MACHINE LEARNING

Business report

S Nitin kumar

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.

1.1) **Read the dataset. Describe the data briefly. Interpret the inferences for each. Initial steps like head() .info(), Data Types, etc . Null value check, Summary stats, Skewness must be**

Head 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

Shape of the data set

```
df.shape
```

```
(1525, 9)
```

Information of the data set

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1525 entries, 0 to 1524
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: 107.4+ KB
```

We have 2 object in the data sets and the unique data set for these object data sets are

1.2) **Perform EDA (Check the null values, Data types, shape, Univariate, bivariate analysis). Also check for outliers (4 pts). Interpret the inferences for each (3 pts) Distribution plots(histogram) or similar plots for the continuous columns. Box plots, Correlation plots. Appropriate plots for categorical variables. Inferences on each plot. Outliers proportion should be discussed, and inferences from above used plots should be there. There is no restriction on how the learner wishes to implement this but the code should be able to represent the correct output and inferences should be logical and correct.**

There are no null values in the data set

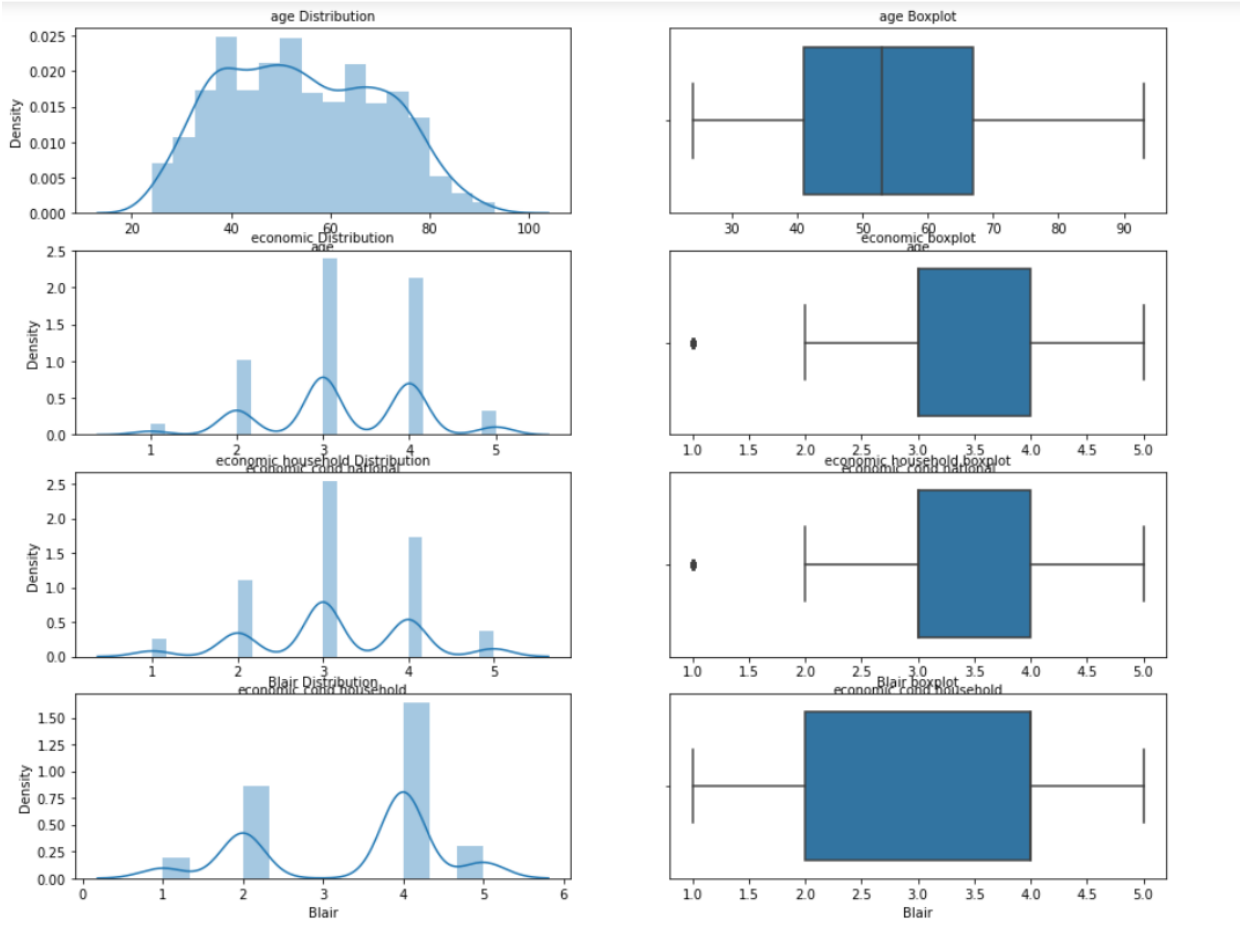
```
df.isnull().sum()
```

```
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
```

```
df.dtypes
```

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

```
VOTE  2
Conservative     462
Labour          1063
Name: vote, dtype: int64
GENDER  2
male      713
female    812
Name: gender, dtype: int64
```



Age:

- The data is normally distributed.
- Maximum number of people are aged between 40 and 70.
- Outliers are not present.
- The minimum value is 24 and the maximum value is 93.

Count plot of 'economic.cond.national':

- The top 2 variables are 3 and 4.
- 1 has the least value which is 37.
- 3 has the highest value which is 604.
- 3 is slightly higher than the 2nd highest variable 4 whose value is 538.

Count plot of 'economic.cond.household':

- The top 2 variables are 3 and 4.
- 1 has the least value which is 65.
- 3 has the highest value which is 645.
- 3 is moderately higher than the 2nd highest variable 4 whose value is 435.

Count plot of 'Blair':

- The top 2 variables are 2 and 4.

- 3 has the least value which is 1.
- 4 has the highest value which is 833.
- 4 is much higher than the 2nd highest variable 2 whose value is 434.

'Hague':

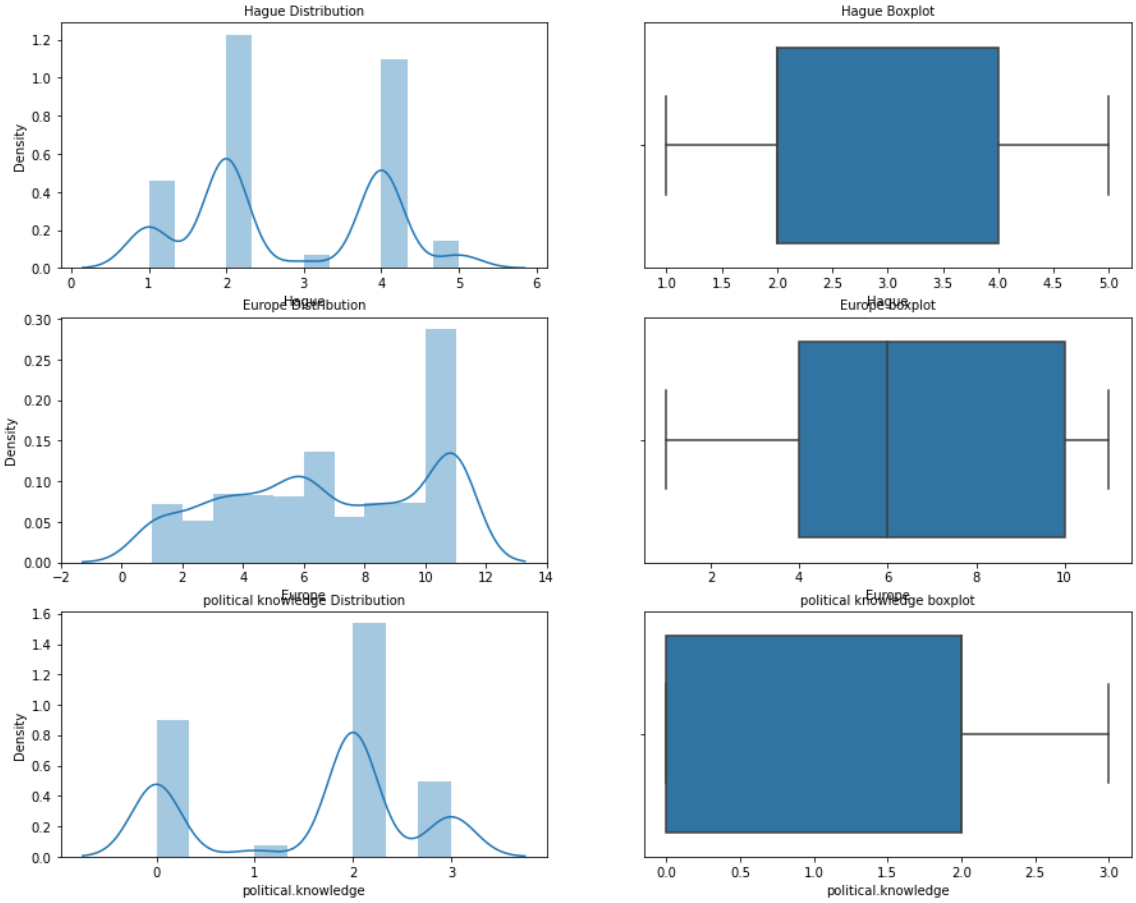
- The top 2 variables are 2 and 4.
- 3 has the least value which is 37.
- 2 has the highest value which is 617.
- 2 is slightly higher than the 2nd highest variable 4 whose value is 557.

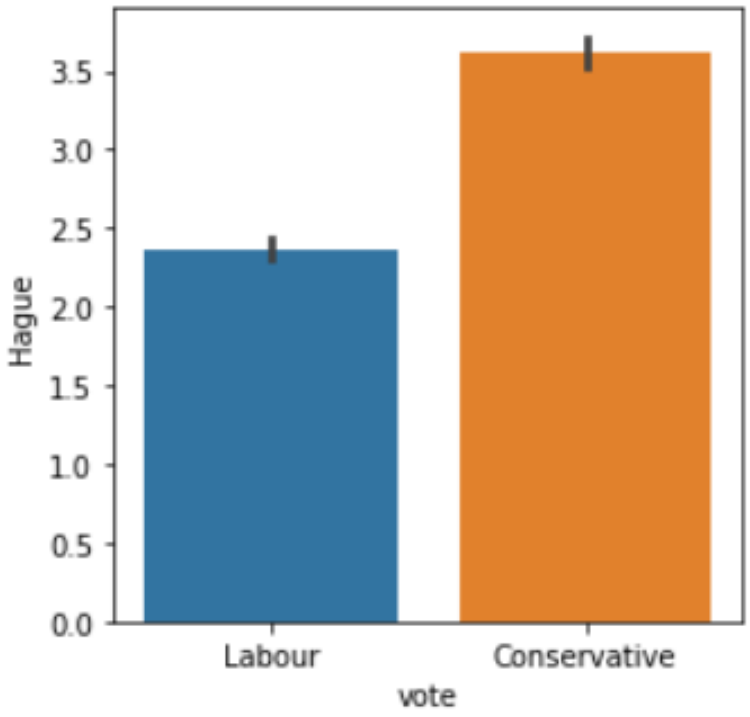
'Europe':

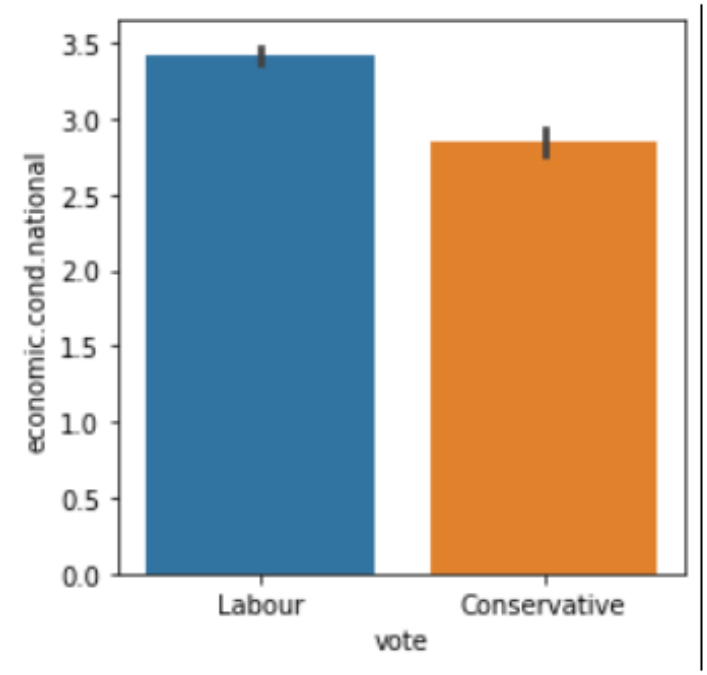
- The top 2 variables are 11 and 6. • 2 has the least value which is 77.
- 11 has the highest value which is 338.
- 11 is moderately higher than the 2nd highest variable 6 whose value is 207.

'political.knowledge':

- The top 2 variables are 2 and 0.
- 1 has the least value which is 38.
- 2 has the highest value which is 776.
- 2 is much higher than the 2nd highest variable 0 whose value is 454.







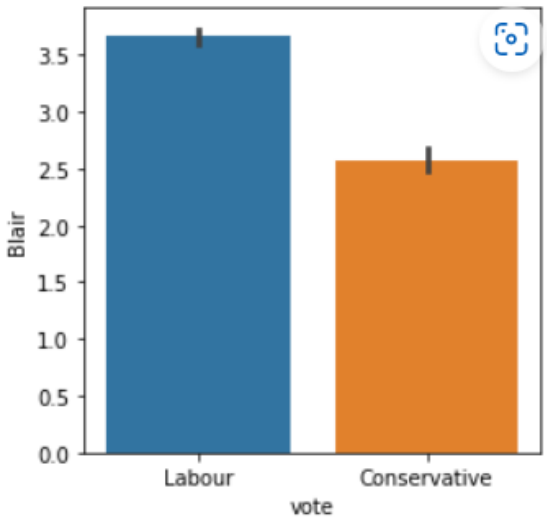
'vote' with respect to 'economic.cond.national':

- Labour party has higher votes overall.
- Out of 82 people who gave a score of 5, 73 people have voted for the labour party.
- Out of 538 people who gave a score of 4, 447 people have voted for the labour party. This is the highest set of people in the labour party.

- Out of 604 people who gave a score of 3, 405 people have voted for the labour party. This is the 2nd highest set of people in the labour party. The remaining 199 people who have voted for the conservative party is the highest set of people in that party.
- Out of 256 people who gave a score of 2, 116 people have voted for the labour party. 140 people have voted for the conservative party. This is the instance where the conservative party has got more votes than the labour party.
- Out of 37 people who gave a score of 1, 16 people have voted for the labour party. 21 people have voted for the conservative party. • The score of 3, 4 and 5 have more votes in the labour party.
- The score of 1 and 2 have more votes in the conservative party.

'vote' with respect to 'economic.cond.household':

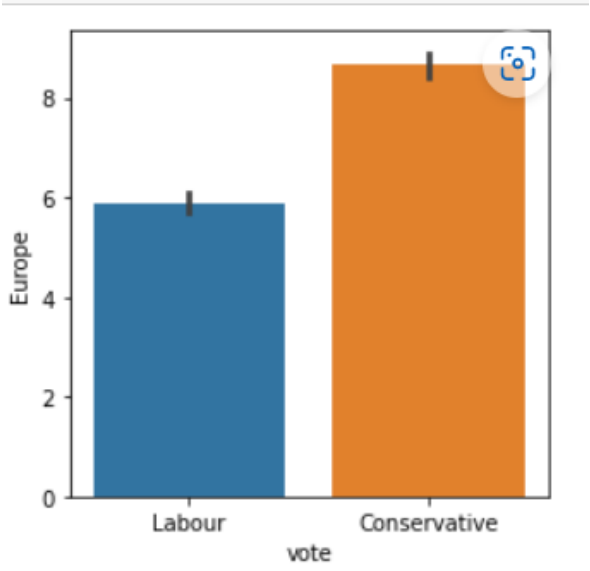
- Labour party has higher votes overall.
- Out of 92 people who gave a score of 5, 69 people have voted for the labour party.
- Out of 435 people who gave a score of 4, 349 people have voted for the labour party. This is the 2nd highest set of people in the labour party.
- Out of 645 people who gave a score of 3, 448 people have voted for the labour party. This is the highest set of people in the labour party. The remaining 197 people who have voted for the conservative party is the highest set of people in that party.
- Out of 280 people who gave a score of 2, 154 people have voted for the labour party. 126 people have voted for the conservative party.
- Out of 65 people who gave a score of 1, 37 people have voted for the labour party. 28 people have voted for the conservative party.



'vote' and 'Blair':

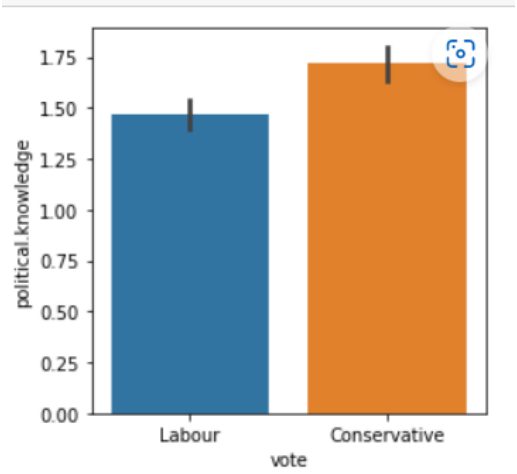
- Labour party has higher votes overall.

- Out of 152 people who gave a score of 5, 149 people have voted for the labour party. The remaining 3 people, despite giving a score of 5 to the labour leader, have chosen to vote for the conservative party.
- Out of 833 people who gave a score of 4, 676 people have voted for the labour party. The remaining 157 people, despite giving a score of 4 to the labour leader, have chosen to vote for the conservative party.
- Only 1 person has given a score of 3 and that person has voted for the conservative party.



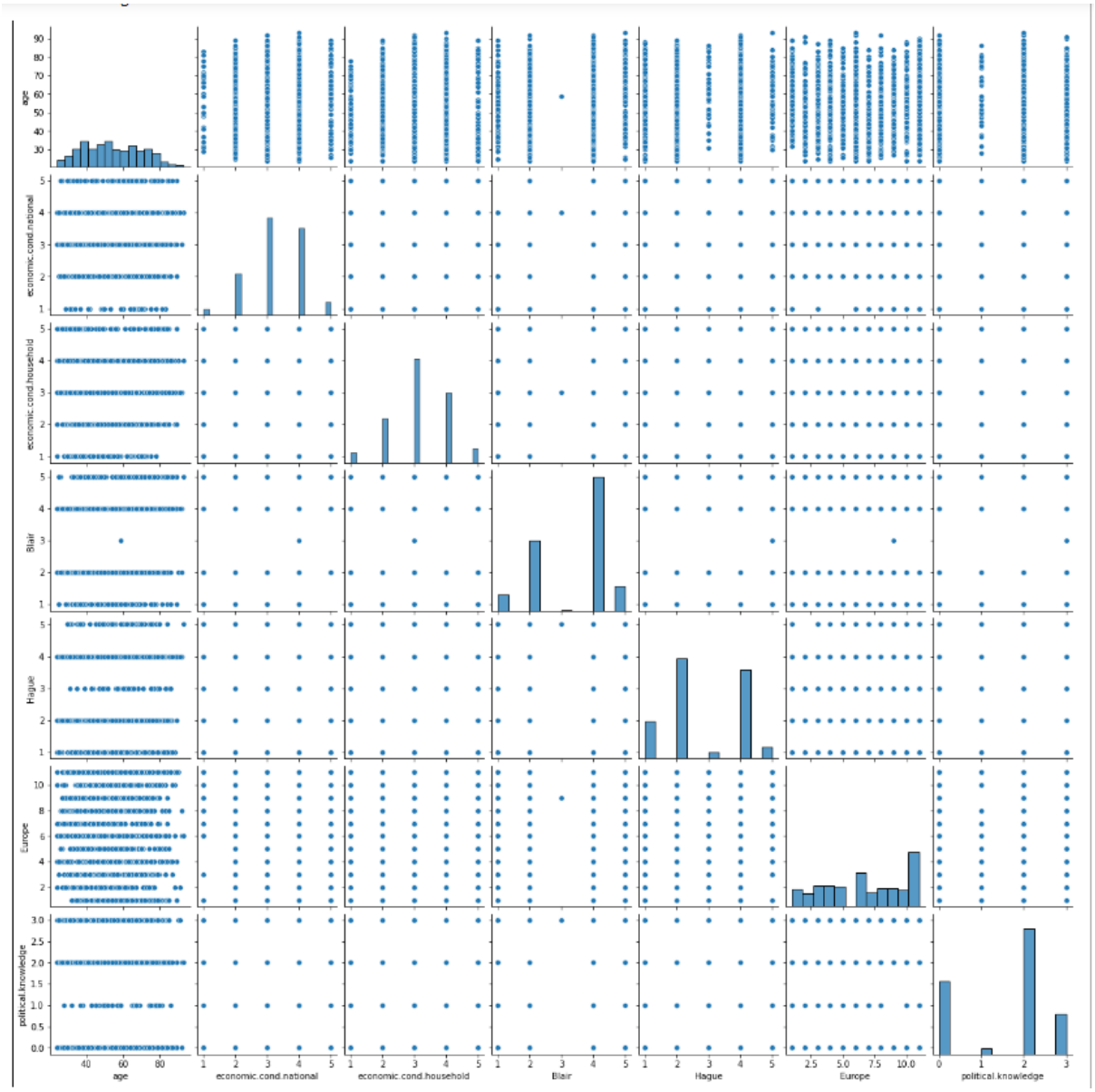
'vote' and 'Europe':

- Out of 338 people who gave a score of 11, 166 people have voted for the labour party and 172 people have voted for the conservative party.
- People who gave score of 7 to 10 have voted for labour and conservative almost equally. Conservative party seem to be slightly higher in these instances.
- Out of 207 people who gave a score of 6, 172 people have voted for the labour party and 35 people have voted for the conservative party.



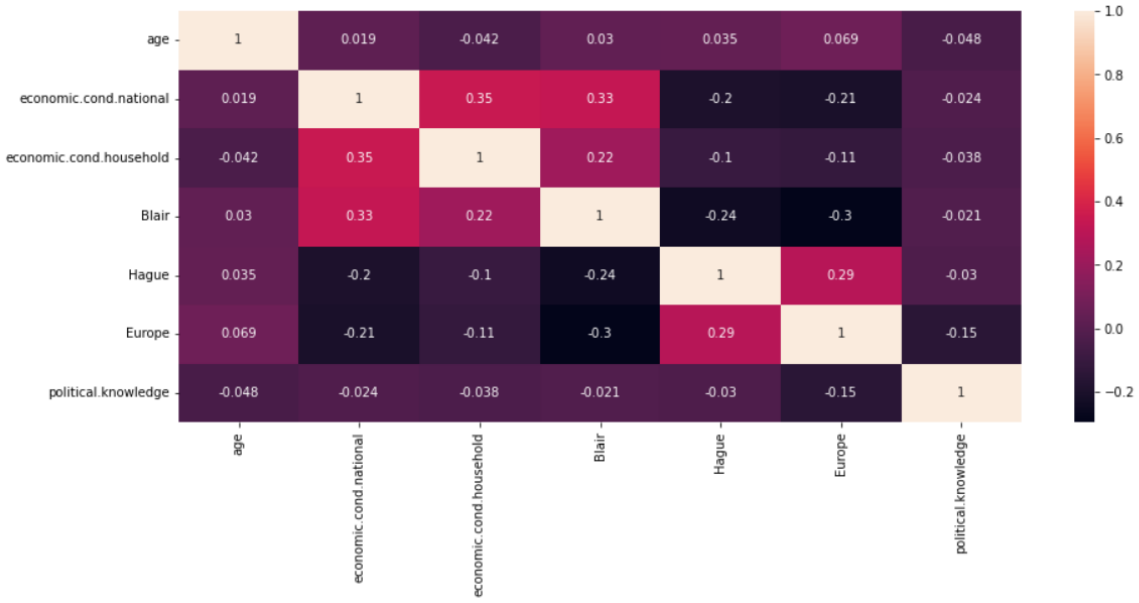
'vote' and 'political.knowledge':

- Out of 249 people who gave a score of 3, 177 people have voted for the labour party and 72 people have voted for the conservative party.
- Out of 776 people who gave a score of 2, 493 people have voted for the labour party and 283 people have voted for the conservative party.
- Out of 38 people who gave a score of 1, 27 people have voted for the labour party and 11 people have voted for the conservative party.
- Out of 454 people who gave a score of 0, 360 people have voted for the labour party and 94 people have voted for the conservative party.



Observation:

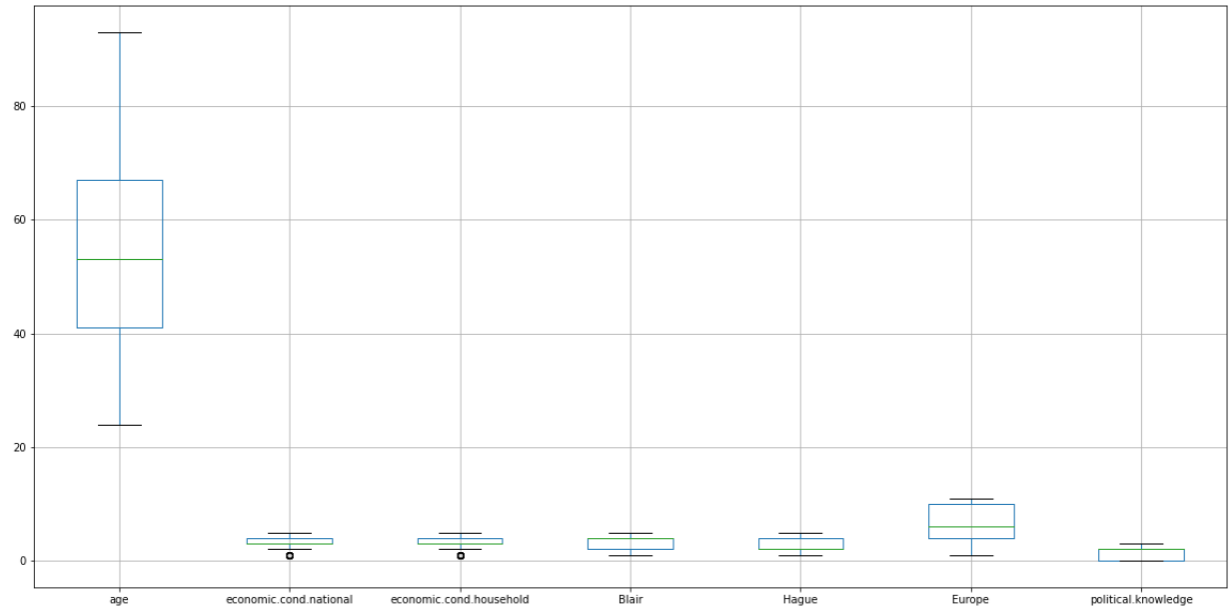
- Pair plot is a combination of histograms and scatter plots.
- From the histogram, we can see that, the 'Blair','Europe' and 'political.knowledge' variables are slightly left skewed.
- All other variables seem to be normally distributed.
- From the scatter plots, we can see that, there is mostly no correlation between the variables.
- We can use the correlation matrix to view them more clearly.



Observation:

- We can see that, mostly there is no correlation in the dataset through this matrix. There are some variables that are moderately positively correlated and some that 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 moderate negative correlation.

'Europe' with 'economic.cond.national' and 'Blair' have moderate negative correlation.



1.3) **Encode the data (having string values) for Modelling. Is Scaling necessary here or not?(2 pts), Data Split: Split the data into train and test (70:30) (2 pts). The learner is expected to check and comment about the difference in scale of different features on the bases of appropriate measure for example std dev, variance, etc. Should justify whether there is a necessity for scaling. Object data should be converted into categorical/numerical data to fit in the models. (pd. categorical(). codes(), pd.get_dummies(drop_first=True)) Data split, ratio defined for the split, train-test split should be discussed.**

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 set. Earlier, we have extracted the target variable 'vote_Labour' in a separate vector for subsets. Random state chosen as 1.

- **Training Set:** 70percent of data.
- **Testing Set:** 30 percent of the data.

Why scaling?:

- The dataset contains features highly varying in magnitudes, units and range between the 'age' column and other columns.
- But since, most of the machine learning algorithms use Eucledian distance between two data points in their computations, this is a problem.
- If left alone, these algorithms only take in the magnitude of features neglecting the units.
- The results would vary greatly between different units, 1 km and 1000 metres.
- The features with high magnitudes will weigh in a lot more in the distance calculations than features with low magnitudes.
- To supress this effect, we need to bring all features to the same level of magnitudes. This can be acheived by scaling.

in this case, we have a lot of encoded, ordinal, categorical and continuous variables. So, we use the minmaxscaler technique to scale the data.

Before scaling

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

After scaling

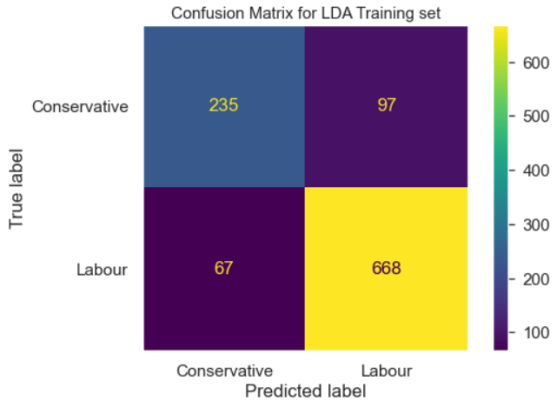
	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender_male
0	-0.711973	-0.279218	-0.150948	0.566716	-1.419886	-1.434426	0.422643	-0.937059
1	-1.157661	0.856268	0.924730	0.566716	1.018544	-0.524358	0.422643	1.067169
2	-1.221331	0.856268	0.924730	1.418187	-0.607076	-1.131070	0.422643	1.067169
3	-1.921698	0.856268	-1.226625	-1.136225	-1.419886	-0.827714	-1.424148	-0.937059
4	-0.839313	-1.414704	-1.226625	-1.987695	-1.419886	-0.221002	0.422643	1.067169
5	-0.457295	-0.279218	0.924730	0.566716	1.018544	-0.827714	0.422643	1.067169
6	0.179402	-1.414704	-1.226625	0.566716	1.018544	1.295778	0.422643	1.067169
7	1.452797	-0.279218	0.924730	0.566716	-1.419886	-1.737782	-1.424148	1.067169
8	-0.966652	-0.279218	-0.150948	0.566716	1.018544	1.295778	-1.424148	-0.937059
9	1.007109	-0.279218	-1.226625	1.418187	-1.419886	1.295778	0.422643	1.067169

1.4) **Apply Logistic Regression and LDA (Linear Discriminant Analysis) (2 pts). Interpret the inferences of both model s (2 pts). Successful implementation of each model. Logical reason behind the selection of different values for the parameters involved in each model. Calculate Train and Test Accuracies for each model. Comment on the validness of models (over fitting or under fitting)**

```
logit_train_accu  0.84
logit_train_precision  0.77
logit_train_recall  0.7
logit_train_f1  0.73
```

Classification report for training data set

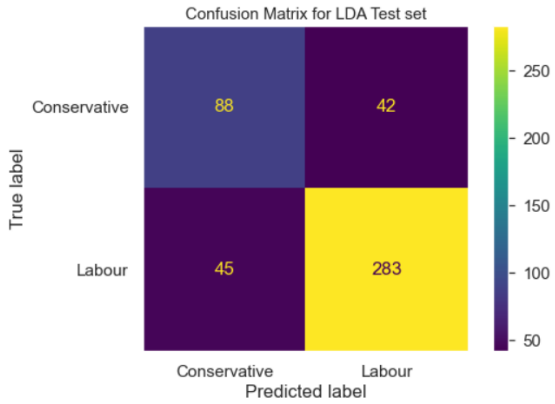
```
              precision    recall  f1-score   support

           0       0.78      0.71      0.74       332
           1       0.87      0.91      0.89       735

    accuracy                           0.85      1067
   macro avg       0.83      0.81      0.82      1067
weighted avg       0.84      0.85      0.84      1067
```




```
logit_test_accu  0.82
logit_test_precision  0.68
logit_test_recall  0.7
logit_test_f1  0.69
```

Classification report for testing data set

	precision	recall	f1-score	support
0	0.66	0.68	0.67	130
1	0.87	0.86	0.87	328
accuracy			0.81	458
macro avg	0.77	0.77	0.77	458
weighted avg	0.81	0.81	0.81	458




```
AUC for Training data = 0.8983402999754119
AUC for Test data = 0.8824695121951219
```

Validness of the model:

- The model is not over-fitted or under-fitted.
- The error in the test data is slightly higher than the train data, which is absolutely fine because the error margin is low and the error in both train and test data is not too high. Thus, the model is not over-fitted or under-fitted.

- ## Linear Discriminant Analysis Model:

- There are no outliers present in the continuous variable 'age'. The remaining variables are categorical in nature. Our model will use all the variables and 'vote_Labour' is the target variable.

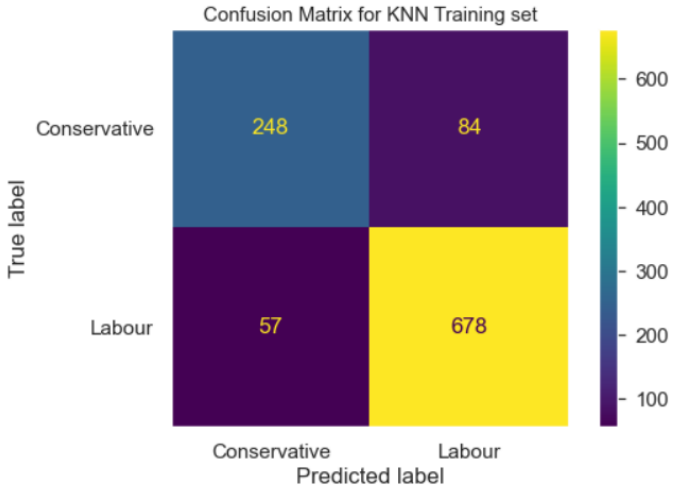
1.5) **Apply KNN Model and Naïve Bayes Model (2pts). Interpret the inferences of each model (2 pts). Successful implementation of each model. Logical reason behind the selection of different values for the parameters involved in each model. Calculate Train and Test Accuracies for each model. Comment on the vapidness of models (over fitting or under fitting)**

K-Nearest Neighbor Model:

2. There are no outliers present in the continuous variable 'age'. The remaining variables are categorical in nature. Our model will use all the variables and 'vote_Labour' is the target variable. We take K value as 7.

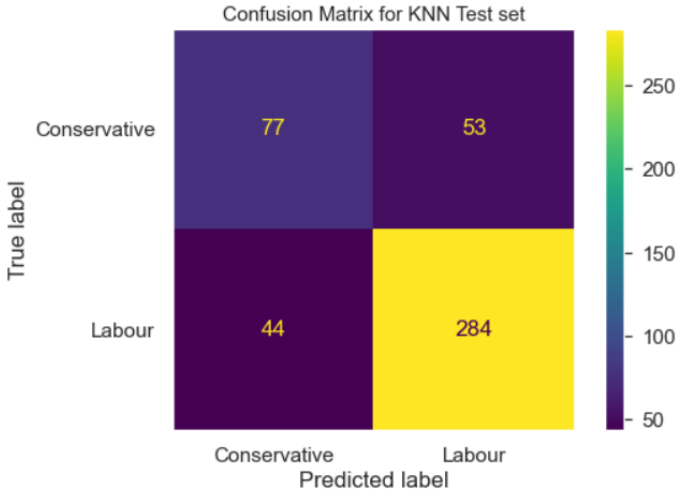
Classification report for training data set

	precision	recall	f1-score	support
0	0.81	0.75	0.78	332
1	0.89	0.92	0.91	735
accuracy			0.87	1067
macro avg	0.85	0.83	0.84	1067
weighted avg	0.87	0.87	0.87	1067



Classification report for testing data set

	precision	recall	f1-score	support
0	0.64	0.59	0.61	130
1	0.84	0.87	0.85	328
accuracy			0.79	458
macro avg	0.74	0.73	0.73	458
weighted avg	0.78	0.79	0.79	458



AUC for Training data = 0.9287435456110154
AUC for Test data = 0.8277908067542215

Validness of the model:

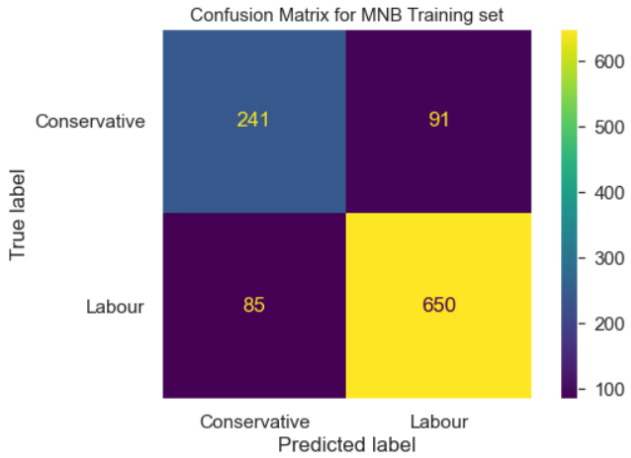
- The model is over-fitted.

Naïve Bayes Model:

There are no outliers present in the continuous variable 'age'. The remaining variables are categorical in nature. Our model will use all the variables and 'vote_Labour' is the target variable.

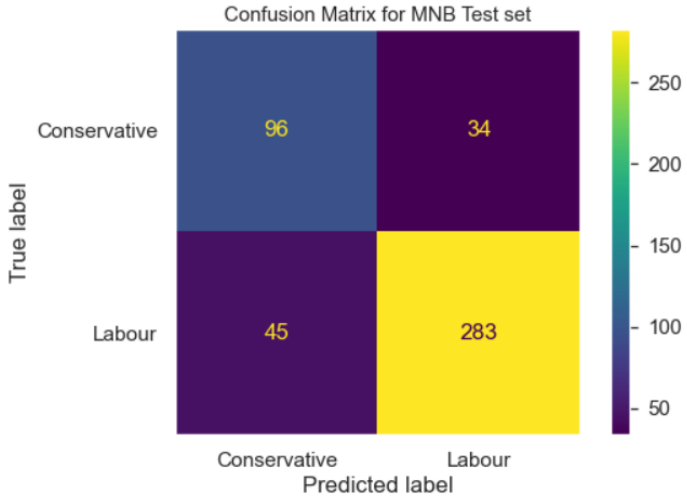
Classification report for training data set

	precision	recall	f1-score	support
0	0.74	0.73	0.73	332
1	0.88	0.88	0.88	735
accuracy			0.84	1067
macro avg	0.81	0.81	0.81	1067
weighted avg	0.83	0.84	0.83	1067



Classification report for testing data set

	precision	recall	f1-score	support
0	0.68	0.74	0.71	130
1	0.89	0.86	0.88	328
accuracy			0.83	458
macro avg	0.79	0.80	0.79	458
weighted avg	0.83	0.83	0.83	458




```
AUC for Training data = 0.8953733300549136
AUC for Test data = 0.8885201688555346
```

Validness of the model:

- The model is not over-fitted or under-fitted.

- The error in the test data is slightly higher than the train data, which is absolutely fine because the error margin is low and the error in both train and test data is not too high. Thus, the model is not over-fitted or under-fitted.

2.1) **Model Tuning (4 pts) , Bagging (1.5 pts) and Boosting (1.5 pts). Apply grid search on each model (include all models) and make models on best_params. Define a logic behind choosing particular values for different hyper-parameters for grid search. Compare and comment on performances of all. Comment on feature importance if applicable. Successful implementation of both algorithms along with inferences and comments on the model performances.**

Linear Regression with SMOTE

	precision	recall	f1-score	support
0	0.689922	0.684615	0.687259	130.000000
1	0.875380	0.878049	0.876712	328.000000
accuracy	0.823144	0.823144	0.823144	0.823144
macro avg	0.782651	0.781332	0.781986	458.000000
weighted avg	0.822739	0.823144	0.822937	458.000000

	precision	recall	f1-score	support
0	0.684211	0.700000	0.692015	130.000000
1	0.880000	0.871951	0.875957	328.000000
accuracy	0.823144	0.823144	0.823144	0.823144
macro avg	0.782105	0.785976	0.783986	458.000000
weighted avg	0.824427	0.823144	0.823747	458.000000

LDA with SMOTE

Gradient Boosting Classifier

```
GBC_test_accu  0.79        GBC_train_accu  0.99
GBC_test_precision  0.62   GBC_train_precision  0.98
GBC_test_recall  0.69      GBC_train_recall  0.99
GBC_test_f1  0.65          GBC_train_f1  0.99
```

Bagging

```
Bagging_train_accu  0.88          Bagging_test_accu  0.82
Bagging_train_precision  0.79     Bagging_test_precision  0.65
Bagging_train_recall  0.86        Bagging_test_recall  0.78
Bagging_train_f1  0.82            Bagging_test_f1  0.71
```

Bagging using RandomForest

```
BaggingClassifier(base_estimator=RandomForestClassifier(class_weight={0: 4,
                                                                       1: 1.5},
                                                        min_samples_leaf=2,
                                                        min_samples_split=4),
                  n_estimators=50, random_state=1)
```

```
Bagging_train_accu  0.88
Bagging_train_precision  0.79
Bagging_train_recall  0.86
Bagging_train_f1  0.82
```

```
Bagging_test_accu  0.82
Bagging_test_precision  0.65
Bagging_test_recall  0.78
Bagging_test_f1  0.71
```

Boost

```
XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
              colsample_bynode=1, colsample_bytree=1, gamma=0, gpu_id=-1,
              importance_type='gain', interaction_constraints='',
              learning_rate=0.01, max_delta_step=0, max_depth=5,
              min_child_weight=3, missing=nan, monotone_constraints='()',
              n_estimators=1000, n_jobs=8, num_parallel_tree=1, random_state=0,
              reg_alpha=0, reg_lambda=1, scale_pos_weight=1, subsample=1,
              tree_method='exact', validate_parameters=1, verbosity=None)
```

```
XGB_test_accu  0.81
XGB_test_precision  0.67
XGB_test_recall  0.67
XGB_test_f1  0.67
```

```
XGB_train_accu  0.89
XGB_train_precision  0.86
XGB_train_recall  0.77
XGB_train_f1  0.81
```

Hyperparameter tuning using GridsearchCV

Logistic Regression with GridSearchCV

```
Best Parametres from LogisticRegression(C=6.158482110660261, class_weight={0: 2, 1: 1},
                   penalty='none', solver='sag')
```

```
logit_train_accu  0.84
logit_train_precision  0.7
logit_train_recall  0.84
logit_train_f1  0.76
```

```
logit_test_accu  0.82
logit_test_precision  0.68
logit_test_recall  0.7
logit_test_f1  0.69
```

Linear Discriminant Analysis with GridsearchCV

```
Best Parameters from LDA  {'solver': 'svd', 'tol': 0.0001}
```

```
LDA_train_accu  0.85
LDA_train_precision  0.78
LDA_train_recall  0.71
LDA_train_f1  0.74
```

```
LDA_test_accu  0.81
LDA_test_precision  0.66
LDA_test_recall  0.68
LDA_test_f1  0.67
```

KNN Model with GridsearchCV

```
Best Parameters from KNN Model {'metric': 'minkowski', 'n_neighbors': 11, 'weights': 'uniform'}
```

```
KNN_train_accu  0.85
KNN_train_precision  0.79
KNN_train_recall  0.71
KNN_train_f1  0.75

KNN_test_accu  0.81
KNN_test_precision  0.69
KNN_test_recall  0.6
KNN_test_f1  0.64
```

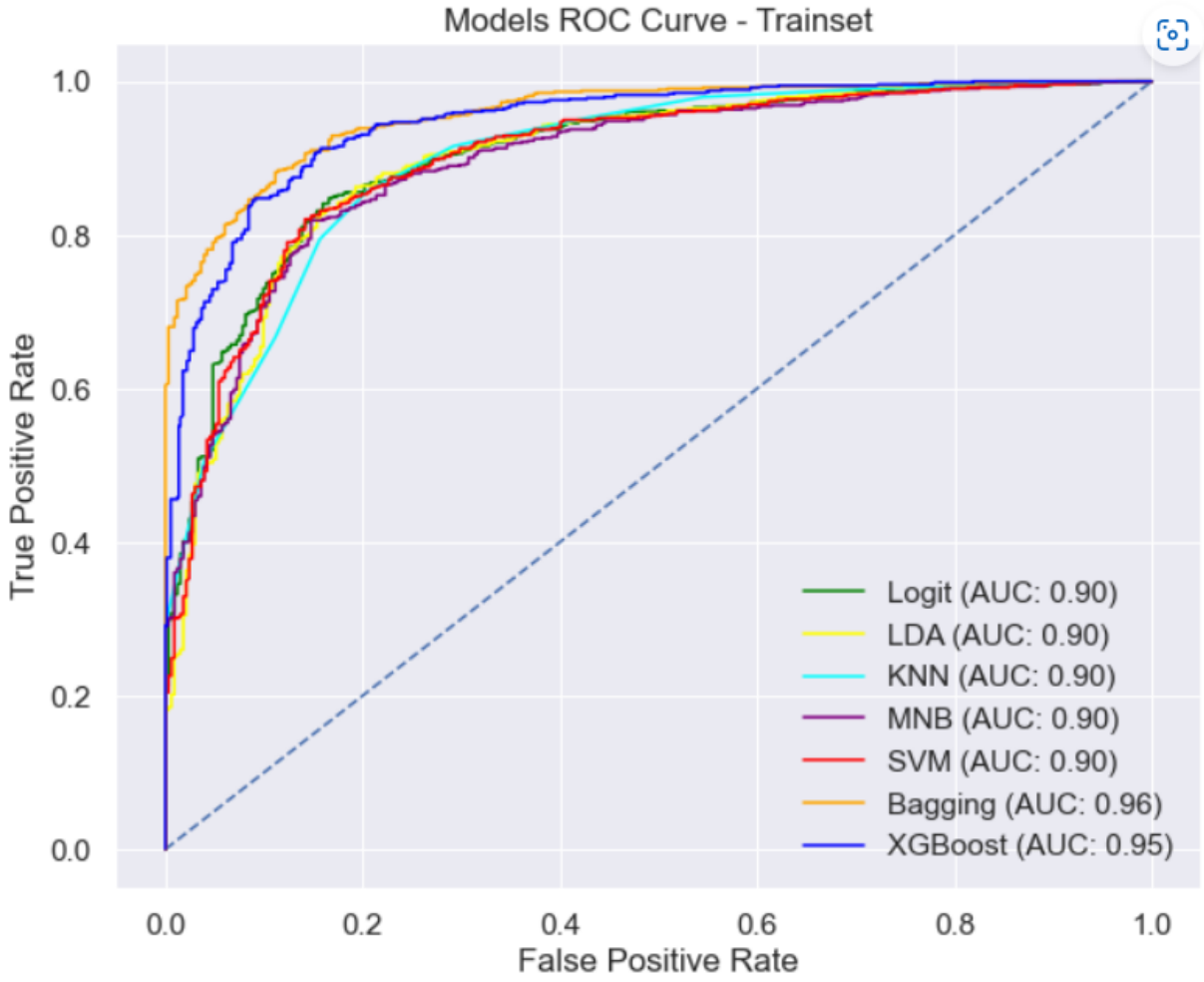
Support Vector Machine with GridsearchCV

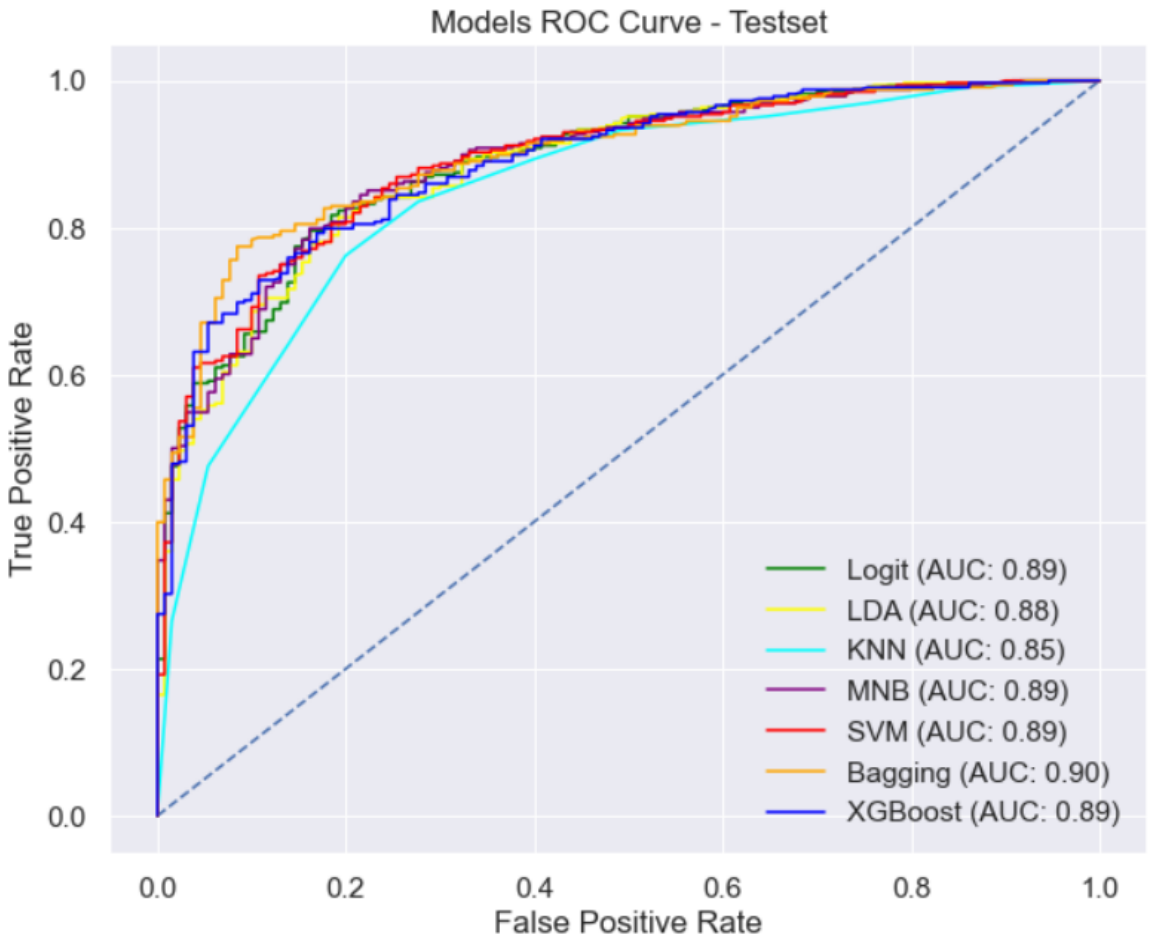
```
Best Parameters from SVM Model {'C': 0.1, 'kernel': 'linear'}

SVM_train_accu  0.83
SVM_train_precision  0.68
SVM_train_recall  0.86
SVM_train_f1  0.76

SVM_test_accu  0.8
SVM_test_precision  0.61
SVM_test_recall  0.82
SVM_test_f1  0.7
```

2.2) **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, classification report (4 pts) Final Model - Compare and comment on all models on the basis of the performance metrics in a structured tabular manner. Describe on which model is best/optimized, After comparison which model suits the best for the problem in hand on the basis of different measures. Comment on the final model.(3 pts)**

	Logit Train	LDA Train	KNN Train	MNB Train	SVM Train	Bagging Train	XGB Train
Accuracy	0.84	0.85	0.85	0.84	0.83	0.88	0.89
AUC	0.90	0.90	0.90	0.90	0.90	0.96	0.95
Recall-0	0.84	0.71	0.71	0.73	0.86	0.86	0.77
Recall-1	0.84	0.91	0.92	0.88	0.81	0.90	0.94
Precision-0	0.70	0.78	0.79	0.74	0.68	0.79	0.86
Precision-1	0.92	0.87	0.87	0.88	0.93	0.93	0.90
F1 Score-0	0.77	0.74	0.75	0.73	0.76	0.82	0.81
F1 Score-1	0.88	0.89	0.89	0.88	0.87	0.91	0.92

	Logit Test	LDA Test	KNN Test	MNB Test	SVM Test	Bagging Test	XGB Test
Accuracy	0.81	0.81	0.81	0.83	0.80	0.82	0.81
AUC	0.89	0.88	0.85	0.89	0.89	0.90	0.89
Recall-0	0.82	0.68	0.60	0.74	0.82	0.78	0.67
Recall-1	0.80	0.86	0.89	0.86	0.79	0.84	0.87
Precision-0	0.62	0.66	0.69	0.68	0.61	0.65	0.67
Precision-1	0.92	0.87	0.85	0.89	0.92	0.90	0.87
F1 Score-0	0.71	0.67	0.64	0.71	0.70	0.71	0.67
F1 Score-1	0.86	0.87	0.87	0.88	0.85	0.87	0.87


Models ROC Curve - Trainset
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logit (AUC: 0.90)
LDA (AUC: 0.90)
KNN (AUC: 0.90)
MNB (AUC: 0.90)
SVM (AUC: 0.90)
Bagging (AUC: 0.96)
XGBoost (AUC: 0.95)



- There is no under-fitting or over-fitting in any of the tuned models.
- All the tuned models have high values and every model is good. But as we can see, the most consistent tuned model in both train and test data is the Gradient Boost model.
- The tuned gradient boost model performs the best with 88.31% accuracy score in train and 87.28% accuracy score in test. Also it has the best AUC score of 94% in both train and test data which is the highest of all the models.

It also has a precision score of 88% and recall of 94% which is also the highest of all the models. So, we conclude that Gradient Boost Tuned model is the best/optimized model.

2.3) **Based on your analysis and working on the business problem, detail out appropriate insights and recommendations to help the management solve the business objective. There should be at least 3-4 Recommendations and insights in total. Recommendations should be easily understandable and business specific, students should not give any technical suggestions. Full marks should only be allotted if the recommendations are correct and business specific.**

- Labour party has more than double the votes of conservative party.
- Most number of people have given a score of 3 and 4 for the national economic condition and the average score is 3.245221

- Most number of people have given a score of 3 and 4 for the household economic condition and the average score is 3.137772
- Blair has higher number of votes than Hague and the scores are much better for Blair than for Hague.
- The average score of Blair is 3.335531 and the average score of Hague is 2.749506. So, here we can see that, Blair has a better score.
- On a scale of 0 to 3, about 30% of the total population has zero knowledge about politics/parties.
- People who gave a low score of 1 to a certain party, still decided to vote for the same party instead of voting for the other party. This can be because of lack of political knowledge among the people.
- People who have higher Eurosceptic sentiment, has voted for the conservative party and lower the Eurosceptic sentiment, higher the votes for Labour party.
- Out of 454 people who gave a score of 0 for political knowledge, 360 people have voted for the labour party and 94 people have voted for the conservative party.
- All models performed well on training data set as well as test dat set. The tuned models have performed better than the regular models.
- There is no over-fitting in any model except Random Forest and Bagging regular models.
- Gradient Boosting model tuned is the best/optimized model.

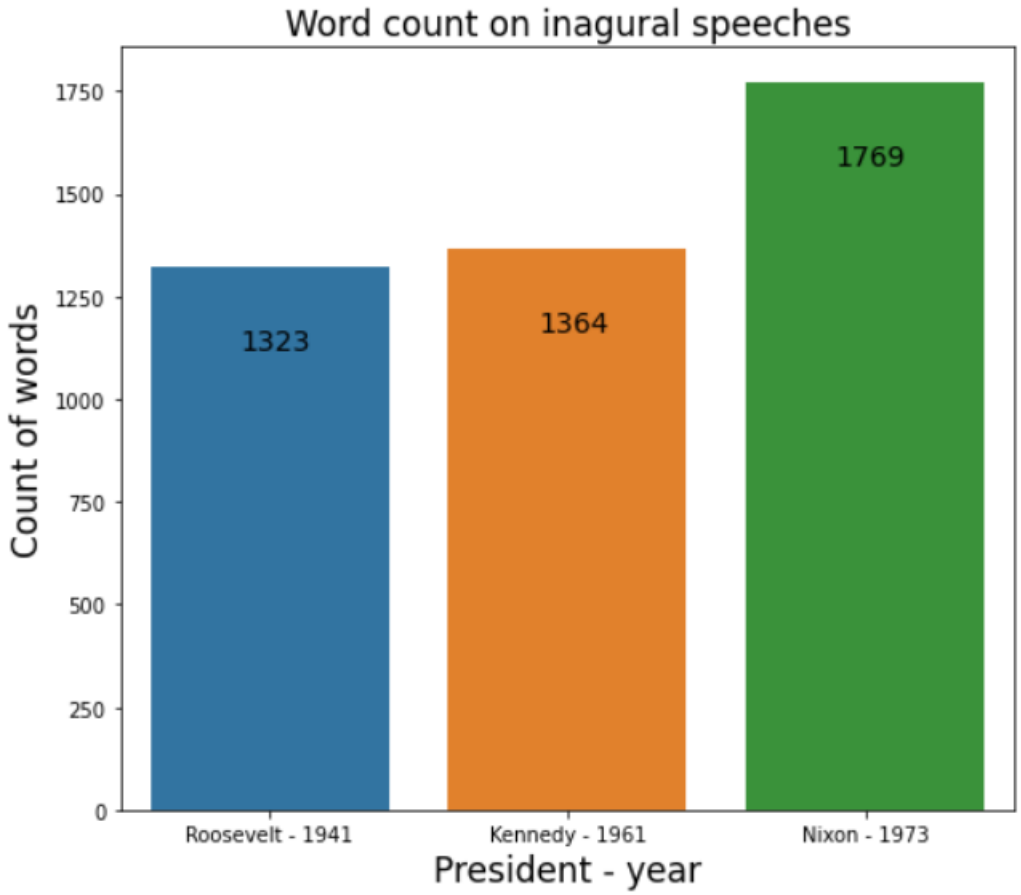
Problem 2:

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

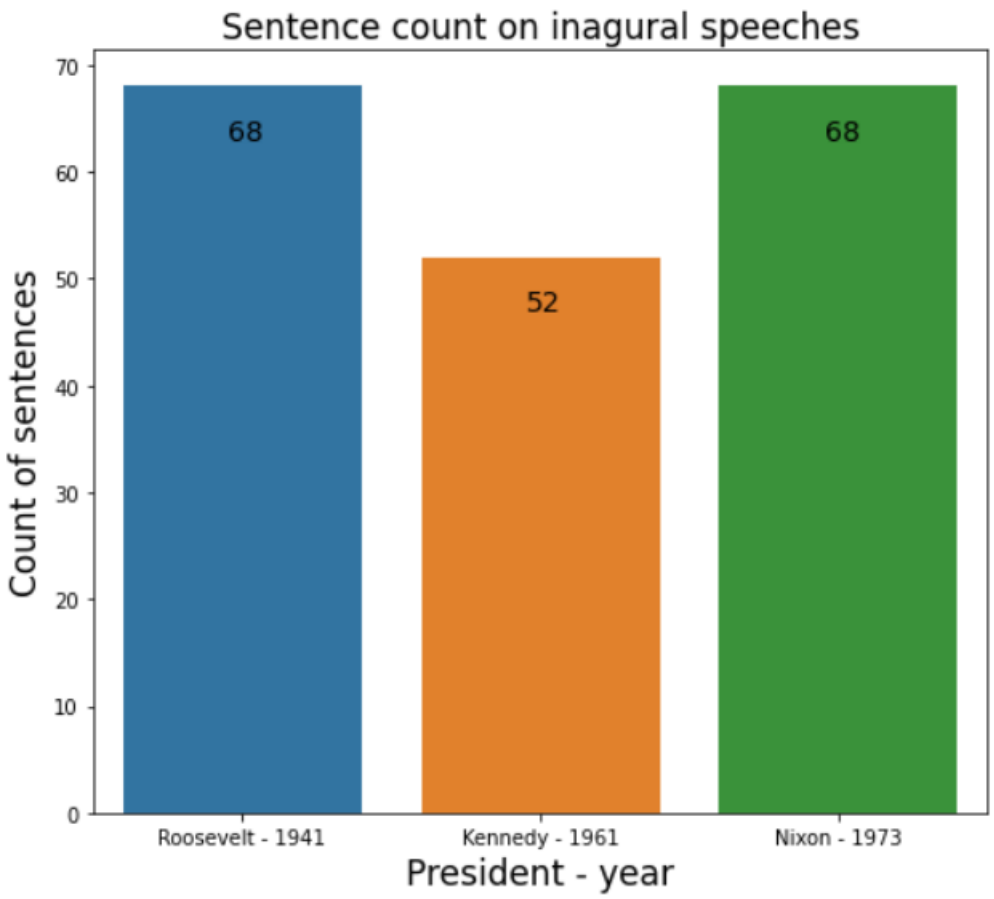
2.1) Find the number of characters, words and sentences for the mentioned documents. (Hint: use .words(), .raw(), .sent() for extracting counts)

	president	text	word_count
1941-Roosevelt	Roosevelt - 1941	On each national day of inauguration since 178...	1323
1961-Kennedy	Kennedy - 1961	Vice President Johnson, Mr. Speaker, Mr. Chief...	1364
1973-Nixon	Nixon - 1973	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	1769



- Number of word count of president Roosevelt-1941 is 1329
- Number of word count of president Kennedy-1961 is 1364
- Number of word count of president Nixon-1973 is 1769

	president	text	word_count	char_count	sents_count
1941-Roosevelt	Roosevelt - 1941	On each national day of inauguration since 178...	1323	7571	68
1961-Kennedy	Kennedy - 1961	Vice President Johnson, Mr. Speaker, Mr. Chief...	1364	7618	52
1973-Nixon	Nixon - 1973	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	1769	9991	68



- Number of word count of president Roosevelt-1941 is 68
- Number of word count of president Kennedy-1961 is 52
- Number of word count of president Nixon-1973 is 68

2.2) Remove all the stopwords **from the three speeches. Show the word count before and after the removal of stopwords. Show a sample sentence after the removal of stopwords.**

After removing all the stop words

	president	text	word_count	char_count	sents_count
1941-Roosevelt	Roosevelt - 1941	national day inauguration since 1789 people re...	1323	7571	68
1961-Kennedy	Kennedy - 1961	vice president johnson speaker chief justice p...	1364	7618	52
1973-Nixon	Nixon - 1973	vice president speaker chief justice senator c...	1769	9991	68

Speech of president Roosevelt without stopwords

['national day inauguration since 1789 people renewed sense dedication united states washingtons day task people create weld together nation lincolns day task people preserve nation disruption within day task people save nation institutions disruption without come time midst swift happenings pause moment take stock recall place history rediscover may risk real peril inaction lives nations determined count years lifetime human spirit life man threescore years ten little little less life nation fullness measure live men doubt men believe democracy form government frame life limited measured kind mystical artificial fate unexplained reason tyranny slavery become surging wave future freedom ebbing tide americans know true eight years ago life republic seemed frozen fatalistic terror proved true midst shock acted acted quickly boldly decisively later years living years fruitful years people democracy brought greater security hope better understanding lifes ideals measured material things vital present future experience democracy successfully survived crisis home put away many evil things built new structures enduring lines maintained fact democracy action taken within threeway framework constitution united states coordinate branches government continue freely function bill rights remains inviolate freedom elections wholly maintained prophets downfall american democracy seen dire predictions come naught democracy dying know seen reviveand grow know cannot die built unhampered initiative individual men women joined together common enterprise enterprise undertaken carried free expression free majority know democracy alone forms government enlists full force mens enlightened know democracy alone constructed unlimited civilization capable infinite progress improvement human life know look surface sense still spreading every continent humane advanced end unconquerable forms human society nation like person bodya body must fed clothed housed invigorated rested manner measures objectives time nation like person mind mind must kept informed alert must know understands hopes needs neighbors nations live within narrowing circle world nation like person something deeper something permanent something larger sum parts something matters future calls forth sacred guarding present thing find difficult even impossible hit upon single simple word yet understand spirit faith america product centuries born multitudes came many lands high degree mostly plain people sought early late find freedom freely democratic aspiration mere recent phase human history human history permeated ancient life early peoples blazed anew middle ages written magna charta americas impact irresistible america new world tongues peoples continent newfound land came believed could create upon continent new life life new freedom vitality written mayflower compact declaration independence constitution united states gettysburg address first came carry longings spirit millions followed stock sprang moved forward constantly consistently toward ideal gained stature clarity generation hopes republic cannot forever tolerate either undeserved poverty selfserving wealth know still far go must greatly build security opportunity knowledge every citizen measure justified resources capacity land enough achieve purposes alone enough clothe feed body nation instruct inform mind also spirit three greatest spirit without body mind men know nation could live spirit america killed even though nations body mind constricted alien world lived america know would perished spirit faith speaks daily lives ways often unnoticed seem obvious speaks capital nation speaks processes governing sovereignties 48 states speaks counties cities towns villages speaks nations hemisphere across seas enslaved well free sometimes fail hear heed voices freedom privilege freedom old old story destiny america proclaimed words prophecy spoken first president first inaugural 1789 words almost directed would seem year 1941 preservation sacred fire liberty destiny republican model government justly considered deeply finally staked experiment intrusted hands american people lose sacred fireif smothered doubt fear reject destiny washington strove valiantly triumphantly establish preservation spirit faith nation furnish highest justification every sacrifice may make cause national defense face great perils never encountered strong purpose protect perpetuate integrity democracy muster spirit america faith america retreat content stand still americans go forward service country god']

Speech of president Kennedy without stopwords

['vice president johnson speaker chief justice president eisenhower vice president nixon president truman reverend clergy fellow citizens observe today victory party celebration freedom symbolizing end well beginning signifying renewal well change sworn almighty god solemn oath forebears l prescribed nearly century three quarters ago world different man holds mortal hands power abolish forms human poverty forms human life yet revolutionary beliefs forebears fought still issue around globe belief rights man come generosity state hand god dare forget today heirs first revolution word go forth time pl ace friend foe alike torch passed new generation americans born century tempered war disciplined hard bitter peace proud ancient heritage unwilling witness permit slow undoing human rights nation always committed committed today home around world every nation know whether wishes well ill pay price bear burden meet hardship support friend oppose foe order assure survival success liberty much pledge old allies whose cultural spiritual origins share pledge loyalty faithful friends united little cannot host cooperative ventures divided little dare meet powerful challenge odds split asunder new states welcome ranks free pledge word one form colonial control passed away merely replaced far iron tyranny always expect find supporting view always hope find strongly supporting freedom remember past foolishly sought power riding back tiger ended inside peoples huts villages across globe struggling break bonds mass misery pledge best efforts help help whatever period required communists may seek votes right free society cannot help many poor cannot save rich sister republics south border offer special pledge convert good words good deeds new alliance progress assist free men free governments casting chains poverty peaceful revolution hope cannot become prey hostile powers neighbors know join oppose aggression subversion anywhere americas every power know hemisphere intends remain master house world assembly sovereign states united nations last best hope age instruments war far outpaced instruments peace renew pledge supportto prevent becoming merely forum invective strengthen shield new weak enlarge area writ may run finally nations would make adversary offer pledge request sides begin anew quest peace dark powers destruction unleashed science engulf humanity planned accidental selfdestruction dare tempt weakness arms sufficient beyond doubt certain beyond doubt never employed neither two great powerful groups nations take comfort present course sides overburdened cost modern weapons rightly alarmed steady spread deadly atom yet racing alter uncertain balance terror stays hand mankinds final war begin anew remembering sides civility sign weakness sincerity always subject proof never negotiate fear never fear negotiate sides explore problems unite instead belaboring problems divide sides first time formulate serious precise proposals inspection control arms bring absolute power destroy nations absolute control nations sides seek invoke wonders science instead terrors together explore stars conquer deserts eradicate disease tap ocean depths encourage arts commerce sides unite heed corners earth command isaiah undo heavy burdens oppressed go free beachhead cooperation may push back jungle suspicion sides join creating new endeavor new balance power new world law strong weak secure peace preserved finished first 100 days finished first 1000 days life administration even perhaps lifetime planet begin hands fellow citizens mine rest final success failure course since country founded generation americans summoned give testimony national loyalty graves young americans answered call service surround globe trumpet summons call bear arms though arms need call battle though embattled call bear burden long twilight struggle year year rejoicing hope patient tribulation struggle common enemies man tyranny poverty disease war forge enemies grand global alliance north south east west assure fruitful life mankind join historic effort long history world generations granted role defending freedom hour maximum danger shrink responsibility welcome believe would exchange places people generation energy faith devotion bring endeavor light country serve glow fire truly light world fellow americans ask country ask country fellow citizens world ask america together freedom man finally whether citizens america citizens world ask high standards strength sacrifice ask good conscience sure reward history final judge deeds go forth lead land love asking blessing help knowing earth gods work must truly']

```
Speech of president Nixon without stopwords
['vice president speaker chief justice senator cook mrs eisenhower fellow citizens great good country share together met four years ago america bleak s
pirit depressed prospect seemingly endless war abroad destructive conflict home meet today stand threshold new era peace world central question use pea
ce resolve era enter postwar periods often time retreat isolation leads stagnation home invites new danger abroad resolve become time great responsibil
ities greatly borne renew spirit promise america enter third century nation past year saw farreaching results new policies peace continuing revitalize
traditional friendships missions peking moscow able establish base new durable pattern relationships among nations world americas bold initiatives 1972
long remembered year greatest progress since end world war ii toward lasting peace world peace seek world flimsy peace merely interlude wars peace endu
re generations come important understand necessity limitations americas role maintaining peace unless america work preserve peace peace unless america
work preserve freedom freedom clearly understand new nature americas role result new policies adopted past four years respect treaty commitments suppor
t vigorously principle country right impose rule another force continue era negotiation work limitation nuclear arms reduce danger confrontation great
powers share defending peace freedom world expect others share time passed america make every nations conflict make every nations future responsibility
presume tell people nations manage affairs respect right nation determine future also recognize responsibility nation secure future americas role indis
pensable preserving worlds peace nations role indispensable preserving peace together rest world resolve move forward beginnings made continue bring wa
lls hostility divided world long build place bridges understanding despite profound differences systems government people world friends build structure
peace world weak safe strong respects right live different system would influence others strength ideas force arms accept high responsibility burden gl
adly gladly chance build peace noblest endeavor nation engage gladly also act greatly meeting responsibilities abroad remain great nation remain great
nation act greatly meeting challenges home chance today ever history make life better america ensure better education better health better housing bett
er transportation cleaner environment restore respect law make communities livable insure godgiven right every american full equal opportunity range ne
eds great reach opportunities great bold determination meet needs new ways building structure peace abroad required turning away old policies failed bu
ilding new era progress home requires turning away old policies failed abroad shift old policies new retreat responsibilities better way peace home shi
ft old policies new retreat responsibilities better way progress abroad home key new responsibilities lies placing division responsibility lived long c
onsequences attempting gather power responsibility washington abroad home time come turn away condescending policies paternalism washington knows best
person expected act responsibly responsibility human nature encourage individuals home nations abroad decide locate responsibility places measure other
s today offer promise purely governmental solution every problem lived long false promise trusting much government asked deliver leads inflated expecta
tions reduced individual effort disappointment frustration erode confidence government people government must learn take less people people remember am
erica built government people welfare work shirking responsibility seeking responsibility lives ask government challenges face together ask government
help help national government great vital role play pledge government act act boldly lead boldly important role every one must play individual member c
ommunity day forward make solemn commitment heart bear responsibility part live ideals together see dawn new age progress america together celebrate 20
0th anniversary nation proud fulfillment promise world americas longest difficult war comes end learn debate differences civility decency reach one pre
cious quality government cannot provide new level respect rights feelings one another new level respect individual human dignity cherished birthright e
very american else time come renew faith america recent years faith challenged children taught ashamed country ashamed parents ashamed americas record
home role world every turn beset find everything wrong america little right confident judgment history remarkable times privileged live americas record
century unparalleled worlds history responsibility generosity creativity progress proud system produced provided freedom abundance widely shared system
history world proud four wars engaged century including one bringing end fought selfish advantage help others resist aggression proud bold new initiati
ves steadfastness peace honor made breakthrough toward creating world world known structure peace last merely time generations come embarking today era
presents challenges great nation generation ever faced answer god history conscience way use years stand place hallowed history think others stood thin
k dreams america think recognized needed help far beyond order make dreams come true today ask prayers years ahead may gods help making decisions right
america pray help together may worthy challenge pledge together make next four years best four years americas history 200th birthday america young vita
l began bright beacon hope world go forward confident hope strong faith one another sustained faith god created striving always serve purpose']
```

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)

word which occurs the most number of times in his inaugural address for each president is

inaugural speech for Roosevelt

```
nation      11
know        10
spirit       9
democracy    9
life         8
dtype: int64
```

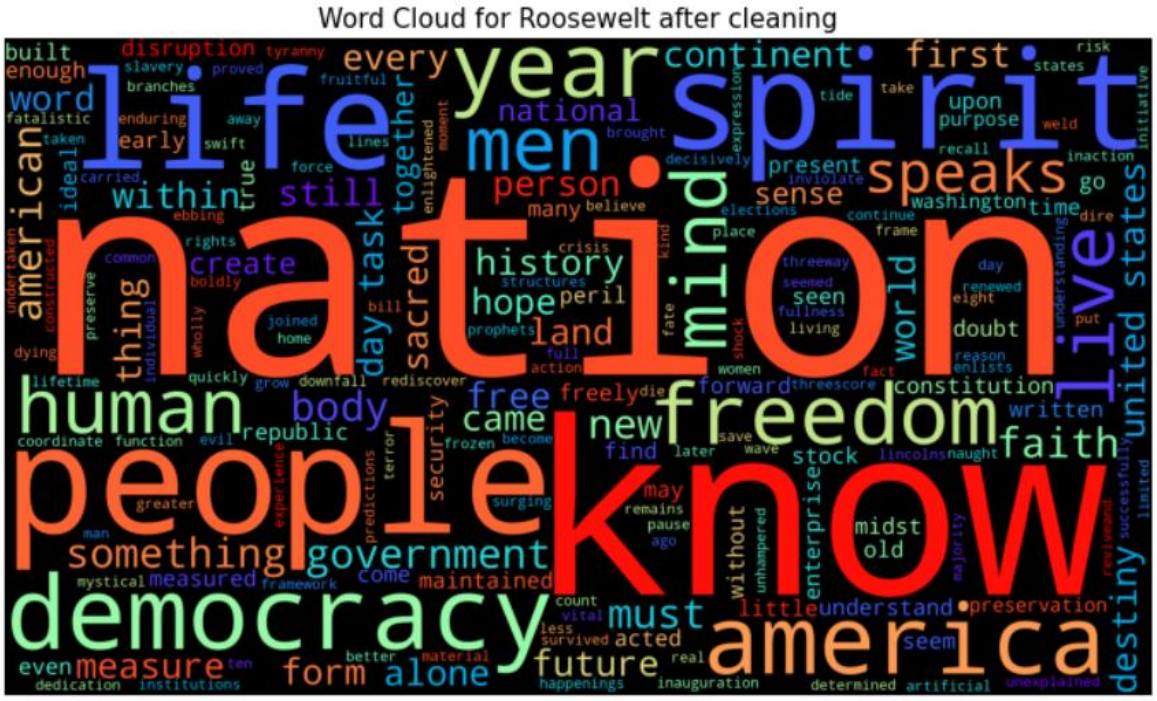
inaugural speech for Kennedy

```
world     8
sides     8
new       7
pledge    7
power     5
dtype: int64
```

inaugural speech for Nixon

```
peace          19
world          16
new            15
america        13
responsibility 11
dtype: int64
```

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



Word Cloud for Kennedy after cleaning

Word Cloud for Nixon after cleaning