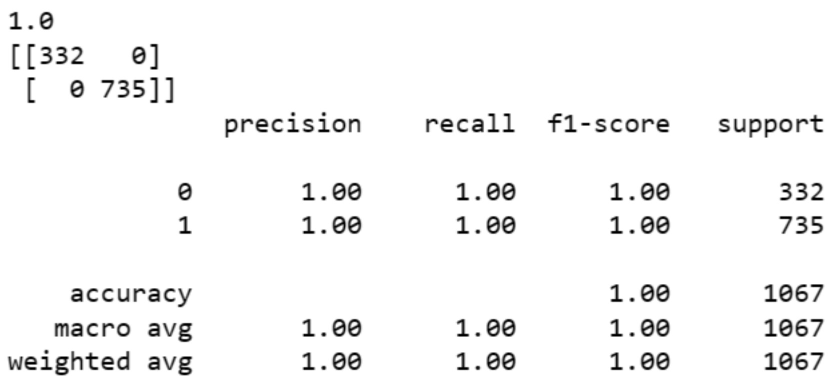


-AUC CURVE OF TEST DATA(0.883)

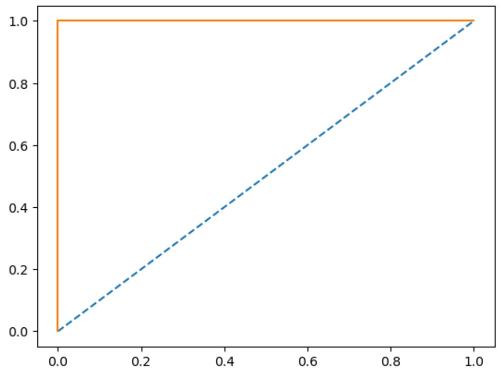
Random Forest – Bagging:

RF model with bagging applied, performs similar to the normal RF as they are not different. The model has good recall and precision also

TRAIN ACCURACY, CONFUSION MATRIX AND CLASSIFICATION OF BAGGING MODEL:

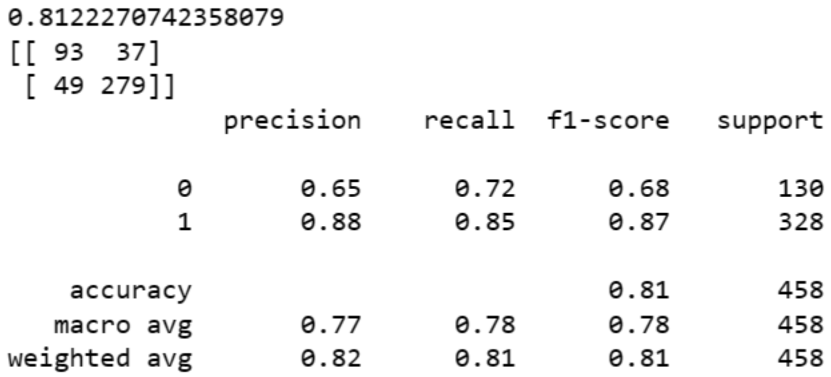


```
1.0
[[332   0]
 [  0 735]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       332
           1       1.00      1.00      1.00       735

    accuracy                           1.00      1067
   macro avg       1.00      1.00      1.00      1067
weighted avg       1.00      1.00      1.00      1067
```

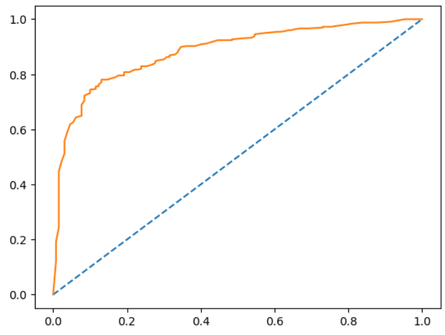


-AUC CURVE OF TRAIN DATA(1 AUC)

TEST ACCURACY, CONFUSION MATRIX AND CLASSIFICATION



```
0.8122270742358079
[[ 93  37]
 [ 49 279]]
              precision    recall  f1-score   support

           0       0.65      0.72      0.68       130
           1       0.88      0.85      0.87       328

    accuracy                           0.81       458
   macro avg       0.77      0.78      0.78       458
weighted avg       0.82      0.81      0.81       458
```

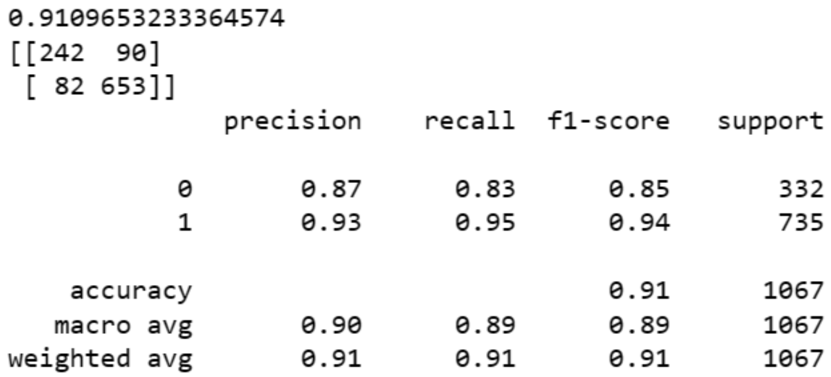


- AUC CUVRE OF TEST DATA(0.882 AUC)

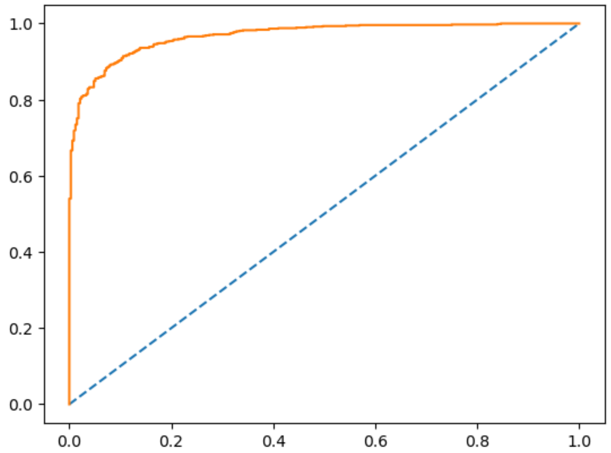
Model Comparison and Best Model :

The gradient Boosting model performs the best with 86% train accuracy. And also has 91% precision and 94% recall which is better than any other models that we have performed here with the Election dataset.

GRADIENT BOOSTING MODEL TRAIN:

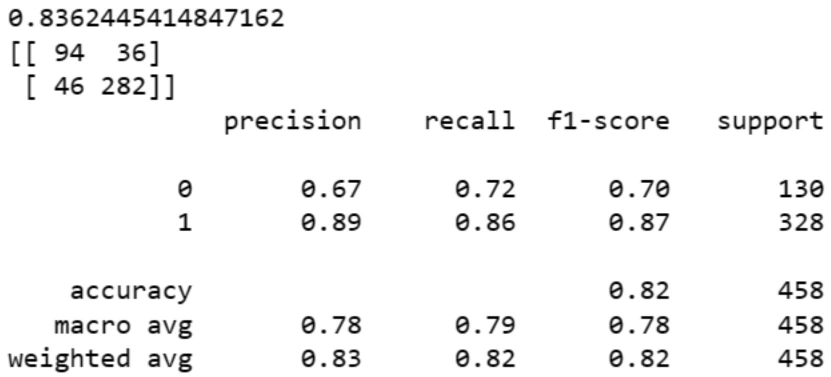


```
0.9109653233364574
[[242  90]
 [ 82 653]]
              precision    recall  f1-score   support

           0       0.87      0.83      0.85       332
           1       0.93      0.95      0.94       735

    accuracy                           0.91      1067
   macro avg       0.90      0.89      0.89      1067
weighted avg       0.91      0.91      0.91      1067
```

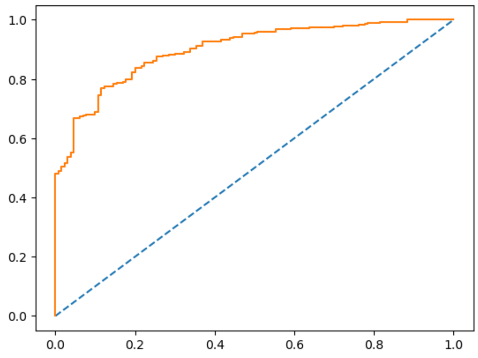


-AUC CUVRVE OF TRAIN DATA(0.96 AUC)

TRAIN ACCURACY OF GRADIENT BOOSTING MODEL:



```
0.8362445414847162
[[ 94  36]
 [ 46 282]]
              precision    recall  f1-score   support

           0       0.67      0.72      0.70       130
           1       0.89      0.86      0.87       328

    accuracy                           0.82       458
   macro avg       0.78      0.79      0.78       458
weighted avg       0.83      0.82      0.82       458
```



-AUC CUVRE OF TEST DATA (0.91 AUC)

Q8) Based on these predictions, what are the insights

- Comparing all the performance measure, Naïve Bayes model from second iteration is performing best. Although there are some other models such as SVM and Extreme Boosting which is performing almost same as that of Naïve Bayes. But Naïve Bayes model is very consistent when train and test results are compared with each other. Along with other parameters such as Recall value, AUC_SCORE and AUC_ROC_Curve, those results were pretty good is this model.

- Labour party is performing better than the Conservative from a huge margin.

- Female voters turnout is greater than that male voters.

- Those who have better national economic conditions prefer to vote for the Labour party.

- Persons having higher Eurosceptic sentiments conservative party prefer to vote for Conservative Party.

- Those who have higher political knowledge have voted for the Conservative party

- Looking at the assessment for both the leaders, the Labour Leader is performing well as he has got better ratings in the assessment.

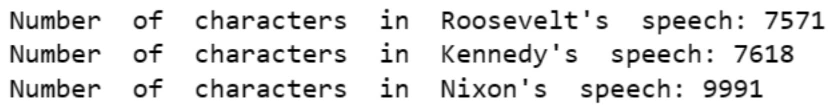
Problem 2: In this particular project, we are going to work on the inaugural corpora from the nltk in Python. We will be looking at the following speeches of the Presidents of the United States of America:

1. President Franklin D. Roosevelt in 1941

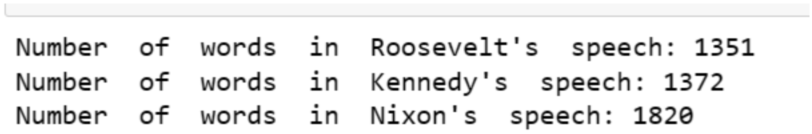
2. President John F. Kennedy in 1961

3. President Richard Nixon in 1973

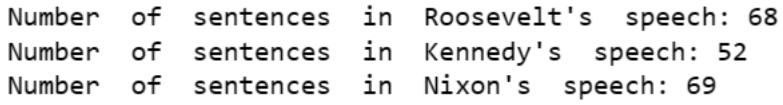
2.1 Find the number of characters, words and sentences for the mentioned documents. (Hint: use .words(), .raw(), .sent() for extracting counts):



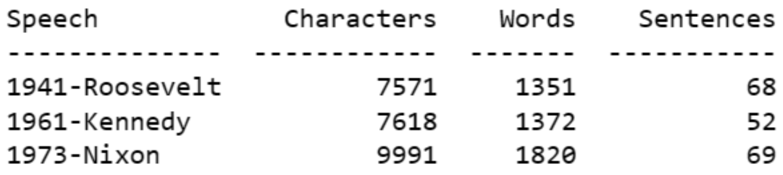
```
Number  of  characters  in  Roosevelt's  speech: 7571
Number  of  characters  in  Kennedy's  speech: 7618
Number  of  characters  in  Nixon's  speech: 9991
```



```
Number  of  words  in  Roosevelt's  speech: 1351
Number  of  words  in  Kennedy's  speech: 1372
Number  of  words  in  Nixon's  speech: 1820
```

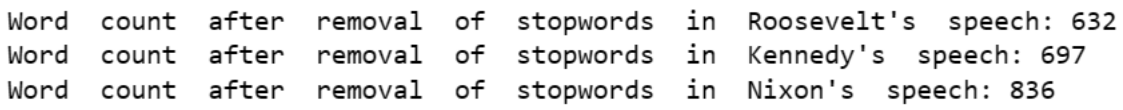


```
Number  of  sentences  in  Roosevelt's  speech: 68
Number  of  sentences  in  Kennedy's  speech: 52
Number  of  sentences  in  Nixon's  speech: 69
```



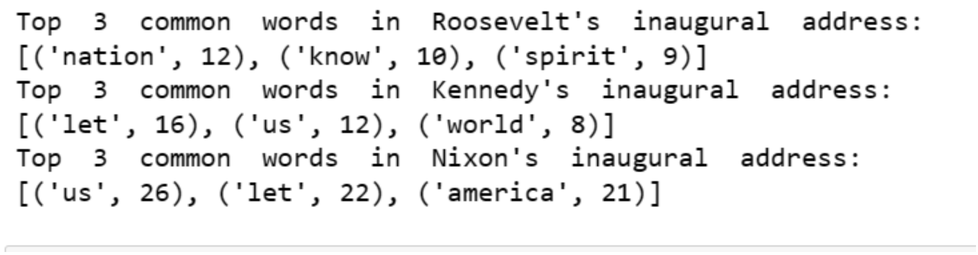
```
Speech            Characters    Words    Sentences
--------------  ------------  -------  -----------
1941-Roosevelt          7571     1351           68
1961-Kennedy            7618     1372           52
1973-Nixon              9991     1820           69
```

2.2 Remove all the stop words from all the three speeches.



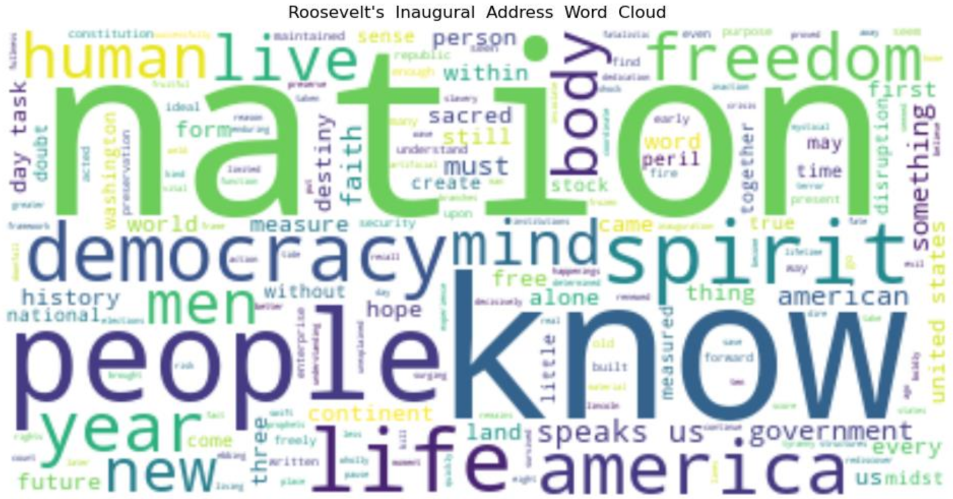
```
Word  count  after  removal  of  stopwords  in  Roosevelt's  speech: 632
Word  count  after  removal  of  stopwords  in  Kennedy's  speech: 697
Word  count  after  removal  of  stopwords  in  Nixon's  speech: 836
```

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 stop words)

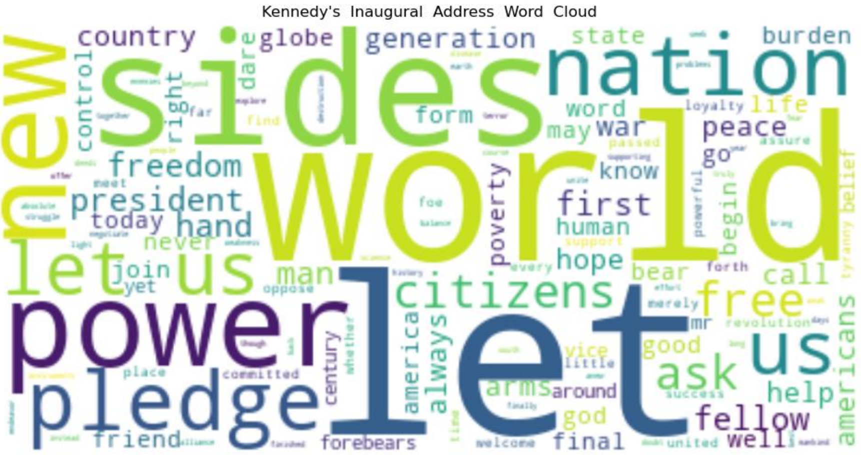


```
Top 3 common words in Roosevelt's inaugural address:
[('nation', 12), ('know', 10), ('spirit', 9)]
Top 3 common words in Kennedy's inaugural address:
[('let', 16), ('us', 12), ('world', 8)]
Top 3 common words in Nixon's inaugural address:
[('us', 26), ('let', 22), ('america', 21)]
```

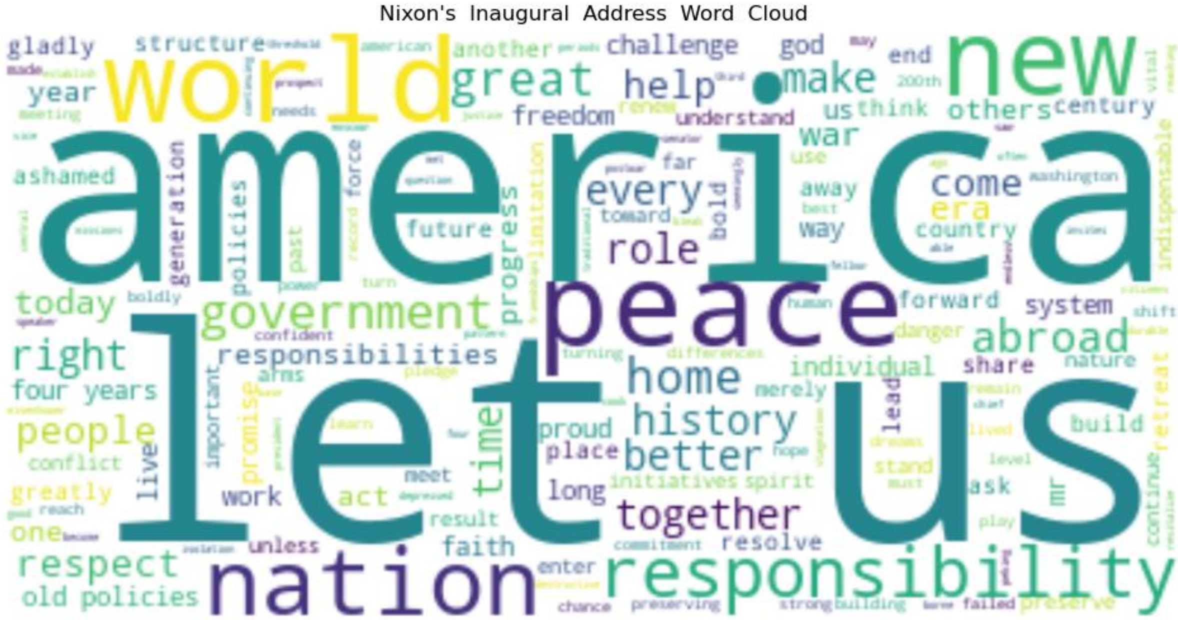
2.4 Plot the word cloud of each of the speeches of the variable. (after removing the stop words):



<u>Word Cloud for President Franklin D. Roosevelt's speech (after cleaning)</u>



__Word Cloud for President John F. Kennedy's Speech (after cleaning)__



__Word Cloud for President Richard Nixon's Speech (after cleaning)__