1.1) Read the dataset. Do the descriptive statistics and do null value condition check.

Below is the data structure for the given election data set with vote as the dependent target variable with the rest of them as independent variables.

```
#   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
```

Below is the description of the dataset across categorical and continuous variables.

	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
count	1525	1525.000000	1525.000000	1525.000000	1525.000000	1525.000000	1525.000000	1525.000000	1525
unique	2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2
top	Labour	NaN	NaN	NaN	NaN	NaN	NaN	NaN	female
freq	1063	NaN	NaN	NaN	NaN	NaN	NaN	NaN	812
mean	NaN	54.182295	3.245902	3.140328	3.334426	2.746885	6.728525	1.542295	NaN
std	NaN	15.711209	0.880969	0.929951	1.174824	1.230703	3.297538	1.083315	NaN
min	NaN	24.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.000000	NaN
25%	NaN	41.000000	3.000000	3.000000	2.000000	2.000000	4.000000	0.000000	NaN
50%	NaN	53.000000	3.000000	3.000000	4.000000	2.000000	6.000000	2.000000	NaN
75%	NaN	67.000000	4.000000	4.000000	4.000000	4.000000	10.000000	2.000000	NaN
max	NaN	93.000000	5.000000	5.000000	5.000000	5.000000	11.000000	3.000000	NaN

It could be noticed that there isn't much outlier considering the spread on the either side of median for continuous variables and that the target variable vote has only 2 distinct values indicating only two classes from classification perspective with Labour holding higher frequency. Also it could be noticed that gender has 2 unique values with females holding a higher frequency of rows.

Below gives the exact distribution of categorical variables:

```
1. VOTE :  2
     Conservative     462
     Labour          1063
     Name: vote, dtype: int64


2. GENDER :  2
```

```
male      713
female    812
Name: gender, dtype: int64
```

Also there are no null values present in any of the variables in the dataset as listed below:

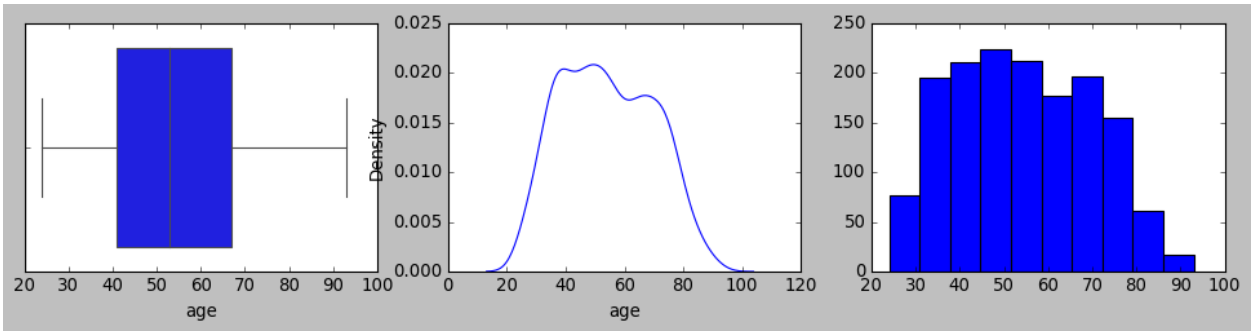
```
vote
age
economic.cond.national
economic.cond.household
Blair
Hague
Europe
political.knowledge
gender
```

Dataset contains 1525 rows and 9 columns.

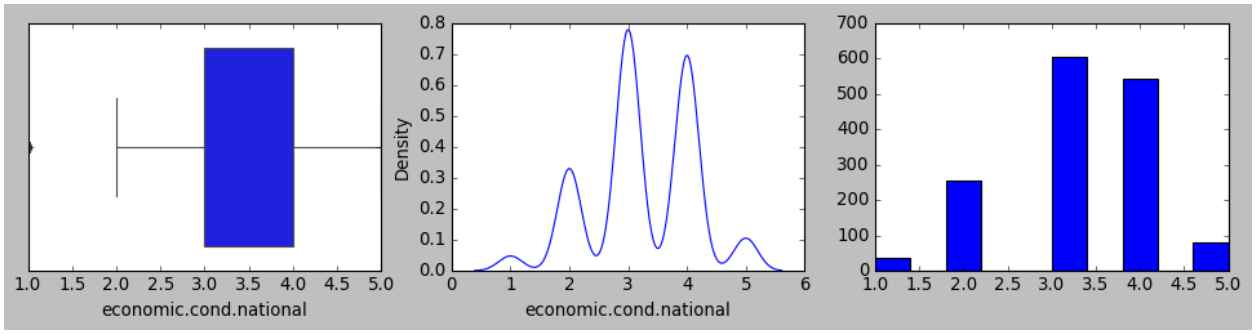
1.2. Perform EDA (Check the null values, Data types, shape, Univariate, bivariate analysis). Also check for outliers . Interpret the inferences for each.

Univariate Analysis for unscaled dataset

```
1. Univariate analysis for age

Mean is 54.182295, Median is 53.000000, Mode(s) are 37.0000
Column age does not have outliers
word will result in an error or misinterpretation.
```



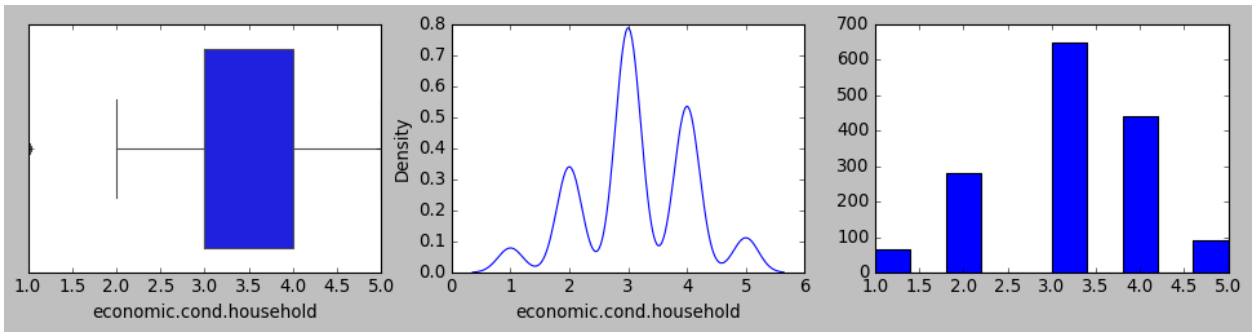

```
Column age is not normally distributed

 2. Univariate analysis for economic.cond.national

Mean is 3.245902, Median is 3.000000, Mode(s) are 3.0000
Column economic.cond.national has outliers
```



Column economic.cond.national is not normally distributed

3. Univariate analysis for economic.cond.household

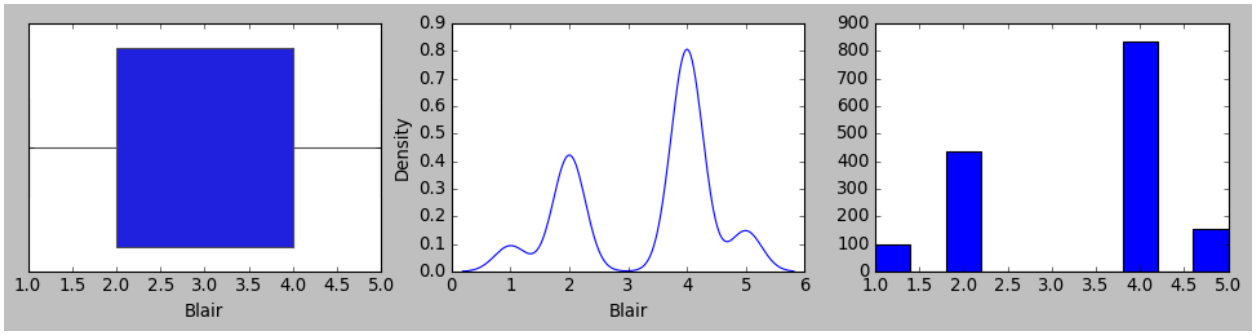
Mean is 3.140328, Median is 3.000000, Mode(s) are 3.0000
Column economic.cond.household has outliers



Column economic.cond.household is not normally distributed

4. Univariate analysis for Blair

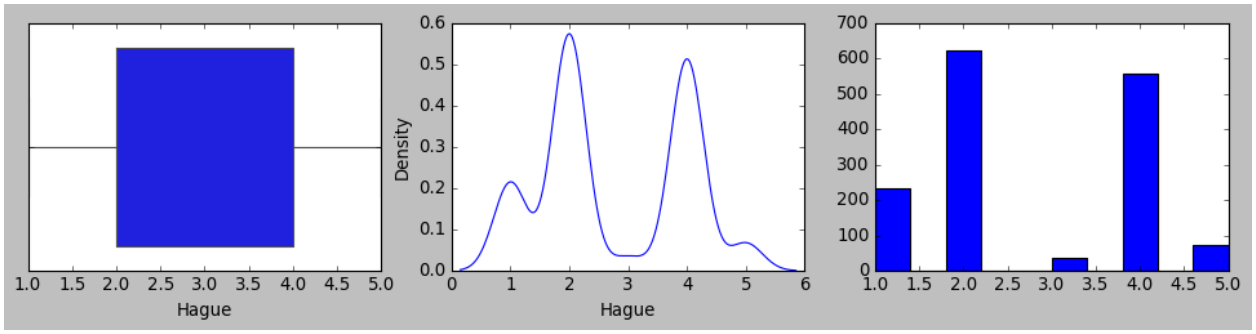
Mean is 3.334426, Median is 4.000000, Mode(s) are 4.0000
Column Blair does not have outliers



Column Blair is not normally distributed

5. Univariate analysis for Hague

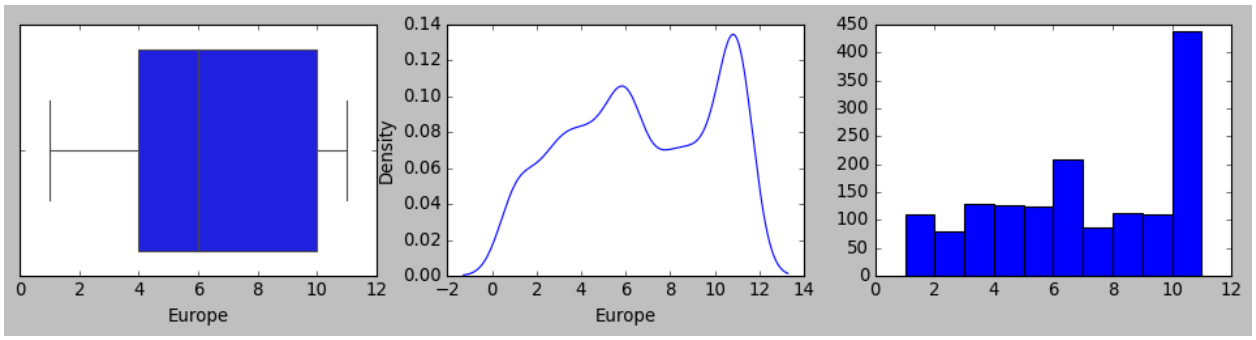
Mean is 2.746885, Median is 2.000000, Mode(s) are 2.0000
Column Hague does not have outliers



Column Hague is not normally distributed

6. Univariate analysis for Europe

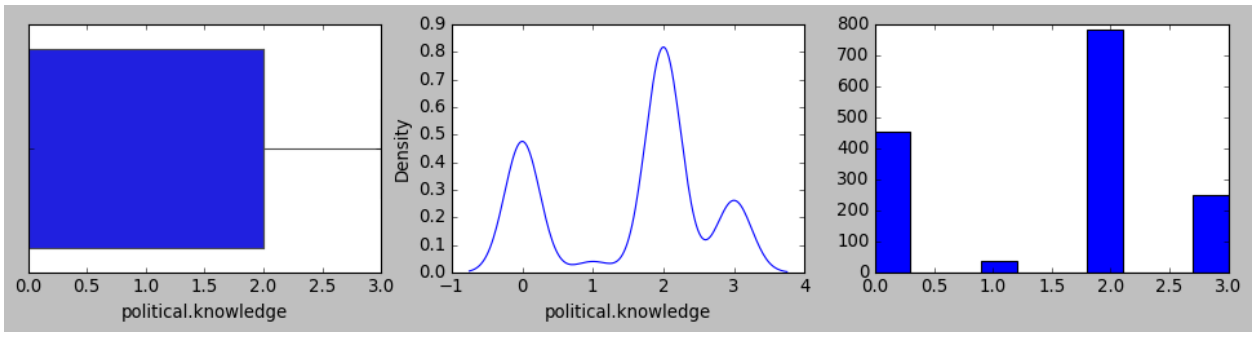
Mean is 6.728525, Median is 6.000000, Mode(s) are 11.0000
Column Europe does not have outliers



Column Europe is not normally distributed

7. Univariate analysis for political.knowledge

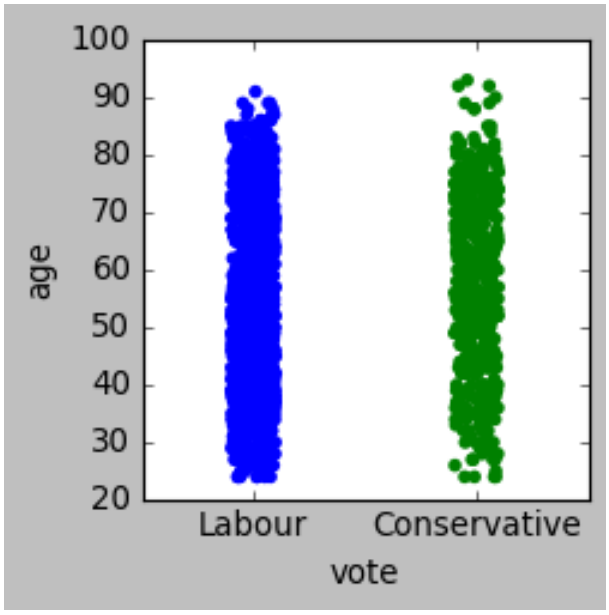
Mean is 1.542295, Median is 2.000000, Mode(s) are 2.0000
Column political.knowledge does not have outliers



Column political.knowledge is not normally distributed

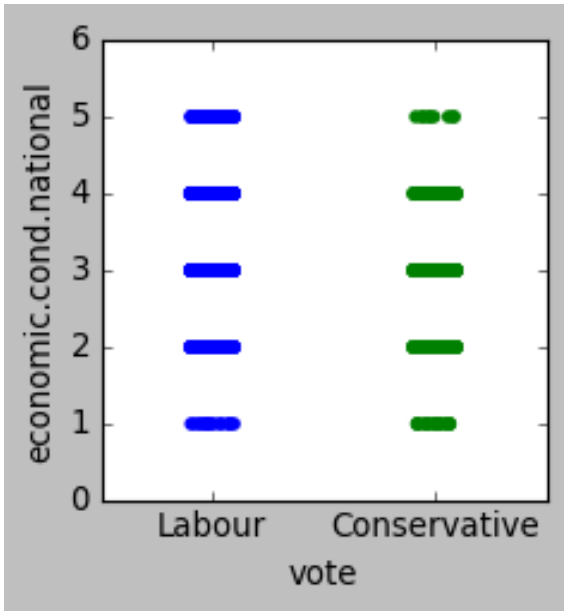
Bivariate analysis

1. bivariate analysis for age



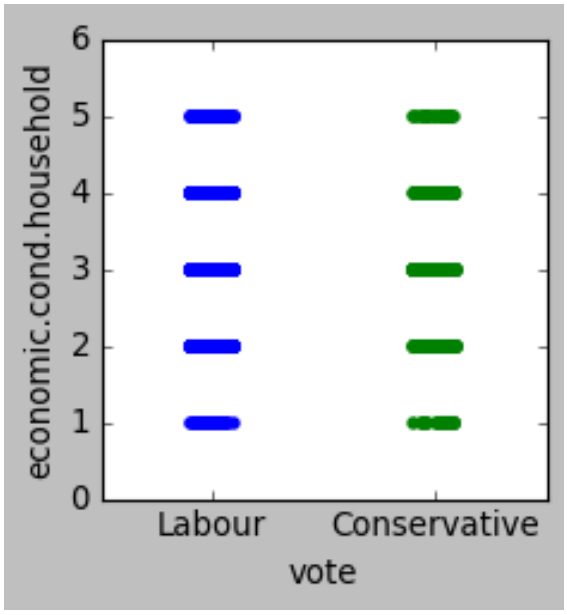
Data reveals that lower age group tend to vote for Labour while Labour also retains a better vote share for higher age group comparatively.

2. bivariate analysis for economic.cond.national



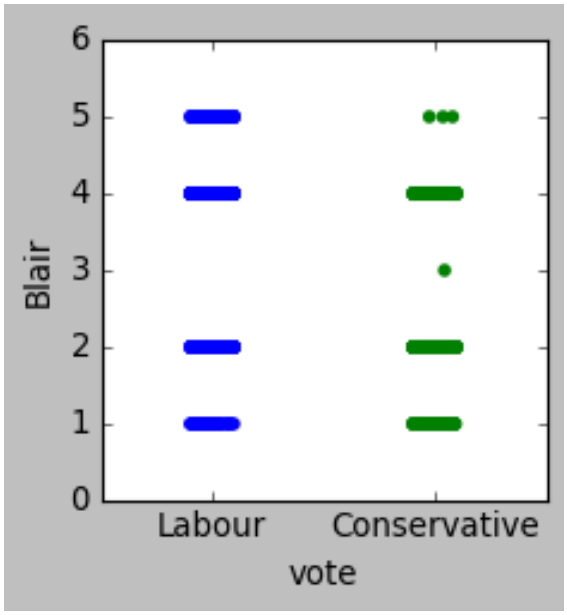
Trend suggests that voters among higher economic conditions at national level tend to vote Labour and comparatively lower economic levels in the same category also tend to vote Labour higher than Conservative.

3. bivariate analysis for economic.cond.household



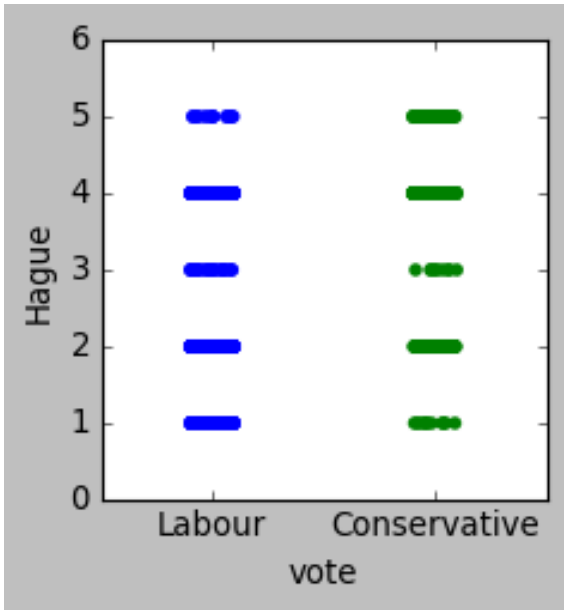
No noticeable difference across both voter groups across the levels of household economic condition.

4. bivariate analysis for Blair



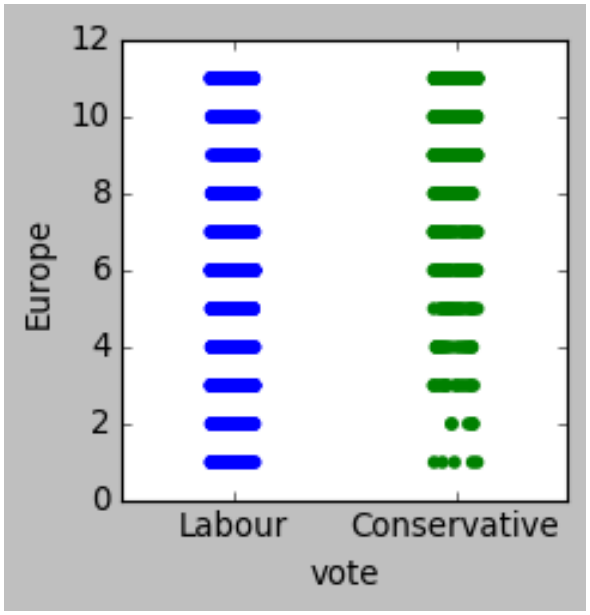
Trend suggests that Blair has received better higher scale assessment ratings among Labour voters compared to Conservative.

5. bivariate analysis for Hague



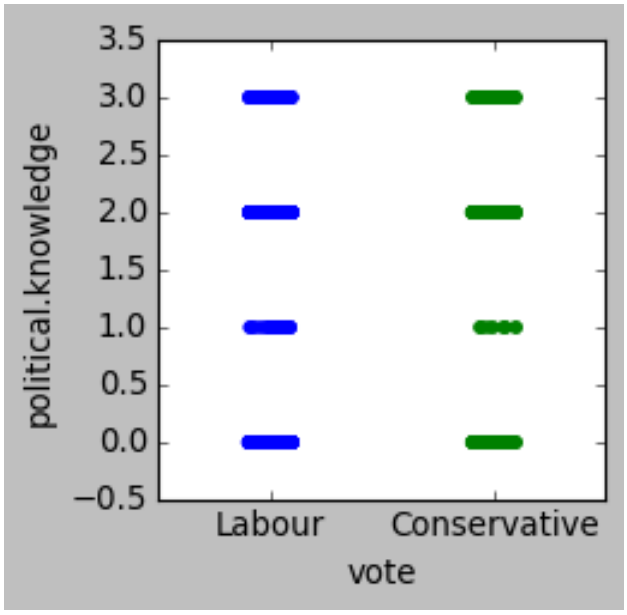
Trend suggests that Hague received higher scale ratings from Conservative voters while increased frequency in lower scale ratings from Labour

6. bivariate analysis for Europe



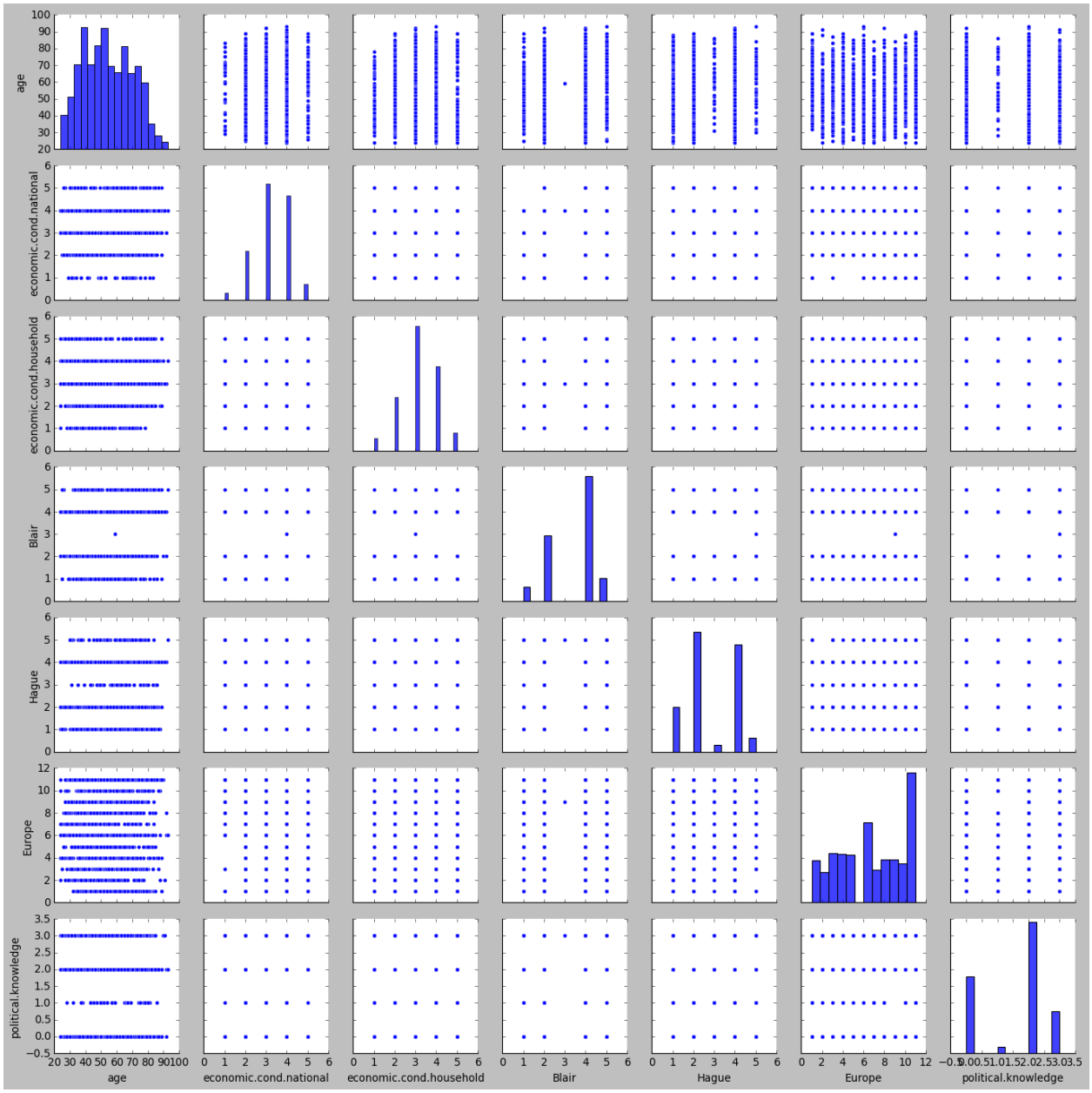
Trend suggests that conservative voters have comparatively lesser number of voters with low eurosceptic sentiments.

7. bivariate analysis for political.knowledge

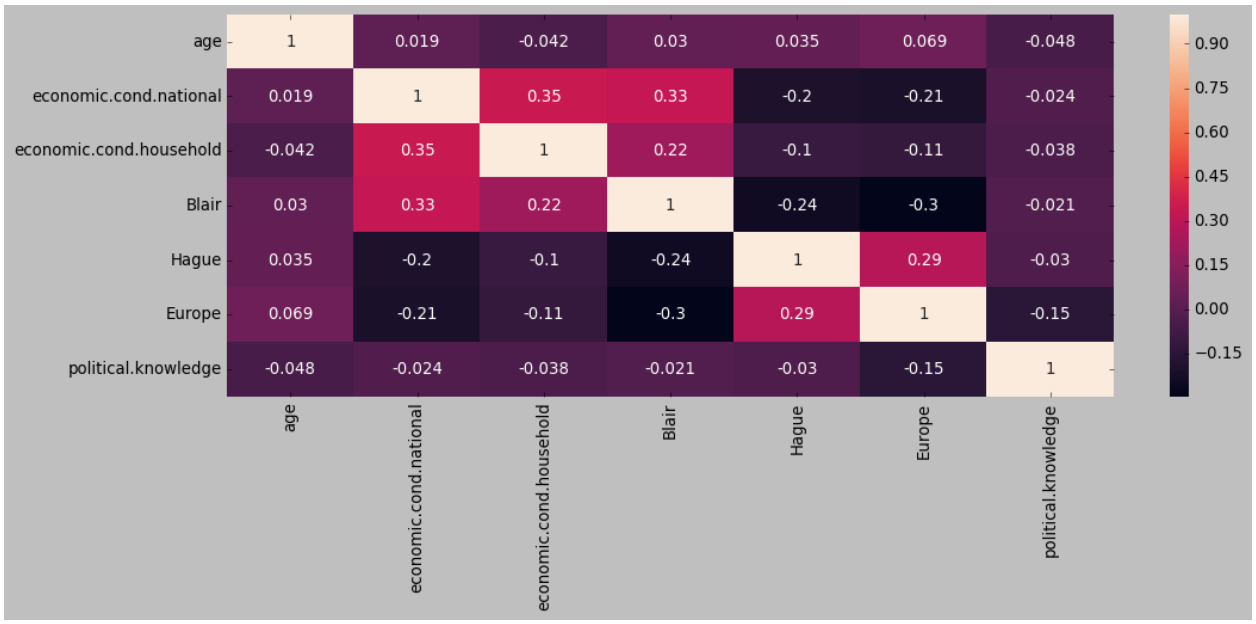


The Smaller number of voters for Conservative falls under lower scale knowledge regarding party position in european integrations.

Multivariate analysis:
Pairplot for unscaled dataset


age
economic.cond.national
economic.cond.household
Blair
Hague
Europe
political.knowledge

Heatmap for unscaled dataset:



It is noticed that at lower age group is are less in frequency within lower scale range for national level economic condition. Also folks with highest political knowledge tend to dip with the highest age group.

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) .

Scaling:

For logistic regression since it goes by linear equations scaling needs to be assessed.

Also since KNN goes by distance computation (Euclidean) between data points it is highly dependent on scaling needs.

For LDA, scaling standardizes the co efficients of independent variables which helps in clear separation of classes as comparison of coefficients happens on standardized data.

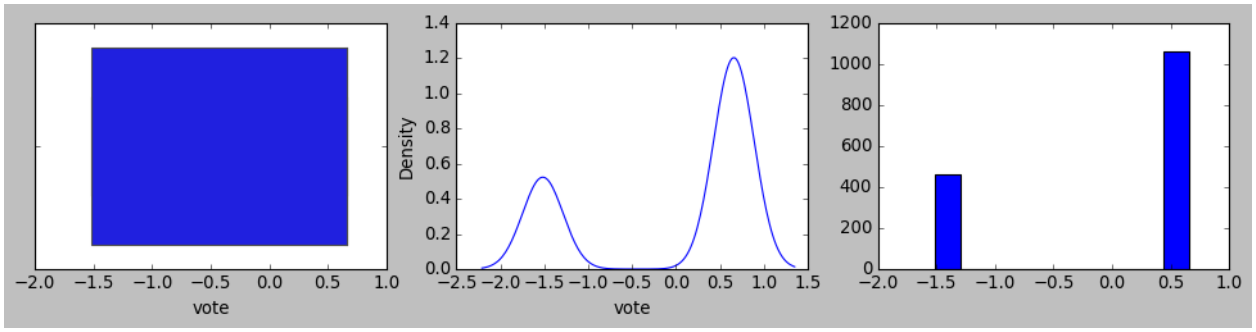
However Naïve bayes is unaffected by scaling. Going by the fact that independent variables have different units, scaling becomes necessary to remove the units the variables are associated with so that the linear equation can be formed on the independent variables post standardization of them.

Z scoring based scaling of data would change the coefficient, neutralize/remove the intercept while the accuracy score remains the same before and after. MSE would get scaled too.

Accordingly, below is the univariate analysis for the dataset that is standardized on z scores

Univariate analysis for scaled and outlier treated dataset:

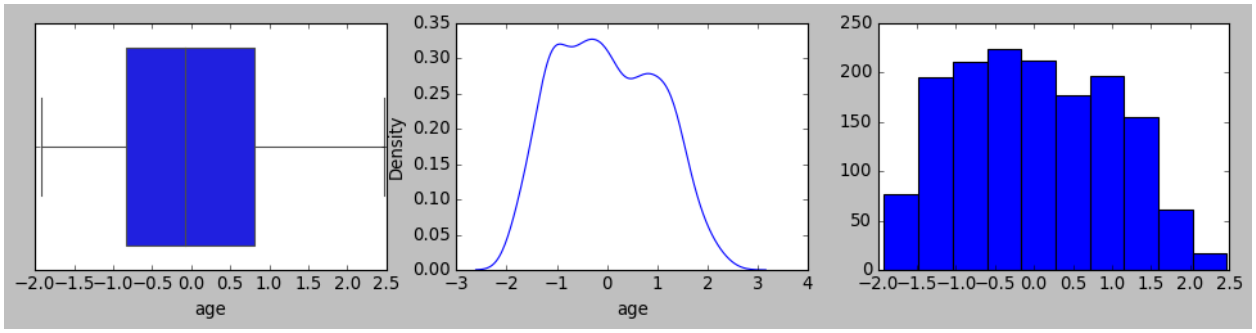
```
1. Univariate analysis for vote

Mean is -0.000000, Median is 0.659256, Mode(s) are 0.6593
Column vote does not have outliers
```



Column vote is not normally distributed

2. Univariate analysis for age

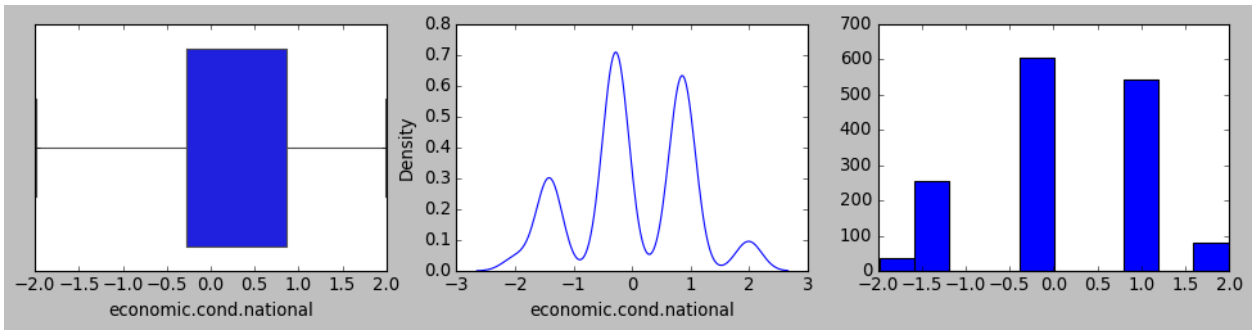
Mean is 0.000000, Median is -0.075276, Mode(s) are -1.0940
Column age does not have outliers



Column age is not normally distributed

3. Univariate analysis for economic.cond.national

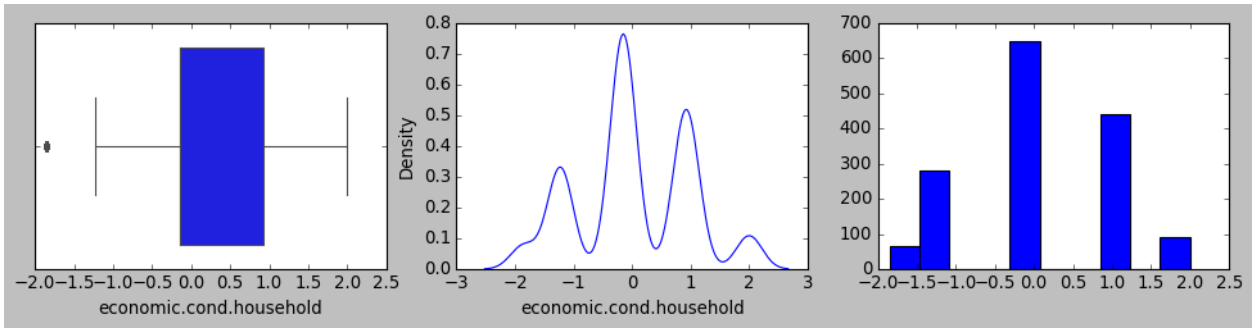
Mean is 0.013775, Median is -0.279218, Mode(s) are -0.2792
Column economic.cond.national does not have outliers



Column economic.cond.national is not normally distributed

4. Univariate analysis for economic.cond.household

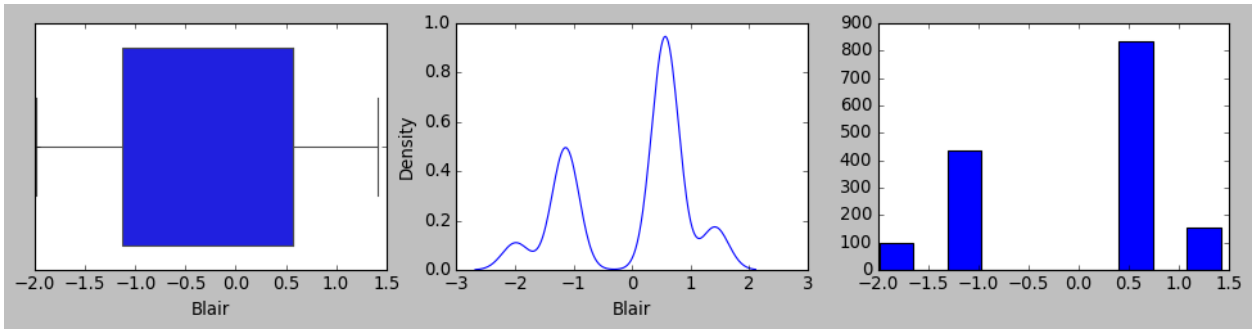
Mean is 0.019100, Median is -0.150948, Mode(s) are -0.1509
Column economic.cond.household has outliers



Column economic.cond.household is not normally distributed

5. Univariate analysis for Blair

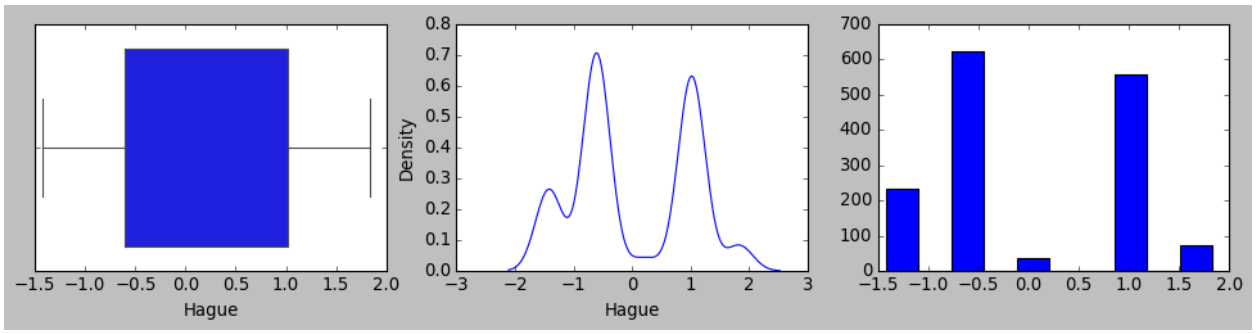
Mean is 0.000000, Median is 0.566716, Mode(s) are 0.5667
Column Blair does not have outliers



Column Blair is not normally distributed

6. Univariate analysis for Hague

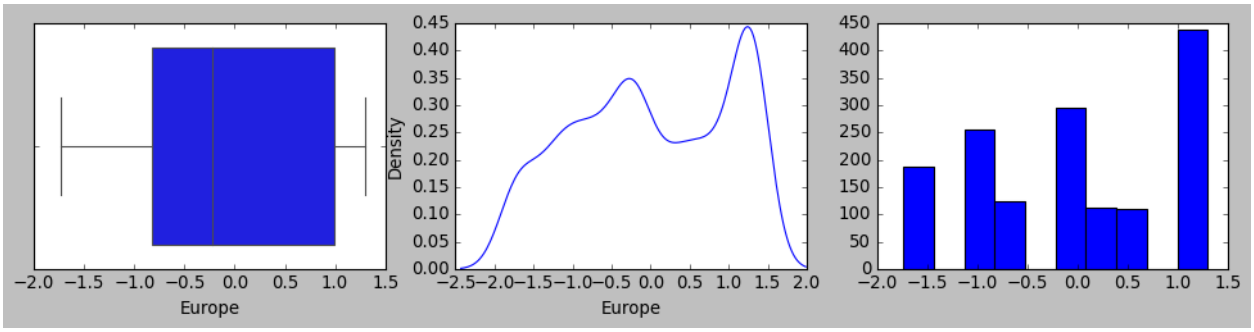
Mean is -0.000000, Median is -0.607076, Mode(s) are -0.6071
Column Hague does not have outliers



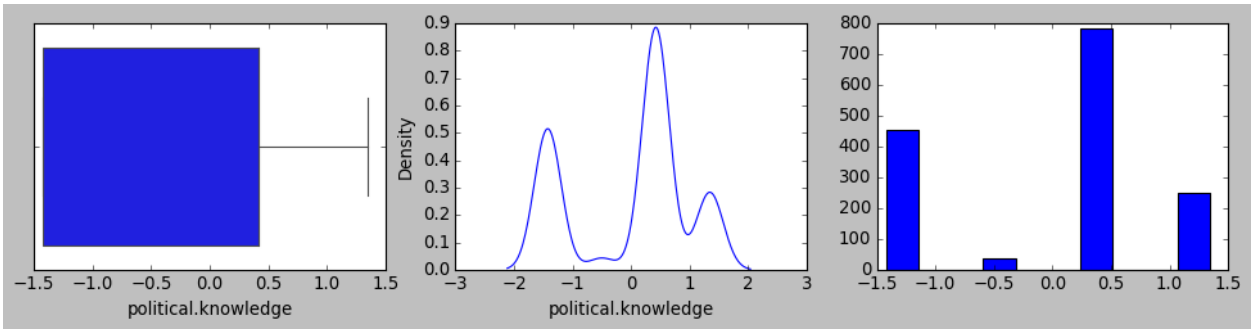
Column Hague is not normally distributed

7. Univariate analysis for Europe

Mean is -0.000000, Median is -0.221002, Mode(s) are 1.2958
Column Europe does not have outliers




```
Column Europe is not normally distributed

 8. Univariate analysis for political.knowledge

Mean is -0.000000, Median is 0.422643, Mode(s) are 0.4226
Column political.knowledge does not have outliers
```




```
Column political.knowledge is not normally distributed

 9. Univariate analysis for gender_male

Mean is -0.000000, Median is -0.937059, Mode(s) are -0.9371
Column gender_male does not have outliers

Column gender_male is not normally distributed
```

Encoding the categorical variables:

Below gives the exact distribution of categorical variables:

```
1. VOTE :  2
     Conservative     462
     Labour          1063
     Name: vote, dtype: int64


2. GENDER :  2
     male       713
     female     812
     Name: gender, dtype: int64
```

Target variable vote has been encoded with Labour=1 and Conservative=0

One hot encoding has been applied to gender to create binary variable depicting the binary status for each of the gender (Male/Female) and to optimize the first one has been dropped as the 0 or 1 in the other is suffice to represent both.

Below is the data structure after encoding the dataset.

```
<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   int64
 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_male             1525 non-null   uint8
dtypes: int64(8), uint8(1)
memory usage: 108.7 KB
```

Data split into training and testing by 70:30 ratio.

Based on the splitting of training and testing data below are the respective dataset properties for its shape and 5 number summary depicting the split.

```
Number of rows and columns of the training set for the independent
variables: (1067, 8)
Number of rows and columns of the training set for the dependent variable:
(1067,)
Number of rows and columns of the test set for the independent variables:
(458, 8)
Number of rows and columns of the test set for the dependent variable:
(458,)
```

5 number summary for training data set

	age	economic_cond_national	economic_cond_household	Blair	Hague	Europe	political_knowledge	gender_male
count	1067.000000	1067.000000	1067.000000	1067.000000	1067.000000	1067.000000	1067.000000	1067.000000
mean	-0.017634	0.014497	0.017156	0.009709	0.034336	-0.019144	-0.007468	-0.011019
std	1.001376	0.973226	0.957571	0.997633	0.993164	0.990621	0.997315	0.999691
min	-1.921698	-1.982446	-1.854176	-1.987695	-1.419886	-1.737782	-1.424148	-0.937059
25%	-0.839313	-0.279218	-0.150948	-1.136225	-0.607076	-0.827714	-1.424148	-0.937059
50%	-0.075276	-0.279218	-0.150948	0.566716	-0.607076	-0.221002	0.422643	-0.937059
75%	0.816100	0.856268	0.924730	0.566716	1.018544	0.992422	0.422643	1.067169
max	2.471512	1.991754	2.000408	1.418187	1.831354	1.295778	1.346038	1.067169

5 number summary for testing data set

	age	economic_cond_national	economic_cond_household	Blair	Hague	Europe	political_knowledge	gender_male
count	458.000000	458.000000	458.000000	458.000000	458.000000	458.000000	458.000000	458.000000
mean	0.041081	0.012091	0.023632	-0.022620	-0.079992	0.044600	0.017397	0.025671
std	0.997764	0.958450	0.965372	1.007309	1.013406	1.022290	1.008193	1.002435
min	-1.921698	-1.982446	-1.854176	-1.987695	-1.419886	-1.737782	-1.424148	-0.937059
25%	-0.839313	-0.279218	-0.150948	-1.136225	-0.607076	-0.827714	-1.424148	-0.937059
50%	-0.011607	-0.279218	-0.150948	0.566716	-0.607076	0.082354	0.422643	-0.937059
75%	0.816100	0.856268	0.924730	0.566716	1.018544	0.992422	0.422643	1.067169
max	2.216833	1.991754	2.000408	1.418187	1.831354	1.295778	1.346038	1.067169

1.4. Apply Logistic Regression and LDA (Linear Discriminant Analysis). Interpret the inferences of both models.

Based on the regression model that has been built based on the above training and testing data set below are the statistical summary using the statsmodel library.

OLS Regression Results

Dep. Variable:	vote	R-squared:	0.379
Model:	OLS	Adj. R-squared:	0.374
Method:	Least Squares	F-statistic:	80.62
Date:	Sat, 05 Dec 2020	Prob (F-statistic):	6.12e-104
Time:	18:56:22	Log-Likelihood:	-430.21
No. Observations:	1067	AIC:	878.4
Df Residuals:	1058	BIC:	923.2
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.6977	0.011	62.569	0.000	0.676	0.720
age	-0.0482	0.011	-4.299	0.000	-0.070	-0.026
political_knowledge	-0.0642	0.012	-5.576	0.000	-0.087	-0.042
economic_cond_national	0.0452	0.013	3.530	0.000	0.020	0.070
economic_cond_household	0.0120	0.013	0.944	0.346	-0.013	0.037
Blair	0.0994	0.012	8.100	0.000	0.075	0.124
Hague	-0.1568	0.012	-13.148	0.000	-0.180	-0.133
Europe	-0.0992	0.012	-8.143	0.000	-0.123	-0.075
gender_male	0.0102	0.011	0.892	0.373	-0.012	0.033

Omnibus:	25.379	Durbin-Watson:	1.938
Prob(Omnibus):	0.000	Jarque-Bera (JB):	26.556
Skew:	-0.375	Prob(JB):	1.71e-06
Kurtosis:	2.817	Cond. No.	1.84

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Based on the null hypothesis that states that the independent variables have no co relationship with target variables from the universe, please note the coefficients of the predictor variables from the sample should be having p value lesser than the significance value of 5% . Based on the stats model based statistical observation all the predictor variables have their p value is much lesser than 5%. And

hence the null hypothesis will be rejected for each of the predictor variables and all of the predictors will be part of the linear equation.

Subsequently VIF has been evaluated to check multi collinearity among predictor variables. Any multi collinearity among predictor variables would mean that despite prediction itself need not be impacted with any low accuracy, the interpretation of the coefficients would be wrong without treating multi collinearity. Any VIF score between 1 to 5 is an acceptable range to retain the predictor variables in linear equation. Please find below the VIF for all the predictor variables.

```
VIF for age is 1.0139198118441526
VIF for gender_male is 1.0340695795969062
VIF for political_knowledge is 1.0584898802222538
VIF for Hague is 1.1370817306128767
VIF for economic_cond_household is 1.1556640850266833
VIF for Europe is 1.2096030388419405
VIF for Blair is 1.2372034397053333
VIF for economic_cond_national is 1.2565812563307652
```

Based on the above observation none of the independent variables have correlation with other independent variables and hence there is no multi collinearity.

Please find below the accuracy metrics/score:

- `R Squared: 0.379  (Same as adjusted R Squared)`

Accordingly, based on the scaled (normalized) data followed by evaluation of tuning opportunities of the model looking at sample vs universe hypothesis testing, it could be narrowed down that the below formula goes well towards building both Logistic regression as well as LDA models.

vote ~ age+political_knowledge+economic_cond_national+economic_cond_household+Blair+Hague+Europe +gender_male

Accordingly, both models have been built based on the below linear equation that depicts individual coefficients of independent variables along with intercept value.

```
(0.7) * Intercept + (-0.05) * age + (-0.06) * political_knowledge + (0.05)
* economic_cond_national + (0.01) * economic_cond_household + (0.1) *
Blair + (-0.16) * Hague + (-0.1) * Europe + (0.01) * gender_male
```

Based on the model built, below are the count comparison of prediction of test samples against the original classification.
Note: Vote refers to original classification in the dataset and predicted vote refers to model output. Age column refers to count grouped by vote and predicted vote to give a comparison. It enables us to see the real mapping of classification across original and prediction in terms of accuracy by each model.

For Logistical regression:

		age
vote	predicted_vote	
0	0.0	9
	1.0	24
1	0.0	33
	1.0	82

For LDA:

		age
vote	predicted_vote	
0	0.0	8
	1.0	25
1	0.0	33
	1.0	82

1.5. Apply KNN Model and Naïve Bayes Model. Interpret the inferences of each model.

KNN model is a distance based mechanism mostly using Euclidean measure. And in this case we are looking for a classification need and KNN can help build classification based on feature similarity. Since KNN continues to retain the training dataset instead of learning to create a model using the learning, it is a lazy learner and simplest of all models for classification. K in KNN refers to the number of nearest neighbors. A small value of K indicates higher influence of noise over result while larger value is cost heavy for compute. So usually a standardized approach is to adopt K=sqrt(N)/2 where N is the size of the training data set. Also, K has to be an odd number to avoid ties between predicting classes.

Bayes theorem is an extension of conditional probability that is based on the knowledge of prior probability values of something that has occurred. It provides a mechanism to compute posterior probability of class for the given predictor based on prior probability of the given predictor for the same class. Idea is to factor all available evidence in the form of predictors into naïve Bayes rule to obtain more accurate probability for class prediction. Naïve Bayes classifier works on the principle of Bayes theorem with the assumption that input features are independent of each other. Usually not ideal for data sets with a large number of numerical attributes, however, does well with noisy and missing data even with low training samples. For those independent variables that are continuous it is assumed to be distributed normally and the estimates go by mean and standard deviation of continuous variables.

Based on the model built below are the count comparison of prediction of test samples against the original classification.
For Naïve Bayes:

		age
vote	predicted_vote	
0	0.0	10
	1.0	23
1	0.0	30
	1.0	85

For KNN:

		age
vote	predicted_vote	
0	0.0	8
	1.0	25
1	0.0	31
	1.0	84

1.6. Model Tuning , Bagging and Boosting.

For KNN,
Based on the standardized approach to adopt K=sqrt(N)/2 where N is the size of the training data set and that K has to be an odd number to avoid tie between predicting classes we could with K value as 39 for KNN.

For Logistical regression and LDA we have made our investigations to arrive at below conclusions:
Based on the null hypothesis that states that the independent variables have no co relationship with target variables from the universe, please note the coefficients of the predictor variables from the sample should be having p value lesser than the significance value of 5% . Based on the stats model based statistical observation all the predictor variables have their p value is much lesser than 5%. And hence the null hypothesis will be rejected for each of the predictor variables and all of the predictors will be part of the linear equation.
Subsequently VIF has been evaluated to check multi collinearity among predictor variables. Any multi collinearity among predictor variables would mean that despite prediction itself need not be impacted with any low accuracy, the interpretation of the coefficients would be wrong without treating multi collinearity. Any VIF score between 1 to 5 is an acceptable range to retain the predictor variables in linear equation. Please find below the VIF for all the predictor variables.

Also the max_iteration for Logistical regression has been kept at 100 to strike a balance to optimize as well as control over fitment by increasing to more iterations with “newton-cg” solver to arrive at optimal prediction for train and test convergence.
Bagging:
For Bagging we are using random forest that utilizes bootstrapping and ensemble methods to create multiple trees out which final classification goes by voting mechanism across the models.
And in this case we are also tuning the model in terms of the number of estimators (trees) to be limited at 50 followed by maximum features to 50% of total features which is 4. Accordingly bagging spawns many trees as strong learners in parallel and hence over-fitment of individual trees should not be that much of a concern if in rare cases as the overall classification goes by voting across trees and the individual datasets across multiple such models running parallelly are only a subset sample.

Accordingly based on the model execution below are the comparative counts between original classification in the dataset versus predicted vote by the random forest classifier.

		age
vote	predicted_vote	
0	0.0	9
	1.0	24
1	0.0	33
	1.0	82

Boosting:
In contrast to Bagging, boosting trains a large number of weak learners and that too sequentially by focusing on misclassified data by improving their weights further towards training it for accuracy. Boosting then combines all weak learners into a single strong learner to provide an overall classification and hence it has increased aggregate complexity while using simple low complex individual models.
In the project we have used an over-sampling technique using SMOTE (synthetic minority oversampling technique) towards improving the under-represented class of the target variable to avoid false positive predictions. SMOTE creates more samples of the minority class, however not by replicating the existing data points but by creating new data points within the range of possibility.

Accordingly based on the model execution below are the comparative count between original classification in the dataset versus predicted vote on the test dataset for each of the four models which clearly depicts increased accuracy and prevent false positives:
For logistical regression:

		age
vote	predicted_vote	
0	0.0	12
	1.0	21
1	0.0	41
	1.0	74

For LDA:

		age
vote	predicted_vote	
0	0.0	12
	1.0	21
1	0.0	41
	1.0	74

For Naïve Bayes:

		age
vote	predicted_vote	
0	0.0	12
	1.0	21
1	0.0	40
	1.0	75

For KNN

		age
vote	predicted_vote	
0	0.0	13
	1.0	20
1	0.0	43
	1.0	72

1.7. Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model .Final Model - Compare all models on the basis of the performance metrics in a structured tabular manner. Describe on which model is best/optimized

Accuracy metrics below:

	Without boosting	
Model	**Training score**	**Test score**
Logistical regression	0.830365511	0.849344978
LDA	0.826616682	0.847161572
Naïve Bayes	0.826616682	0.847161572
KNN	0.999062793	0.858078603

	Boosting	
Model	**Training score**	**Test score**
Logistical regression	0.819892473	0.825328
LDA	0.819892473	0.825328
Naïve Bayes	0.8125	0.825328
KNN	0.891801075	0.790393

Performance accuracy

	Original	
Model	**Training score**	**Test score**
Logistical regression	0.830365511	0.849344978
LDA	0.826616682	0.847161572
Naïve Bayes	0.826616682	0.847161572
KNN	0.999062793	0.858078603
Random Forest classifier	0.999062793	0.829694323

	Boosting	
Model	**Training score**	**Test score**
Logistical regression	0.819892473	0.825327511
LDA	0.819892473	0.825327511

Naïve Bayes	0.8125	0.825327511
KNN	0.891801075	0.790393013

Confusion matrix:

Logistic regression:

Without boosting		
Logistic regression/Training	**Predicted Negative**	**Predicted Positive**
Actual Negative	212	111
Actual Positive	70	674

Without b	
Logistic regression/Testing	**Predic**
Actual Negative	
Actual Positive	

Boosting		
LDA/Training	**Predicted Negative**	**Predicted Positive**
Actual Negative	610	134
Actual Positive	134	610

Boosting	
LDA/Testing	**Predicted Negative**
Actual Negative	114
Actual Positive	55

LDA:

Without boosting		
LDA/Training	**Predicted Negative**	**Predicted Positive**
Actual Negative	217	106
Actual Positive	79	665

Without b	
LDA/Testing	**Predicted Nega**
Actual Negative	
Actual Positive	

Boosting		
LDA/Training	**Predicted Negative**	**Predicted Positive**
Actual Negative	610	134
Actual Positive	134	610

Boosting	
LDA/Testing	**Predicted Negative**
Actual Negative	114
Actual Positive	55

Naïve Bayes:

Without boosting		
Naïve Bayes/Training	**Predicted Negative**	**Predicted Positive**
Actual Negative	226	97
Actual Positive	88	656

Without b	
Naïve Bayes/Testing	**Predicted I**
Actual Negative	
Actual Positive	

Boosting		
Naïve Bayes/Training	**Predicted Negative**	**Predicted Positive**
Actual Negative	600	144
Actual Positive	135	609

Boosting	
Naïve Bayes/Testing	**Predicted Negative**
Actual Negative	112
Actual Positive	53

KNN:

Without boosting		
KNN/Training	**Predicted Negative**	**Predicted Positive**
Actual Negative	323	0
Actual Positive	1	743

Without b	
KNN/Testing	**Predicted Nega**
Actual Negative	
Actual Positive	

Boosting		
KNN/Training	**Predicted Negative**	**Predicted Positive**
Actual Negative	715	29
Actual Positive	132	612

Boosting	
KNN/Testing	**Predicted Negative**
Actual Negative	114
Actual Positive	71

Bagging through Random Forest:

Bagging		
Random Forest/Training	**Predicted Negative**	**Predicted Positive**
Actual Negative	323	0
Actual Positive	1	743

Baggi	
Random Forest/Testing	**Predicte**
Actual Negative	
Actual Positive	

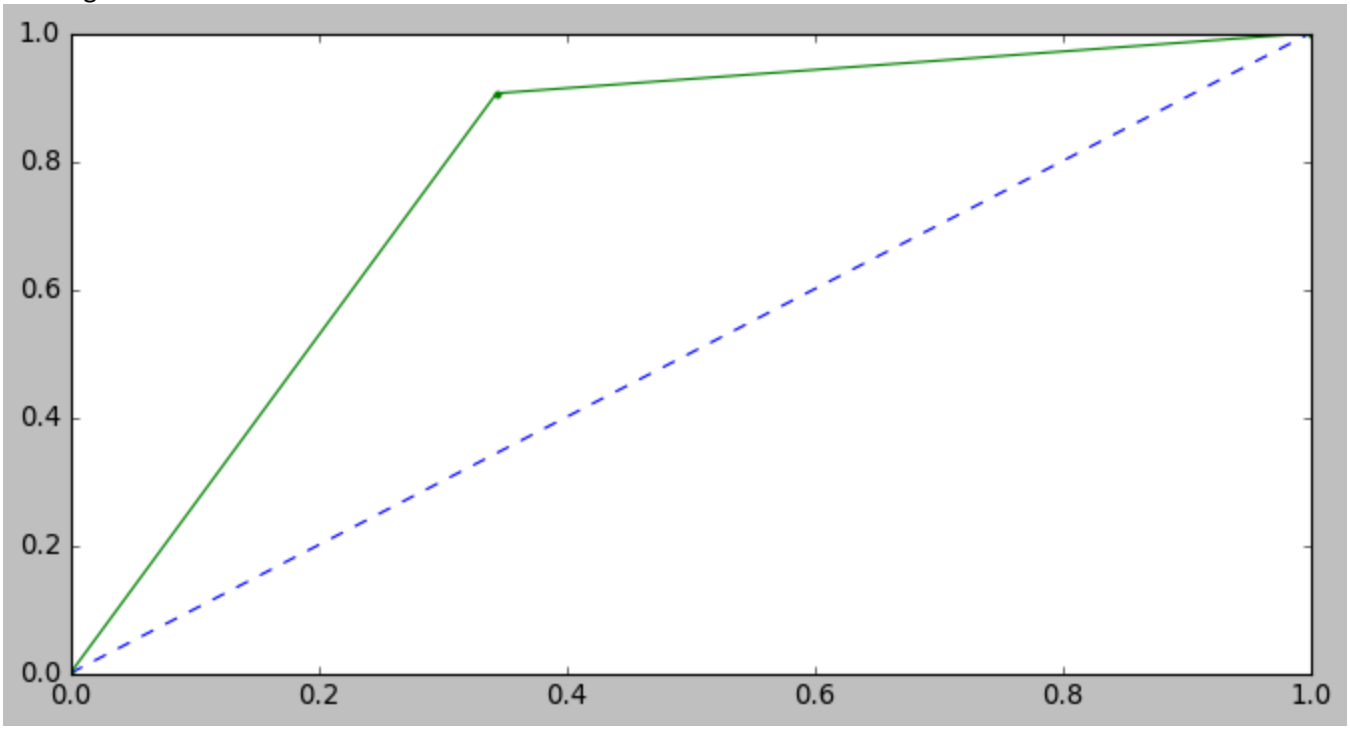
AUC score:

AUC	**Without boosting**	
Model	**Training score**	**Test score**
Logistical regression	0.781	0.801
LDA	0.877	0.916
Naïve Bayes	0.875	0.91
KNN	1	0.917
Random Forest classifier	1	0.899

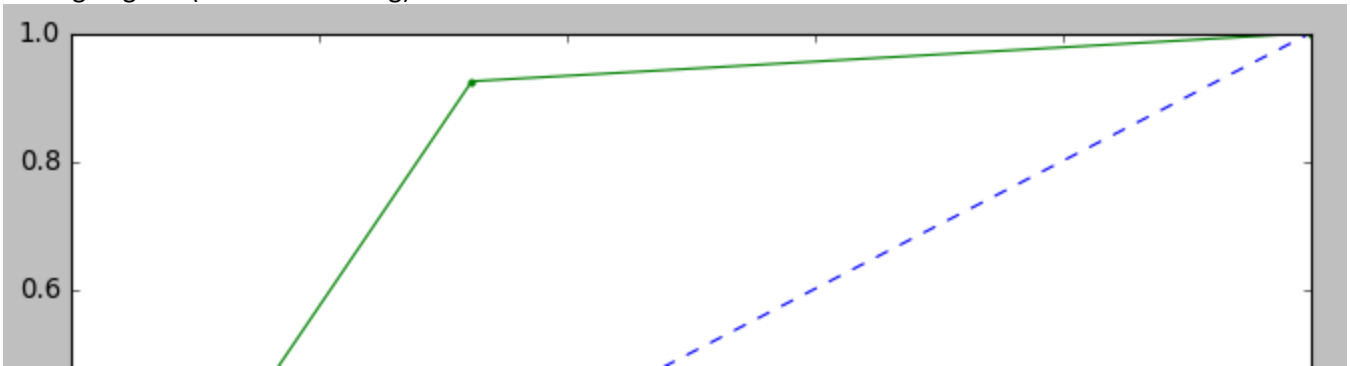
AUC	**Boosting**	
Model	**Training score**	**Test score**
Logistical regression	0.884	0.914
LDA	0.884	0.914
Naïve Bayes	0.886	0.908
KNN	0.965	0.863

ROC curve by model without boosting.

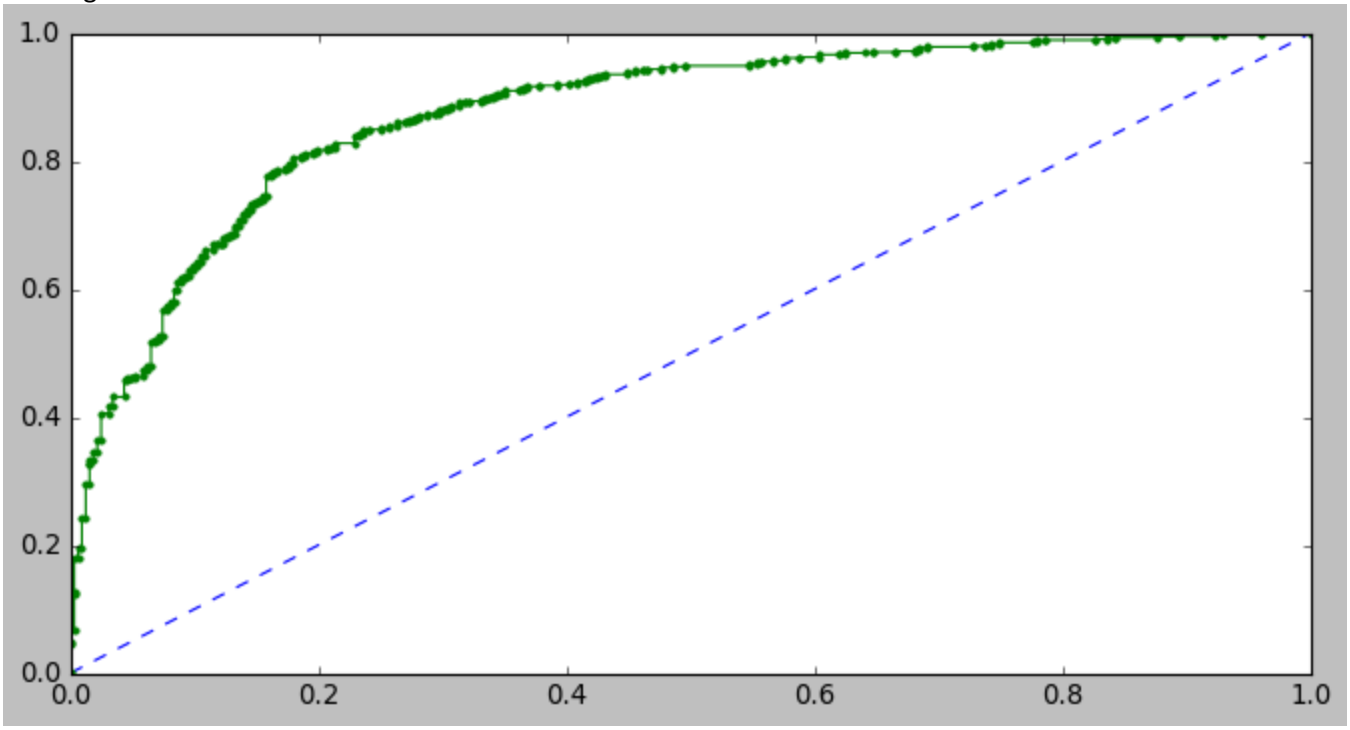
Logistic (without boosting)
Training



Testing-Logistic (without boosting)

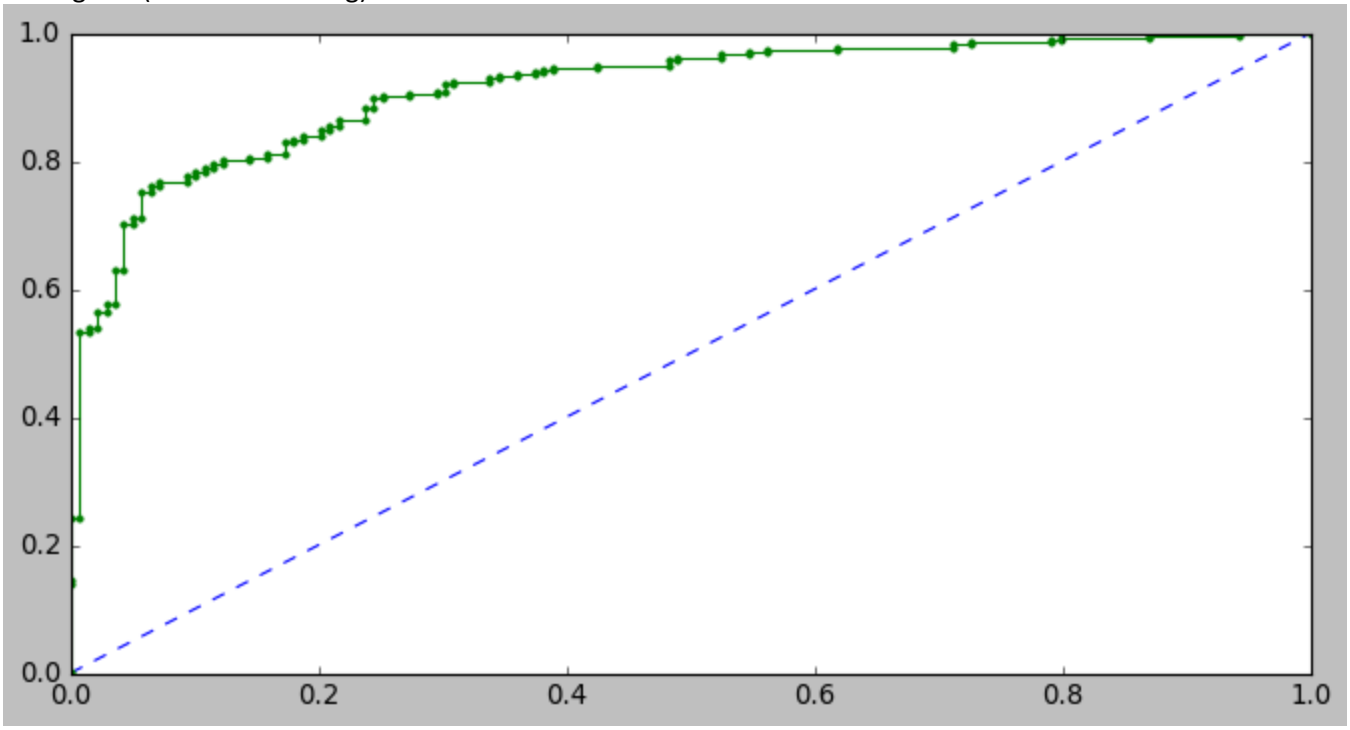


LDA(Without boosting)
Training



LDA(
Train

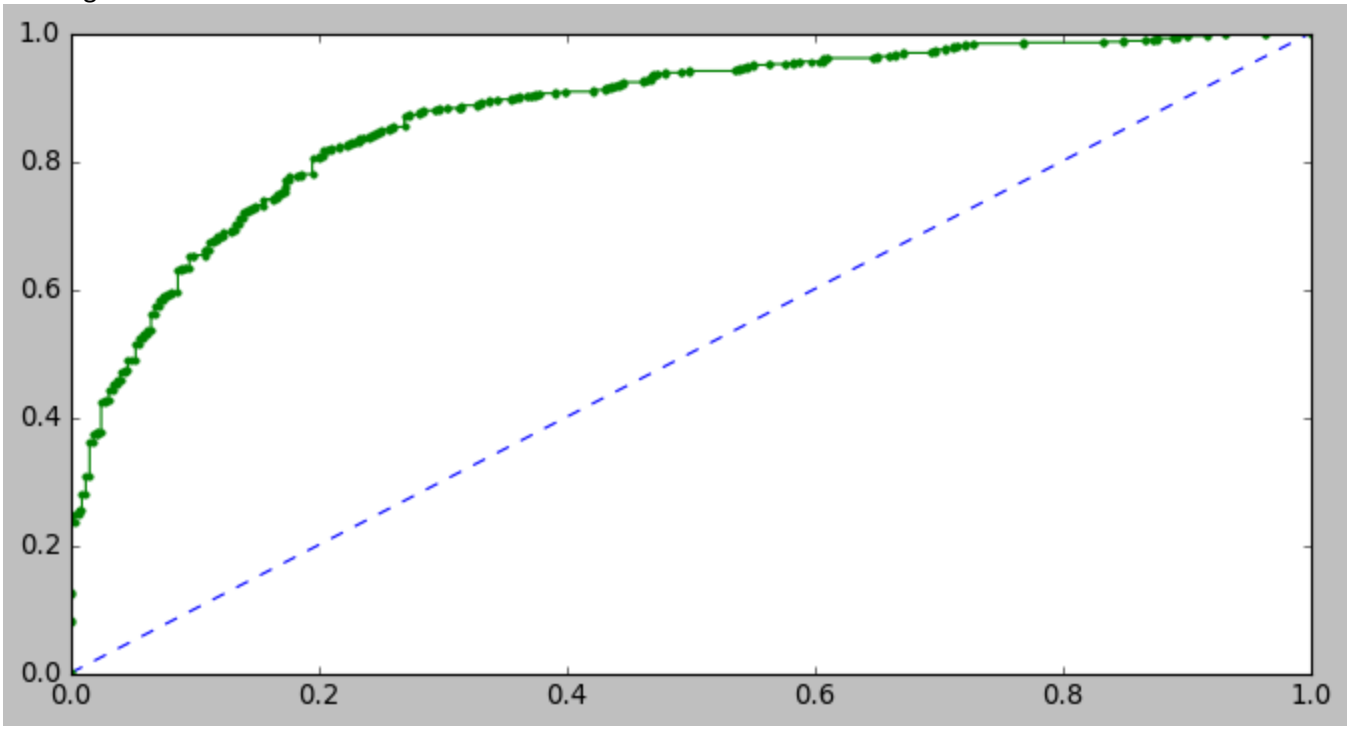
Testing-LDA(Without boosting)



Testi

Naïve Bayes (Without boosting)

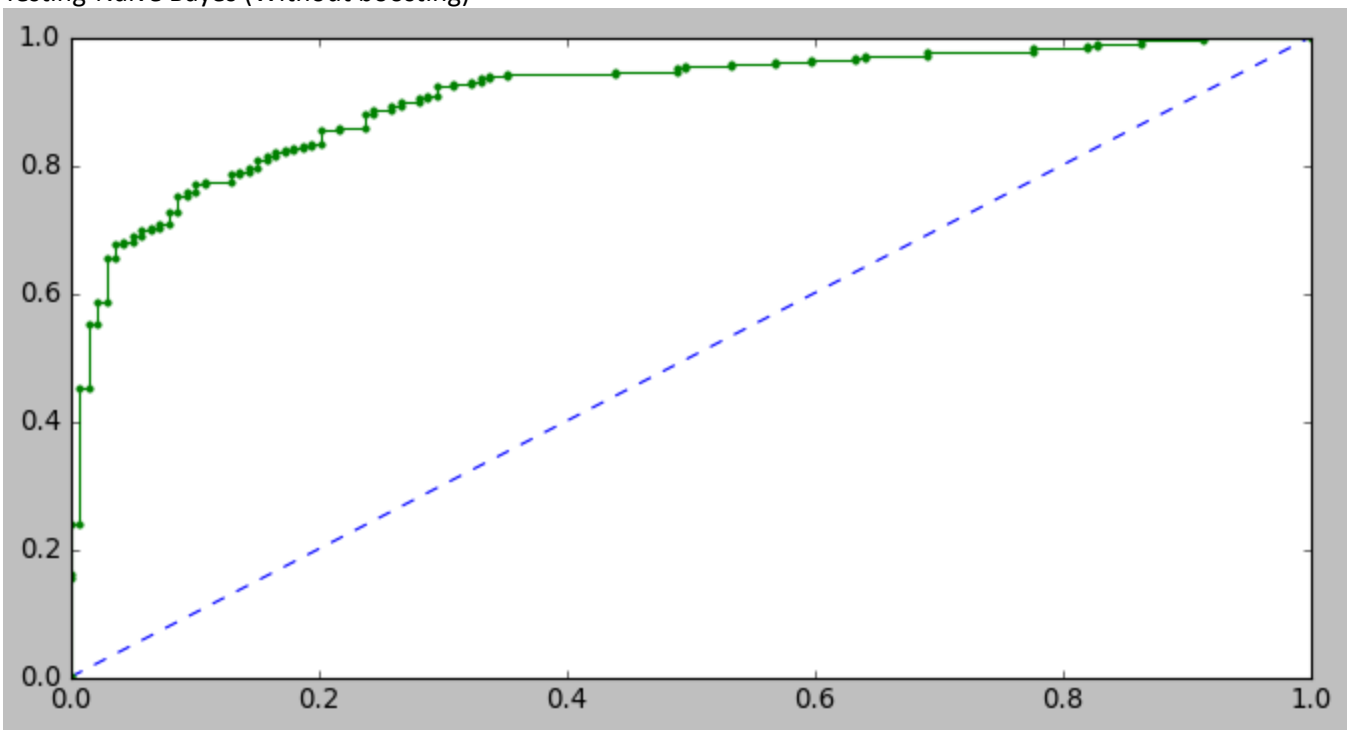
Training



Naïve

Train

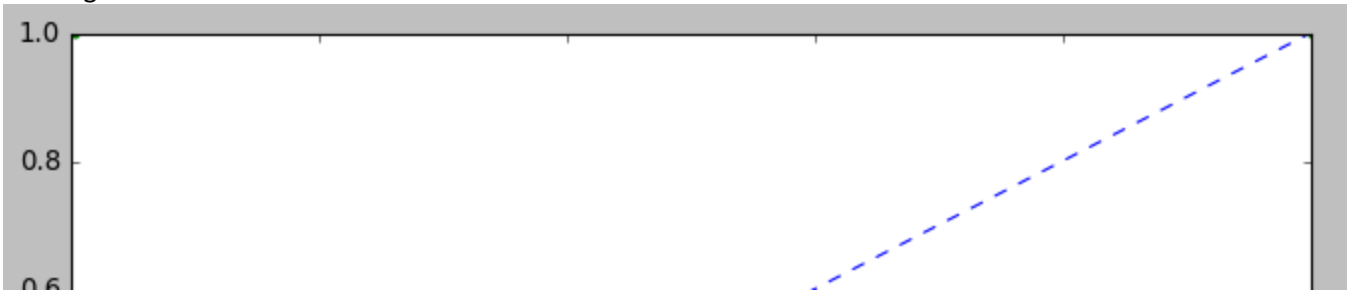
Testing-Naïve Bayes (Without boosting)



Testi

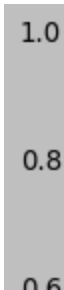
KNN (without boosting)

Training

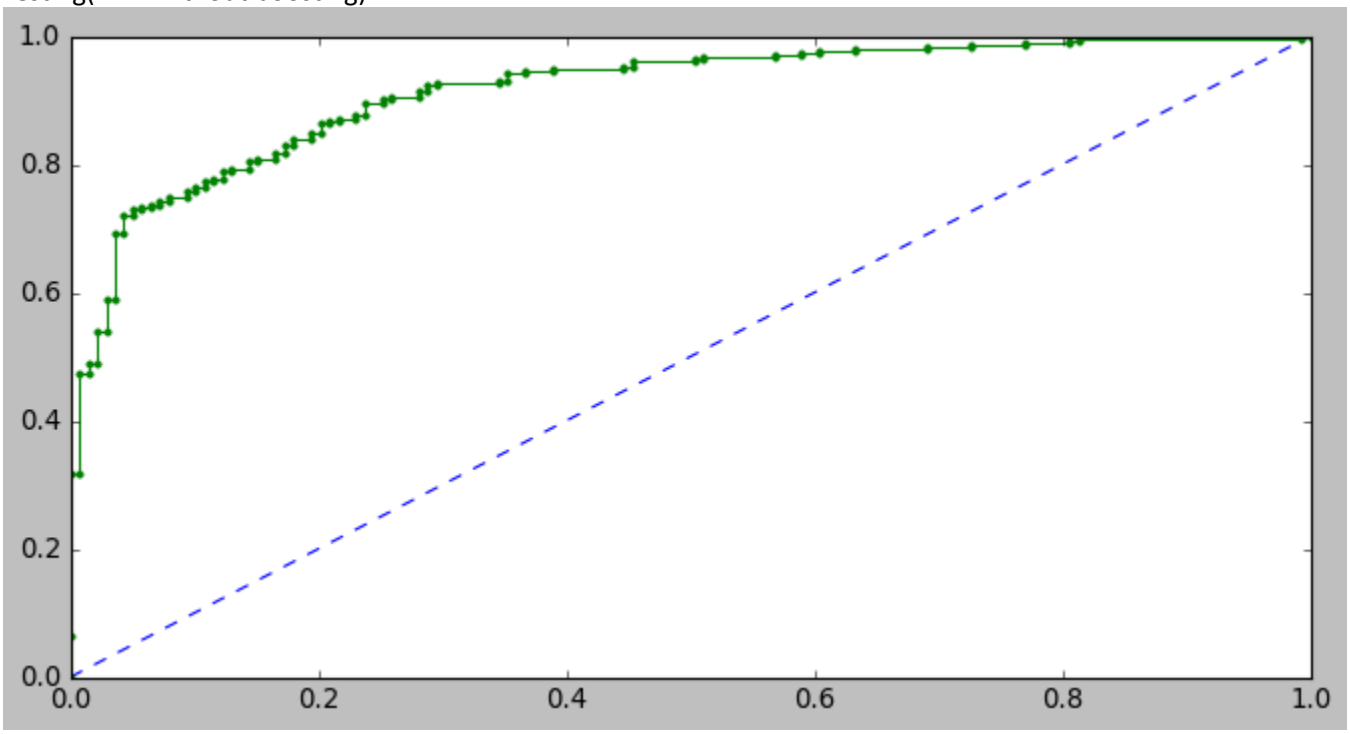


KNN

Train



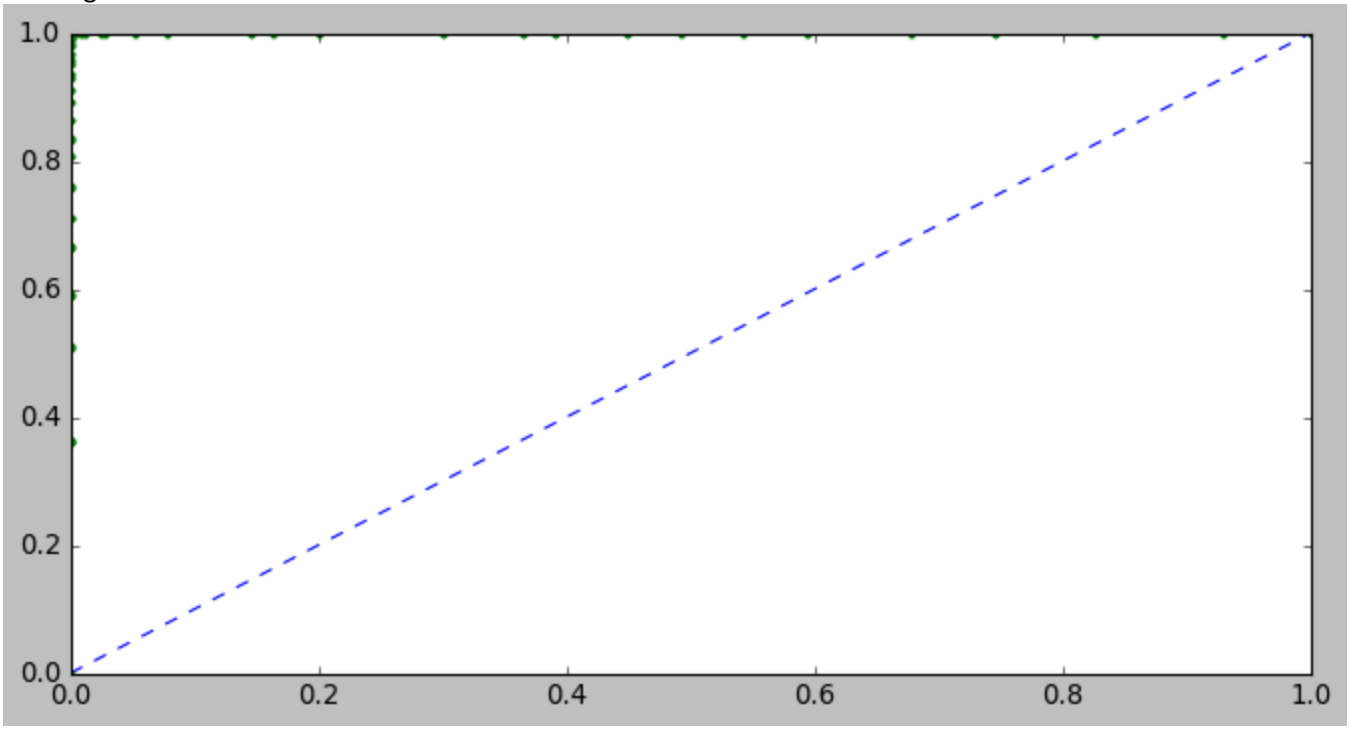
Testing(KNN without boosting)



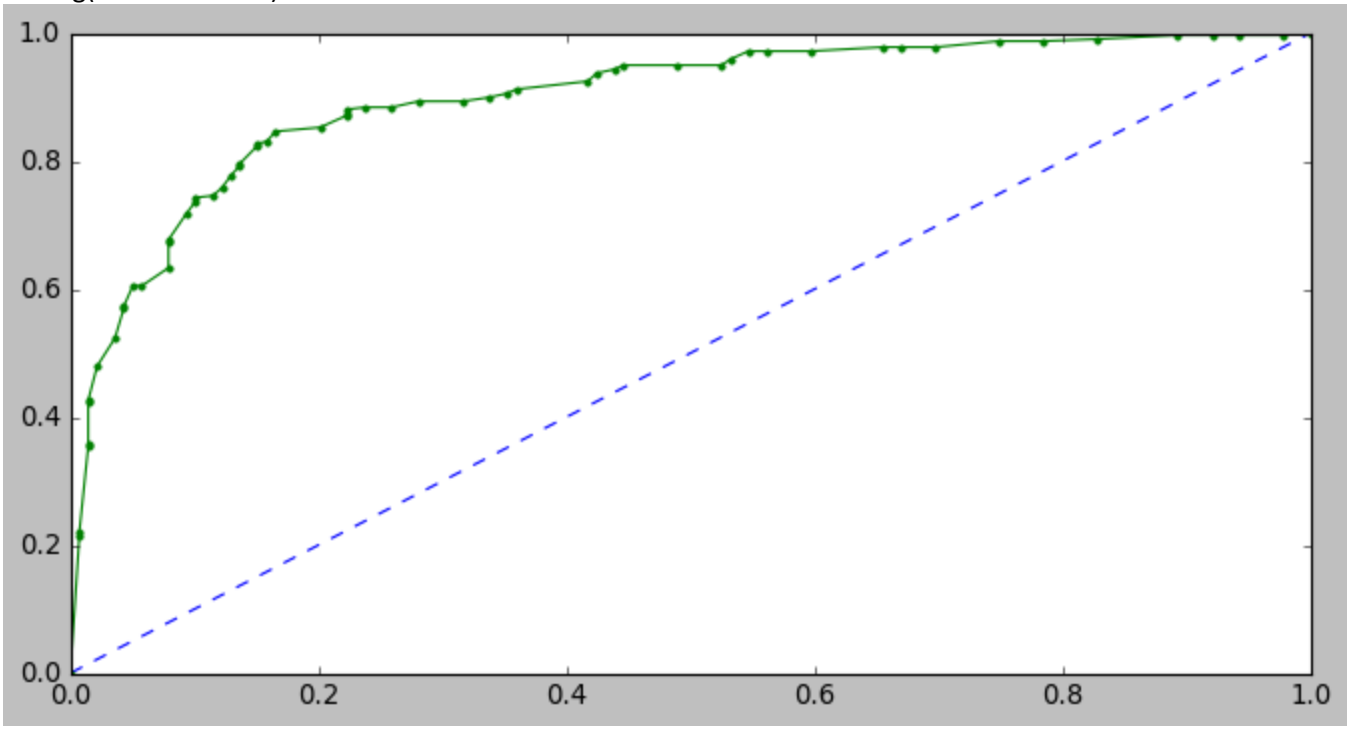
Testi

Random forest classifier

Training



Testing(Random forest)



Classification matrix is as below:

Logistic regression without boosting

classification report for training data set:

	precision	recall	f1-score	support
0	0.75	0.66	0.70	323
1	0.86	0.91	0.88	744
accuracy			0.83	1067
macro avg	0.81	0.78	0.79	1067
weighted avg	0.83	0.83	0.83	1067

classification report for testing data set:

	precision	recall	f1-score	support
0	0.80	0.68	0.73	139
1	0.87	0.92	0.90	319
accuracy			0.85	458
macro avg	0.83	0.80	0.81	458
weighted avg	0.85	0.85	0.85	458

Logistic regression with b

Classification Repor

	preci
0	0
1	0
accuracy	
macro avg	0
weighted avg	0

Classification Repor

	preci
0	0
1	0
accuracy	
macro avg	0
weighted avg	0

LDA with boosting

Classification Repor

	preci
0	0
1	0
accuracy	
macro avg	0
weighted avg	0

LDA without boosting

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.73	0.67	0.70	323
1	0.86	0.89	0.88	744
accuracy			0.83	1067
macro avg	0.80	0.78	0.79	1067
weighted avg	0.82	0.83	0.82	1067

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.77	0.71	0.74	139
1	0.88	0.91	0.89	319
accuracy			0.85	458
macro avg	0.82	0.81	0.81	458
weighted avg	0.84	0.85	0.85	458

Naïve Bayes without boosting

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.72	0.70	0.71	323
1	0.87	0.88	0.88	744
accuracy			0.83	1067
macro avg	0.80	0.79	0.79	1067

Naïve Bayes with boosting

Classification Repor

	preci
0	0
1	0
accuracy	

KNN without boosting

Classification Report of the training data:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	323
1	1.00	1.00	1.00	744
accuracy			1.00	1067
macro avg	1.00	1.00	1.00	1067
weighted avg	1.00	1.00	1.00	1067

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.80	0.71	0.75	139
1	0.88	0.92	0.90	319
accuracy			0.86	458
macro avg	0.84	0.82	0.83	458
weighted avg	0.86	0.86	0.86	458

KNN with boosting

Classification Repor

	prec
0	
1	
accuracy	
macro avg	
weighted avg	

Classification Repor

	prec
0	
1	
accuracy	
macro avg	
weighted avg	

Random Forest

Classification Report of the training data:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	323
1	1.00	1.00	1.00	744
accuracy			1.00	1067
macro avg	1.00	1.00	1.00	1067
weighted avg	1.00	1.00	1.00	1067

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.74	0.68	0.71	139
1	0.87	0.89	0.88	319
accuracy			0.83	458
macro avg	0.80	0.79	0.79	458
weighted avg	0.83	0.83	0.83	458

1.8. Based on your analysis and working on the business problem, detail out appropriate insights and recommendations to help the management solve the business objective.

Since the classification has to focus on the probability of favorable votes across both the parties of labour and conservative, the business question is to find who will vote (True positives) more than who did not get the vote (True negative). And hence it is important to strike a balance on the true positives across prediction metrics (precision) and actual metrics(recall). That leads to the fact that both Logistic regression and LDA has done better without boosting even though naïve bayes have better scores under boosting wherein the training and testing convergence is much better in the earlier case of LDA and logistic regression.

Also based on the model we could derive below linear equation where we can see that Blair assessment has the most significant impact on the voting pattern, meaning every unit of increase in his assessment results in better prediction of voting by 10%.

```
(0.7) * Intercept + (-0.05) * age + (-0.06) * political_knowledge + (0.05)
* economic_cond_national + (0.01) * economic_cond_household + (0.1) *
Blair + (-0.16) * Hague + (-0.1) * Europe + (0.01) * gender_male
```

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:
Total characters in the speech from President Roosevelt:7571
Total characters in the speech from President Kennedy:7571
Total characters in the speech from President Nixon:7571
```

Number of words:

```
Total words spoken by President Roosevelt:1360
Total words spoken by President Kennedy:1390
Total words spoken by President Nixon:1819
```

Number of sentences:

```
Total sentences in the speech from President Roosevelt:38
Total sentences in the speech from President Kennedy:27
Total sentences in the speech from President Nixon:51
```

2.2) Remove all the stopwords from the three speeches.
Stop words are common words that are not useful in providing value or context and hence would have to be processed to eliminate them as a part of text cleaning. Ex: 'the', 'an', 'in'
Following are the top 10 common words by it's frequency spoken in president Roosevelt's speech

```
[('the', 104),
 ('of', 81),
 ('and', 41),
 ('to', 35),
 ('in', 30),
 ('a', 28),
 ('is', 24),
 ('--', 22),
 ('we', 22),
 ('that', 21)]
```

After stop words have been removed with words converted to lower case following are the top 10 common words spoken in president Roosevelt's speech by it's frequency.

```
[('us', 25),
 ('let', 22),
 ('--', 17),
 ('new', 15),
 ('peace', 11),
 ('great', 9),
 ('america', 9),
 ('world.', 8),
 ("america's", 8),
 ('shall', 7)]
```

Following are the top 10 common words by it's frequency spoken in president Kennedy's speech

```
[('the', 86),
 ('of', 65),
 ('to', 42),
 ('and', 41),
 ('we', 30),
 ('a', 29),
 ('in', 26),
 ('--', 24),
 ('our', 21),
 ('not', 19)]
```

After stop words have been removed with words converted to lower case following are the top 10 common words spoken in president Kennedy ‘s speech by it’s frequency.

```
[('--', 24),
 ('let', 16),
 ('us', 11),
 ('new', 7),
 ('pledge', 7),
 ('sides', 7),
 ('shall', 5),
 ('ask', 5),
 ('president', 4),
 ('fellow', 4)]
```

Following are the top 10 common words by it’s frequency spoken in president Nixon’s speech

```
[('the', 83),
 ('of', 68),
 ('to', 65),
 ('in', 58),
 ('and', 50),
 ('we', 47),
 ('a', 35),
 ('that', 33),
 ('our', 32),
 ('for', 32)]
```

After stop words have been removed with words converted to lower case following are the top 10 common words spoken in president Nixon’s speech by it’s frequency.

```
[('us', 25),
 ('let', 22),
 ('--', 17),
 ('new', 15),
 ('peace', 11),
 ('great', 9),
 ('america', 9),
 ('world.', 8),
 ("america's", 8),
 ('shall', 7)]
```

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)

After stop words have been removed following are the top 3 words spoken in President Roosevelt’s speech by it’s frequency.

```
[('us', 25),
 ('let', 22),
 ('--', 17)]
```

```
Subsequently after removing the punctuations and performing stemming the
top 3 words stands corrected as below for President Roosevelt.
[('nation', 10),
 ('know', 9),
```

```
 ('us', 8)]
```

After stop words have been removed following are the top 10 common words spoken in President `Kennedy` ‘s speech by it’s frequency.

```
[('--', 24),
 ('let', 16),
 ('us', 11)]
```

```
Subsequently after removing the punctuations and performing stemming the
top 3 words stands corrected as below for President Kennedy.
[('let', 16),
 ('us', 11),
 ('power', 7)]
```

After stop words have been removed following are the top 10 common words spoken in President `Nixon`’s speech by it’s frequency.

```
[('us', 25),
 ('let', 22),
 ('--', 17)]
```

```
Subsequently after removing the punctuations and performing stemming the
top 3 words stands corrected as below for President Nixon.
[('us', 25),
 ('let', 22),
 ('new', 15)]
```

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

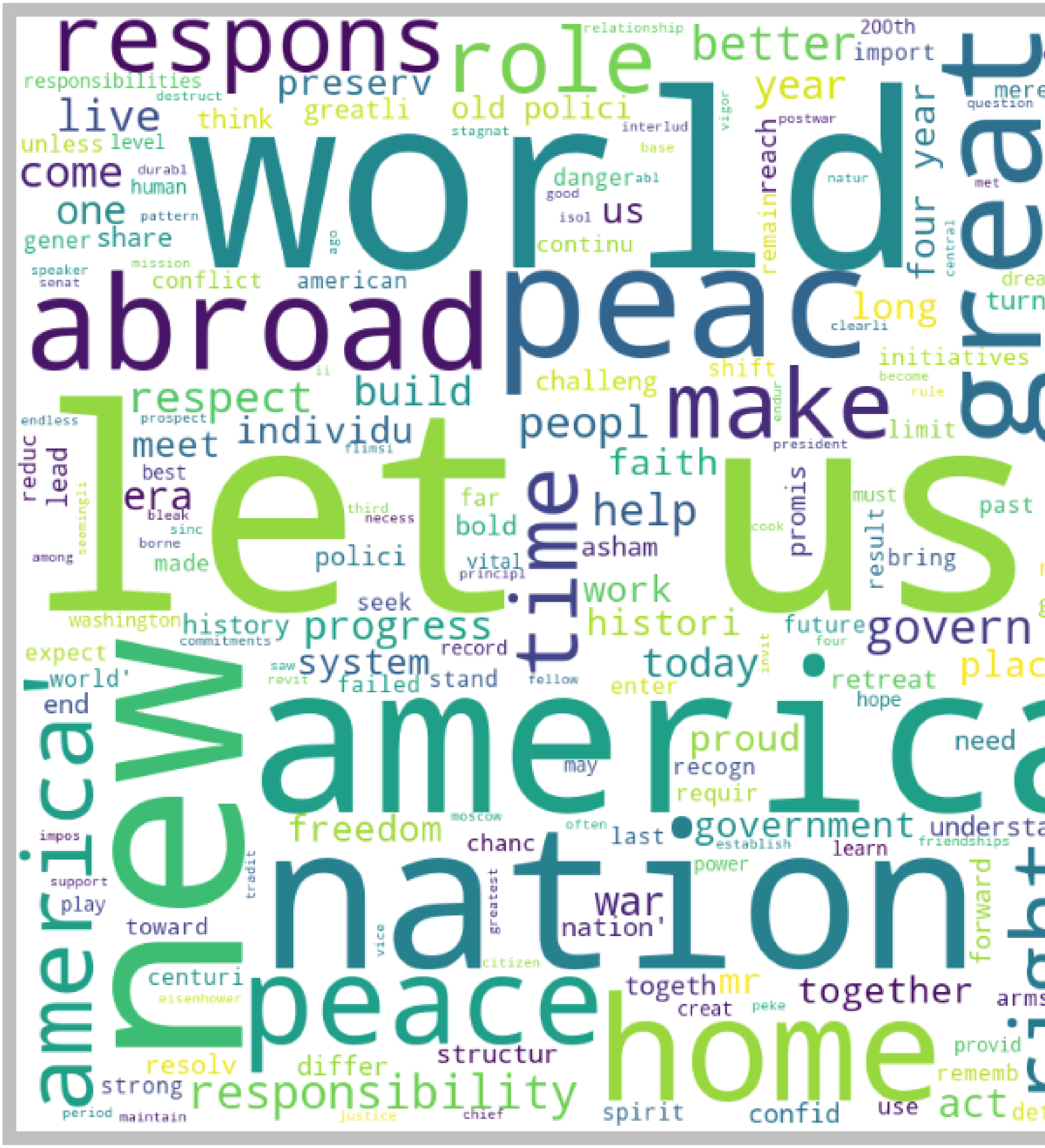
Following is the word cloud from President Roosevelt’s speech which can be correlated to the top words spoken after stop words and punctuation removals and stemming based on the font sizing.


nation
spirit
year
know
live
life
americ
freedo
human
democraci
democracy
mind
american
faith
free
speak us
new
people
must
someth
hope
seem
thing
unit states
disrupt
paus
recal
success
maintain
weld
take
will
happenings
upon
creat
littl
time
terror
moment
lifetim
washington'
inform
artifici
tide
action
boldly
reason
believ
contin
even
secur
without
count
swift
sens
mani
person
kind
midst
doubt
old
fate
word
enterpris
forward
built
present
form
day
task
experi
carri
come
full
men
slaveri
true
may
earli
determin
lincoln'
risk
real
us
find
acted
surg
everi
bodi
enough
prove
fatalist
way
sacr
frozen
dedic
togeth
purpos
act
fire
go
quickly
less
surviv
govern
task peopl
constitut
scor
naught
downfal
stock
cloth
sinc
eight
inaction
peril
history
rediscov
states
mystic
ideal
seen
unexplain
elect
prophet
world
republ
destini
limit
came
peopl
frame
tyranni
framework
lines
unhamp
maintained
common
later
dire
bill
life'
fruit
better
women
predict
initi
ten
save
grow
fact
evil
structur
institut
endur
branch
join
continu
individu
shock
right
coordin
eb
dying
remain
wave
place
revive
greater
put
die
function
taken
home
histori
freeli
wholli
away
materi

Following is the word cloud from President Kennedy’s speech which can be correlated to the top words spoken after stop words and punctuation removals and stemming based on the font sizing..


begin
new
never
wish
days
truli
life
unit
form
though
cooper
support
solemn
commit
old
hope
faith
togeth
place
poverty
forget
poverti
absolut
scienc
cultur
rememb
century
friend
state
earth
dare
word
symbol
truman
now
whose
summon
mr
come
price
order
issu
prescrib
alike
nixon
endeavor
peac
still
mere
oath
light
control
god
yet
mortal
doubt
around
generos
almighti
renewal
hardship
johnson
problem
littl
know
terror
ill
instrument
origin
celebr
clergy
freedom
call
free
pay
quarter
abolish
time
best
change
disciplin
today
balanc
everi
heritag
long
ago
seek
whether
struggl
help
histori
hand
disease
anew
deed
nearli
differ
offer
weak
observ
effort
pledg
chief
much
revolution
may
assur
bring
tyranny
end
belief
beyond
success
gener
unwil
find
l
bitter
instead
fought
right
fellow
peace
torch
proud
countri
peopl
becom
committed
temper
ancient
signifi
oppos
explor
hold
join
enemi
globe
back
first
forebear
man
heir
foe
well
justice
slow
alli
vice
wit
weakness
hard
victori
centuri
finish
citizen
ask
arm
sworn
year
undo
home
three
south
reverend
spiritu
eisenhower
surviv

Following is the word cloud from President Nixon’s speech which can be correlated to the top words spoken after stop words and punctuation removals and stemming based on the font sizing..


respons
role
better
200th
import
relationship
world
responsibilities
destruct
preserv
year
mere
question
live
greatli
old polici
think
interlud
postwar
unless
level
stagnat
vigor
base
come
durabl
human
danger
abl
natur
met
good
reach
one
pattern
isol
us
remain
four year
gener
share
continu
ago
speaker
senat
mission
central
abroad
peac
conflict
american
long
turn
clearli
initiatives
shift
challeng
become
ii
endur
rule
respect
build
make
endless
prospect
individu
peopl
limit
let us
meet
flimsi
president
reduc
lead
faith
best
era
time
far
must
third
necess
help
promis
past
bleak
seemingli
sinc
bold
cook
result
borne
asham
among
bring
made
polici
vital
principl
work
seek
washington
history
histori
progress
commitments
govern
future
four
expect
record
system
today
invit
retreat
saw
revit
failed
stand
fellow
world'
enter
hope
end
new
america
proud
need
america'
may
recogn
requir
freedom
moscow
government
understa
often
impos
chanc
last
establish
friendships
learn
support
tradit
power
nation
play
greatest
war
forward
toward
vice
nation'
citizen
centuri
mr
together
togeth
arms
eisenhower
creat
peke
peace
home
provid
resolv
differ
structur
rememb
strong
responsibility
period
maintain
justice
chief
spirit
confid
use
act