Project Machine learning

Parnoshree Chatterjee

A. Election_Data.xlsx Dataset:

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

Answer:

Table 1.1: Reading data set

	Data Dictionary
0	1. vote: Party choice: Conservative or Labour
1	2. age: in years
2	3. economic.cond.national: Assessment of curre...
3	4. economic.cond.household: Assessment of curr...
4	5. Blair: Assessment of the Labour leader, 1 t...

Table 1.2: We are converting the xlsx file to excel and checking info.

```
Out[6]: <bound method DataFrame.info of      Unnamed: 0          vote  age  economic.cond.national  \
0              1        Labour   43                       3
1              2        Labour   36                       4
2              3        Labour   35                       4
3              4        Labour   24                       4
4              5        Labour   41                       2
...          ...           ...  ...                     ...
1520        1521  Conservative   67                       5
1521        1522  Conservative   73                       2
1522        1523        Labour   37                       3
1523        1524  Conservative   61                       3
1524        1525  Conservative   74                       2

      economic.cond.household  Blair  Hague  Europe  political.knowledge  \
0                           3      4      1       2                    2
1                           4      4      4       5                    2
2                           4      5      2       3                    2
3                           2      2      1       4                    0
4                           2      1      1       6                    2
...                       ...    ...    ...     ...                  ...
1520                        3      2      4      11                    3
1521                        2      4      4       8                    2
1522                        3      5      4       2                    2
1523                        3      1      4      11                    2
1524                        3      2      4      11                    0

      gender
0     female
1       male
2       male
3     female
4       male
...      ...
1520    male
1521    male
1522    male
1523    male
1524  female

[1525 rows x 10 columns]>
```

We have converted the file to excel and checking the head.

Table 1.3: Reading the excel file head

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

Descriptive Statistics for the dataset

Table 1.4: Descriptive Statistics for Cubic Zirconia dataset

	Unnamed: 0	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge
count	1525.000000	1525.000000	1525.000000	1525.000000	1525.000000	1525.000000	1525.000000	1525.000000
mean	763.000000	54.182295	3.245902	3.140328	3.334426	2.746885	6.728525	1.542295
std	440.373894	15.711209	0.880969	0.929951	1.174824	1.230703	3.297538	1.083315
min	1.000000	24.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.000000
25%	382.000000	41.000000	3.000000	3.000000	2.000000	2.000000	4.000000	0.000000
50%	763.000000	53.000000	3.000000	3.000000	4.000000	2.000000	6.000000	2.000000
75%	1144.000000	67.000000	4.000000	4.000000	4.000000	4.000000	10.000000	2.000000
max	1525.000000	93.000000	5.000000	5.000000	5.000000	5.000000	11.000000	3.000000

Checking for Null values

```
Unnamed: 0                0
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
```

There are no null values in the data set which shows we have a clean dataset.

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)

Answer: EDA

Table 1.5: EDA for Election data set

Shape of the data set:

```
(1525, 10)
```

Checking if the data set has all integers or object type values:

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

We will drop the unnamed column from the data set as we don't have any use of the column.

Checking for Duplicates:

```
No. of duplicates rows = 8
```

There are 8 duplicates in the data we will remove those to make sure there is no duplication in the dataset. The reason we are removing these duplicates is also because the number of duplicates is less and won't cause any problem to the data set for further EDA.

Univariate Analysis:

Fig. 1.1 is Bar plot and Box plot for vote and age.

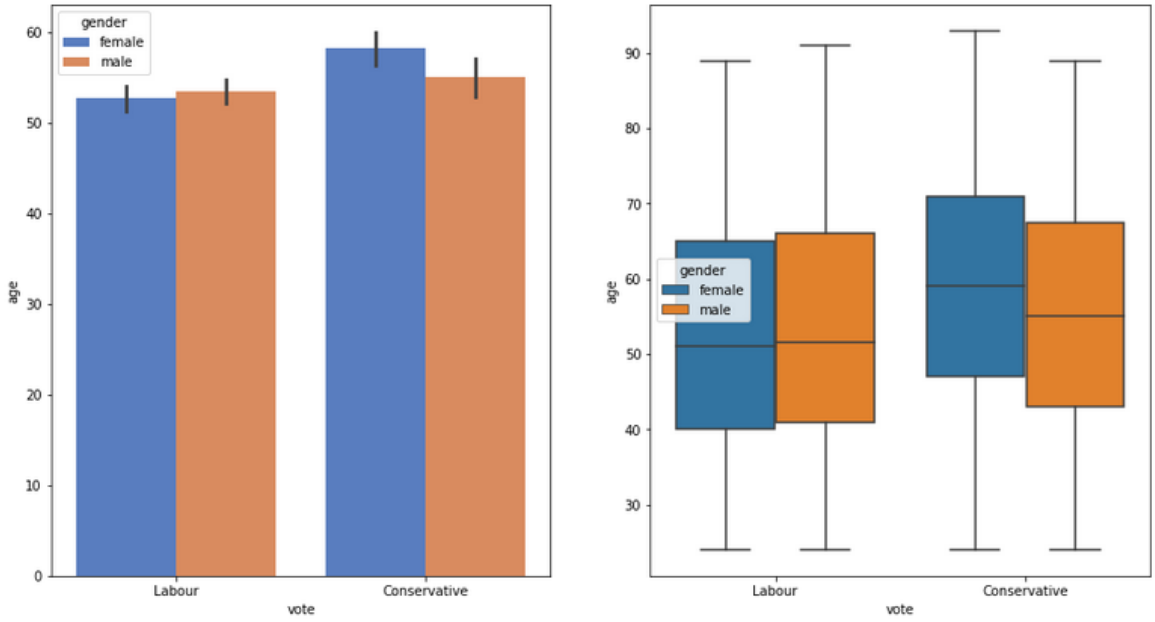


Fig. 1.2 is Bar plot and Box plot for economic condition national and economic condition household.

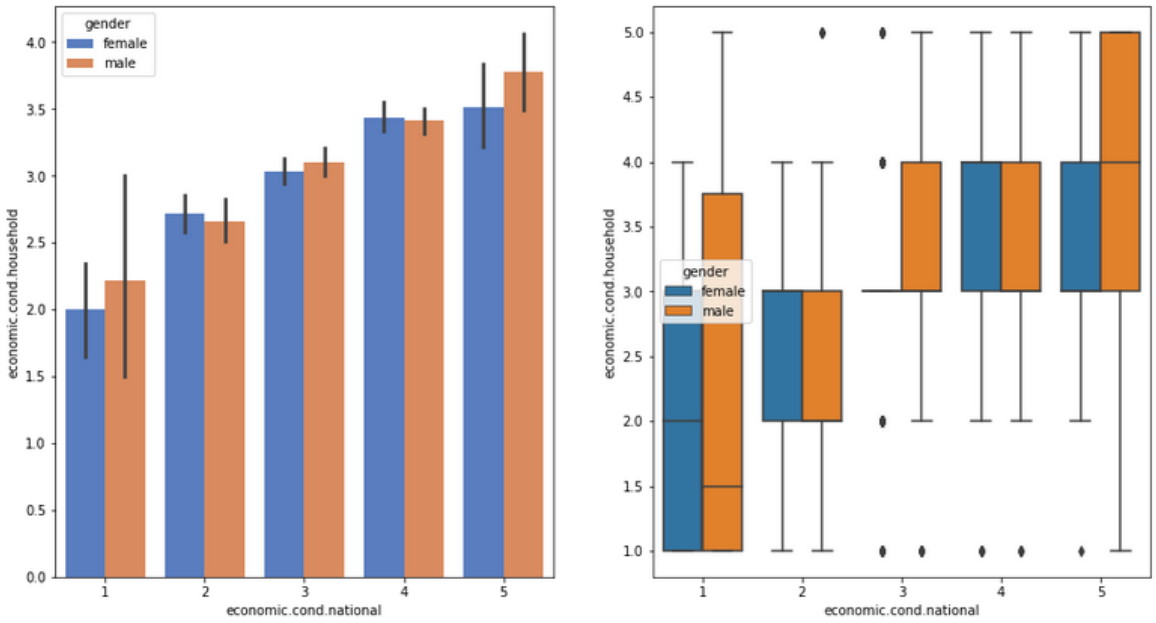


Fig. 1.3 is Bar plot and Box plot for Europe and Hauge household.

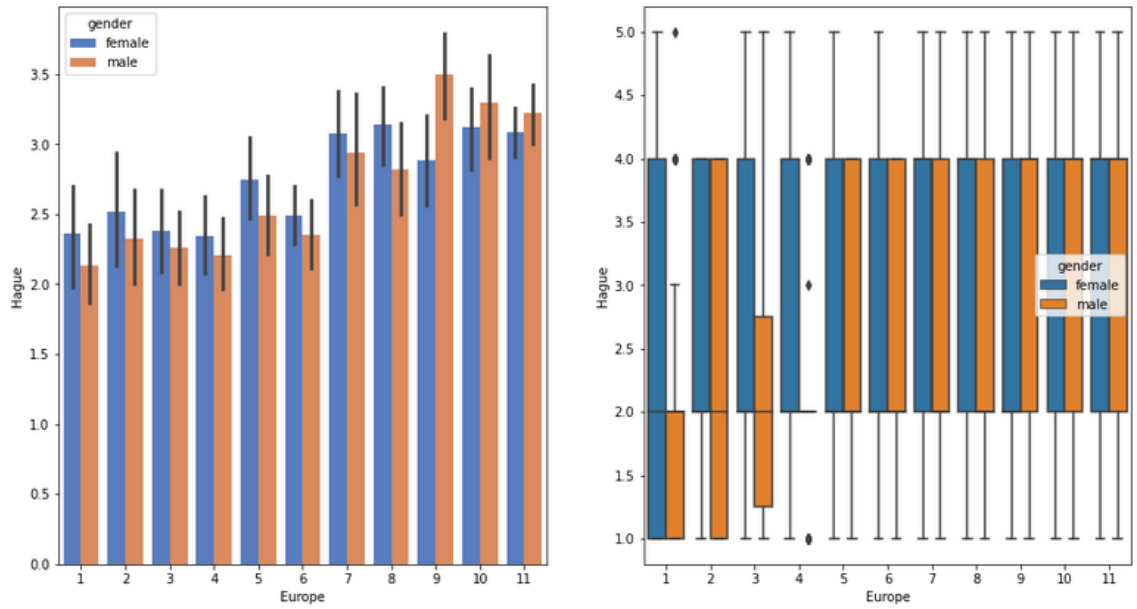
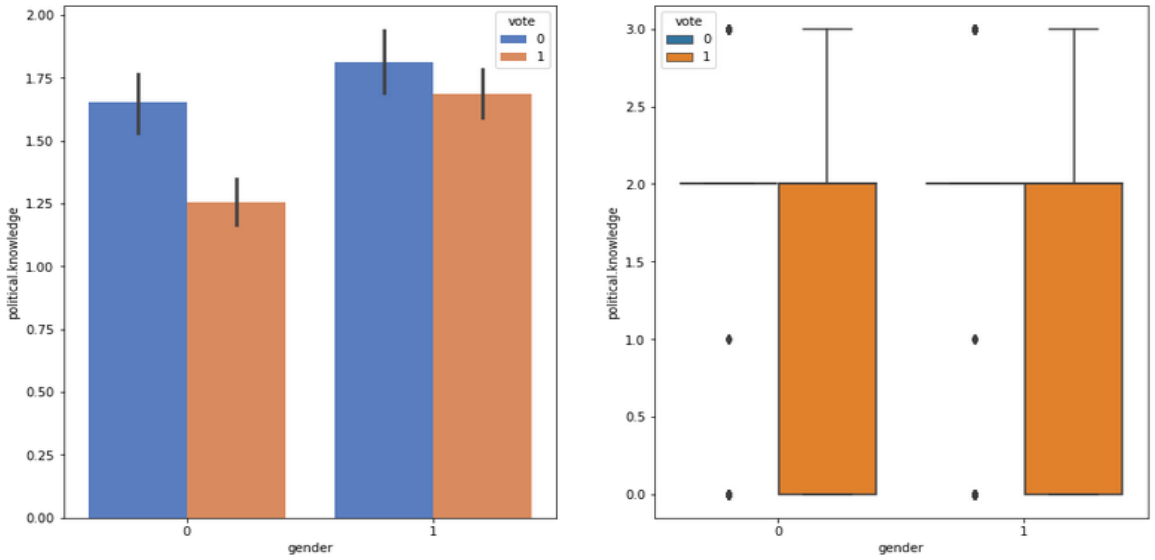


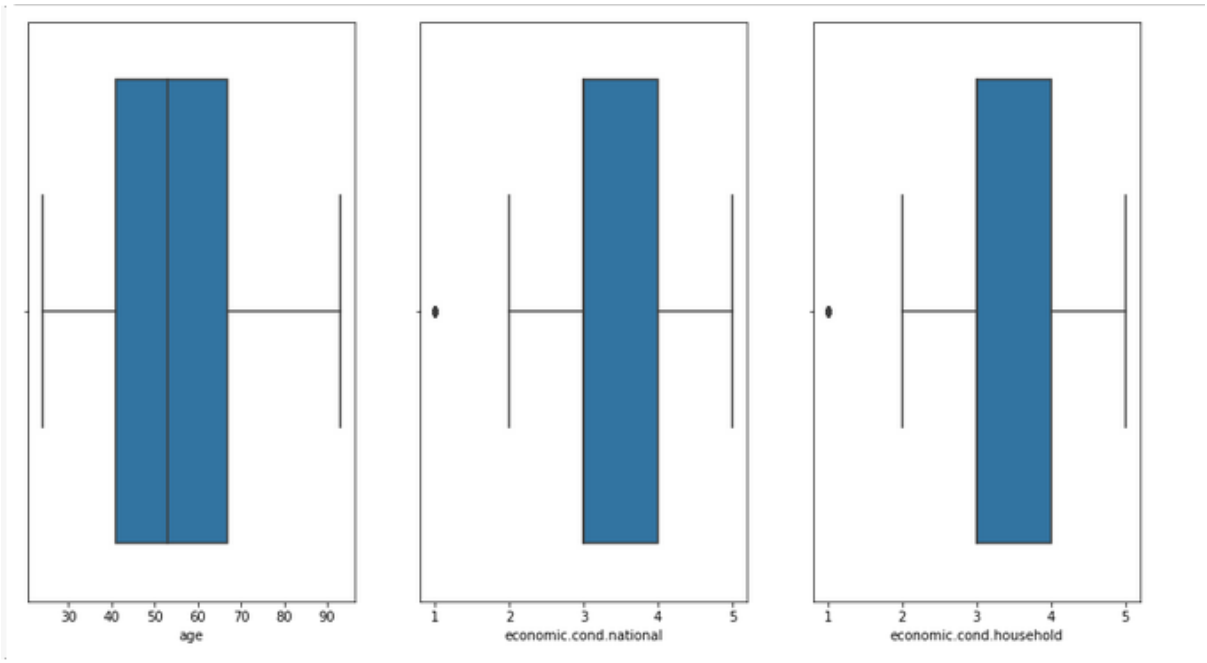
Fig. 1.4 is Bar plot and Box plot for Europe and political knowledge.



- From the above box plots we understand the data set has very minimal outliers hence uniformly distributed and clean data.
- We see Europe and economic cond national has few outliers which can be ignored due to the number.
- We have performed the plots keeping the Gender, male and female as hue.
- We see the distribution in bar plot is almost the same for Male and Female.
- In the last plot we are using gender as hue and variable we can see political knowledge is high for men when compared to women, however there is not much difference.
- The plot for Hauge and Europe the females are in the higher end when compared to men. But it varies in few plots.
- The box plots show's a clear picture of the comparison between the variables considering the gender as hue.

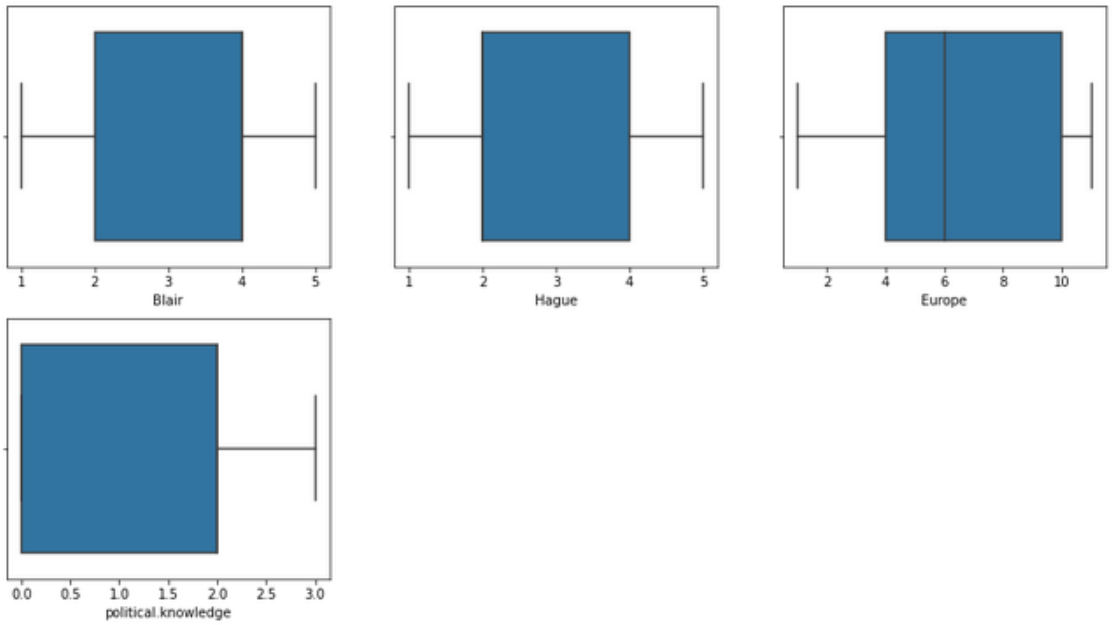
Checking for outliers:

Figure 1.5: Box plots of age, economic.cond.national, economic.cond.household



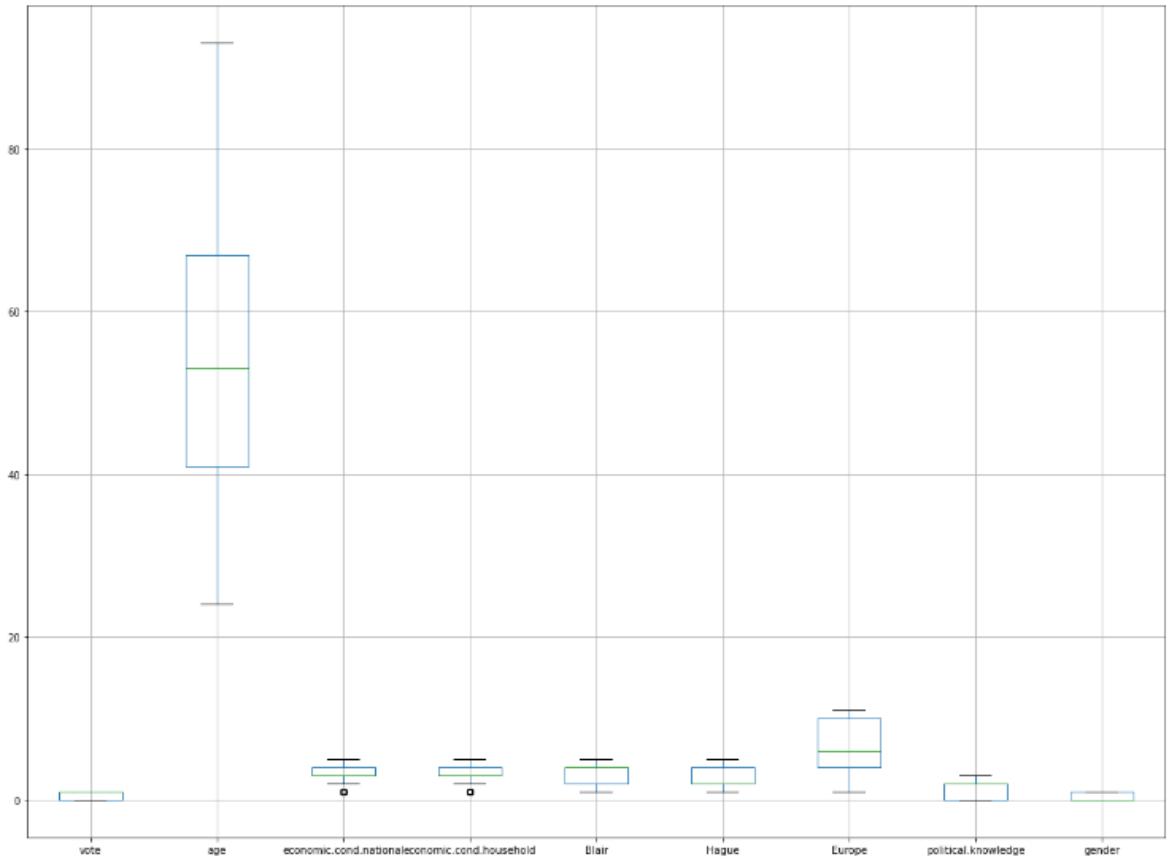
- While we check outliers for these variables as we see can see there are only few outliers in the variables Economic condition national and Economic condition household.

Figure 1.5: Box plots for Blair, Hague, Europe and political.knowledge



- There are no outliers in these variables Blair, Hague, Europe and political.knowledge hence we can understand the data set is clean.
- We can also see the age variable has no outliers.
- The above histogram and box plots represent positive skewness.
- Political knowledge has the highest skewness and zero negative skewness.
- There is hardly negative skewness in the data set.

Removing outliers:

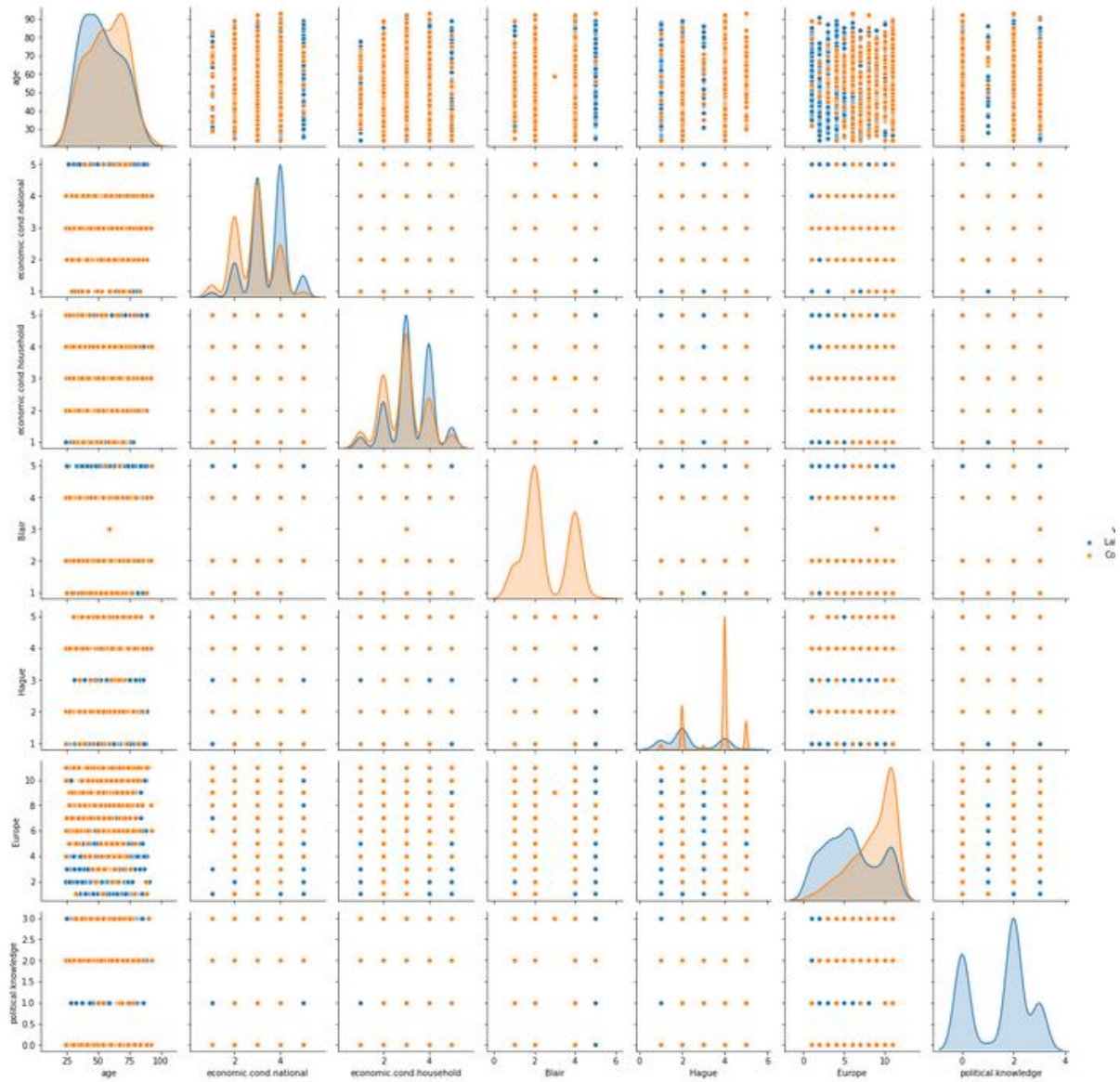


- Here we are treating the outliers from the variables economic.cond.national, economic.cond.household.

- We see now the data is almost equally skewed and there are not much of differences.

Multivariate Analysis

Pair plot:



- In the above diagram scatter plots are plotted for all the numerical columns in the dataset. A scatter plot is a visual representation of the degree of correlation between any two columns. The pair plot function in seaborn makes it very easy to generate joint scatter plots for all the columns in the data.

- Pair plot shows scatter plot as well as the histogram between all the variables of the dataset.
- Pair plot makes it visually easier to understand if the data is highly co related to each other or not.
- In the above pair plot we are using hue as **vote** as it is the target variable.

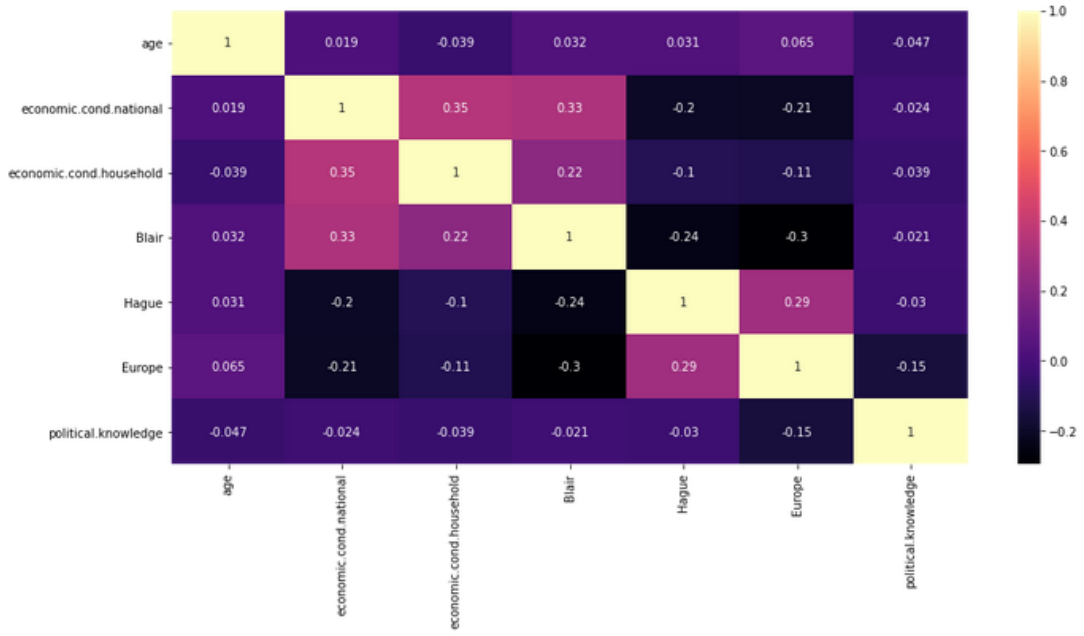
Correlation Matrix

Table 1.5: Correlation matrix of the Election data set.

	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge
age	1.000000	0.018687	-0.038868	0.032084	0.031144	0.064562	-0.046598
economic.cond.national	0.018687	1.000000	0.347687	0.326141	-0.200790	-0.209150	-0.023510
economic.cond.household	-0.038868	0.347687	1.000000	0.215822	-0.100392	-0.112897	-0.038528
Blair	0.032084	0.326141	0.215822	1.000000	-0.243508	-0.295944	-0.021299
Hague	0.031144	-0.200790	-0.100392	-0.243508	1.000000	0.285738	-0.029906
Europe	0.064562	-0.209150	-0.112897	-0.295944	0.285738	1.000000	-0.151197
political.knowledge	-0.046598	-0.023510	-0.038528	-0.021299	-0.029906	-0.151197	1.000000

- A correlation matrix is simply a table which displays the correlation. The measure is best used in variables that demonstrate a linear relationship between each other.
- To summarize a large amount of data where the goal is to see patterns. In our example above, the observable pattern is that all the variables are not highly correlate with each other.
- To input into other analyses. For example, people commonly use correlation matrixes as inputs for exploratory factor analysis, confirmatory factor analysis, structural equation models, and linear regression when excluding missing values pairwise.

Heat map:



- A heat map (or heatmap) is a graphical representation of data where values are depicted by color. Heat maps make it easy to visualize complex data and understand it at a glance:
- As a visual tool, heat maps help you make informed, data-based decisions for A/B testing, updating, or (re)designing your website. And they are also useful on a wider business scale: heat maps let you show team members and stakeholders what's happening and get their buy-in more easily when changes are needed.
- In the above heatmap we can see political knowledge is not a correlated data to age.
- The highest correlation is between economic cond national and economic cond household.

1.3) Encode the data (having string values) for Modelling. Is Scaling necessary here or not? (3 pts), Data Split: Split the data into train and test (70:30) (2 pts).

Answer:

Encoding:

Table 1.6: Encoding of the Election data set.

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

- **Encoding** keeps your data safe since the files are not readable unless you have access to the algorithms that were used to encode it. This is a good way to protect your data from theft since any stolen files would not be usable.

- Encoded data is easy to organize, even if the original data was mostly unstructured.
- We are encoding the data using Label Encoder we have used the age and vote variables for encoding the data set.

Splitting data into Train and test 70:30 ratio:

Table 1.7: Splitting data into train and test of the Election data set.

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

- We have split the data into train and test using Vote as it is the target variable.
- We see the data set vote column is removed and we have the remaining columns.

Scaling:

Table 1.8: Scaling of the train and test Election data set.

x_train_ml

```
array([[-1.29671043, -1.45558149,  0.90210034, ...,  1.33208942,
         0.45223123, -0.93695043],
       [-0.91033745,  0.87730667, -0.16374427, ..., -0.20215599,
        -1.4075259 ,  1.06729232],
       [ 0.44196795,  0.87730667, -0.16374427, ...,  0.10469309,
         0.45223123, -0.93695043],
       ...,
       [ 1.2791094 ,  0.87730667, -0.16374427, ...,  1.33208942,
        -1.4075259 , -0.93695043],
       [-1.48989691, -0.28913741, -0.16374427, ..., -0.20215599,
        -1.4075259 , -0.93695043],
       [ 2.24504183, -0.28913741,  1.96794496, ..., -1.7364014 ,
        -1.4075259 ,  1.06729232]])
```

x_test_ml

```
array([[ 1.08592291, -0.28913741, -0.16374427, ...,  0.41154217,
         0.45223123, -0.93695043],
       [-0.71715097, -0.28913741, -1.22958889, ...,  0.41154217,
         1.3821098 ,  1.06729232],
       [ 2.24504183,  2.04375075,  1.96794496, ..., -1.7364014 ,
         0.45223123,  1.06729232],
       ...,
       [ 2.18064633, -0.28913741, -0.16374427, ..., -0.20215599,
        -1.4075259 , -0.93695043],
       [-0.07319601,  0.87730667, -0.16374427, ...,  1.02524033,
        -1.4075259 , -0.93695043],
       [-1.16791943, -0.28913741, -0.16374427, ..., -0.20215599,
         1.3821098 ,  1.06729232]])
```

- We will be scaling the data Feature scaling is essential for machine learning algorithms that calculate distances between data. ... Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions do not work correctly without normalization.

- Here we have taken the train and test data and scaled them by renaming the variable as x_train_ml and x_test_ml

- We will be using the scaled data only for KNN model as other models don't need scaled data set.

1.4) Apply Logistic Regression and LDA (Linear Discriminant Analysis) (3 pts). Interpret the inferences of both models (2 pts)

Answer:

Logistics regression:

In statistics, the logistic model is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc.

Table 1.8: Predicting on Training and Test dataset

ytrain_predict_prob

	0	1
0	0.931825	0.068175
1	0.096984	0.903016
2	0.298416	0.701584
3	0.110210	0.889790
4	0.017223	0.982777

ytest_predict_prob

	0	1
0	0.424284	0.575716
1	0.148426	0.851574
2	0.007187	0.992813
3	0.836350	0.163650
4	0.068407	0.931593

- The a above tables talk about Ytrain and Ytest predictions.
- We can see the train mode is predicting in 0.93 for 0 and 0.98 for 1 as the best prediction.
- In the other hand we have ytest predicting 0.83 for 0 and 0.99 for 1 as the best prediction.
- We can see the model is an overfit as the values are higher than +-10.

LDA: Linear Discriminant Analysis:

LDA is a simple model in both preparation and application. There is some interesting statistics behind how the model is setup and how the prediction equation is derived, but is not covered in this post.

pred_prob_train

	0	1
0	0.949216	0.050784
1	0.078241	0.921759
2	0.307389	0.692611
3	0.078963	0.921037
4	0.012161	0.987839

pred_prob_test

	0	1
0	0.462093	0.537907
1	0.133955	0.866045
2	0.006414	0.993586
3	0.861210	0.138790
4	0.056545	0.943455

- These statistical properties are estimated from your data and plug into the LDA equation to make predictions. These are the model values that you would save to file for your model.
- In the above model we see prediction for prob train is performing at 0.94 for 0 and 0.98 for 1 are the highest predictions.
- The Prediction prob test is performing at 0.86 for 0 and 0.99 for 1 as the highest predictions.
- We can see the LDA and logistics regression models are peforming almost at the same range there is not much of difference hence both the models are performing extremly good.
- The LDA is preferably better than the Logistic regression as the test data is better.

1.5) Apply KNN Model and Naïve Bayes Model(5 pts). Interpret the inferences of each model (2 pts)

Answer:

KNN Model:

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems.

A **supervised machine learning** algorithm (as opposed to an unsupervised machine learning algorithm) is one that relies on labelled input data to learn a function that produces an appropriate output when given new unlabelled data.

Xtrain model:

```
0.29029217719132894
```

Xtest model:

```
0.3399122807017544
```

- The trainning model for KNN is performing at 0.29 and test model is performing at 0.33.
- For a very low value of k (suppose k=1), the model overfits on the training data, which leads to a high error rate on the validation set. On the other hand, for a high value of k, the model performs poorly on both train and validation set. If you observe closely, the validation error curve reaches a minima at a value of k = 9. This value of k is the optimum value of the model (it will vary for different datasets).
- This curve is known as an '**elbow curve**' (because it has a shape like an elbow) and is usually used to determine the k value.

Naïve Bayes Model:

In statistics, naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naïve) independence assumptions between the features. They are among the simplest Bayesian network models,[1] but coupled with kernel density estimation, they can achieve higher accuracy levels.

Naïve Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem.

Xtrain model:

```
0.8350612629594723
```

Xtest model:

```
0.8223684210526315
```

- The trainning model for Naïve Bayes model is performing at 0.83 and test model is performing at 0.82.
- Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.
- Naive Bayes uses a similar method to predict the probability of different class based on various attributes. This algorithm is mostly used in text classification and with problems having multiple classes.
- As we see both train and test values are near to 1 the model is predicting very well.
- When we compare both the models Naïve Bayes model is performing comparatively better than the k-nearest neighbors.

1.6) Model Tuning (2 pts) , Bagging (2.5 pts) and Boosting (2.5 pts)

Answer:

Model Tuning:

First we will be performing GridSearchCV:

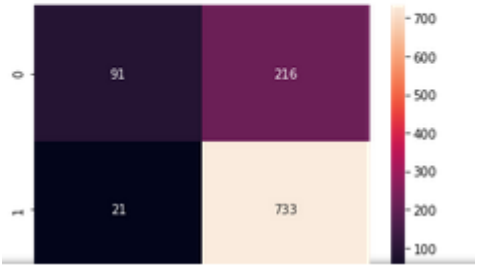
Grid search is an approach to parameter tuning that will methodically build and evaluate a model for each combination of algorithm parameters specified in a grid. The recipe below evaluates different alpha values for the Ridge Regression algorithm on the standard diabetes dataset. This is a one-dimensional grid search.

```
GridSearchCV(cv=3, estimator=LogisticRegression(max_iter=10000, n_jobs=2),
             n_jobs=-1,
             param_grid={'penalty': ['l2', 'none'],
                         'solver': ['newton-cg', 'none'],
                         'tol': [0.001, 0.0001]},
             scoring='f1')
```

Model Tunning for Logistic Regression:

```
0.1

Accuracy Score 0.7465
F1 Score 0.8476

Confusion Matrix
```


```
0.2

Accuracy Score 0.7766
F1 Score 0.8608

Confusion Matrix
```



0.3

Accuracy Score 0.7983
F1 Score 0.8709

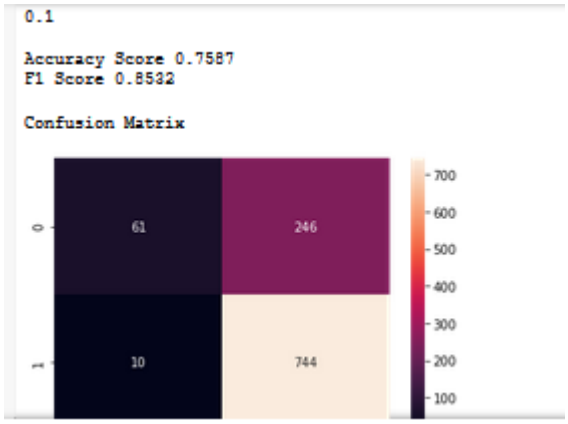
0.4

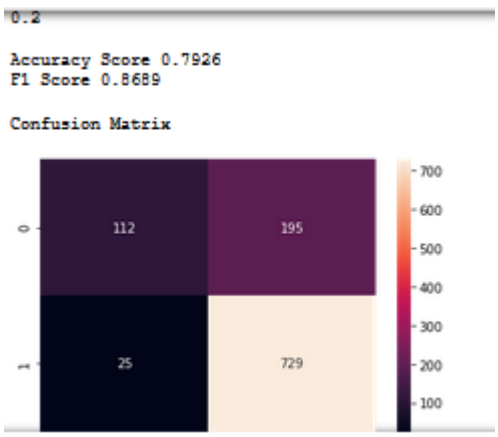
Accuracy Score 0.8294
F1 Score 0.8872

0.9

Accuracy Score 0.6711
F1 Score 0.707

LDA model tunning:





0.3

Accuracy Score 0.8134
F1 Score 0.8787

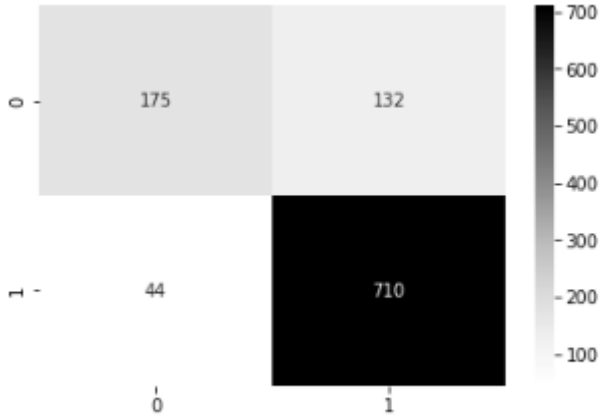
Confusion Matrix

```
0.4

Accuracy Score 0.8341
F1 Score 0.8897

Confusion Matrix
```

```
0.9

Accuracy Score 0.6975
F1 Score 0.7384
```

- We see the model tuning for Logistic regresstion and LDA are quite similar the accuracy score is high for 0.1 and least for 0.9 in both the models.
- The accuracy score is varying only in one or two percent for 0.3, 0.4 and the mid range values.
- The F1 score is increasing for the values 0.1 to 0.2, however it has decrease to 0.73 for LDA and 0.70 for Logistics regression for 0.9.
- The highest F1 score for 0.4 0.8846 for logistic regression and 0.8897 for 0.4 for LDA model.
- Hence we can see the best F1 score are achived for both the models at 0.4.

Confusion matrix for y_train and data_predict_custom_cutoff_train_lda

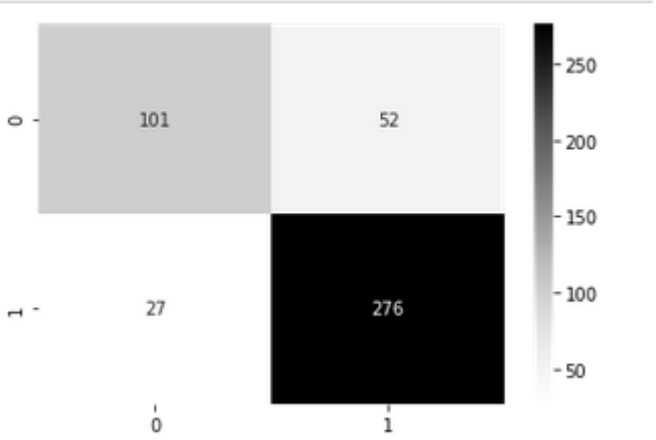


KNN Model tunning:

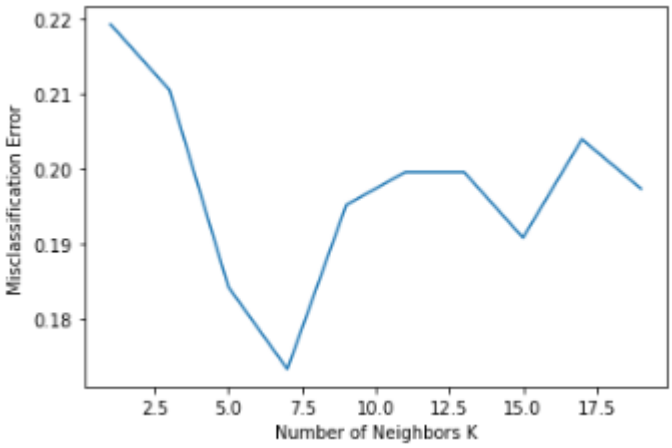
Checking for accuracy matrix:

```
[0.2192982456140351,
 0.21052631578947367,
 0.1842105263157895,
 0.17324561403508776,
 0.19517543859649122,
 0.19956140350877194,
 0.19956140350877194,
 0.1907894736842105,
 0.20394736842105265,
 0.19736842105263153]
```

y_test,data_pred_custom_cutoff heatmap confsuion matrix



Ploting misclassification error vs KNN



- In the above heatmap for LDA we can see the values for Confusion matrix for y_train and data_predict_custom_cutoff_train_lda are 175 and 44 for 0 and 132 and 44 for 1.
- In the above heatmap for KNN we can see the values for y_test,data_pred_custom_cutoff heatmap confsuion matrix are 101 and 27 for 0 and 52 and 276 for 1.
- The miss classification error vs KNN graph represents what is the percentage where the pepole are voting.
- We are using scaled data for model tunning on KNN model.

Boosting Model Tunning for Gradient:

The accuracy score is 0.89 for y_train_predict.

```
0.8925541941564562
[[239  68]
 [ 46 708]]
```

The accuracy score is 0.89 for y_test_predict.

```
0.8355263157894737
[[105  48]
 [ 27 276]]
```

Boosting Model Tunning for ADA:

The accuracy score is 0.85 for y_train_predict.

```
0.8501413760603205
[[214  93]
 [ 66 688]]
```

The accuracy score is 0.81 for y_test_predict.

```
0.8135964912280702
[[103  50]
 [ 35 268]]
```

- If we compare the Gradient and Ada model for Boosting we can see the Ada model is performing well as the values are nearer to +_10% and the model is on overfitting.
- The gradient model is also performing fine as there is not much of a difference between both the models.
- We are using the normal data for the model tunning.

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 (4 pts) Final Model - Compare all models on the basis of the performance metrics in a structured tabular manner. Describe on which model is best/optimized (3 pts)

Answer:

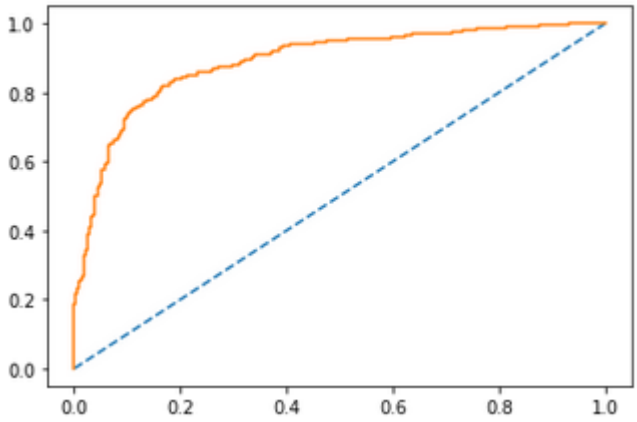
Logistics regresstion performance matrix:

Model score for X_train, y_train:

```
0.8312912346842601
```

AUC Curve for X_train, y_train

```
AUC: 0.890
```

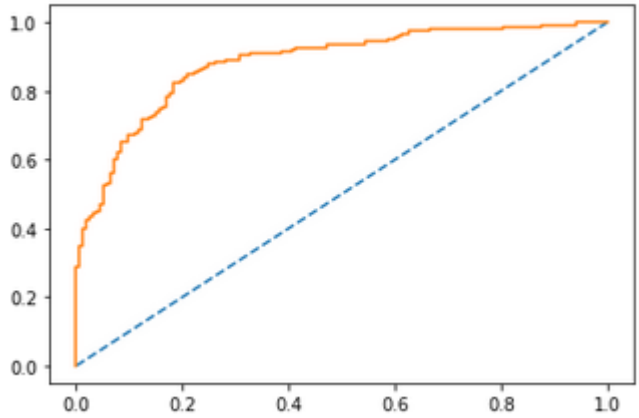


Model score for X_test, y_test

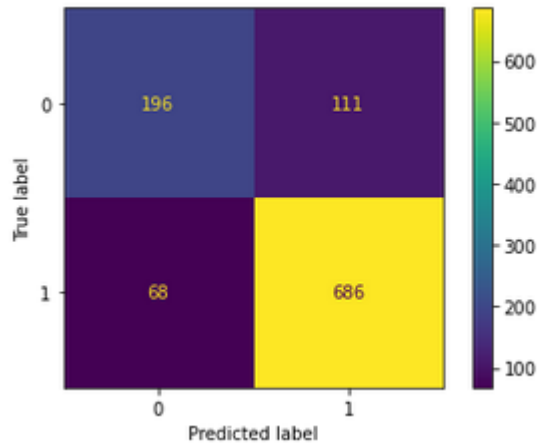
```
0.8355263157894737
```

AUC Curve for X_test, y_test

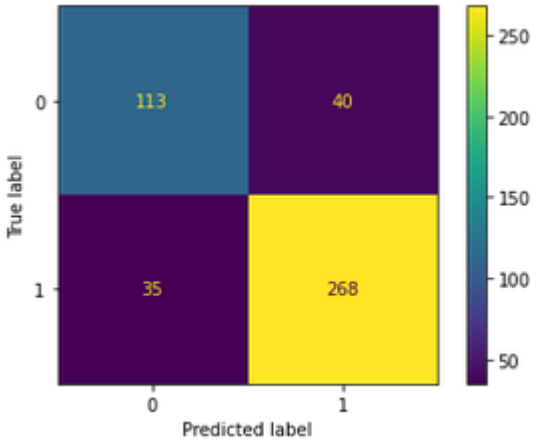
```
AUC: 0.890
```



Confusion matrix for model,X_train,y_train



Confusion matrix for model, X_test, y_test



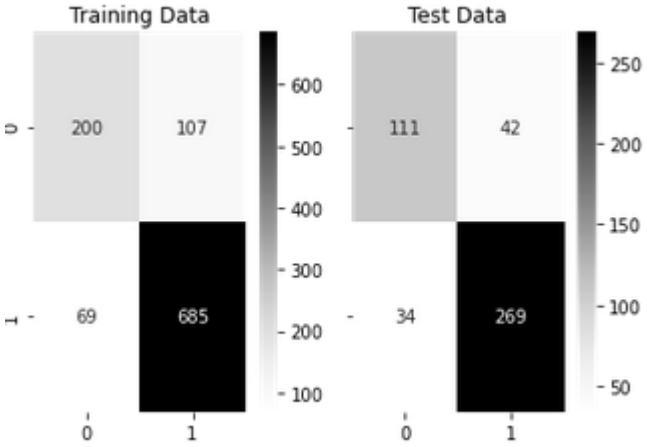
Performance matrix for train and test data:

	precision	recall	f1-score	support
0	0.74	0.64	0.69	307
1	0.86	0.91	0.88	754
accuracy			0.83	1061
macro avg	0.80	0.77	0.79	1061
weighted avg	0.83	0.83	0.83	1061

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

LDA(Linear Discriminant Analysis) performance matrix:

Confusion matrix for train and test data:



Performance matrix for train and test data:

Classification Report of the training data:

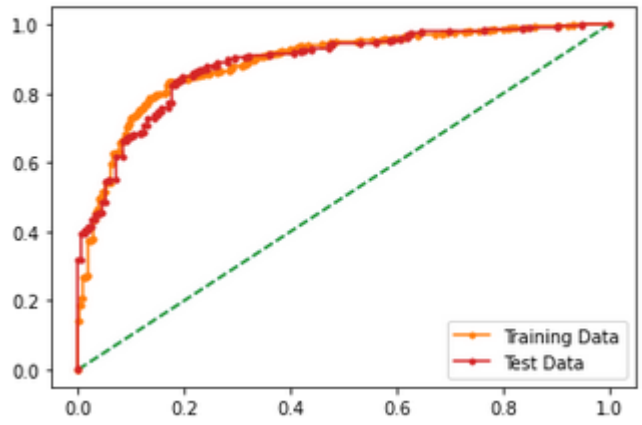
	precision	recall	f1-score	support
0	0.74	0.65	0.69	307
1	0.86	0.91	0.89	754
accuracy			0.83	1061
macro avg	0.80	0.78	0.79	1061
weighted avg	0.83	0.83	0.83	1061

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.77	0.73	0.74	153
1	0.86	0.89	0.88	303
accuracy			0.83	456
macro avg	0.82	0.81	0.81	456
weighted avg	0.83	0.83	0.83	456

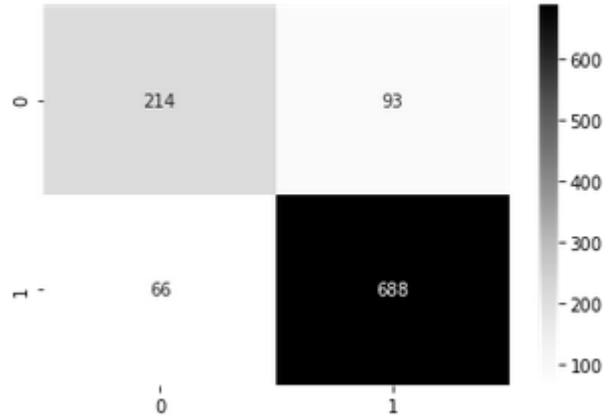
AUC ROC plot for Train and Test.

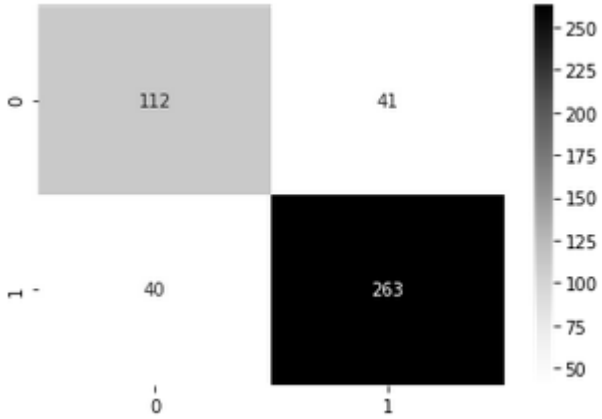
```
AUC for the Training Data: 0.889
AUC for the Test Data: 0.888
```



KNN Model Classification report:

Confusion matrix for train and test scaled data:





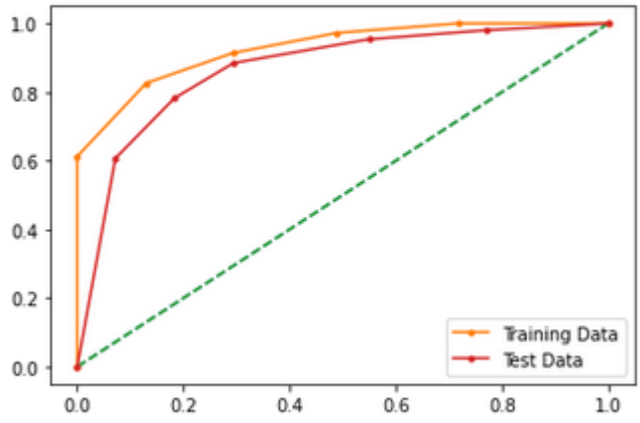
Performance matrix for train and test scaled data:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	307
1	0.71	1.00	0.83	754
accuracy			0.71	1061
macro avg	0.36	0.50	0.42	1061
weighted avg	0.51	0.71	0.59	1061

	precision	recall	f1-score	support
0	0.00	0.00	0.00	153
1	0.66	1.00	0.80	303
accuracy			0.66	456
macro avg	0.33	0.50	0.40	456
weighted avg	0.44	0.66	0.53	456

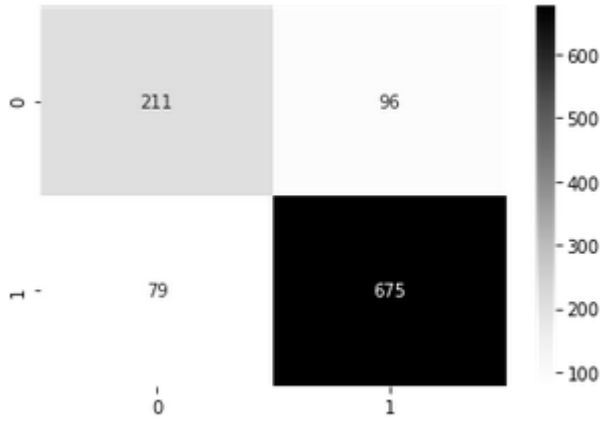
AUC ROC plot for Train and Test scaled data.

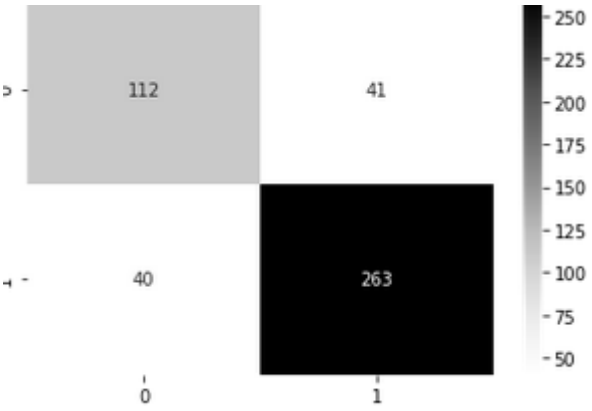
```
AUC for the Training Data: 0.928
AUC for the Test Data: 0.867
```



Naïve Bayes Model Classification report:

Confusion matrix for train and test data:



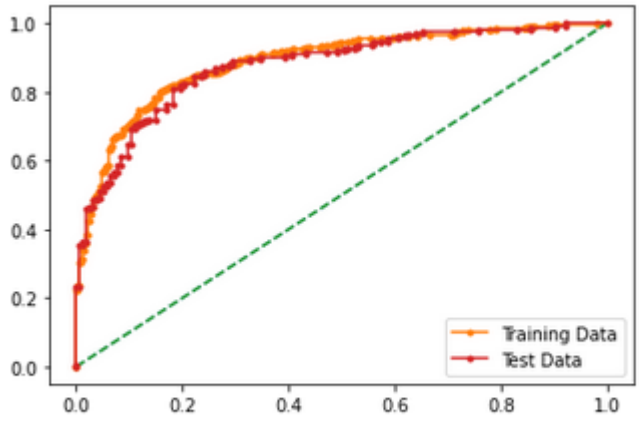


Performance matrix for train and test data:

	precision	recall	f1-score	support
0	0.73	0.69	0.71	307
1	0.88	0.90	0.89	754
accuracy			0.84	1061
macro avg	0.80	0.79	0.80	1061
weighted avg	0.83	0.84	0.83	1061

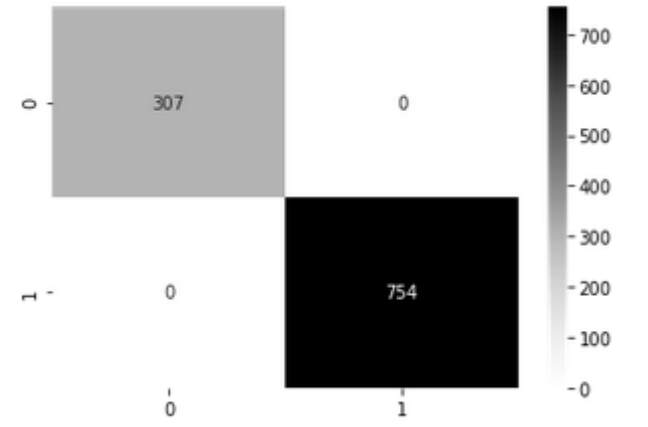
	precision	recall	f1-score	support
0	0.74	0.73	0.73	153
1	0.87	0.87	0.87	303
accuracy			0.82	456
macro avg	0.80	0.80	0.80	456
weighted avg	0.82	0.82	0.82	456

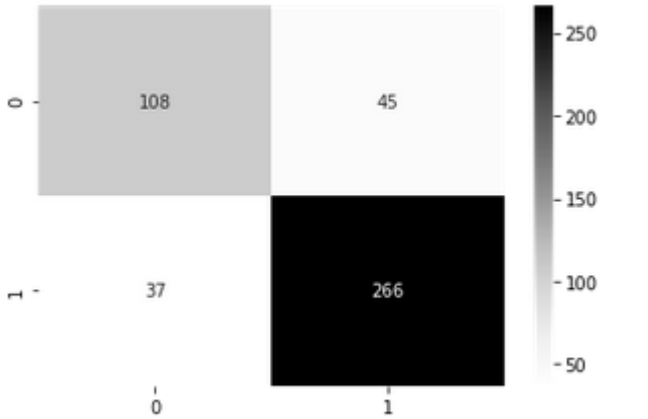
AUC ROC plot for Train and Test scaled data.



Bagging classification report:

Confusion matrix for train and test scaled data:





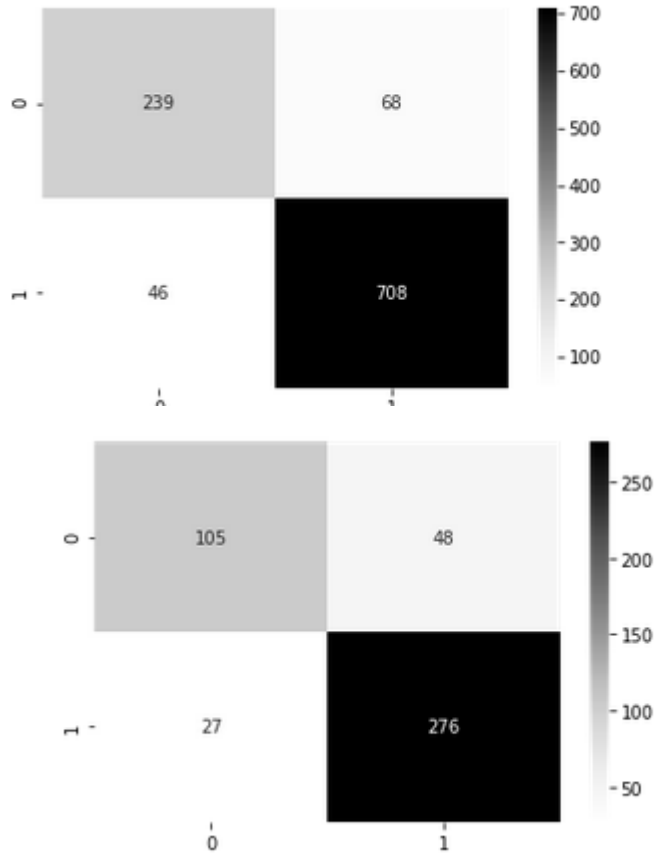
Performance matrix for train and test data:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	307
1	1.00	1.00	1.00	754
accuracy			1.00	1061
macro avg	1.00	1.00	1.00	1061
weighted avg	1.00	1.00	1.00	1061

	precision	recall	f1-score	support
0	0.74	0.71	0.72	153
1	0.86	0.88	0.87	303
accuracy			0.82	456
macro avg	0.80	0.79	0.80	456
weighted avg	0.82	0.82	0.82	456

Boosting Gradient Classification report:

Confusion matrix for train and test scaled data:



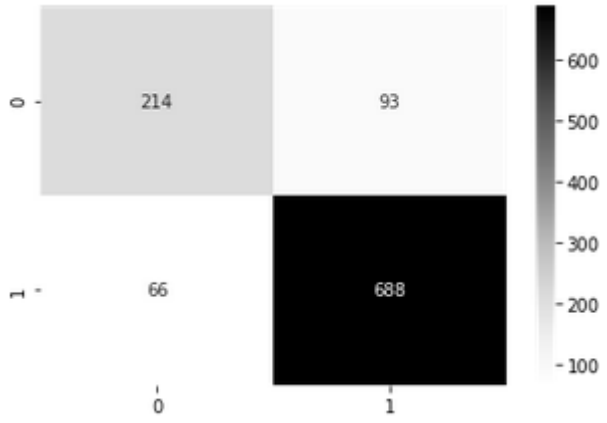
Performance matrix for train and test data:

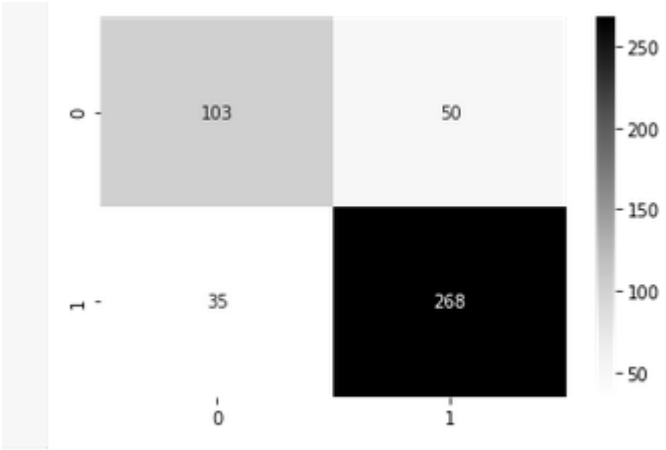
	precision	recall	f1-score	support
0	0.73	0.69	0.71	307
1	0.88	0.90	0.89	754
accuracy			0.84	1061
macro avg	0.80	0.79	0.80	1061
weighted avg	0.83	0.84	0.83	1061

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

Boosting Ada:

Confusion matrix for train and test scaled data:





Performance matrix for train and test data:

	precision	recall	f1-score	support
0	0.73	0.69	0.71	307
1	0.88	0.90	0.89	754
accuracy			0.84	1061
macro avg	0.80	0.79	0.80	1061
weighted avg	0.83	0.84	0.83	1061

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

- Above are the models Logistics regression, LDA, KNN, Naïve Bayes, Boosting(Gradient and ADA) and Bagging
- Based on the analysis with all the models we can prove the Gradient Boost model is performing the best with acuracy score 0.84 and F1 score of 0.71 and 0.89 for 0 and 1 respectively.
- We see all models are peforming in the same range.
- The logistic regresstion mode is giving accuracy 0.83 and F1 score for 0 and 1 0.69 and 0.88 respectively.
- The above Tabs have clear indication of accuracy and F1 score to check which model is performing the best and it is not a overfit or underfit.
- We can also see Bagging is a overfit model as it has the higest values for accuracy.

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.

- Based on the analysis with all the models we can prove the Gradient Boost model is performing the best with acuracy score 0.84 and F1 score of 0.71 and 0.89 for 0 and 1 respectively.
- We can see most of the models are performing on the same lines.
- Bagging is a overfit model as we can the accuracy is 1 for the train data and 0.82 for the test data.

- Looking at the models the recommendations we can make for the leading news channel CNBE who wants to analyse recent elections. They can use the Gradient boosting model to analyse the data and understand where should be focusing.

- The variables in the first data set are Vote, age, economic.cond.national, economic.cond.household, Blair,Hague, Europe, political.knowledge and gender While performing EDA we have seen the age and gender play a huge role.
- In India we know the age to vote is 18 hence the age didn't have any outliers.
- The , , economic.cond.national, economic.cond.household where also the factors which effected votes.
- We also found out the ratio of Male and Female to vote was not very different. As there was not huge difference.
- After going through the data set my recommendation would be to focus on the , economic.cond.national, economic.cond.household to understand how can we make use of the pepole under this category.
- We need to focus on pepole from these categories who will come forward to vote regardless of the gender or how rich they are.
- The house hold pepole are normally home makers who can take the time out to go and vote.
- The economical condition house hold are the working class, who should be given options to vote at the nearest booth by carying a digital card.
- The easier we can make the voting system pepole can come forward to vote and this will help the particular party.
- Working for the pepole not only during elections but through out is also a option we should be looking at.
- Understanding the problem the normal public is going through and acknowledging through local chanels or visitng the areas will promote public to vote for that party.

- Concluding my recommendations by saying for a party to Win the leader has to strong and discplined to help the public and we can do the same by using the Gradient boosting mode to check which areas can be focused.

B. Data set 2 text Mining

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

Answer:

We have found the raw data and details for character, word and sentence are given for President Roosevely, Kennedy and Nixon for the same is given below.

```
1941_Roosevelt President document contains: 7571 characters
1941_Roosevelt President document contains: 1536 words
1941_Roosevelt President document contains: 68 sentences

1961_Kennedy President document contains: 7618 characters
1961_Kennedy President document contains: 1917 words
1961_Kennedy President document contains: 52 sentences

1973_Nixon President document contains: 9991 characters
1973_Nixon President document contains: 2028 words
1973_Nixon President document contains: 69 sentences
```

2.2) Remove all the stopwords from the three speeches.

Answer:

We can skip this line since we have already converted all the words to lowercase

Converting all the words to lower case

Only keeping the words which are not the 'stop words

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)

Answer:

The top 3 words for President Roosevely, Kennedy and Nixon

```
[',', '--', 'nation', 'know', 'spirit']
```

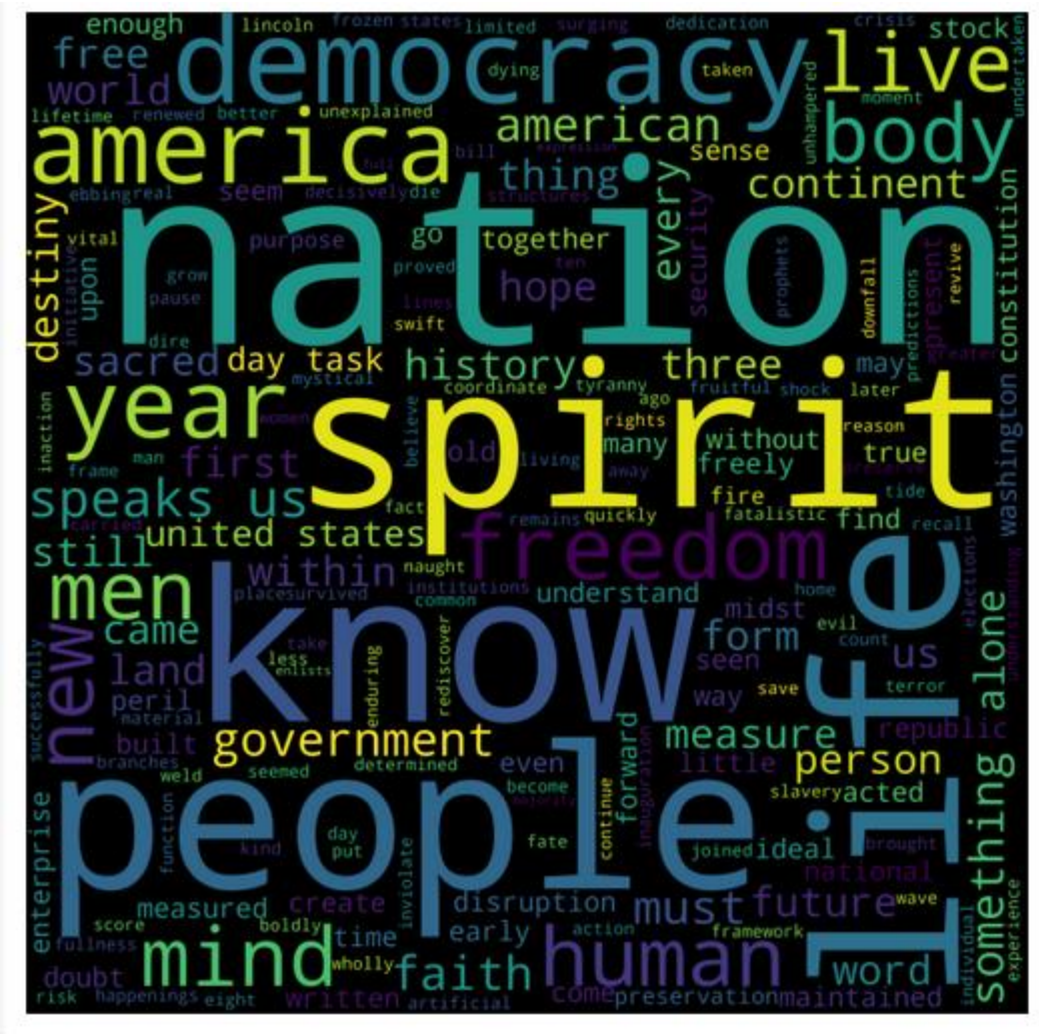
```
[',', '--', 'world', 'may', 'nations']
```

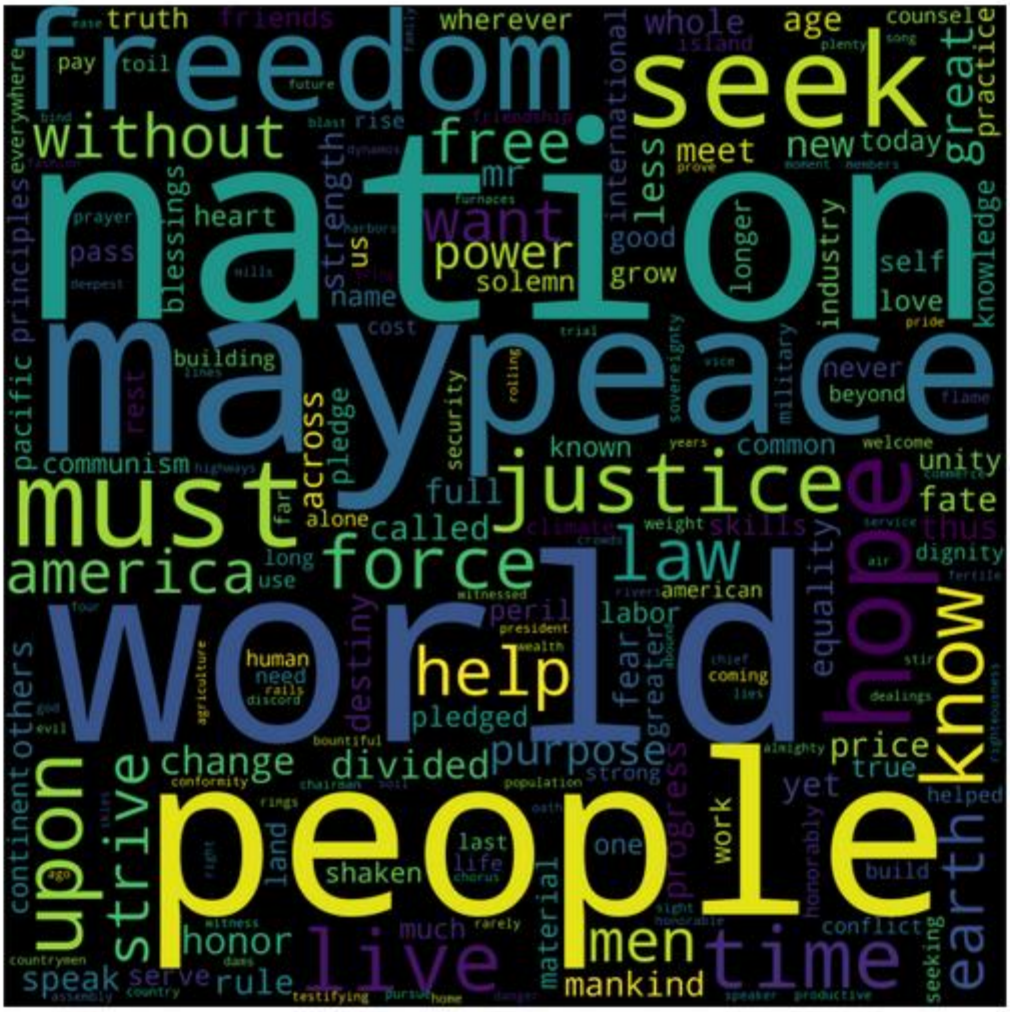
```
[',', 'us', 'let', 'america', 'peace']
```

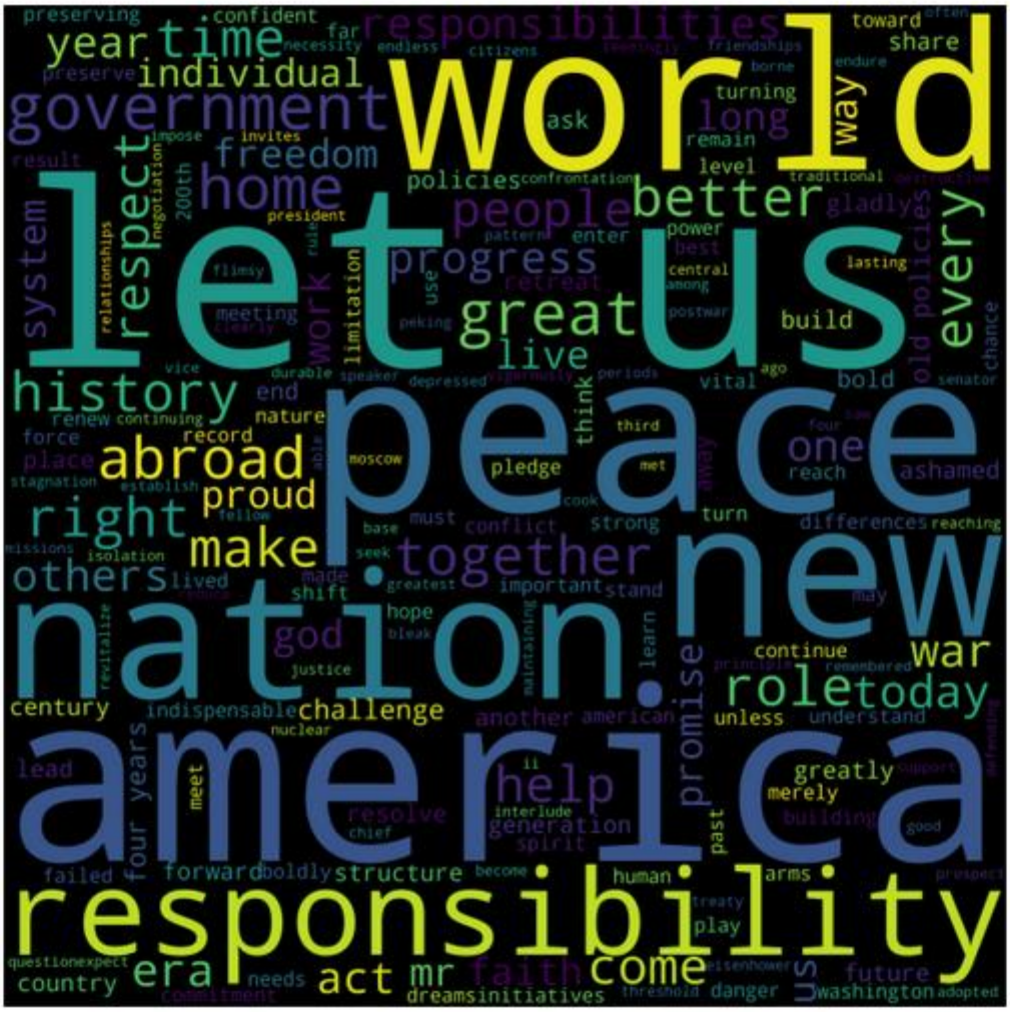
2.4) Plot the word cloud of each of the three speeches. (after removing the stopwords)

Answer:

Word cloud for President Roosevely, Kennedy and Nixon the speeches are:




nation
may
peace
world
people
freedom
seek
must
justice
hope
force
law
help
america
without
free
know
upon
strive
earth
time
live
men
want
power
great
new
meet
today
age
whole
truth
wherever
counsel
practice
pay
toil
heart
pass
blessings
strength
name
solemn
grow
good
less
longer
international
industry
self
love
knowledge
building
across
pledge
security
known
common
unity
fate
communism
called
full
alone
long
use
peril
labor
american
equality
dignity
human
need
pledged
change
divided
purpose
price
true
yet
helped
continent
others
last
one
work
shaken
land
honor
much
material
mankind
conflict
seeking
speak
serve
rule
build
never
beyond
military
sovereignty
principles
pacific
everywhere
destiny
fear
greater
progress
strong
thus
skills
mr
us


world
let us
peace
new
nation
america
responsibility
responsibilities
government
time
year
individual
home
freedom
respect
people
better
progress
great
every
history
abroad
right
proud
make
together
others
role
today
war
help
promise
challenge
century
era
act
mr
come
country
future
washington
live
system
work
one
long
way
share
toward
confident
preserving

ssss