PROJECT ON MACHINE LEARNING

M VALLI RAJA SEKAR

PGPDSBA.O.APR22.C

EXECUTIVE SUMMARY

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

1. ***Vote*: Party choice: Conservative or Labour**
2. ***Age*: in years**
3. ***Economic.cond.national*: Assessment of current national economic conditions, 1 to 5.**
4. ***Economic.cond.household*: Assessment of current household economic conditions, 1 to 5.**
5. ***Blair*: Assessment of the Labour leader, 1 to 5.**
6. ***Hague*: Assessment of the Conservative leader, 1 to 5.**
7. ***Europe*: an 11-point scale that measures respondents' attitudes toward European integration. High scores represent 'Eurosceptic' sentiment.**
8. ***Political*.knowledge: Knowledge of parties' positions on European integration, 0 to 3.**
9. ***Gender*: female or male.**

Data Ingestion

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

This survey was conducted on 1525 voters with 9 variables.The header of the dataset are as follow

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

The shape of the document is 1525,10.We need to drop unnamed:0 as it not convince anything in the data . Now the Header of the Document look alike

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

Now the Shape of the Document is 1525,9.

The info of the Document as follow

```
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   vote                        1525 non-null   object
 1   age                         1525 non-null   int64
 2   economic.cond.national      1525 non-null   int64
 3   economic.cond.household     1525 non-null   int64
 4   Blair                       1525 non-null   int64
 5   Hague                       1525 non-null   int64
 6   Europe                      1525 non-null   int64
 7   political.knowledge         1525 non-null   int64
 8   gender                      1525 non-null   object
dtypes: int64(7), object(2)
memory usage: 107.4+ KB
```

- From info of the Document, we can clearly see that there is no missing Value present in the dataset.
- Expect vote and Gender, all other variable present in the data are int which means that there is no bad data available in the data

```
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
```

No Missing value present in the Data. Hence there is missing 0s present in the Varaible.As only Vote and gender are Categorically after verifying those category we can say that no bad data present in the Data

- Hence we can come to conclusion that there is there is no missing Value and bad data present in the Data
- Next need to Verify Anomalies is present in the Data

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

- In above Data on verifying with Mean, Median,Max and Min.We can conclude that there is no anomalies present in the Dataset.
- Hence there is no null Value present in the Dataset on the Basis of Bad data, Anamolies and Missing Values

Now the most Important Parameter is to check is there any whether any duplicated variable present in the Data or not

On Verifying, the Duplicated Variable there are total of 8 duplicated present in the Dataset after removing those variable the shape of the Document is (1517,9)

The Below mention the Skewness of the Document

```
vote                         0.857014
age                          0.139800
economic.cond.national      -0.238474
economic.cond.household     -0.144148
Blair                       -0.539514
Hague                        0.146191
Europe                      -0.141891
political.knowledge         -0.422928
gender                       0.130929
dtype: float64
```

Inference based on computing statistics and and the Null Value condition check is that

- From above Dataset, We can say that all the variable even present in numerical represent only the Categorical Value.
- The Variable of Europe values vary from 0 to 11
- Blair and Hague vary their values from 1 to 5
- No Missing Value present in data
- Bad Data are verified and there is Bad data available in the Data
- In terms of Anomalies, there is no anomalies present in the Data and verified using the Mean, Median and Business related verification

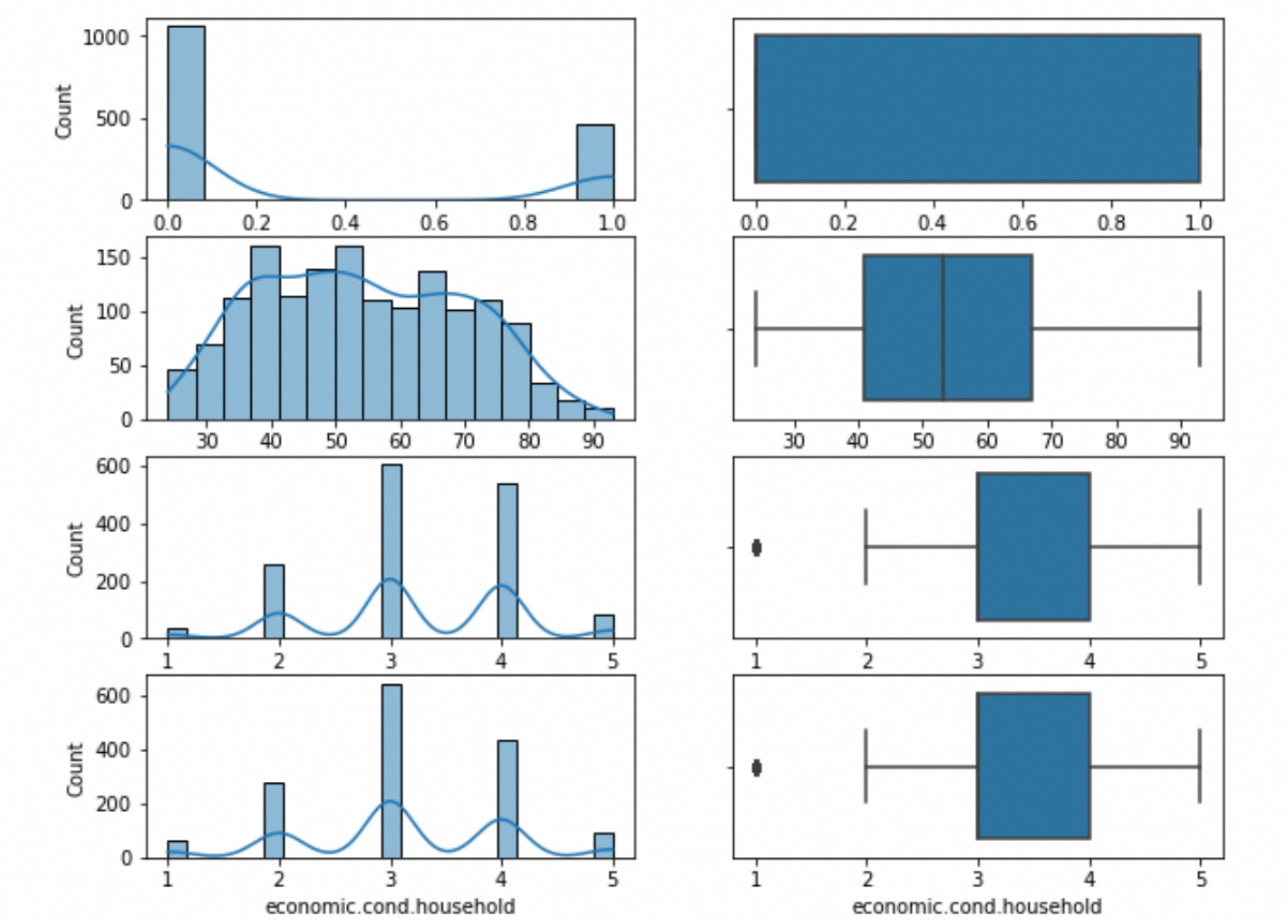
Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers

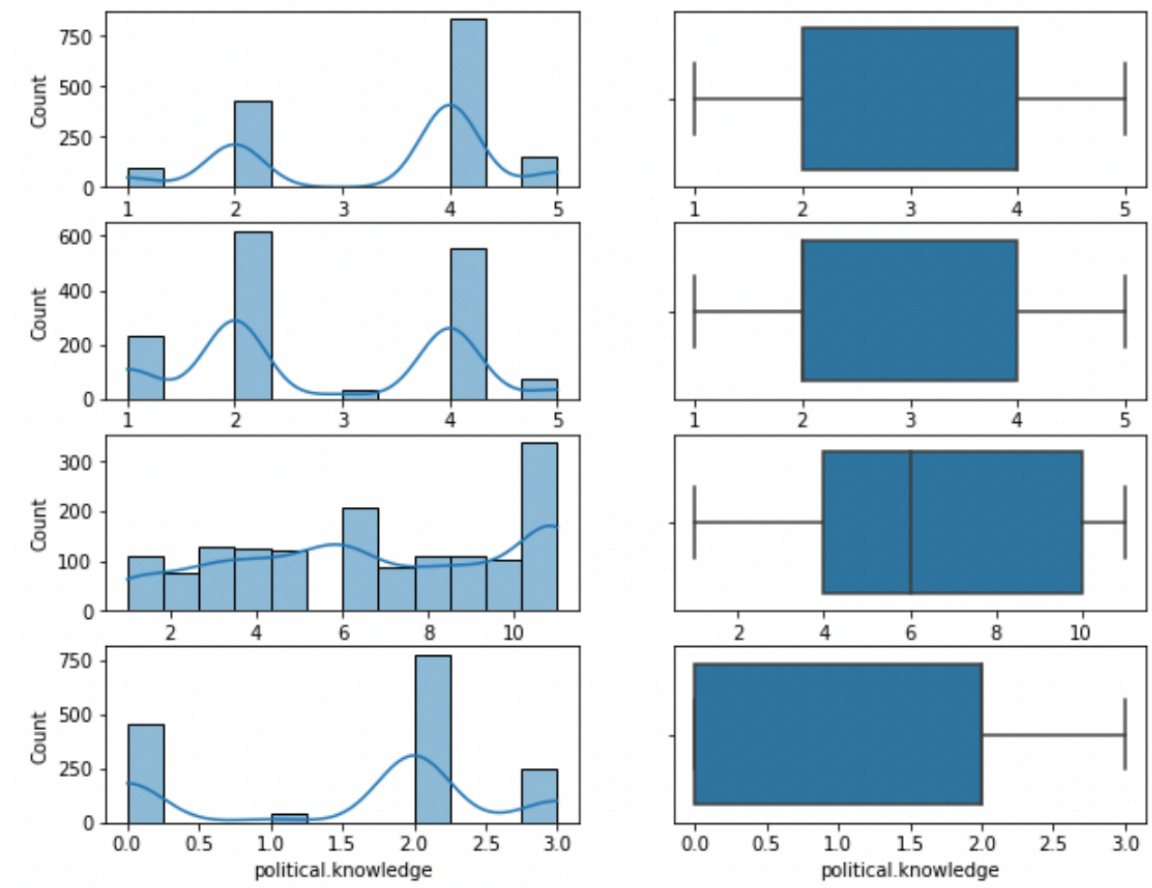
Data Visualization

Univariate Analysis

Univariate data requires to analyze each variable separately. Uni means one, so in other words the data has only one variable.In Univariate Analysis, we can use Histogram and Box plot for Numerical type where as countplot for Categorical type The Histogram and Box Plot Analysis of the eight numeric data are given below

The Histogram and Box Plot Analysis of the eight numeric data are given below

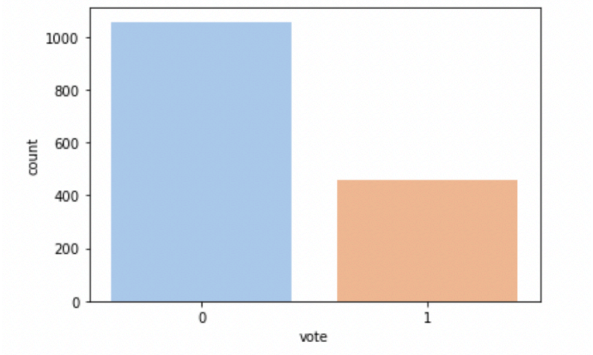




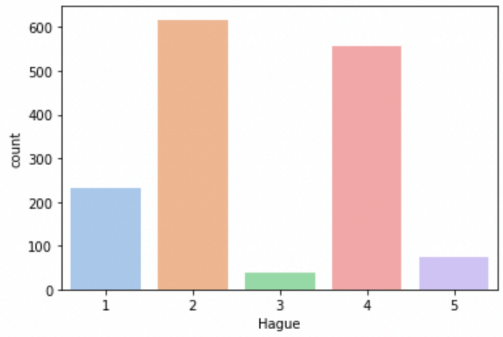
The Eight Numerical Data is presented in the form of Histogram and Boxplot and from the above observation we can clearly see that there is lots of Outlier present in the Data and also Histogram are not much Normally Distributed. Hence Treatment of Outliers and Standardisation is required that to based on the Scenario

Bar graph/count plot for all the variable in Categorical along with their Value count

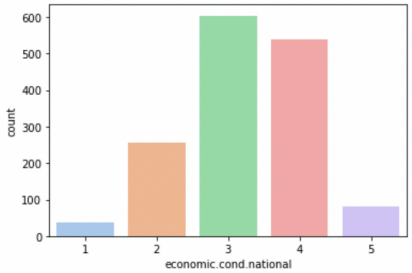
```
0    1057
1     460
Name: vote, dtype: int64
```



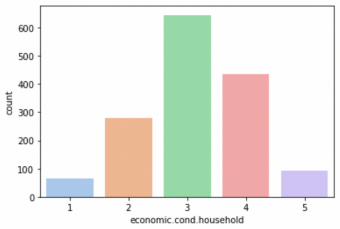
```
2    617
4    557
1    233
5     73
3     37
Name: Hague, dtype: int64
```



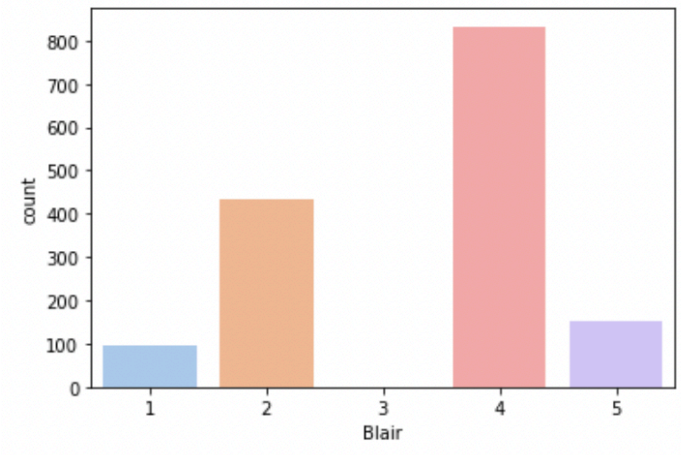

```
3    604
4    538
2    256
5     82
1     37
Name: economic.cond.national, dtype: int64
```



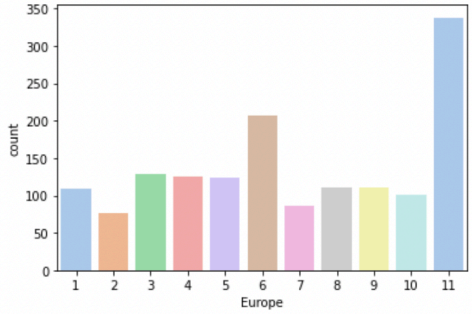

```
3    645
4    435
2    280
5     92
1     65
Name: economic.cond.household, dtype: int64
```



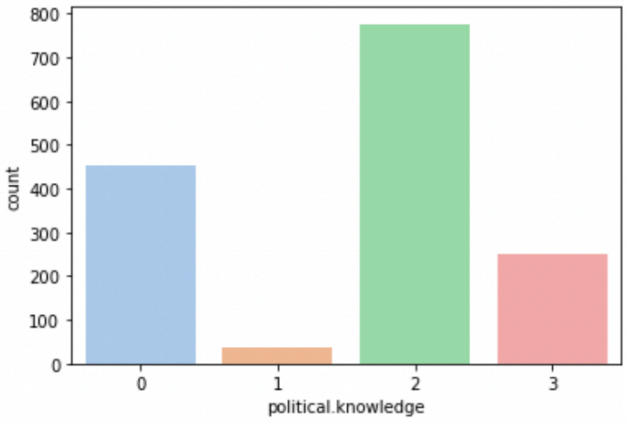

```
4    833
2    434
5    152
1     97
3      1
Name: Blair, dtype: int64
```



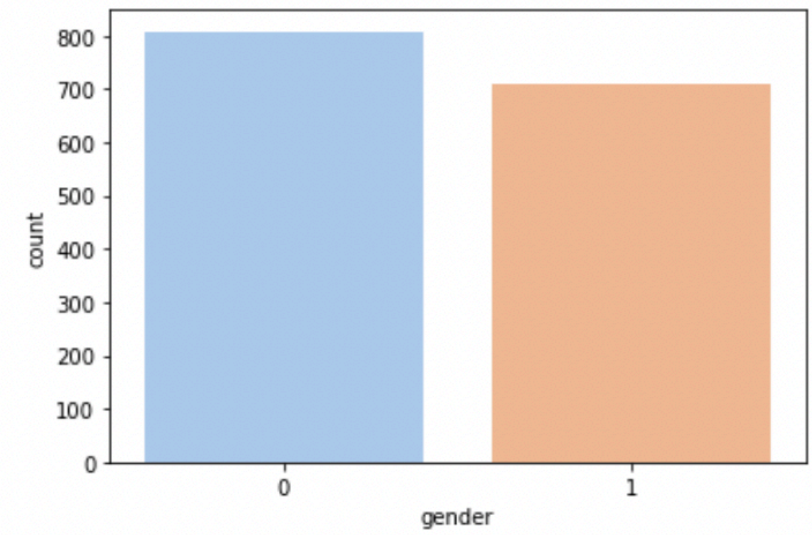
```
11    338
6     207
3     128
4     126
5     123
9     111
8     111
1     109
10    101
7      86
2      77
Name: Europe, dtype: int64
```




```
2    776
0    454
3    249
1     38
Name: political.knowledge, dtype: int64
```

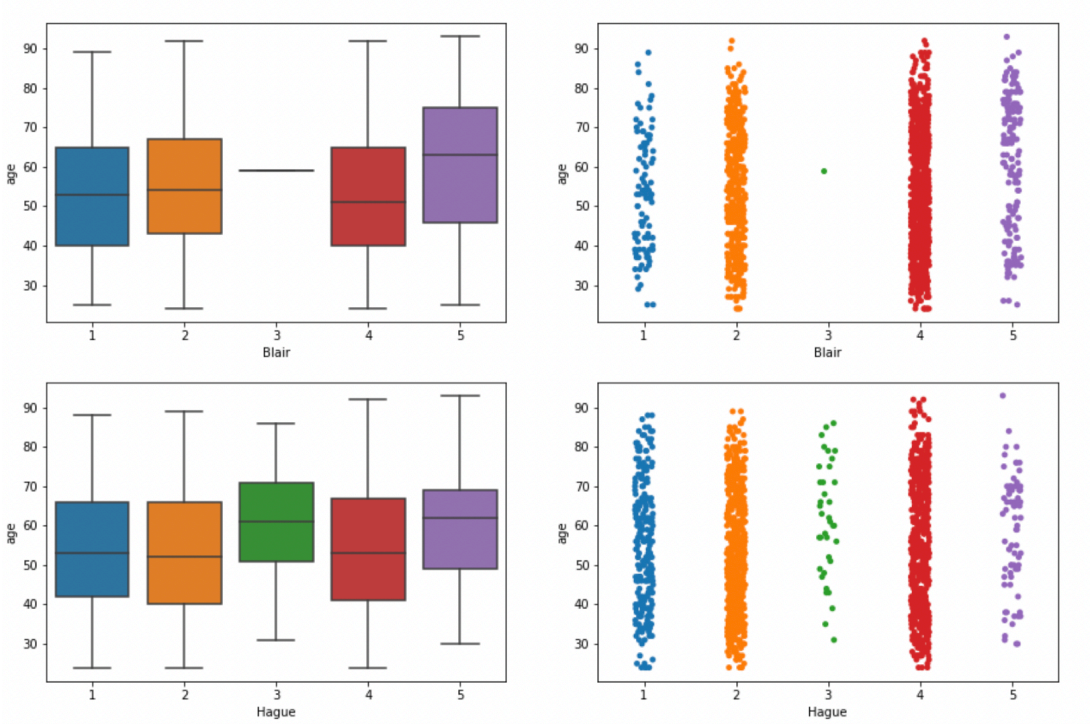


```
0    808
1    709
Name: gender, dtype: int64
```

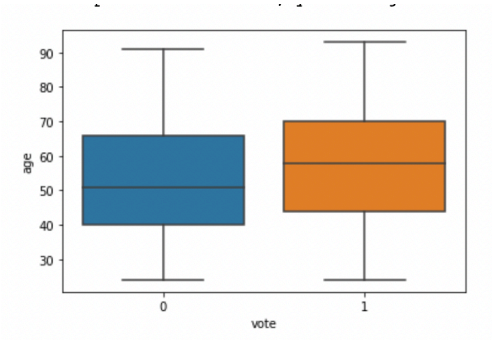


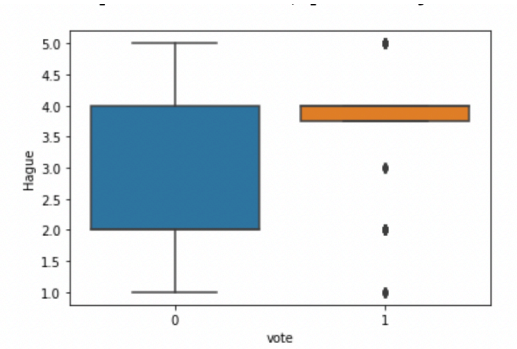
Bivariate Analysis

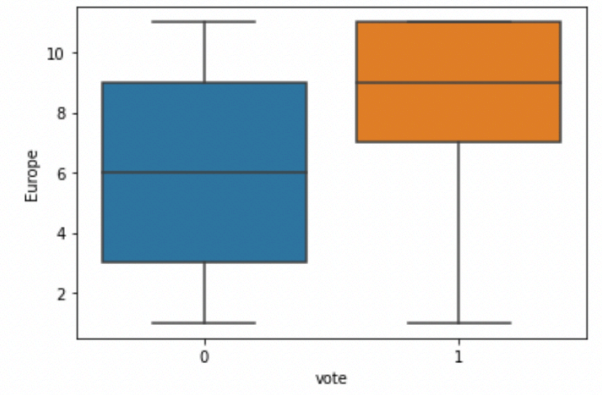
In Bivariate Analysis, there are two variables wherein the analysis is related to cause and the relationship between the two variables.Here the Numeric and Numeric relation are established in the relationship using Pairplot/scatter plot whereas Categorical and Categorical are established with Countplot with Hue

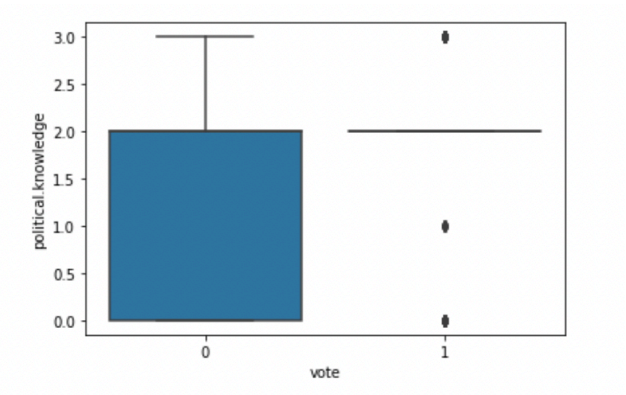


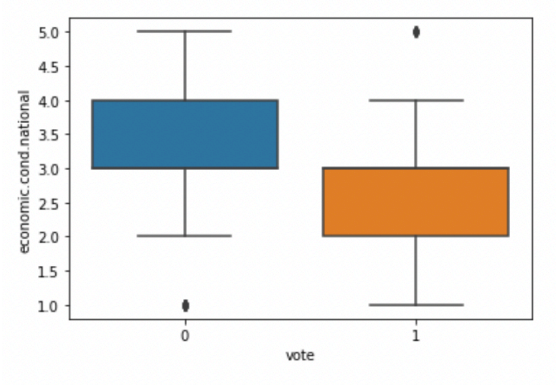
The Bivariate Analysis of the Dependent Variable(Vote) and all other Independent Variable are as follow

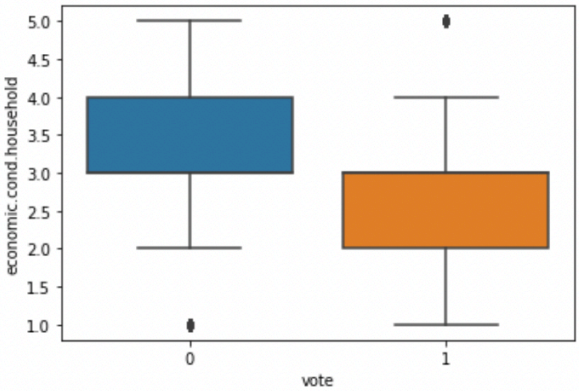










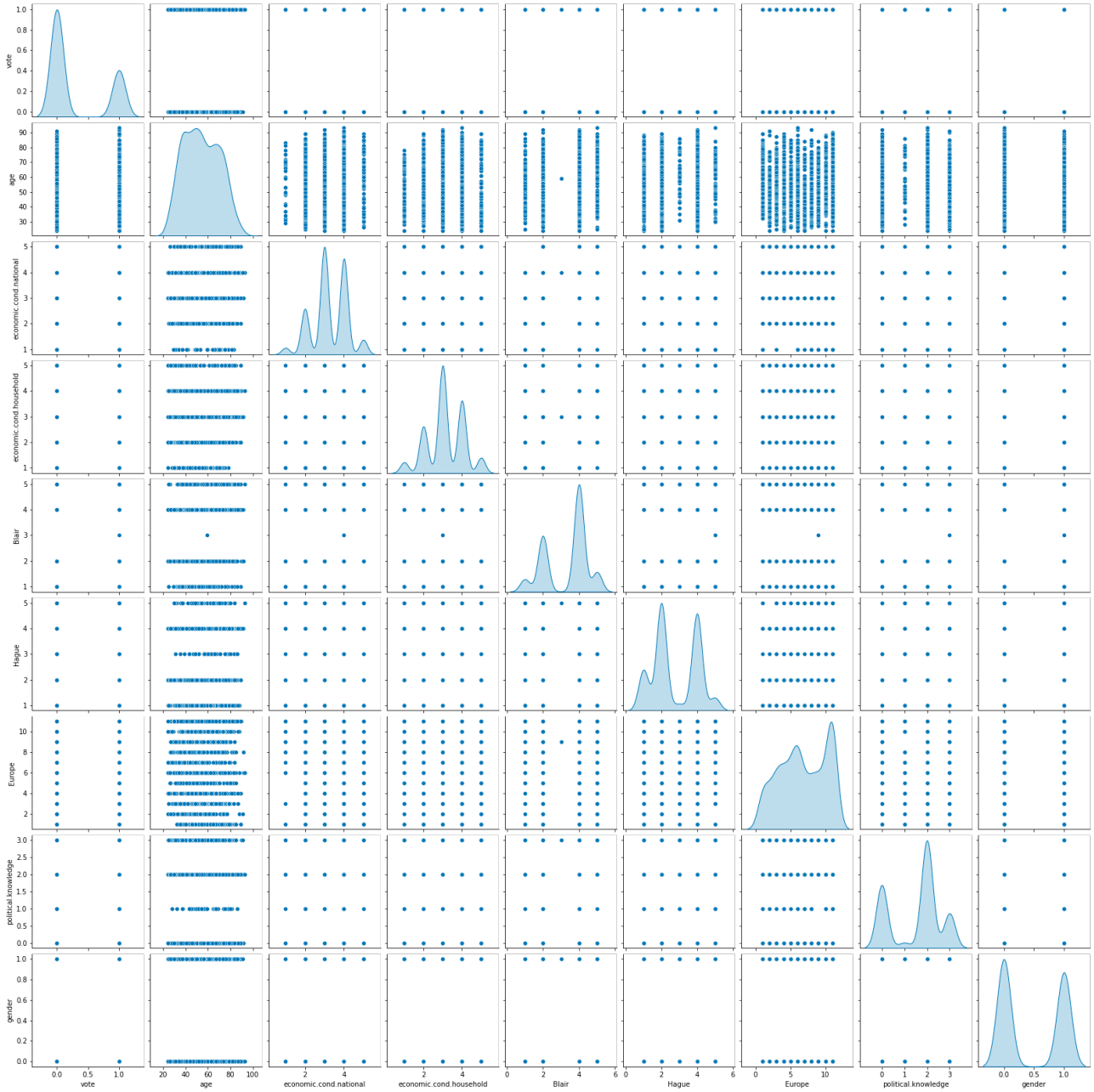


Multivariate Analysis

Multivariate descriptive displays or plots are designed to reveal the relationship among several variables simulataneously.. As was the case when examining relationships among pairs of variables, there are several basic characteristics of the relationship among sets of variables that are of interest. Some of the Interesting Multivariate Analysis are pair plot,Heat Map, Facet etc

Pairwise Analysis

A pairs plot allows us to see both distribution of single variables and relationships between two variable.The Pairwise plots are Mentioned below which determines the Overall relationship between the Dependent Variable and all other Independent variable. From this we Can identify microlevel and we can establish some kind of Relation

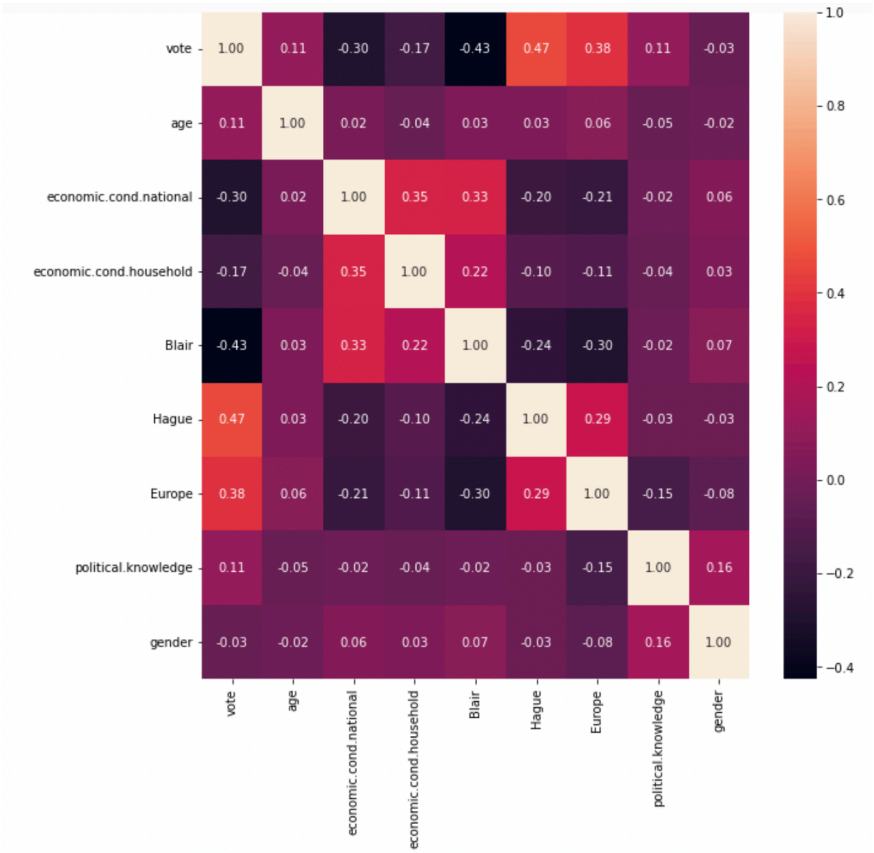


Heat map

The heat map is a data visualization technique that shows magnitude of a phenomenon as color in two dimensions. The variation in color may be by hue or intensity, giving obvious visual cues to the reader about how the phenomenon is clustered or varies over space.

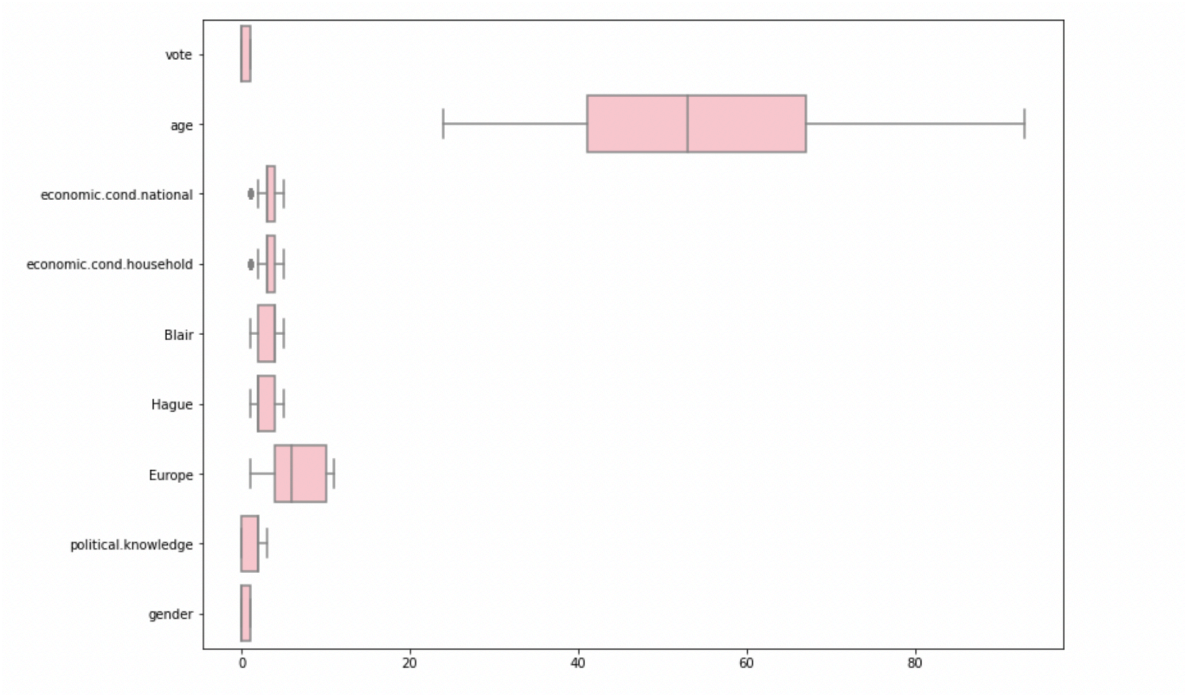
The Data of correlation and Pictorial representation of Heat map are given below

	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
vote	1.000000	0.109274	-0.302280	-0.174688	-0.426606	0.468186	0.384612	0.111589	-0.034464
age	0.109274	1.000000	0.018687	-0.038868	0.032084	0.031144	0.064562	-0.046598	-0.017933
economic.cond.national	-0.302280	0.018687	1.000000	0.347687	0.326141	-0.200790	-0.209150	-0.023510	0.055664
economic.cond.household	-0.174688	-0.038868	0.347687	1.000000	0.215822	-0.100392	-0.112897	-0.038528	0.033102
Blair	-0.426606	0.032084	0.326141	0.215822	1.000000	-0.243508	-0.295944	-0.021299	0.067624
Hague	0.468186	0.031144	-0.200790	-0.100392	-0.243508	1.000000	0.285738	-0.029906	-0.028309
Europe	0.384612	0.064562	-0.209150	-0.112897	-0.295944	0.285738	1.000000	-0.151197	-0.076059
political.knowledge	0.111589	-0.046598	-0.023510	-0.038528	-0.021299	-0.029906	-0.151197	1.000000	0.156923
gender	-0.034464	-0.017933	0.055664	0.033102	0.067624	-0.028309	-0.076059	0.156923	1.000000



From the Heatmap we can identify the Correlationship of data in better way than establishing the relationship in Univariate and Bivariate Method

Checking Of Outliers



In above the Outliers are Present in the Data mainly in Economic condition National and Household but we are not going to remove those Outliers in the Data because of the Data Leakage. So outliers available in the Data are not going to be Treated

Data Preparation

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

The Info of the Data are as follow

```
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   vote                     1525 non-null   object
 1   age                      1525 non-null   int64
 2   economic.cond.national   1525 non-null   int64
 3   economic.cond.household  1525 non-null   int64
 4   Blair                    1525 non-null   int64
 5   Hague                    1525 non-null   int64
 6   Europe                   1525 non-null   int64
 7   political.knowledge      1525 non-null   int64
 8   gender                   1525 non-null   object
dtypes: int64(7), object(2)
memory usage: 107.4+ KB
```

- From above data, we can clearly see that only vote and gender are in Category /object while rest of the Variable are either in Discrete and continuous in Nature.
- Hence we need to Decode the vote and Gender to the Integer Values

Scaling of Data

vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
0	43	3	3	4	1	2	2	0
0	36	4	4	4	4	5	2	1
0	35	4	4	5	2	3	2	1
0	24	4	2	2	1	4	0	0
0	41	2	2	1	1	6	2	1
0	47	3	4	4	4	4	2	1
0	57	2	2	4	4	11	2	1
0	77	3	4	4	1	1	0	1
0	39	3	3	4	4	11	0	0
0	70	3	2	5	1	11	2	1

In term of Scaling, the head of the Data are as follow

In above expect age which is continuous in nature all other variable are in Discrete in Nature.We can check with the Summary

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

In above also we can clearly see that expect age all other variables mean,Median are almost same in Nature. So we can compute without doing the Scaling but then on processing with and without scaling there is not much changes in result. But we had proceed by doing Scaling.Hence the data will be as follow

vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
-0.659692	-0.716161	-0.278185	-0.148020	0.565802	-1.419969	-1.437338	0.423832	-0.936736
-0.659692	-1.162118	0.856242	0.926367	0.565802	1.014951	-0.527684	0.423832	1.067536
-0.659692	-1.225827	0.856242	0.926367	1.417312	-0.608329	-1.134120	0.423832	1.067536
-0.659692	-1.926617	0.856242	-1.222408	-1.137217	-1.419969	-0.830902	-1.421084	-0.936736
-0.659692	-0.843577	-1.412613	-1.222408	-1.988727	-1.419969	-0.224465	0.423832	1.067536

Now the Data are Scaled and mostly the Value lies between the Interval of -3 to 3 and now the data got normalised to standard normal function. The scaling is not nessacary at this time but we had done scaling only to Normal Distribution to avoid the skewness and age factor which is different in nature comparing to the other variable present.

Data Split: Split the data into train and test

The data is split into X_train, X_test, y_train,y_test along with the test size of 0.30 with the random variable of 1

The (X_train, X_test, y_train, y_test) with the test size of 0.30 along with random state of 1 are as follow

X_train

The X_train is mentioned below and size of the variable is (1061,8). It contain data of 70% of Independent Variable

	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
991	34	2	4	1	4	11	2	0
1274	40	4	3	4	4	6	0	1
649	61	4	3	4	4	7	2	0
677	47	3	3	4	2	11	0	1
538	44	5	3	4	2	8	0	1
...	...	...	...	...	...	...	...	...
717	52	3	3	4	1	6	2	0
908	43	3	4	2	2	9	2	0
1100	74	4	3	5	4	11	0	0
236	31	3	3	2	3	6	0	0
1065	89	3	5	4	2	1	0	1

X_test

The X_test is mentioned below and size of the variable is (456,8). It contain data of 30% of Independent Variable

	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
504	71	3	3	2	2	8	2	0
369	43	3	2	4	2	8	3	1
1075	89	5	5	5	2	1	2	1
1031	47	2	3	2	4	8	2	0
1329	33	5	4	4	4	8	0	1
...	...	...	...	...	...	...	...	...
562	37	4	2	4	2	8	1	1
928	42	2	2	1	2	7	2	0
276	88	3	3	4	1	6	0	0
1128	53	4	3	4	2	10	0	0
1151	36	3	3	4	1	6	3	1

Y_train

The Y_train are as mentioned below and the size of the variable is (1061,1). It contain data of 70% of dependent Variable

	vote
991	1
1274	0
649	1
677	0
538	0
...	...
717	0
908	1
1100	0
236	0
1065	0

y_test

The Y_test are as mentioned below and the size of the variable is (456,1) It contain data of 30% of dependent Variable

	vote
504	0
369	0
1075	0
1031	1
1329	0
...	...
562	0
928	0
276	0
1128	0
1151	0

Modeling

1.4 Apply Logistic Regression and LDA (linear discriminant analysis).

Logistic Regression

In the Logistic Regression we had applied intially with default function because there is not much change required in Iteration as our expect iteration of 100 is already available and solver is lgbs and model is fitted with X_train,y_train

LogisticRegression(random_state=1)
model.fit(X_train, y_train)

The Model score of X_train and y_train is 0.8350612629594723

The Model score of X_test and y_test is 0.8267543859649122
The model is neither over fitting or Under fitting because the the Training performer very well and on the other side testing also done very well like training set.

Linear discriminant analysis

In the LDA we had applied intially with default function because there is not much required change required .The solver “svd” and tolerance level with 0.0001 is already in default function.The Model is mentioned as clfLDA

LinearDiscriminantAnalysis()
clfLDA.fit(X_train, y_train)

The Model score of X_train and y_train is 0.8341187558906692

The Model score of X_test and y_test is 0.8333333333333334

The model is neither over fitting or Under fitting because the the Training performer very well and on the other side testing also done very well like training set.

On Comparison both LDA and Logistic has done Predominantly very well. But still we can try with other Model for Better score

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

KNN Model

In the KNN,we had applied intially with default function because there is not much required change required .The n-neighbour is made as 5 and n_job as none is already in default function.The Model is mentioned as KNN_Model

KNeighborsClassifier()
KNeighborsClassifier.fit(X_train,y_train)

The Model score of X_train and y_train is 0.8623939679547596

The Model score of X_test and y_test is 0.8201754385964912

The model is neither over fitting or Under fitting because the the Training performer very well and on the other side testing also done very well like training set.

Naïve Bayes Model

In the Naive Bayes Model,we had dealt with Gaussian Naive Bayers .The Model is mentioned as NB_Model

GaussianNB()
NB_model.fit(X_train, y_train)

The Model score of X_train and y_train is 0.8350612629594723

The Model score of X_test and y_test is 0.8223684210526315

The model is neither over fitting or Under fitting because the the Training performer very well and on the other side testing also done very well like training set.

In term of Naive Bayers and KNN model, In Training set KNN has done Predominately very well compared to Naive Bayes Model but in Test Set both has done recently very well but the difference in Training and Testing is very in KNN Model

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

Model Tuning
Tuning is the process of maximizing a model's performance without overfitting or creating too high of a variance. In machine learning, this is accomplished by selecting appropriate “hyperparameters.

Grid Search is one of the important Model Tuning Process

Bagging With Random Forest

Without Grid Search CV-

The Random Forest with Default function such as n_estimators=100,min_split=2, minimum sample leaf as 1 with criteria gini with none Max depth. Here the random forest is mentioned as RF_model.fit()

RandomForestClassifier()
RF_model.fit(X_train, y_train)

The Model score of X_train and y_train is 1.00

The Model score of X_test and y_test is 0.8333333333333334

Here the model performed exceptionally very well in Training but not in testing as compared with Training Model. So there might be the Problem of Overfitting the Model

Whereas When we apply Bagging here in Random Forest with n_estimators of 100

The Model score of X_train and y_train is 0.9679547596606974

The Model score of X_test and y_test is 0.8289473684210527

Here still the Model Face the issue of Over fitting which means performance is very high in Training Set

With Grid Search CV-

Using Grid Search CV the values based on Trail and Based we found out the Best Value for the below

```
'min_samples_split' : [10,15],
    'min_samples_leaf':[3,5],
    'max_depth':[10],
    'random_state' : [0]
}
```

The Best Estimator Based on Test and Trail is

```
RandomForestClassifier(max_depth=10, min_samples_leaf=5,
min_samples_split=15,
                       random_state=0)
```

The Model score of X_train and y_train is 0.8727615457115928
The Model score of X_test and y_test is 0.8267543859649122

So Using Hyper tuning by Grid Search in Random Forest we had reduced the Over fitting of the Model. But still Training Model Perform way better than Testing Mode

Boosting Model

In Boosting Model, We had applied ada-Boosting and Gradient Boosting

Ada-boosting Model
Without Grid Out Search CV

AdaBoost can be used to boost the performance of any machine learning algorithm. It is best used with weak learners. These are models that achieve accuracy just above random chance on a classification problem

Here the Ada- boost has set with the Default Value of n_estimator of 50 with none Base estimator and algorithm as samme.r

The Model score of X_train and y_train is 0.8463713477851084

The Model score of X_test and y_test is 0.8135964912280702

Here the Model had performed equally in both the Model.There is no issue in Over/under Fitting in the Model

With Grid search CV

The Parameters in Boosting with Grid Search CV are as follow

```
'n_estimators' : [100,500,1000],
'learning_rate' : [0.1,0.01,0.001],
'algorithm' : ['SAMME', 'SAMME.R']
```

We had found the Best parameter

```
AdaBoostClassifier(learning_rate=0.01, n_estimators=1000)
```

Now after the Grid Search CV
The Model score of X_train and y_train is 0.8369462770970783

The Model score of X_test and y_test is 0.8092105263157895

Now the Model Training and testing got reduced further on Applying the Grid Search CV

Gradient Boosting

In Gradient Boosting, the initially it is carried out using the Default Function such as loss is logg loss,n_estimator as 100,max_depth as 3,min sample split as 3,min sample leaf as 2

The Model score of X_train and y_train is 0.8925541941564562

The Model score of X_test and y_test is 0.8355263157894737

The Model has performed very well in Training set and but not much in Testing Set

Till Now on considering above 80 as good model, We can make all that all the Model fared very well with the Data considering the Accuracy score of the Model. Now we need to consider other parameter for Comparison

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 the models and write inference which model is best/optimized

There are various metrics which we can use to evaluate the performance of ML algorithms, classification as well as regression algorithms. We must carefully choose the metrics for evaluating ML performance because

- How the performance of ML algorithms is measured and compared will be dependent entirely on the metric you choose.
- How you weight the importance of various characteristics in the result will be influenced completely by the metric you choose

Accuracy Score

Logistic Regression

Training Model -0.8350612629594723

Testing Model -0.8267543859649122

LDA Model

Training Model -0.8341187558906692

Testing Model -0.8333333333333334

KNN Model

Training Model -0.8623939679547596
Testing Model -0.8201754385964912

Naive Bayers Model

Training Model -0.835061262959472

Testing Model -0.8223684210526315

Bagging Model (Random Forest)

Without Grid Search CV

Training Model 0.9679547596606974

Testing Model 0.8289473684210527

With Grid Search CV

Training Model -0.8727615457115928

Testing Model -0.8267543859649122

Boosting Model

Ada_boosting

Without Grid Search CV

Training Model -0.8463713477851084

Testing Model -0.8135964912280702

With Grid Search CV

Training Model -0.8369462770970783

Testing Model -0.8092105263157895

Gradient Boosting Model

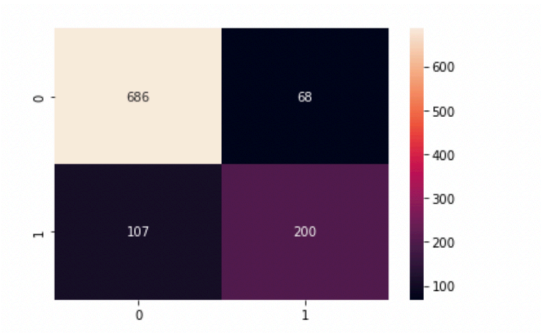
Training Model -0.8925541941564562

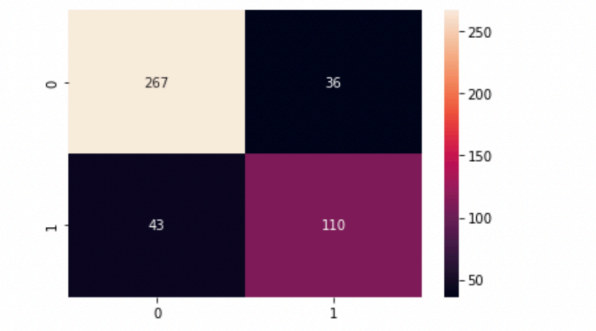
Testing Model -0.8355263157894737

In Terms of Accuracy Random Forest bagging has fared better compared to other in the Model in Training Set.Even though the Testing data set is less compared to training in Bagging but almost all has performed in the Same way

Confusion Matrix

Logistic Regression(training and Testing)





LDA Model (training and Testing)

```
[[685 107]
 [ 69 200]]
```

```
[[269  42]
 [ 34 111]]
```

KNN Model

```
[[705  97]
 [ 49 210]]
```

```
[[274  53]
 [ 29 100]]
```

Naive Bayers Model

```
[[675  96]
 [ 79 211]]
```

```
[[263  41]
 [ 40 112]]
```

Bagging Model (Random Forest)

Without Grid Search CV

```
[[750  30]
 [  4 277]]
```

```
[[274  49]
 [ 29 104]]
```

With Grid Search CV

```
0.872761345
[[707  47]
 [ 88 219]]
```

```
[[277  26]
 [ 53 100]]
```

Boosting Model

Ada_boosting

Without Grid Search CV

```
[[688  97]
 [ 66 210]]
```

```
[[266  48]
 [ 37 105]]
```

With Grid Search CV

```
[[702  52]
 [121 186]]
```

```
[[271  32]
 [ 55  98]]
```

Gradient Boosting Model

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

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

Training Model

- In above overall model, The highest True Positive Prediction is by Bagging Random Forest without Grid CV of 750
- In above overall model, The highest Negative Prediction is also by Bagging Random Forest without Grid CV 750

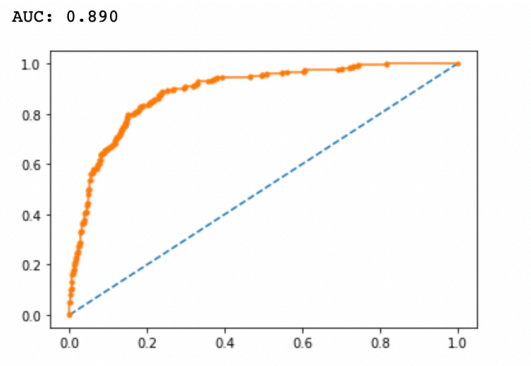
Testing Model

- In above overall model, The highest True Positive Prediction is by Bagging Random Forest with Grid CV of 277 without grid CV of 274
- In above overall model, The highest Negative Prediction is also by Naive Bayers without Grid CV 112 whereas bagging random forest is 104

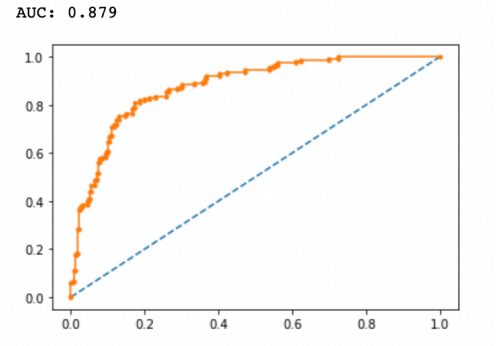
AUC score

Logistics Regression

AUC: 0.890

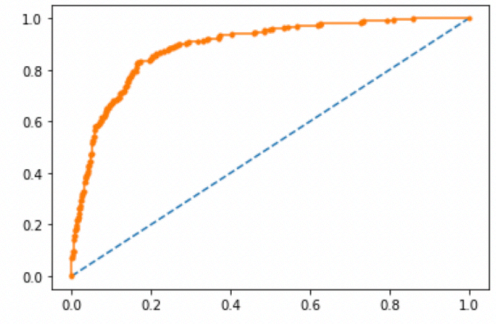


AUC: 0.879

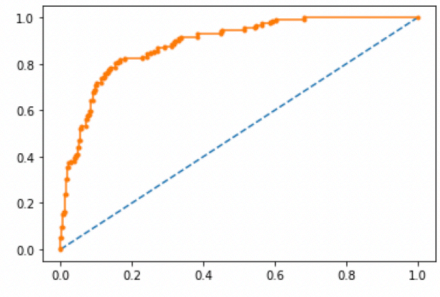


LDA

```
AUC: 0.889
[<matplotlib.lines.Line2D at 0x7f9ce726f250>]
```

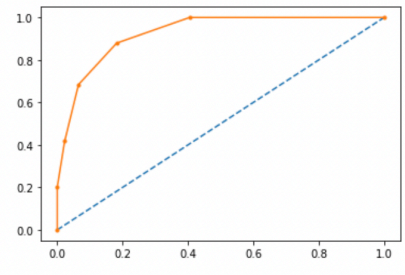



```
AUC: 0.888
```

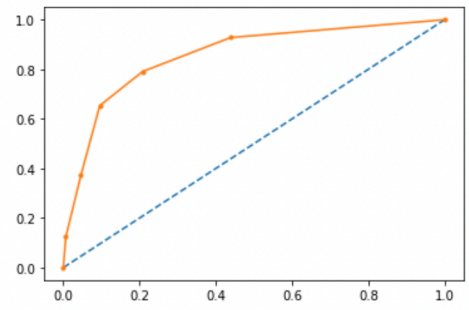


KNN MODEL

```
AUC: 0.926
[<matplotlib.lines.Line2D at 0x7f9ce73e9070>]
```

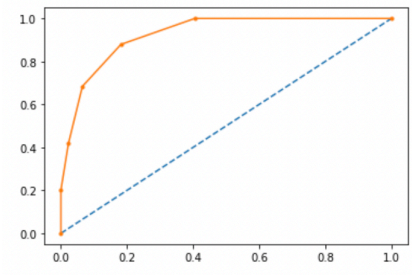



```
AUC: 0.856
```

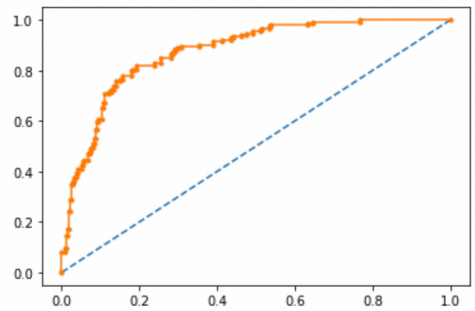


NAIVE Bayers

```
AUC: 0.926
[<matplotlib.lines.Line2D at 0x7f9ce73e9070>]
```



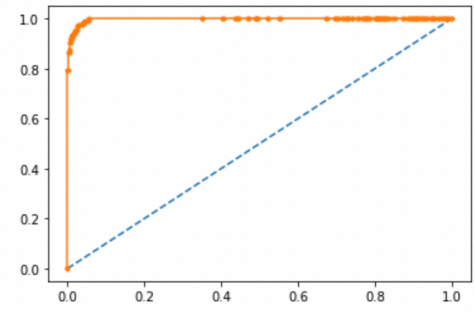

```
AUC: 0.876
```



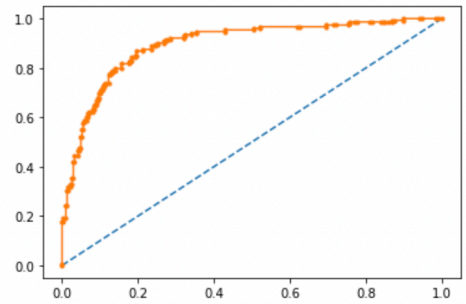
Bagging Method

AUC: 0.997

[<matplotlib.lines.Line2D at 0x7f9cc0b0fa90>]

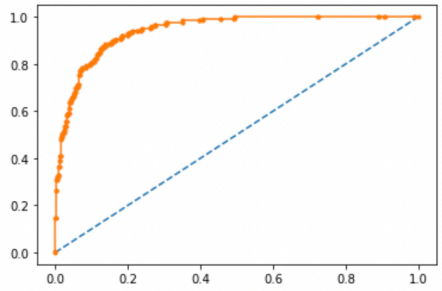


AUC: 0.897

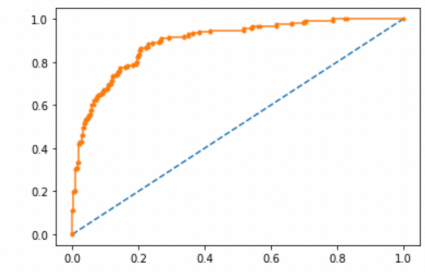


AUC: 0.943

[<matplotlib.lines.Line2D at 0x7f9c8006f700>]



AUC: 0.896

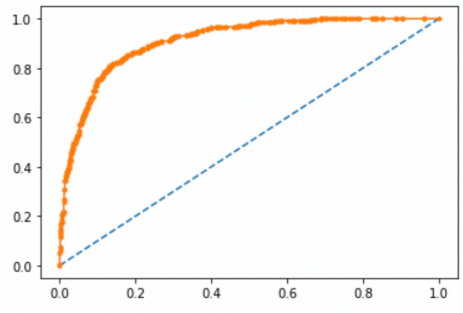


Boosting Method

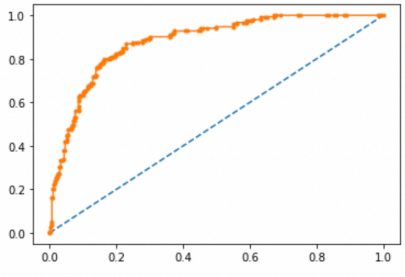
Ada- Boosting

AUC: 0.912

[<matplotlib.lines.Line2D at 0x7f9cd160ee80>]



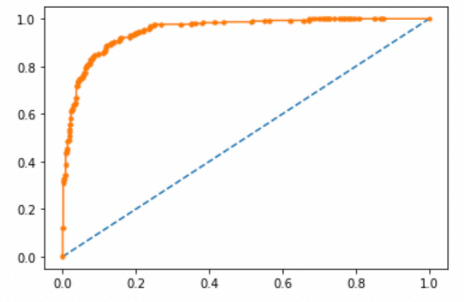
AUC: 0.881



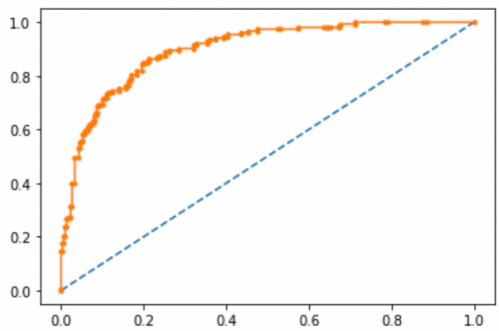
Gradient Boosting Model

AUC: 0.951

[<matplotlib.lines.Line2D at 0x7f9cc0c90730>]



AUC: 0.899



In Terms Traing Set of ROC curve and AUC score , The Best Performance in Training Set is Bagging Method Without Grid Search CV

In Terms Testing Set of ROC curve and AUC score , The Best Performance in Testing Set is Gradient Boosting whereas the Bagging Method Without Grid Search CV is almost equal

So from ROC Curve and AUC score, Bagging Method of Random Forest has fared in Better Way compared to others

So till Bagging in random forest fared better than all other model in various Aspect

F1 score -Training Set - Best Performance- Random Forest in Bagging
F1 score -Testing Set - Best Performance- Gradient Boosting whereas Random Forest in Bagging also performed similar

In this type of Model, F1 score plays a major role compared to other parameters as it deals with Election Result. Hence both Precision and recall are equally Important

1.8 Based on these predictions, what are the insights?

- Overall, the Random Forest in Bagging performed better than other models in the data set in the example above. The Training and Testing Aspects have been equally important for models like Logistic Regression and LDA. Gradient Boosting has performed well in all of the Category tests.
- Voter turnout is higher among women than among men.
- When comparing the ratings for the two leaders, it is clear that the Labour leader is doing well because of his higher ratings.
- Conservative voters are more likely to vote for them if they have stronger anti-European feelings.
- Better national economic conditions are influencing people's decision to vote for the Labour party.
- Conservatives have received the support of those with more political knowledge.
- By a wide margin, the Labour Party is outperforming the Conservatives

Problem 2:

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

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

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

Number of Characters

1. President Franklin D. Roosevelt in 1941-7574
2. President John F. Kennedy in 1961-7619
3. President Richard Nixon in 1973-9992

Number of Words

1. President Franklin D. Roosevelt in 1941-1360
2. President John F. Kennedy in 1961-1390
3. President Richard Nixon in 1973-1819

Number of Sentences

1. President Franklin D. Roosevelt in 1941-68
2. President John F. Kennedy in 1961-52
3. President Richard Nixon in 1973-68

2.2 Remove all the stopwords from all three speeches.

Stop words are a set of commonly used words in any language. For example, in English, “the”, “is” and “and”, would easily qualify as stop words. In NLP and text mining applications, stop words are used to eliminate unimportant words, allowing applications to focus on the important words instead.

Some of English Stopwords are as follow

```
‘ he',
 'him',
 'his',
 'himself',
 'she',
 "she's",
 'her',
 'hers',
 'herself',
 'it',
```

Stemming Process

Stemming is a natural language processing technique that lowers inflection in words to their root forms, hence aiding in the preprocessing of text, words, and documents for text normalization.

Lemmatization

Lemmatization is a text normalization technique used in Natural Language Processing (NLP), that switches any kind of a word to its base root mode. Lemmatization is responsible for grouping different inflected forms of words into the root form, having the same meaning

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) –

The Most number of times words in his inaugural address for each president

President Franklin D. Roosevelt in 1941 are

Nation, America,know ,people,spirit,freedom,faith, histori

President John F. Kennedy in 1961 are

Peace,power, will , world , nation, let, us,

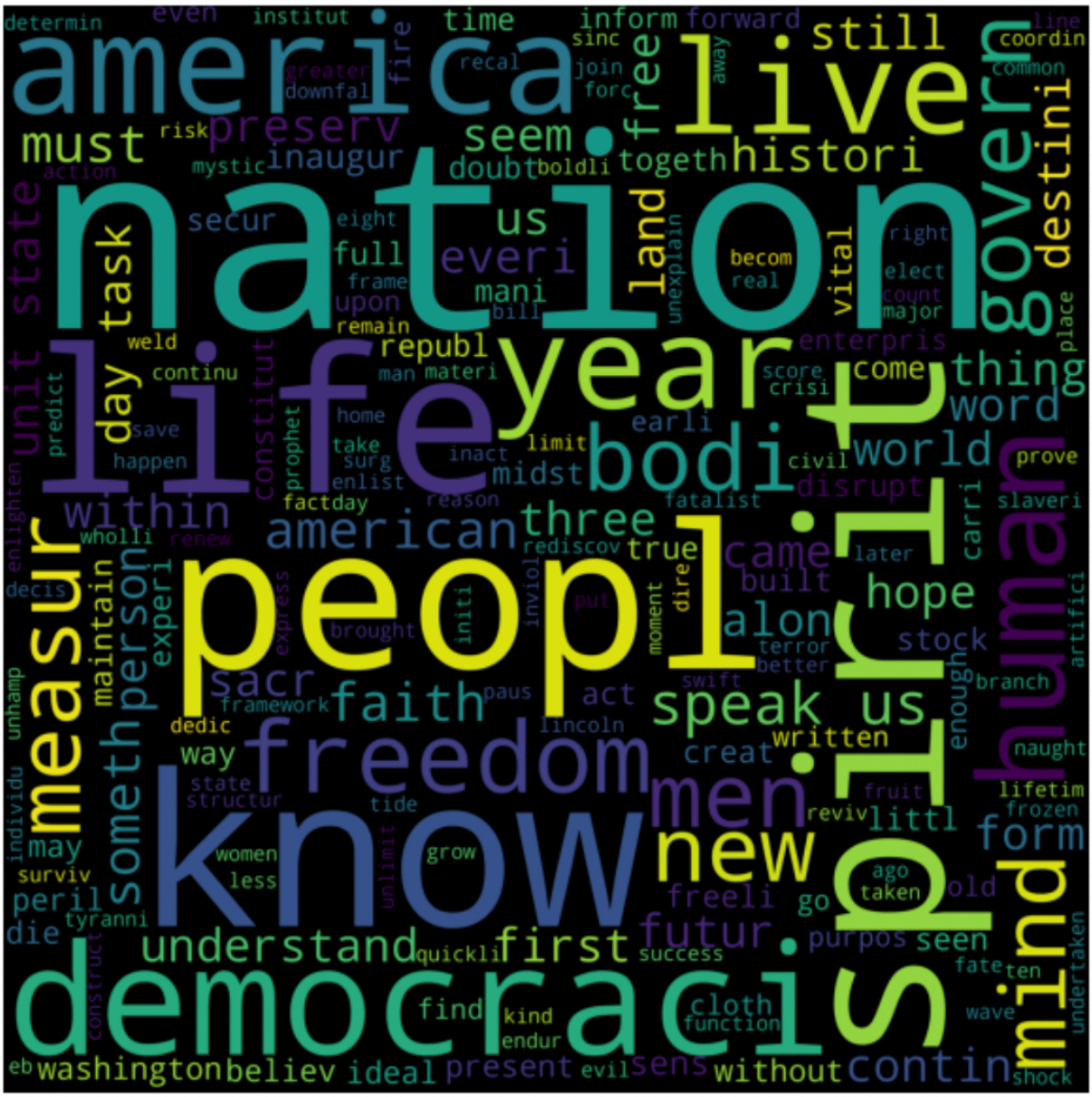
President Richard Nixon in 1973

America, let, peace , world,us, abroad, new, government

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

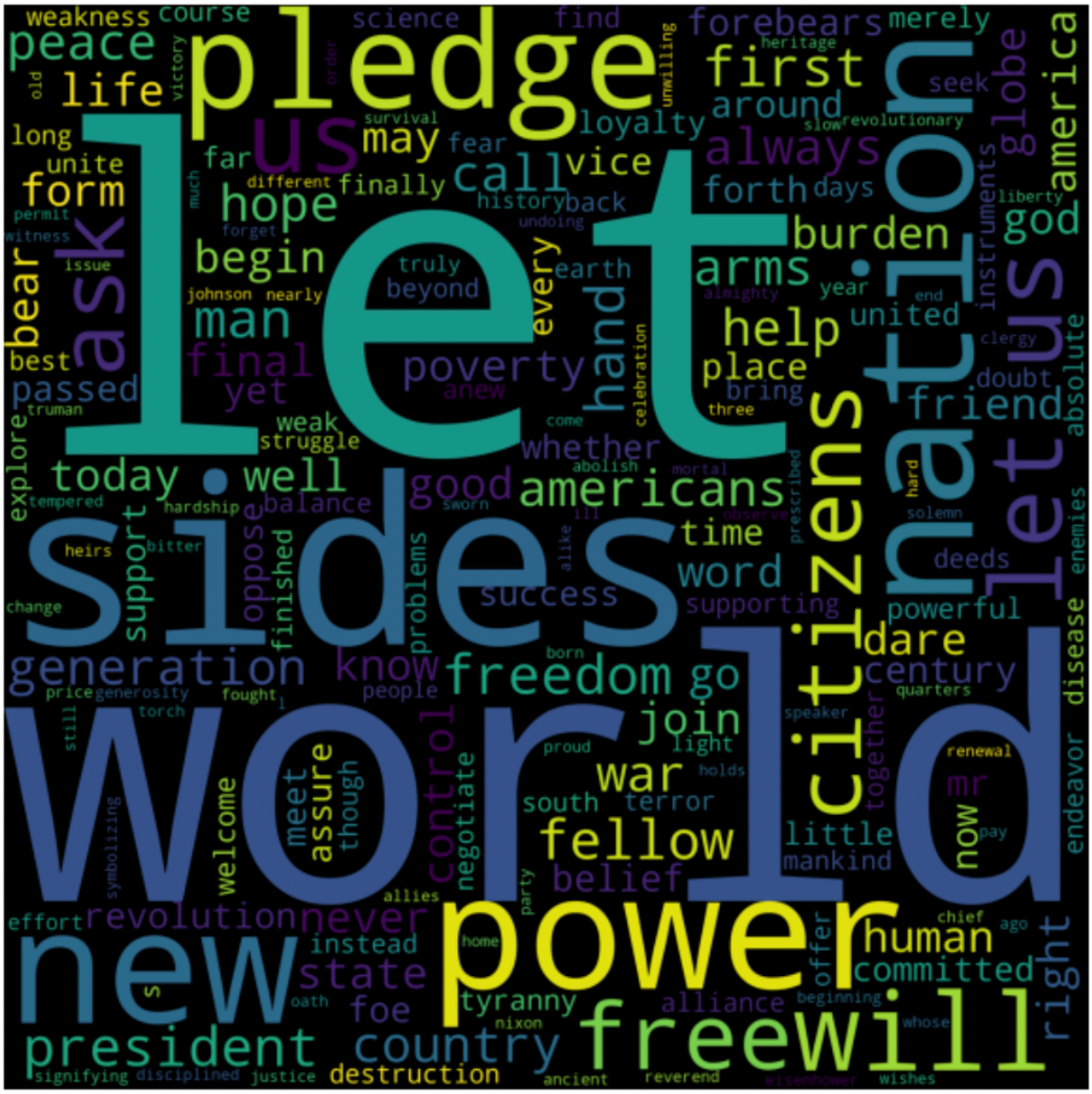
1. ***President Franklin D. Roosevelt in 1941***

```
corpus_rooseveltspeech
```



President John F. Kennedy in 1961

`corpus_Kennedyspeech`



President Richard Nixon in 1973

```
corpus_Nixonspeech
```