Machine Learning

Haresh P Tayade

PGP-DSBA Online

Sept’2021

Date-20/03/2022

Problem : 1

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

Questions

1.1) Read the dataset. Describe the data briefly. Interpret the inferences for each. Initial steps like head() .info(), Data Types, etc . Null value check, Summary stats, Skewness must be discussed.	4
1.2) Perform EDA (Check the null values, Data types, shape, Univariate, bivariate analysis). Also check for outliers (4 pts). Interpret the inferences for each (3 pts) Distribution plots(histogram) or similar plots for the continuous columns. Box plots, Correlation plots. Appropriate plots for categorical variables. Inferences on each plot. Outliers proportion should be discussed, and inferences from above used plots should be there. There is no restriction on how the learner wishes to implement this but the code should be able to represent the correct output and inferences should be logical and correct.	7

1.3) Encode the data (having string values) for Modelling. Is Scaling necessary here or not?(2 pts), Data Split: Split the data into train and test (70:30) (2 pts). The learner is expected to check and comment about the difference in scale of different features on the bases of appropriate measure for example std dev, variance, etc. Should justify whether there is a necessity for scaling. Object data should be converted into categorical/numerical data to fit in the models. (pd.categorical().codes(), pd.get_dummies(drop_first=True)) Data split, ratio defined for the split, train-test split should be discussed.	4
1.4) Apply Logistic Regression and LDA (Linear Discriminant Analysis) (2 pts). Interpret the inferences of both model s (2 pts). Successful implementation of each model. Logical reason behind the selection of different values for the parameters involved in each model. Calculate Train and Test Accuracies for each model. Comment on the validness of models (over fitting or under fitting)	4
1.5) Apply KNN Model and Naïve Bayes Model (2pts). Interpret the inferences of each model (2 pts). Successful implementation of each model. Logical reason behind the selection of different values for the parameters involved in each model. Calculate Train and Test Accuracies for each model. Comment on the validness of models (over fitting or under fitting)	4
1.6) Model Tuning (4 pts) , Bagging (1.5 pts) and Boosting (1.5 pts). Apply grid search on each model (include all models) and make models on best_params. Define a logic behind choosing particular values for different hyper-parameters for grid search. Compare and comment on performances of all. Comment on feature importance if applicable. Successful implementation of both algorithms along with inferences and comments on the model performances.	7
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, classification report (4 pts) Final Model - Compare and comment on all models on the basis of the performance metrics in a structured tabular manner. Describe on which model is best/optimized, After comparison which model suits the best for the problem in hand on the basis of different measures. Comment on the final model.(3 pts)	7
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. There should be at least 3-4 Recommendations and insights in total. Recommendations should be easily understandable and business specific, students should not give any technical suggestions. Full marks should only be allotted if the recommendations are correct and business specific.	5

List of Figures

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.

Introduction

A Leading News channel CNBE wants to analyze & predict elections result of 1525 voters that overall win and seats covered by a particular party.

Question 1.1

Read the dataset. Describe the data briefly. Interpret the inferences for each. Initial steps like head(),tail(),info(),Datatypes,etc.Null value check ,Summary stats, skewness must be discussed .

Answer:

Importing the important Libraries

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
from sklearn.model_selection import train_test_split,GridSearchCV
from sklearn import metrics
from sklearn.neighbors import KNeighborsClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import roc_auc_score,roc_curve,classification_report,confusion_matrix
```

Now we will see the head() and tail() of data.

```
df.head(10)
```

	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
5	6	Labour	47	3	4	4	4	4	2	male
6	7	Labour	57	2	2	4	4	11	2	male
7	8	Labour	77	3	4	4	1	1	0	male
8	9	Labour	39	3	3	4	4	11	0	female
9	10	Labour	70	3	2	5	1	11	2	male

```
df.tail(10)
```

	Unnamed: 0	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
1515	1516	Conservative	82	2	2	2	1	11	2	female
1516	1517	Labour	30	3	4	4	2	4	2	male
1517	1518	Labour	76	4	3	2	2	11	2	male
1518	1519	Labour	50	3	4	4	2	5	2	male
1519	1520	Conservative	35	3	4	4	2	8	2	male
1520	1521	Conservative	67	5	3	2	4	11	3	male
1521	1522	Conservative	73	2	2	4	4	8	2	male
1522	1523	Labour	37	3	3	5	4	2	2	male
1523	1524	Conservative	61	3	3	1	4	11	2	male
1524	1525	Conservative	74	2	3	2	4	11	0	female

Data Description

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.

10. Unnamed: Serial Number

Info and shape of Data

```
df.shape
```

```
(1525, 10)
```

```
df.info()
```

```
<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
```

Null Values

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

```
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
```

Insights

Data consists of both categorical and numerical values

There are total 1525 rows representing voters and 10 columns with 9 variables. Out of 10, 2 columns are of object type and 8 columns are of integer type.

Data does not contain missing values.

The first column is an index ("Unnamed: 0") as these are only serial numbers, we can remove it..

Descriptive

1.Dropping the ‘unnamed :0 ‘ column

2. After dropping "Unnamed: 0", data now contains 1525 rows and 9 columns.

3.confirms presence of unique value counts for variables “vote” and “gender”.

4.There are 2 types of voting parties- Labour and Conservative. From Table 8 top party seems to be the LabourParty.

5.There are 2 types of genders voting- Male and Female with Female being the top most voters.

6.Minimum age of an individual voting is 24 years and maximum age is 93 years. Mean voting age is 54 years.

7.Minimum assessment of current national economic conditions is 1 and a maximum assessment is 5 with an average assessment of 3.

8.Minimum assessment of current household economic conditions 1 and a maximum assessment is 5 with an average assessment of 3.

9.Minimum assessment of the Labour leader Tony Blair is 1 and maximum assessment is 5 with an average assessment of 4.

10.Minimum assessment of the Conservative leader William Hague is 1 and maximum r assessment is 5 with an average assessment of 2.

11-0.75% of the voters on a 11-point scale that measures respondents attitudes toward European integration represent high 'Eurosceptic' sentiment with a maximum scale of 11 and a minimum scale of 1.

12.On an average knowledge of parties positions on European integration is 2. Approximately 25% of parties do not hold positions on European integration with a maximum holding of 3.

13.After dropping unnecessary attribute

-- The number of rows of the dataframe = 1525

--The number of columns of the dataframe = 9

Now we check for Duplicates

In [13]:
```
dups=df.duplicated()
print('Number of duplicate rows = %d' % (dups.sum()))
df[dups]
```

```
Number of duplicate rows = 8
```

Out[13]:

	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
67	Labour	35	4	4	5	2	3	2	male
626	Labour	39	3	4	4	2	5	2	male
870	Labour	38	2	4	2	2	4	3	male
983	Conservative	74	4	3	2	4	8	2	female
1154	Conservative	53	3	4	2	2	6	0	female
1236	Labour	36	3	3	2	2	6	2	female
1244	Labour	29	4	4	4	2	2	2	female
1438	Labour	40	4	3	4	2	2	2	male

There are only 8 duplicated records in the data set, we can remove these records from the data set as they might not be adding any additional value and dropping these may provide distinct records only.

After dropping the duplicated as the shape will also change .

```
df.drop_duplicates(inplace=True)
```

```
df.shape
```

```
(1517, 9)
```

```
df.describe().T
```

	count	mean	std	min	25%	50%	75%	max
age	1517.0	54.241266	15.701741	24.0	41.0	53.0	67.0	93.0
economic.cond.national	1517.0	3.245221	0.881792	1.0	3.0	3.0	4.0	5.0
economic.cond.household	1517.0	3.137772	0.931069	1.0	3.0	3.0	4.0	5.0
Blair	1517.0	3.335531	1.174772	1.0	2.0	4.0	4.0	5.0
Hague	1517.0	2.749506	1.232479	1.0	2.0	2.0	4.0	5.0
Europe	1517.0	6.740277	3.299043	1.0	4.0	6.0	10.0	11.0
political.knowledge	1517.0	1.540541	1.084417	0.0	0.0	2.0	2.0	3.0

It is clear that there are no duplicated records in the data set.

Using shape attribute, it is confirmed that the duplicated records have been dropped from dataset from original data.

Initially data had 1525 records and now data contains 1517 record.

```
print(df.vote.value_counts())
print('\n')
print(df.gender.value_counts())
```

```
Labour          1057
Conservative     460
Name: vote, dtype: int64


female    808
male      709
Name: gender, dtype: int64
```

There are 2 types of voting parties- Labour and Conservative. Vote count for Labour Party is 1057 and vote count for Conservative party is 460(which was also confirmed from Table () top party is Labour).

There are 709 Male voters and 808 Female voters (which was also confirmed from Table () that there are higher count of Females voters).

Skewness

```
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
dtype: float64
```

From above we can see the 'age' & 'Hague' are only variable which skewness is positive remaining all are negative.

Question 1.2

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

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

```
vote                      0
age                       0
economic.cond.national    0
economic.cond.household   0
Blair                     0
Hague                     0
Europe                    0
political.knowledge       0
gender                    0
dtype: int64
```

```
df.dtypes
```

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

From above we can see that we cant see any null value and from dtypes we see that there are 7 numerical datatype and 2 categorical datatype.

Univariate:

• For variable "age": Minimum voting age is 24 years and maximum voting age is 93 years. Mean voting age is 54 years.

• For variable "economic.cond.national" : Minimum assessment of current national economic conditions is 1and a maximum assessment is 5 with an average assessment of 3.

• For variable "economic.cond.household" : Minimum assessment of current household economic conditions 1and a maximum assessment is 5 with an average assessment of 3.

• For variable "Blair": Minimum assessment of the Labour leader Tony Blair is 1 and maximum assessment is 5with an average assessment of 4.

• For variable "Hague": Minimum assessment of the Conservative leader William Hague is 1 and maximum rassessment is 5 with an average assessment of 2

• For variable "Europe": 75% of the voters on a 11-point scale that measures respondent attitudes toward European integration represent high ‘Eurosceptic’ sentiment with a maximum scale of 11 and a minimum scale of 1.

• On an average knowledge of parties positions on European integration is 2.Approximately 25% of parties do not hold positions on European integration with a maximum holding of 3.

• The medians of variables "Blair", "Hague","economic.cond.national" and "economic.cond.household" are identical to the first quartile, which is why there is an overlap in the Boxplot .This could be because data might have identical large proportion of low values.

• We can also confirm presence of outliers in variables "economic.cond.national" and"economic.cond.household".

• Since the lower quartile and middle quartile values are same (i.e. 0), variable "political.knowledge" does not have a lower whisker and middle whisker.

Bivariate and Multivariate Analysis:

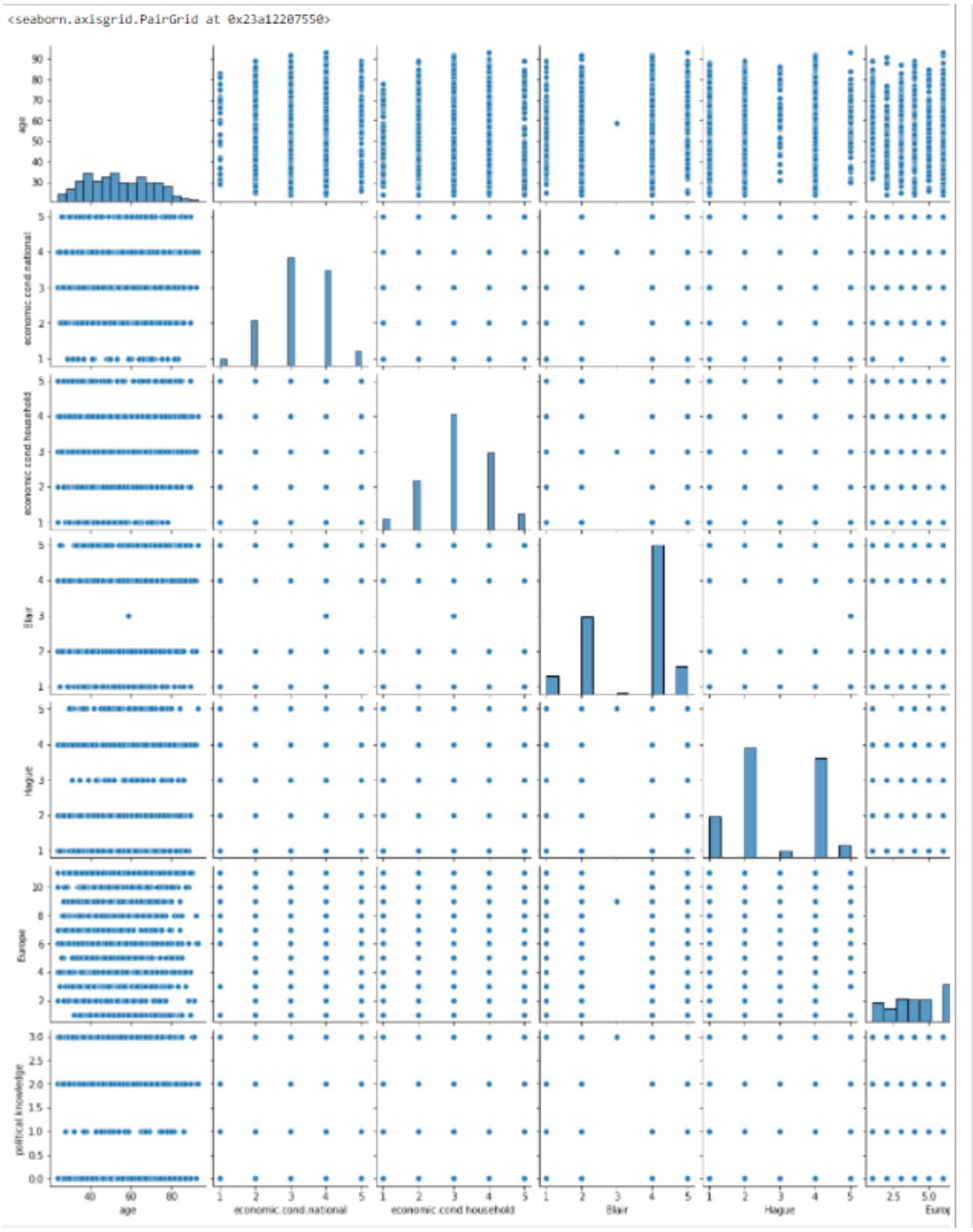


Figure 1: Pairplot for dataframe

Since we need to predict which party a voter will vote on the basis of the given information , we will do a bivariate analysis of variable vote with other variables and also look at the pairwise relationship of variables with dependency on variable vote.

<seaborn.axisgrid.PairGrid at 0x23a103993d0>

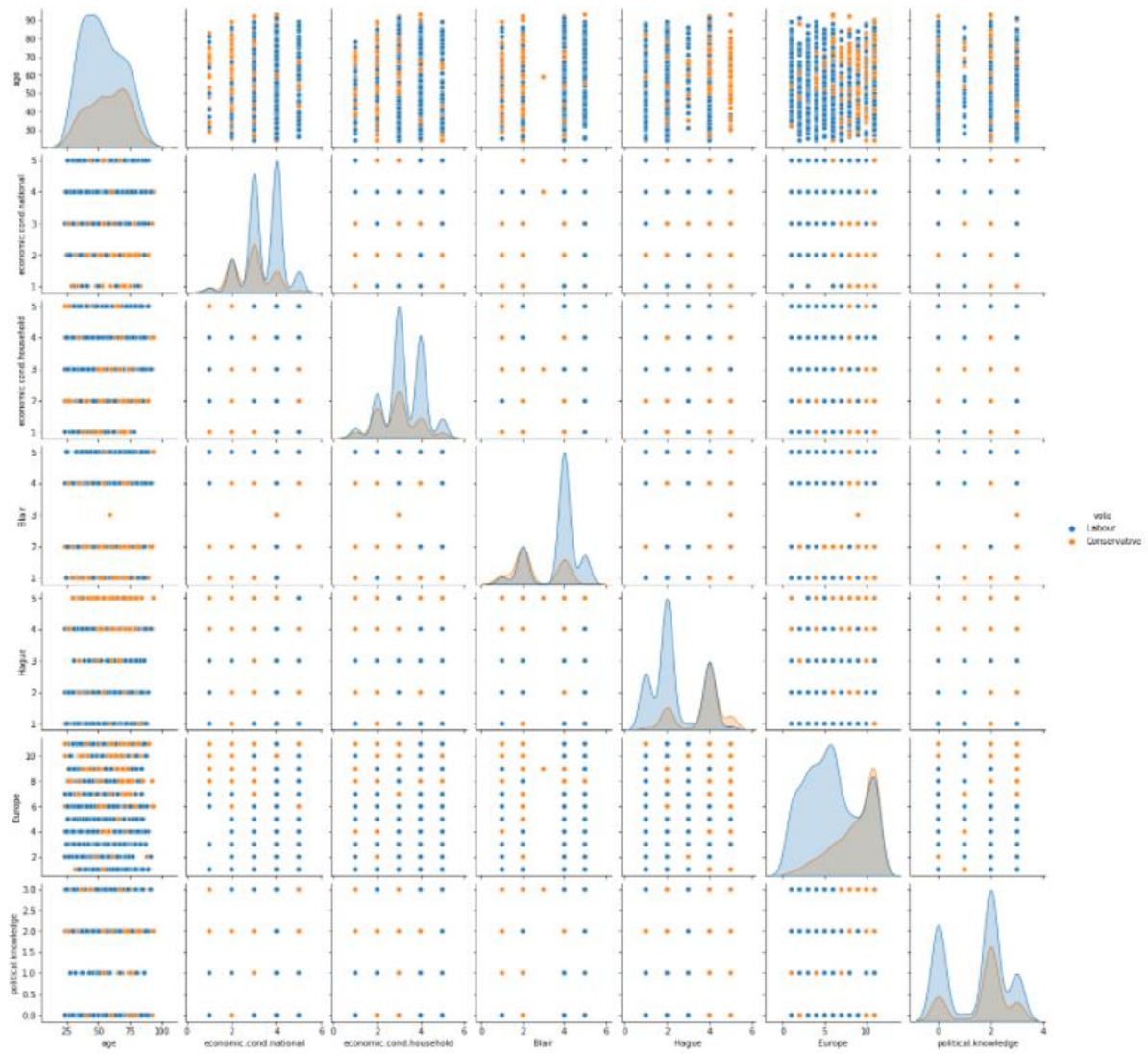


Figure 2: Pairplot with counting for vote

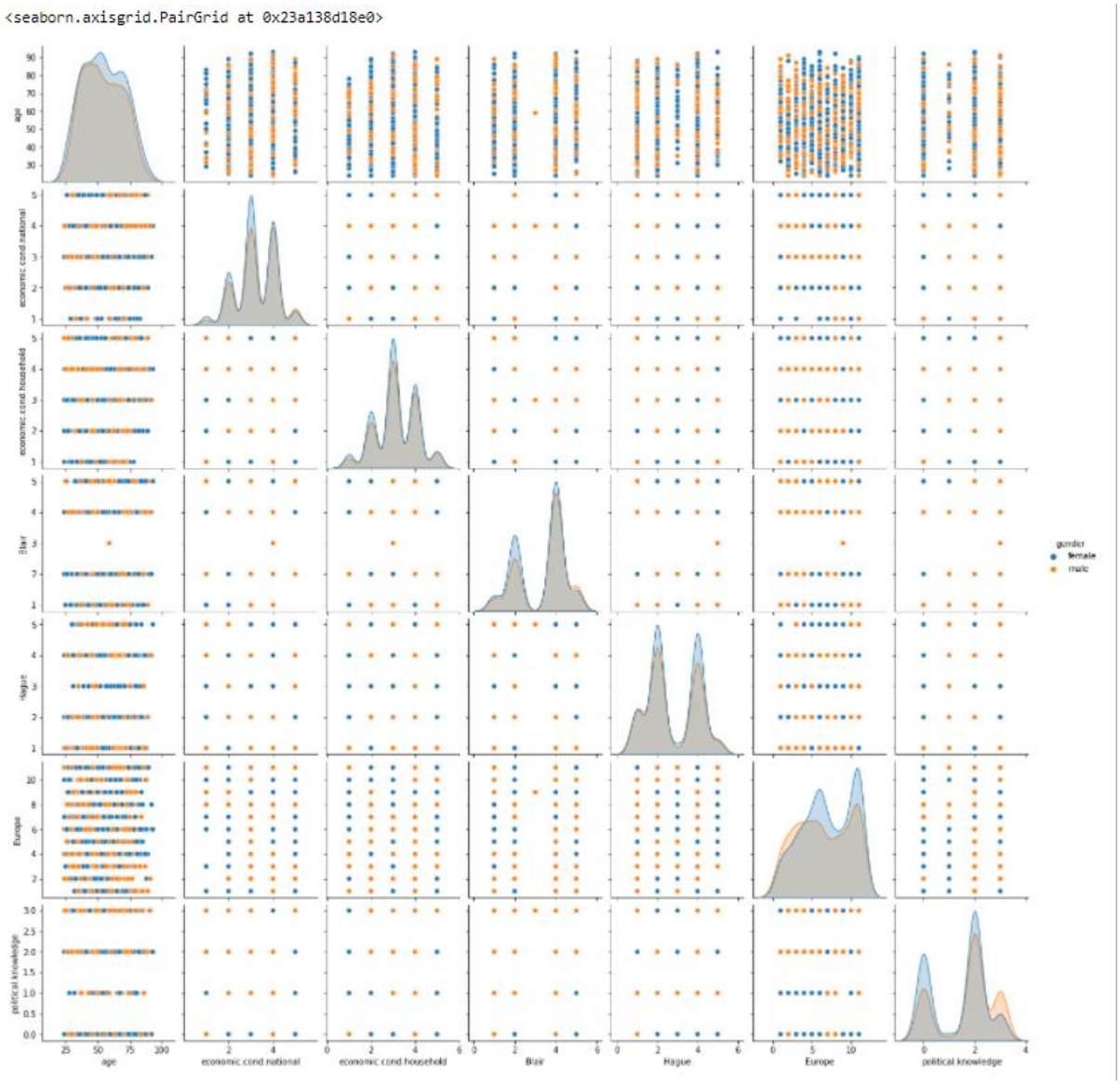


Figure 3: pairplot for counting Gender

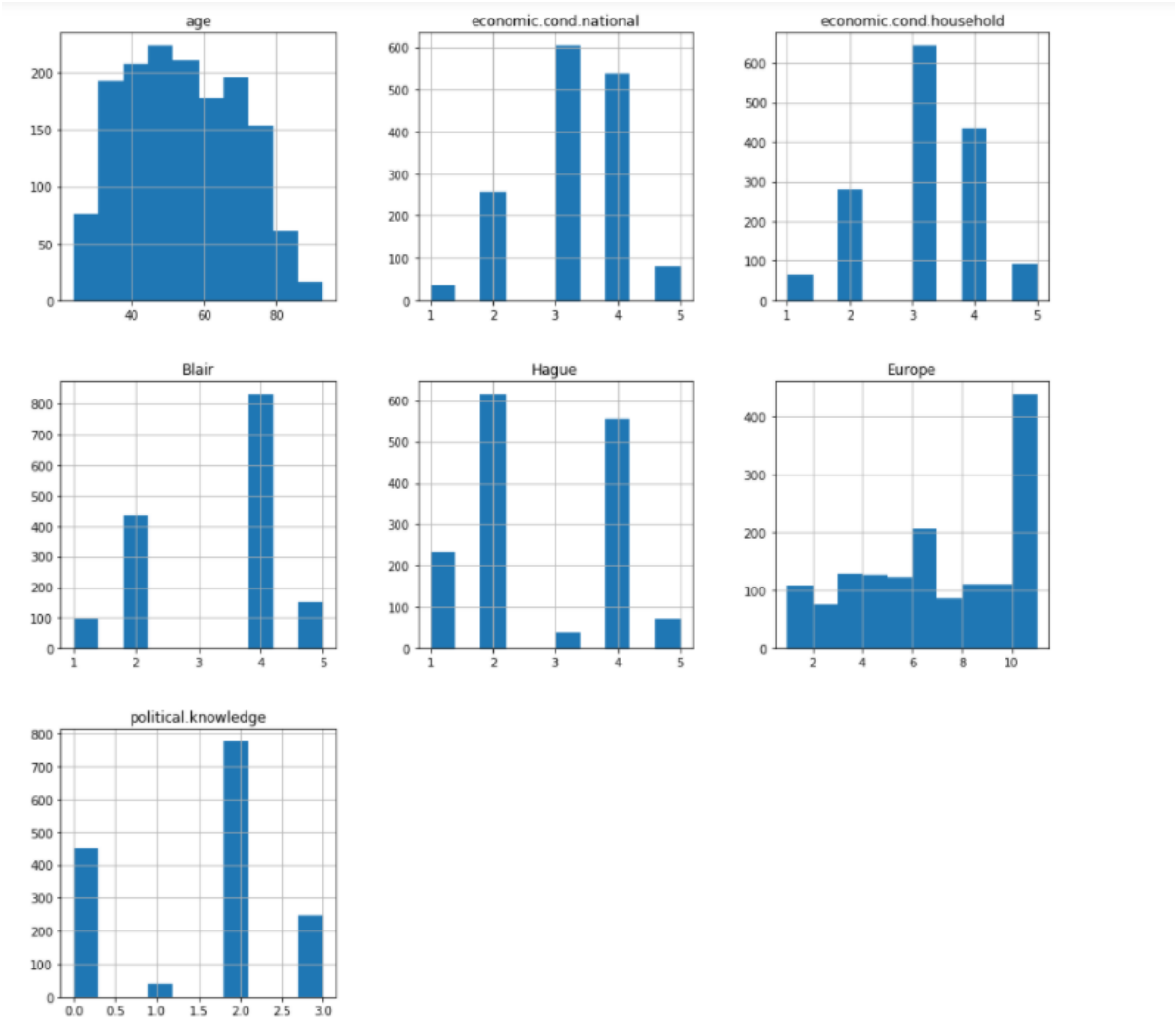


Figure 4: histplot

From above pairplot and histogram we can see that the variable ‘Europe’ has right skewed.

```
df.corr().T
```

	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

```
plt.figure(figsize=(10,8))
sns.heatmap(df.corr(),annot=True,fmt=".2f")
plt.show()
```

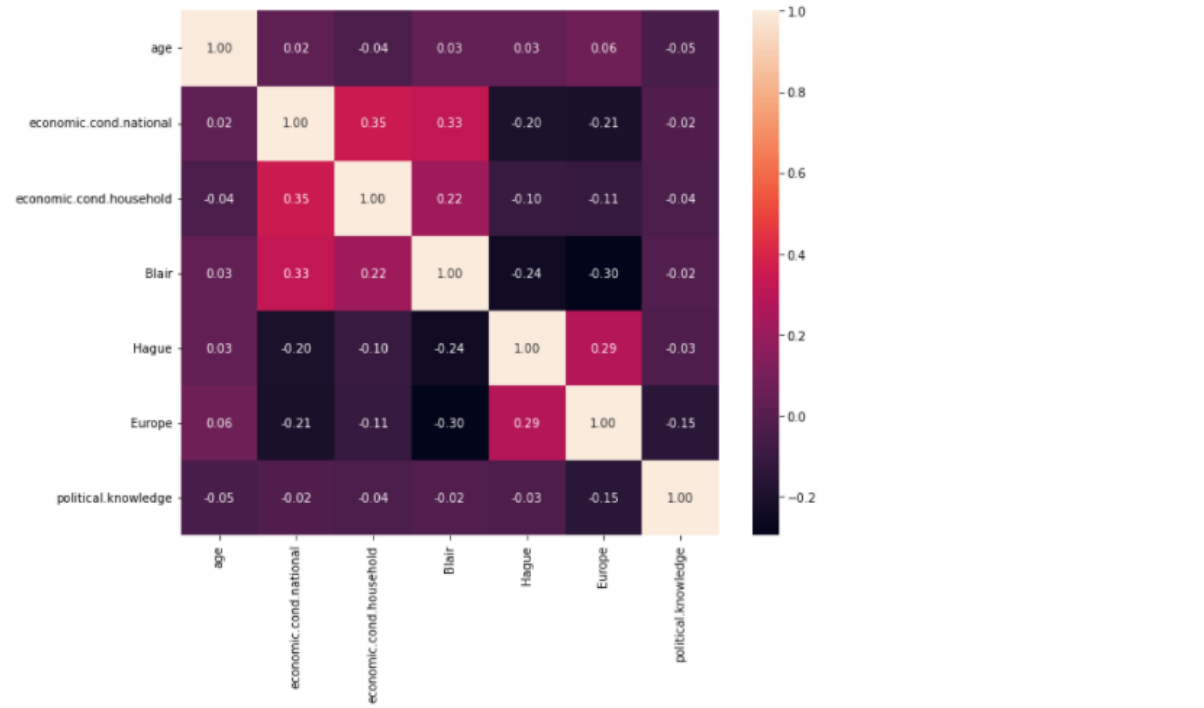


Figure 5: Heatmap for checking coorelation between all variables

Overall the categories in the data do not look very well correlated. Listing down a few observations from Heatmap below:

1. Negative Correlation is an indication that mentioned variables move in the opposite direction who ever is voting for Blair is obviously not voting for Hague. Hence there is a negative correlation between the two indicating cause and effect relationship between the variables.

2. In general, correlation values of -0.30 and + 0.30 represent weak correlation.Variables "Blair" and "Hague" both have weak correlation with national and household economic conditions but "Blair" has slightly better correlation with these parameters (not much of a difference).
3. National economic conditions has very weak correlation with household economic condition.

```
<AxesSubplot:xlabel='vote', ylabel='economic.cond.national'>
```

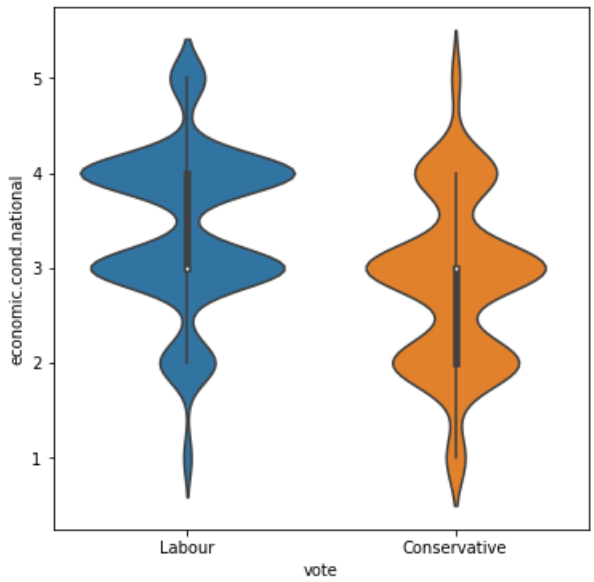


Figure 6.1: violin plot for vote and economic.cond.national

```
<AxesSubplot:xlabel='vote', ylabel='economic.cond.household'>
```

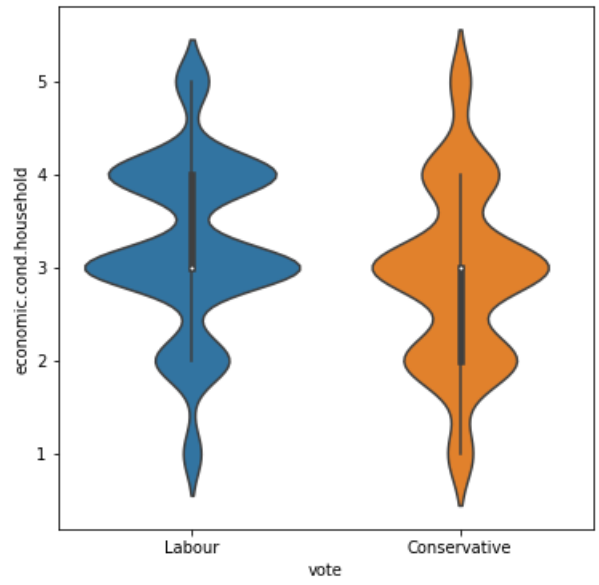


Figure 6.2: violin plot for vote and economic.cond.household

```
<AxesSubplot:xlabel='vote', ylabel='Blair'>
```

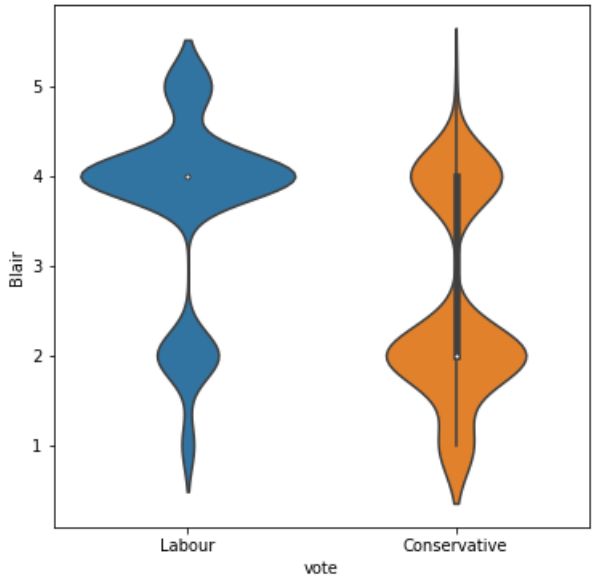


Figure 6.3: violin plot for vote and Blair

```
<AxesSubplot:xlabel='vote', ylabel='Hague'>
```

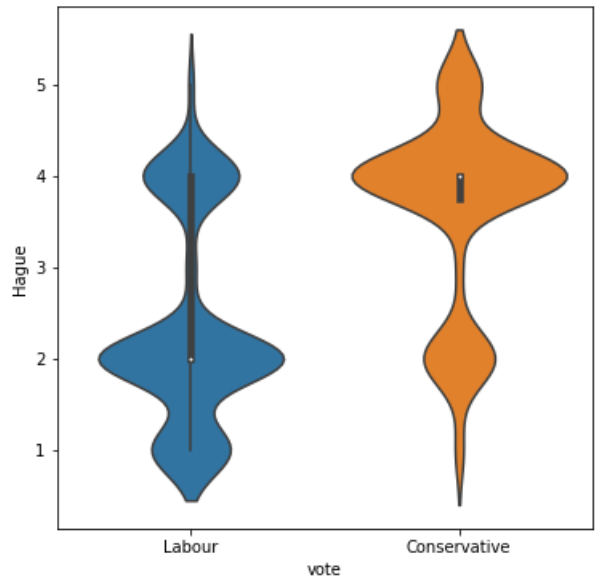


Figure 6.4: violin plot for vote and Hague

<AxesSubplot:xlabel='vote', ylabel='Europe'>

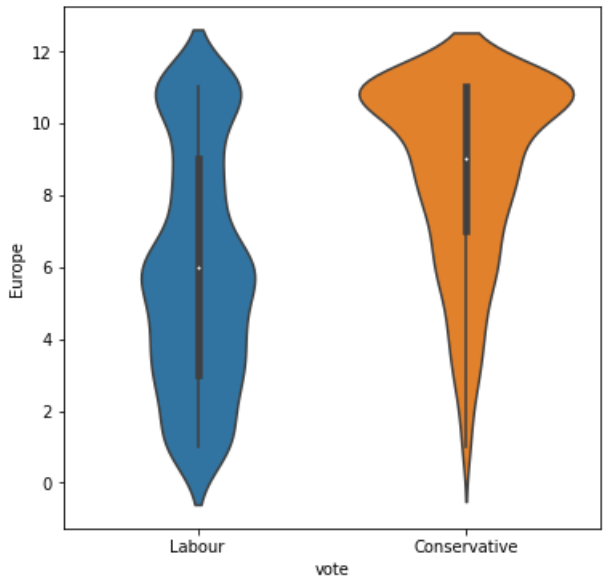


Figure 6.5: violin plot for vote and Europe

- Listing down a few observations from Violin Plots below:

- the white dot represents the median the thick gray bar in the center represents the interquartile range the thin gray line represents the rest of the distribution, except for points that are determined to be “outliers” using a method that is a function of the interquartile range.

- violin is “fatter”, there are more data points in the neighbor hood. And where it is “thinner”, there are less.

Check for outliers

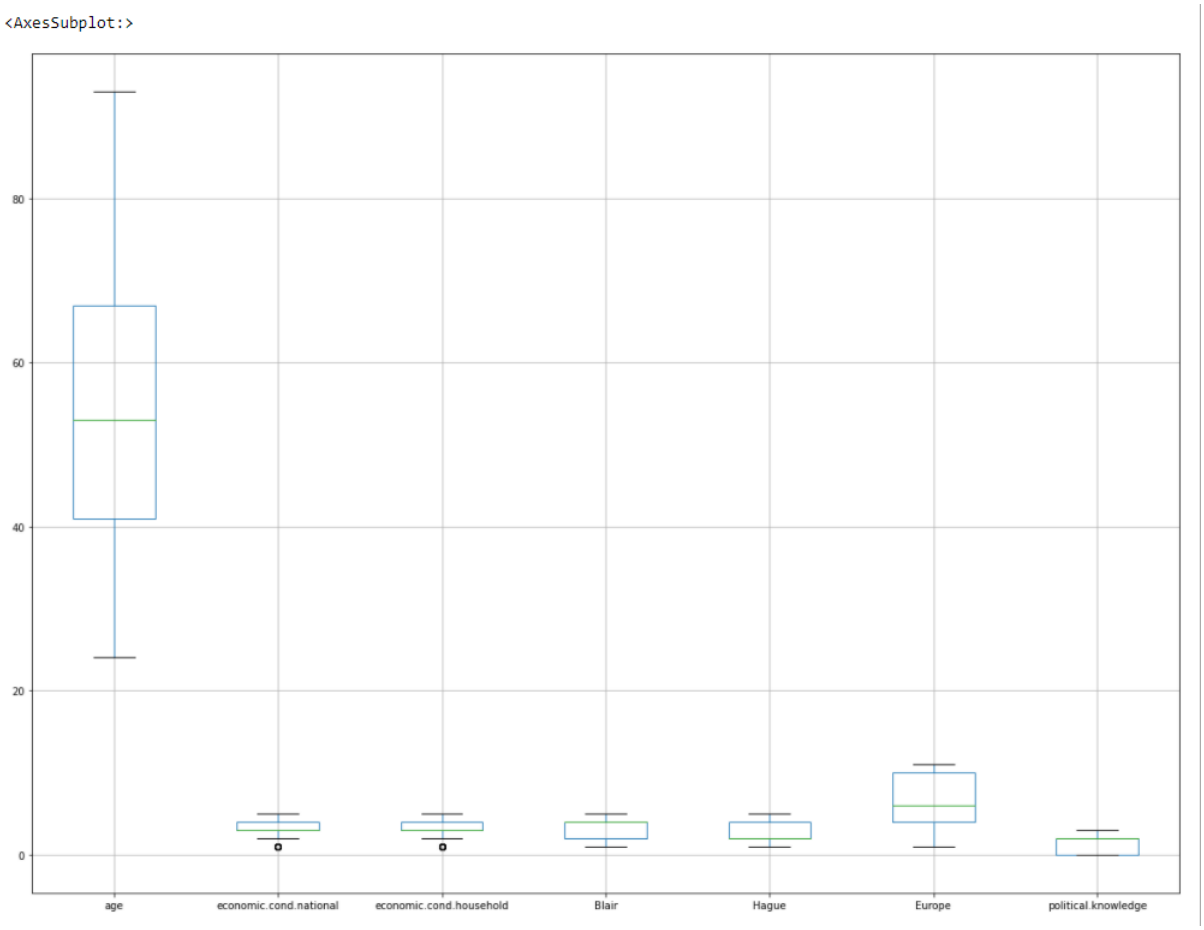


Figure 7: boxplot for checking outliers and before removing outliers

As we can see above that we have outliers in economic.cond.national & economic.cond.household

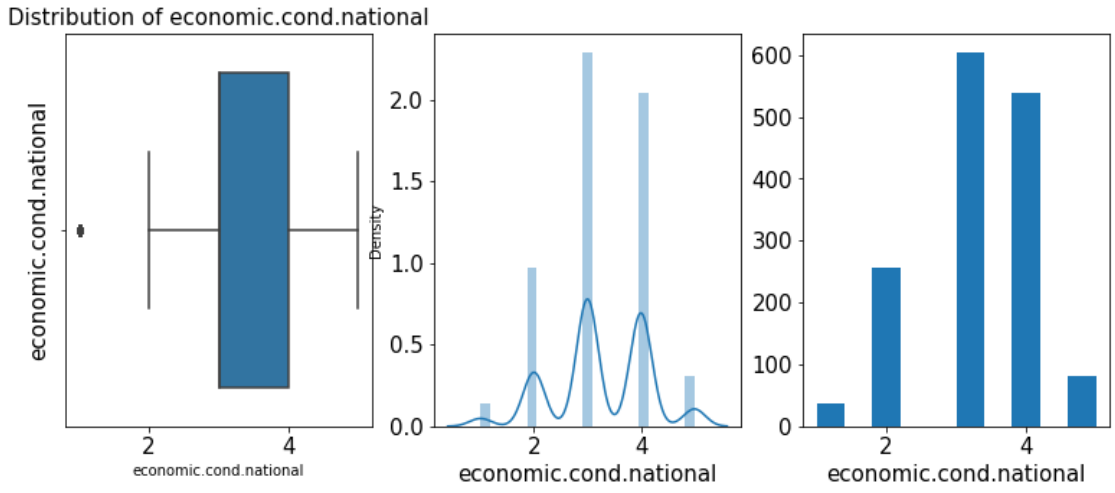


Figure 8.1: Distribution of economic.cond.national

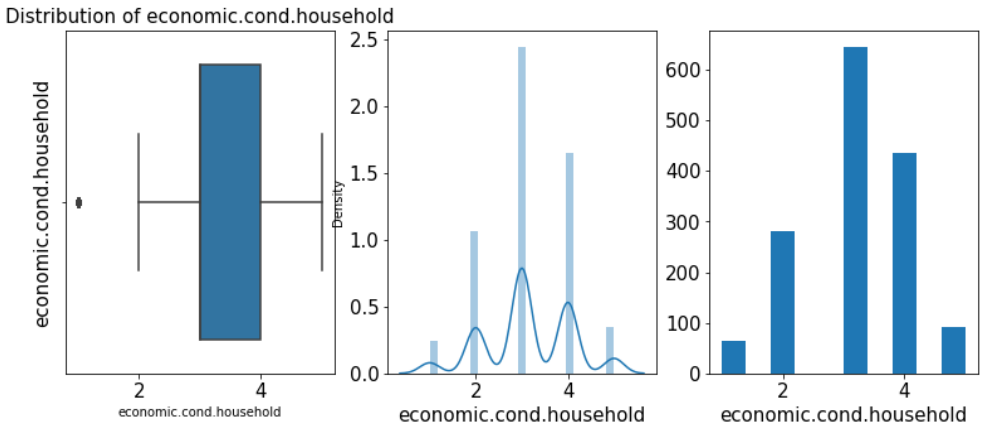


Figure 8.2: Distribution of economic.cond.household

So we need to remove both the outliers

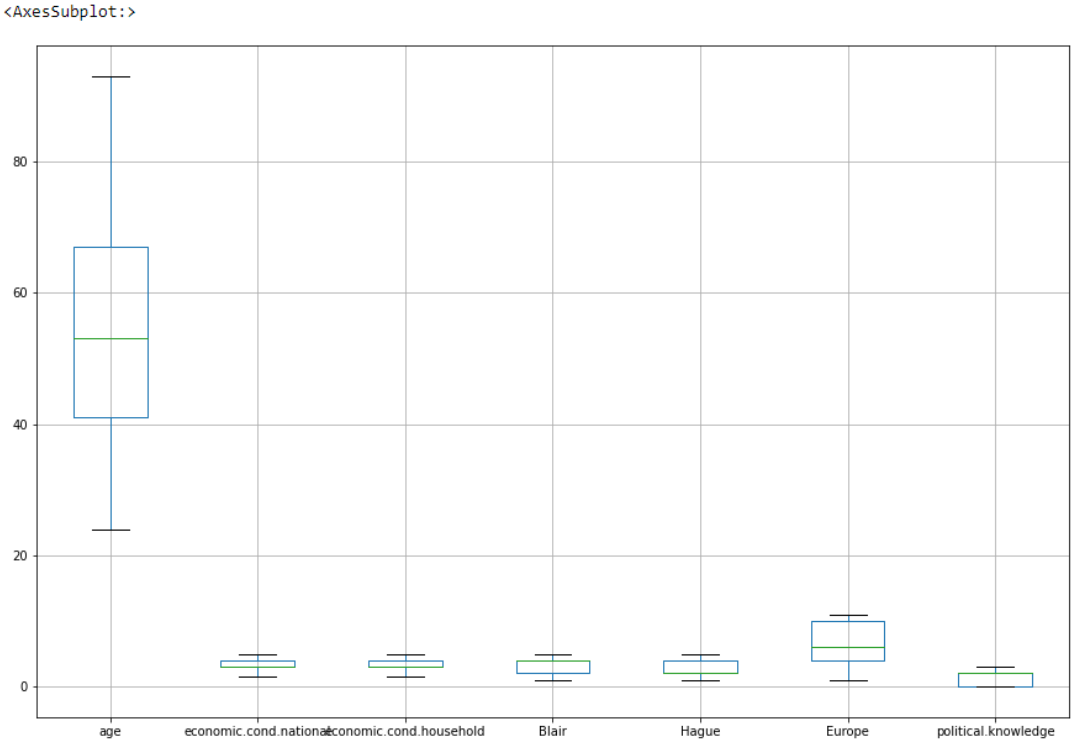


Figure 9: boxplot for checking outliers and after removing outliers

Skewness after removing the outlier

```
Hague                       0.146191
age                         0.139800
economic.cond.household     0.091833
economic.cond.national     -0.069946
Europe                     -0.141891
political.knowledge        -0.422928
Blair                      -0.539514
dtype: float64
```

Distribution and skewness of variables:

Skewness is the measure of how much the probability distribution of a random variable deviates from thenormal distribution, it explains the extent to which the data is normally distributed.Now, looking at the abovevalues we can state the following:

• Distribution is skewed to left tail for all the variable except for variables age and Hague, which has right tail.

• Also, since the skewness is ranging between -0.5 and 0.5 we can say that data is moderately skewed.

• Negative skew refers to a longer or fatter tail on the left side of the distribution, while positive skew refers to alonger or fatter tail on the right. The mean of positively skewed data will be greater than the median.

#Bivariate analysis is one of the simplest forms of quantitative analysis. It involves the analysis of two variables (X, Y), for the purpose of determining the empirical relationship between them.It is the analysis of the relationship between the two variables.

Question 1.3

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

models. (pd.categorical().codes(), pd.get_dummies(drop_first=True)) Data split, ratio defined for the split, train-test split should be discussed.

Answer:

	count	mean	std	min	25%	50%	75%	max
age	1517.0	54.241266	15.701741	24.0	41.0	53.0	67.0	93.0
economic.cond.national	1517.0	3.257416	0.853647	1.5	3.0	3.0	4.0	5.0
economic.cond.household	1517.0	3.159196	0.886279	1.5	3.0	3.0	4.0	5.0
Blair	1517.0	3.335531	1.174772	1.0	2.0	4.0	4.0	5.0
Hague	1517.0	2.749506	1.232479	1.0	2.0	2.0	4.0	5.0
Europe	1517.0	6.740277	3.299043	1.0	4.0	6.0	10.0	11.0
political.knowledge	1517.0	1.540541	1.084417	0.0	0.0	2.0	2.0	3.0

Data Encoding:

There are three common approaches for converting ordinal and categorical variables to numerical values.

Theyare:

- Ordinal Encoding
- One-Hot Encoding
- Dummy Variable Encoding

Here, we will use Dummy Variable Encoding to convert each category into a separate column containing only 0and 1, where 1 indicates presence and 0 indicates absence. In this case:

- gender_male- 1: Male and 0: No Male

- vote_Labour- 1: Voted Labour Party and 0: Not Voted Labour Party.

```
dff=pd.get_dummies(df, columns=cat1,drop_first=True)
dff.head()
```

	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	vote_Labour	gender_male
0	43	3.0	3.0	4	1	2	2	1	0
1	36	4.0	4.0	4	4	5	2	1	1
2	35	4.0	4.0	5	2	3	2	1	1
3	24	4.0	2.0	2	1	4	0	1	0
4	41	2.0	2.0	1	1	6	2	1	1

```
dff.vote_Labour.value_counts()
```

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

```
dff.gender_male.value_counts()
```

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

Here, we have used Drop First as True to ensure that levels of categorical variables are not included as multiple columns in dataset might result in multicoliinearity which in turn land into a dummy trap.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1517 entries, 0 to 1524
Data columns (total 9 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   age                     1517 non-null   int64
 1   economic.cond.national  1517 non-null   float64
 2   economic.cond.household 1517 non-null   float64
 3   Blair                   1517 non-null   int64
 4   Hague                   1517 non-null   int64
 5   Europe                  1517 non-null   int64
 6   political.knowledge     1517 non-null   int64
 7   vote_Labour             1517 non-null   uint8
 8   gender_male             1517 non-null   uint8
dtypes: float64(2), int64(5), uint8(2)
memory usage: 130.1 KB
```

After getting dummies we can see that the object datatype has been converted to float datatype.

confirms that all the categorical data is converted to numerical data now.

We will divide the data into Training and Testing data set, with 70:30 proportion with the fixed random state as1 to ensure uniformity across multiple systems. Before we do the train-test split , we will first separate independent (X and dependent(v)variables to perform Train-Test-split)

Train the data in 70:30 into train and test

```
X= dff.drop("vote_Labour",axis=1)
y= dff[["vote_Labour"]]
```

```
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.3,random_state=1)
```

```
X_train.head()
```

	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender_male
991	34	2.0	4.0	1	4	11	2	0
1274	40	4.0	3.0	4	4	6	0	1
649	61	4.0	3.0	4	4	7	2	0
677	47	3.0	3.0	4	2	11	0	1
538	44	5.0	3.0	4	2	8	0	1

```
y_train.head()
```

	vote_Labour
991	0
1274	1
649	0
677	1
538	1

Scaling:

In general, algorithms that exploit distances or similarities (e.g. in the form of scalar product) between data samples are sensitive to feature transformations i.e. Feature Scaling is performed when we are dealing with Gradient Descent Based algorithms (Linear and Logistic Regression, Neural Network) and Distance-based algorithms (KNN, K-means, SVM) as these are very sensitive to the range of the data points.

Note:

- The Machine Learning algorithms that require the feature scaling are mostly KNN (K-Nearest Neighbours), Neural Networks, Linear Regression, and Logistic Regression.

- The machine learning algorithms that do not require feature scaling is mostly non-linear ML algorithms such as Decision trees, Random Forest, AdaBoost, Naïve Bayes, etc.

Here, we are building a model, to predict which party a voter will vote for on the basis of the given information and to create an exit poll that will help in predicting overall win and seats covered by a particular party. In order to do our analysis we are expected to build model using Logistic Regression, LDA, KNN Model and NaïveBayes Model.For now we are not scaling the data and will do the scaling based on the models we will run ahead. Hence, as mentioned scaling might be necessary for two models and might not be necessary for the other two.

Question 1.4

Apply Logistic Regression and LDA (Linear Discriminant Analysis) (2 pts). Interpret the inferences of both model s (2 pts). Successful implementation of each model. Logical reason behind the selection of different values for the parameters involved in each model. Calculate Train and Test

Accuracies for each model. Comment on the validness of models (over fitting or under fitting)

Answer:

Logistic Regression

Logistic regression is a linear model for classification rather than regression. It is also known as logit regression. In this model, the probabilities describing the possible outcomes of a single trial are modeled using a logistic function.

Note:

- Regularization is applied by default, which is common in machine learning but not in statistics.

- Another advantage of regularization is that it improves numerical stability. No regularization amounts to setting C to a very high value.

There are two methods to solve a Logistic Regression problem:

1. Stats Model

2. Scikit Learn

Here, we will use Grid Search (scikit learn method) to find the optimal hyperparameters of a model which results in the most 'accurate' predictions and get the best parameters.

Note: Grid search is the process of performing hyper parameter tuning in order to determine the optimal values for a given model. This is significant as the performance of the entire model is based on the hyper parameter values specified.

The parameters used in GridsearchCV can be explained as :

• param_grid : requires a list of parameters and the range of values for each parameter of the specified estimator

• estimator: requires the model we are using for the hyper parameter tuning process

• cross-validation (cv):performed in order to determine the hyper parameter value set which provides the best accuracy levels.

• N_jobs: controls the number of cores on which the package will attempt to run in parallel.

• solver is a string ('liblinear' by default) that decides what solver to use for fitting the model. Other options are'newton-cg', 'lbfgs', 'sag', and 'saga'.

• max_iter is an integer (100 by default) that defines the maximum number of iterations by the solver duringmodel fitting.

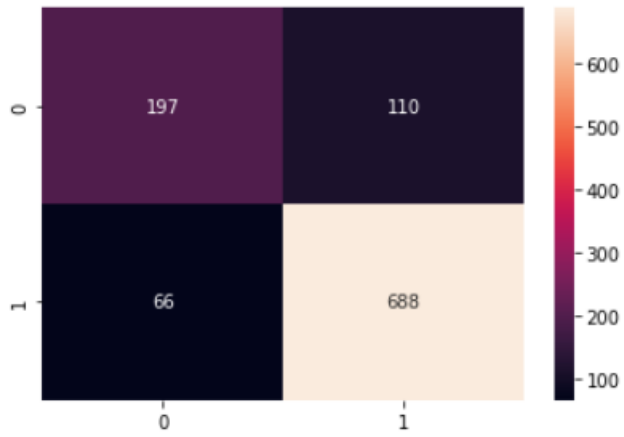
• verbose is a non-negative integer (0 by default) that defines the verbosity for the 'liblinear' and 'lbfgs' solvers.

Logistic Regression Model Without Model Tuning-Train Dataset

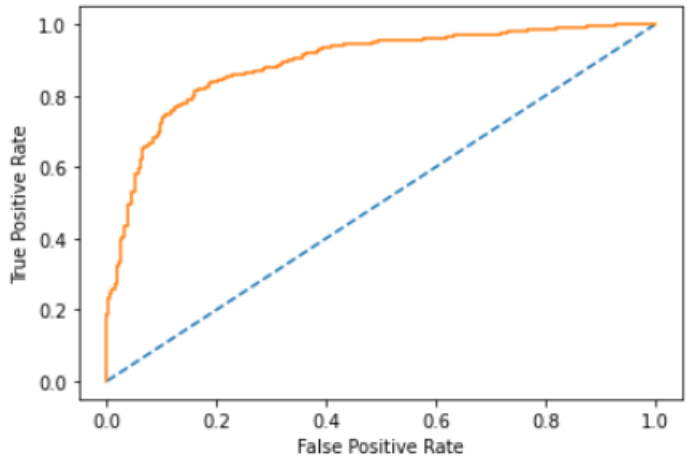
```
0.8341187558906692
[[197 110]
 [ 66 688]]
              precision    recall  f1-score   support

           0       0.75      0.64      0.69       307
           1       0.86      0.91      0.89       754

    accuracy                           0.83      1061
   macro avg       0.81      0.78      0.79      1061
weighted avg       0.83      0.83      0.83      1061
```

```
y_train_prob=Logistic_model.predict_proba(X_train)
pd.DataFrame(y_train_prob).head()
```

	0	1
0	0.933264	0.066736
1	0.095272	0.904728
2	0.293630	0.706370
3	0.112030	0.887970
4	0.016233	0.983767



```
AUC of LR model without Grid Search for Train Data is: 0.890
Logistic Model Score for Test Data 0.8289473684210527
```

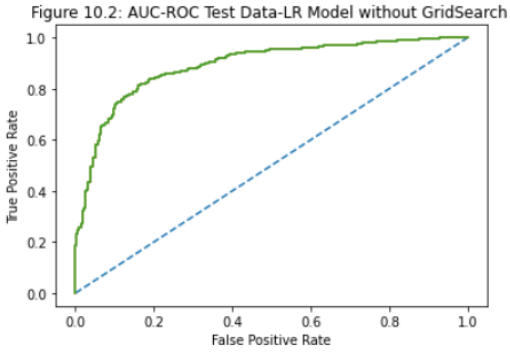


Logistic Regression Model Without Model Tuning-Test Dataset

```
0.8289473684210527
[[111  42]
 [ 36 267]]
              precision    recall  f1-score   support

           0       0.76      0.73      0.74       153
           1       0.86      0.88      0.87       303

    accuracy                           0.83       456
   macro avg       0.81      0.80      0.81       456
weighted avg       0.83      0.83      0.83       456
```

Logistic model score for test data 0.8289473684210527

```
AUC of LR model without Grid Search for Test Data is: 0.883

Text(0.5, 1.0, 'Figure 10.2: AUC-ROC Test Data-LR Model without GridSearch ')
```



Inference of Logistic Regression Model Without GridSearch:

Using the confusion matrix, the True Positive, False Positive, False Negative, and True Negative values can be extracted which will aid in the calculation of the accuracy score, precision score, recall score,

and f1score.Listing below model performance metrics before fine tuning the model**:**

Train Data:

True Positive:197

False Positive:66

False Negative:110

True Negative:688

AUC: 89%

Accuracy: 83%

Precision: 86%

f1-Score: 89%

Recall:91%

Test Data:

True Positive:111

False Positive:36

False Negative:42

True Negative:267

AUC: 88.3%

Accuracy: 83%

Precision: 86%

f1-Score: 87%

Recall:88%

We know that, FPR tells us what proportion of the negative class got incorrectly classified by the classifier. Here, we have higher TNR and a lower FPR which is desirable to classify the negative class.

Here, both Type I Error (False Positives) and Type II Error (False Negatives) are low indicating high Sensitivity/Recall, Precision,Specificity and F1 Score.

Accuracy of the model is more than 70%, which can be considered as a good accuracy score.

Train and Test data scores are mostly in line and the overall performance of model looks good. Hence, it can be inferred that overall this model can be considered as a good model**.**

LDA

Linear Discriminant Analysis Without Tuning-Train Dataset

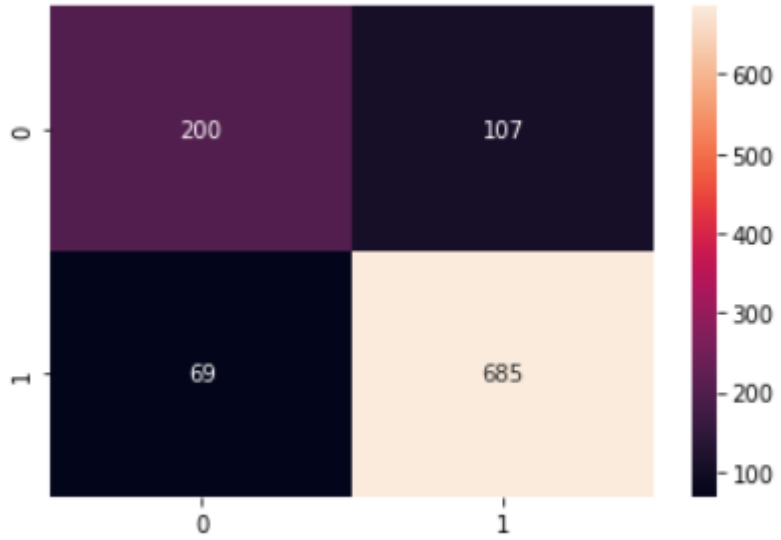
- Training Data Class Prediction with a cut-off value of 0.5

```
LDA Model Score for Training Data without GridSearch is 0.8341187558906692
```

Confusion matrix and classification report

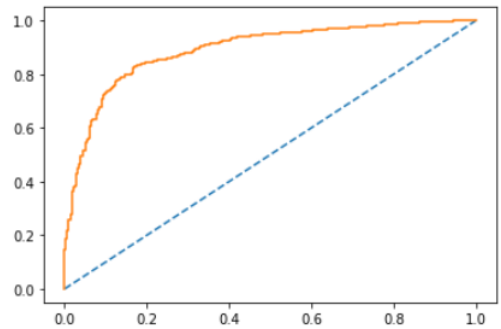
```
[[200 107]
 [ 69 685]]
              precision    recall  f1-score   support

           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
```

Confusion matrix of LDA model on Train data



AUC_ROC Curve LDA Model- Train Data

```
the auc curve for train 0.890

[<matplotlib.lines.Line2D at 0x1d9ad3aed90>]
```



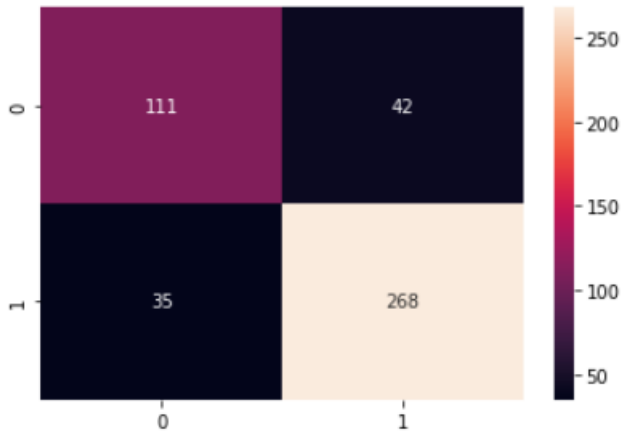
Linear Discriminant Analysis Without Tuning-Test Dataset

```
LDA Model Score for Test Data is 0.831140350877193
```

Confusion matrix and classification report

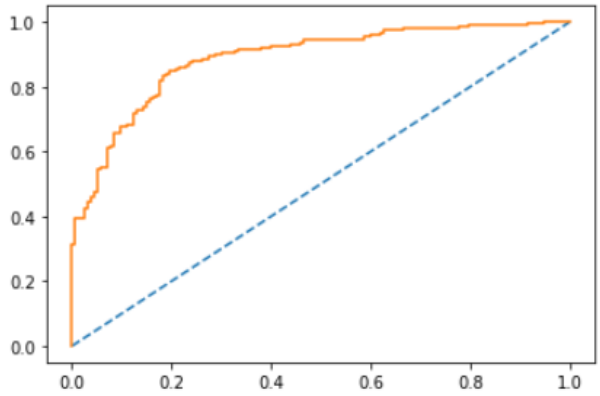
```
[[111  42]
 [ 35 268]]
              precision    recall  f1-score   support

           0       0.76      0.73      0.74       153
           1       0.86      0.88      0.87       303

    accuracy                           0.83       456
   macro avg       0.81      0.80      0.81       456
weighted avg       0.83      0.83      0.83       456
```

Confusion Matrix of LDA Model-Test Data



AUC_ROC Curve LDA Model-Test Data

```
the auc curve for test 0.888

[<matplotlib.lines.Line2D at 0x1d9ad4f9bb0>]
```



Inference of LDA Model Without GridSearch:

Using the confusion matrix, the True Positive, False Positive, False Negative, and True Negative values can be extracted which will aid in the calculation of the accuracy score, precision score, recall score, and f1score.Listing below model performance metrics before fine tuning the model:

Train data:

True positive:200

False Positive:69

False Negative:107

True Negative:685

AUC: 89%

Accuracy: 83%

Precision: 86%

f1-Score: 89%

Recall:91%

Test Data:

True Positive:111

False Positive:35

False Negative:42

True Negative:268

AUC: 88.8%

Accuracy: 83%

Precision: 86%

f1-Score: 87%

Recall:88%

We know that, FPR tells us what proportion of the negative class got incorrectly classified by the classifier .Here, we have higher TNR and a lower FPR which is desirable to classify the negative class.

Here, both Type I Error (False Positives) and Type II Error (False Negatives) are low indicating high Sensitivity/Recall, Precision,Specificity and F1 Score.

Accuracy of the model is more than 70%, which can be considered as a good accuracy score.

Train and Test data scores are mostly in line and the overall performance of model looks good. Hence, it can be inferred that overall this model can be considered as a good model.

Question 1.5

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

Answer:

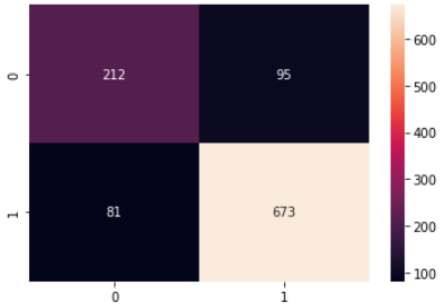
NB Without Tuning-Train Dataset

```
y_train_predict=NB_model.predict(X_train)
NB_model_score=NB_model.score(X_train, y_train)
NB_model_score
```

```
0.8341187558906692
```

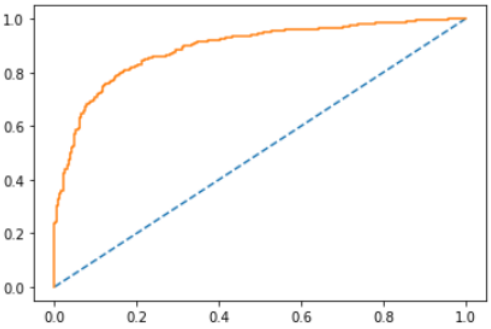
Confusion Matrix of NB Model-Train Data

```
[[212  95]
 [ 81 673]]
              precision    recall  f1-score   support

           0       0.72      0.69      0.71       307
           1       0.88      0.89      0.88       754

    accuracy                           0.83      1061
   macro avg       0.80      0.79      0.80      1061
weighted avg       0.83      0.83      0.83      1061
```



AUC_ROC Curve NB Model-Train Data

```
the auc cure for training set 0.889

[<matplotlib.lines.Line2D at 0x1d9ad60ef10>]
```



NB Model Without Tuning-Test Dataset

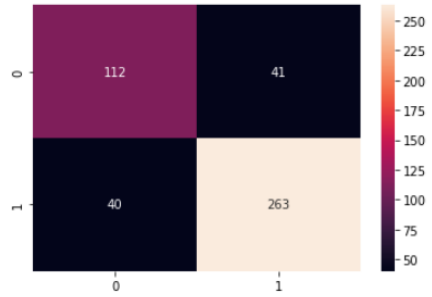
```
y_test_predict=NB_model.predict(X_test)
NB_model_score=NB_model.score(X_test, y_test)
NB_model_score
```

```
0.8223684210526315
```

Confusion Matrix of NB Model Without GridSearch-Test Data

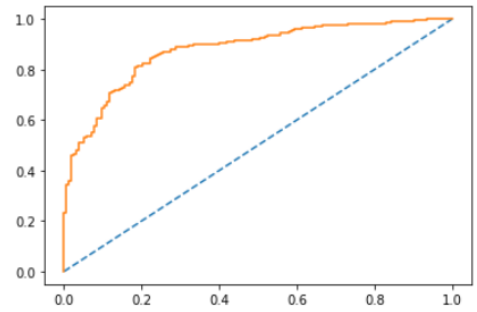
```
[[112  41]
 [ 40 263]]
              precision    recall  f1-score   support

           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
```

```
mt=sns.heatmap(metrics.confusion_matrix(y_test,y_test_predict),annot=True,fmt='d')
```



AUC_ROC Curve NB Model-Test Data

```
the auc cure for testing set 0.876

[<matplotlib.lines.Line2D at 0x1d9ad919700>]
```



Inference of NB Model Without GridSearch:

Using the confusion matrix, the True Positive, False Positive, False Negative, and True Negative values can be extracted which will aid in the calculation of the accuracy score, precision score, recall score, and f1score.Listing below model performance metrics without fine tuning the model:

Train data

True Positive:212

False Positive:81

False Negative:95

True Negative:673

AUC: 88.9%

Accuracy: 83%

Precision: 88%

f1-Score: 88%

Recall:89%

Test data

True Positive:112

False Positive:40

False Negative:41

True Negative:263

AUC: 87.6%

Accuracy: 82%

Precision: 87%

f1-Score: 87%

Recall:87%

We know that, FPR tells us what proportion of the negative class got incorrectly classified by the classifier. Here, we have higher TNR and a lower FPR which is desirable to classify the negative class.

Here, both Type I Error (False Positives) and Type II Error (False Negatives) are low indicating highSensitivity/Recall, Precision,Specificity and F1 Score.

Accuracy of the model is more than 70%, which can be considered as a good accuracy score.

Train and Test data scores are mostly in line and the overall performance of model looks good. Hence, it can be inferred that overall this model can be considered as a good model.

KNN Model

Neighbors-based classification is a type of instance-based learning or non-generalizing learning: it does notattempt to construct a general internal model, but simply stores instances of the training data. Classification iscomputed from a simple majority vote of the nearest neighbors of each point: a query point is assigned thedata class which has the most representatives within the nearest neighbors of the point.

The KNN algorithm uses ‘feature similarity’ to predict the values of any new data points. This means that thenew point is assigned a value based on how closely it resembles the points in the training set.

KNN has thefollowing basic steps:

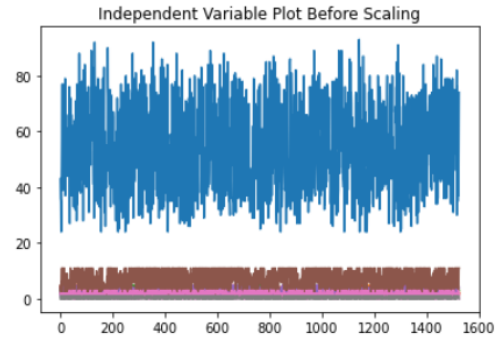
1. Calculate distance

2. Find closest neighbors

3. Vote for labels

Independent Variable Plot Before Scaling

```
Text(0.5, 1.0, 'Independent Variable Plot Before Scaling')
```



Default value of n_neighbors is equal to 5.First we will build KNN Model with k=5.

KNN model score for scaled train data

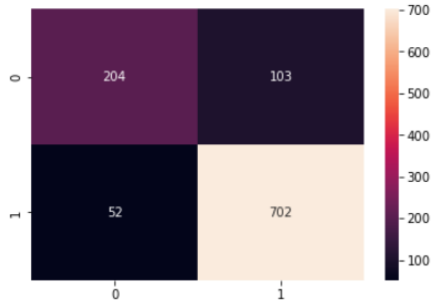
```
y_train_predict=KNN_model.predict(X_train)
KNN_model_score=KNN_model.score(X_train,y_train)
KNN_model_score
```

```
0.8539114043355325
```

Confusion matrix and classification report

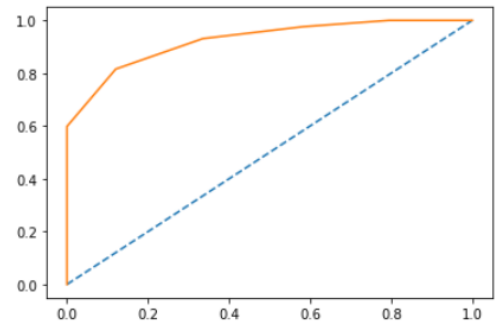
```
[[204 103]
 [ 52 702]]
              precision    recall  f1-score   support

           0       0.80      0.66      0.72       307
           1       0.87      0.93      0.90       754

    accuracy                           0.85      1061
   macro avg       0.83      0.80      0.81      1061
weighted avg       0.85      0.85      0.85      1061
```

```
mx=sns.heatmap(metrics.confusion_matrix(y_train,y_train_predict),annot=True, fmt='d')
```



AUC_ROC Curve KNN Model-Train Data (k=5)

```
the auc 0.923

[<matplotlib.lines.Line2D at 0x26746ccc160>]
```



KNN model score for scaled test data

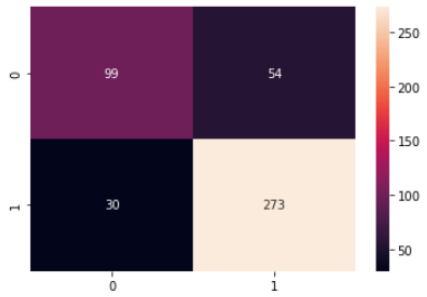
```
y_test_predict=KNN_model.predict(X_test)
KNN_model_score=KNN_model.score(X_test, y_test)
KNN_model_score
```

```
0.8157894736842105
```

Confusion matrix and classification report

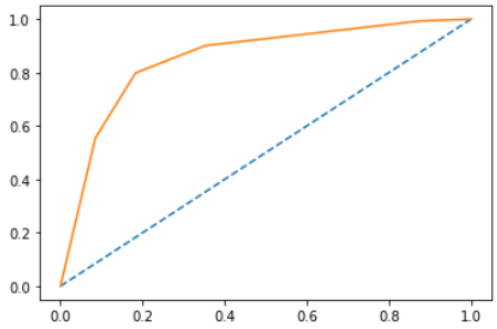
```
[[ 99  54]
 [ 30 273]]
              precision    recall  f1-score   support

           0       0.77      0.65      0.70       153
           1       0.83      0.90      0.87       303

    accuracy                           0.82       456
   macro avg       0.80      0.77      0.78       456
weighted avg       0.81      0.82      0.81       456
```

```
mx=sns.heatmap(metrics.confusion_matrix(y_test,y_test_predict),annot=True,fmt='d')
```



AUC_ROC Curve KNN Model-Test Data (k=5)

```
the auc curve for testinf data 0.853
[<matplotlib.lines.Line2D at 0x26747da1250>]
```



Default value of n_neighbors is equal to 5 has given below model performance:

Train Data:

AUC: 92.3%
Accuracy: 85%
Precision: 87%
f1-Score: 90%
Recall:93%

Test Data:

AUC: 85.3%
Accuracy: 82%
Precision: 83%
f1-Score: 87%
Recall:90%

We can see a considerable difference in model AUC between Train and Test Data while the other parameters are mostly in line.
Lets check the performance of the model for K=7.

Building KNN model with n_neighbours=7

KNN model for scaled train dataset

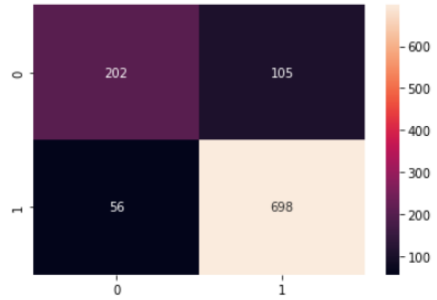
```
y_train_predict=KNN_model.predict(X_train)
KNN_model_score=KNN_model.score(X_train,y_train)
KNN_model_score
```

```
0.8482563619227145
```

Confusion matrix and classification report

```
[[202 105]
 [ 56 698]]
              precision    recall  f1-score   support

           0       0.78      0.66      0.72       307
           1       0.87      0.93      0.90       754

    accuracy                           0.85      1061
   macro avg       0.83      0.79      0.81      1061
weighted avg       0.84      0.85      0.84      1061
```

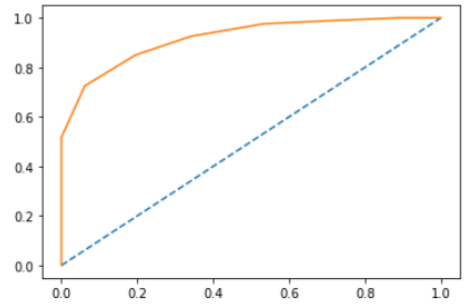
```
mx=sns.heatmap(metrics.confusion_matrix(y_train,y_train_predict),annot=True,fmt='d')
```



AUC_ROC Curve KNN Model-Train Data (K=7)

```
the auc 0.918
```

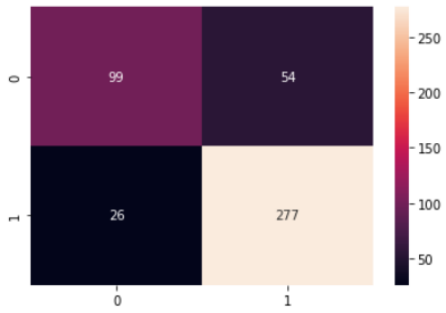
```
[<matplotlib.lines.Line2D at 0x26747ea4b20>]
```



KNN model score for scaled test dataset

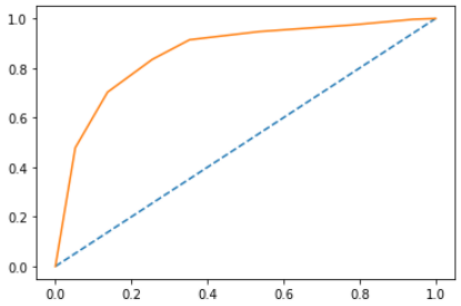
```
[[ 99  54]
 [ 26 277]]
              precision    recall  f1-score   support

           0       0.79      0.65      0.71       153
           1       0.84      0.91      0.87       303

    accuracy                           0.82       456
   macro avg       0.81      0.78      0.79       456
weighted avg       0.82      0.82      0.82       456
```

```
ax=sns.heatmap(metrics.confusion_matrix(y_test,y_test_predict),annot=True, fmt='d')
```



AUC_ROC Curve KNN Model-Test Data (K=7)

```
the auc curve 0.861

[<matplotlib.lines.Line2D at 0x26747fa7d30>]
```



Insights with k as 7 on model performance are as follows:

Train Data:

AUC: 92.3%

Accuracy: 85%

Precision: 87%

f1-Score: 90%

Recall:93%

Test Data:

AUC: 85.3%

Accuracy: 82%

Precision: 84%

f1-Score: 87%

Recall:91%

Inference of K-Nearest Mean (KNN) Model:

KNN Model Score for Scaled Train Data for k=5 is 0.8539

KNN Model Score for Scaled Test Data for k=5 is 0.8157

KNN Model Score for Scaled Train Data with K=7 is 0.8482

KNN Model Score for Scaled Test Data with K=7 is 0.8245

There is a slight improvement in Accuracy Score for Test data with K=7

Accuracy score of 85% is generally considered a good accuracy score

Further, to find the optimal value of k we will look at the K=1,3,5,7....19 and store the train and test scores in aDataframe (ac_score) and using these scores, we will calculate the Misclassification error (MCE) and find themodel with lowest Misclassification error (MCE) using the below mentioned formula: Misclassification error(MCE) = 1 - Test accuracy score.

To find more value of K we need to do (MCE)

Misclassification error(MCE) = 1 - Test accuracy score

```
ac_score=[]
for k in range(1,20,2):
    knn= KNeighborsClassifier(n_neighbors=k)
    knn.fit(X_train, y_train)
    scores=knn.score(X_test,y_test)
    ac_score.append(scores)
MCE=[1-x for x in ac_score]
MCE
```

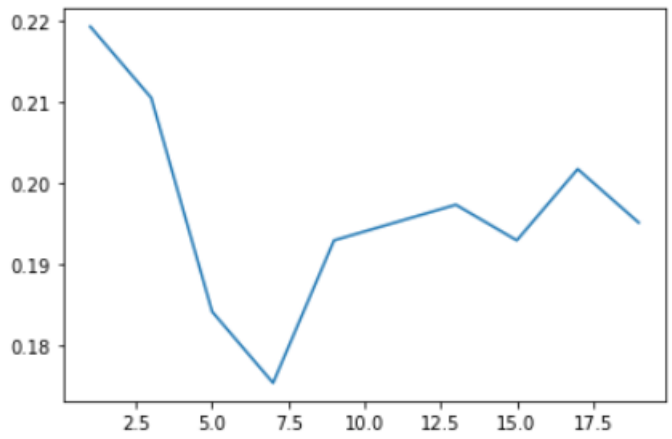
```
[0.2192982456140351,
 0.21052631578947367,
 0.1842105263157895,
 0.17543859649122806,
 0.19298245614035092,
 0.19517543859649122,
 0.19736842105263153,
 0.19298245614035092,
 0.20175438596491224,
 0.19517543859649122]
```

```
ac_score
```

```
[0.7807017543859649,
 0.7894736842105263,
 0.8157894736842105,
 0.8245614035087719,
 0.8070175438596491,
 0.8048245614035088,
 0.8026315789473685,
 0.8070175438596491,
 0.7982456140350878,
 0.8048245614035088]
```

MCE Plot for KNN model

```
[<matplotlib.lines.Line2D at 0x2674801f0a0>]
```



Hence, we can say that the lowest value of Misclassification Error is at k=7. Also, we have seen above that accuracy score for KNN Model at k=7 is 85% which is considered a

good accuracy score and the difference between train and test accuracies is less than 10%, it is a valid model. Therefore, we can say that the optimal value of k is 7 for this particular model.

Question 1.6

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

Answer:

Model Tuning is the process of maximizing a model's performance without overfitting or creating too high of a variance. This is accomplished by selecting appropriate "hyperparameters." which is crucial for model accuracy, but can be computationally challenging. Hyperparameters differ are not learned by the model automatically. Instead, these parameters are set manually. Below mentioned are the three most commonly use approaches:

1. Grid Search- Grid search also known as parameter sweeping.This method involves manually defining a subset of the hyperparametric space and exhausting all combinations of the specified hyperparameter subsets. Each combination's performance is then

evaluated, typically using cross-validation, and the best performing hyperparametric combination is chosen.

2. Random Search- Random search can be said as a basic improvement on grid search. Instead of testing on a predetermined subset of hyperparameters, random search, as its name implies, randomly selects a chosen number of hyperparametric pairs from a given domain and tests only those. This greatly simplifies the analysis without significantly sacrificing optimization. For example, if the region of hyperparameters that are near optimal occupies at least 5% of the grid, then random search with 60 trials will find that region with high probability (95%).

3. Bayesian Optimization- This process builds a probabilistic model for a given function and analyzes this model to make decisions about where to next evaluate the function. It offers an efficient framework for optimising the highly expensive black-box functions without knowing its form. It is an efficient tool for hyperparameter tuning for complex models like deep neural networks.

Here, we will use Grid Search Method for Model tuning.

Naive Bayes Model with Tuning- Grid Search :

Explaining the parameters used to find the optimal combinations :

param_grid_NB: Dictionary that contains all of the parameters to try

var_smoothing : Stability calculation to widen (or smooth) the curve and therefore account for more samples that are further away from the distribution mean.

np.logspace : Returns numbers spaced evenly on a log scale, starting from 0, ending at -9, and generating100 samples

estimator: Machine learning model of interest

verbose is the verbosity: the higher, the more messages; in this case, it is set to 1

cv: cross-validation generator or an iterable, in this case, there is a 10-fold cross-validation.

n_jobs: Maximum number of iterations; in this case, it is set to -1 which implies that all CPUs are used.

Here, we will build the Naive Bayes Model using Gridsearch to find an optimal combination of hyperparameters that minimizes a predefined loss function to give better results.

NB Model With Tuning on Train Data

Here, we will perform model prediction on training and testing data to evaluate the model's accuracy and efficiency after fine tuning the model.

NB model score for gridsearch for train data

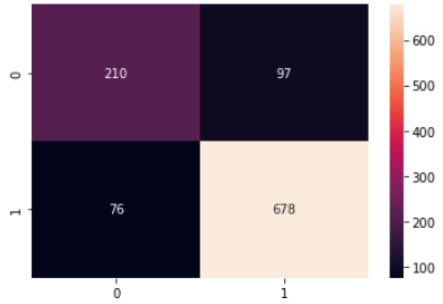
```
y_train_predict=NB_model_grid.predict(X_train)
NB_model_grid_score=NB_model.score(X_train, y_train)
NB_model_grid_score
```

```
0.8341187558906692
```

Confusion Matrix of NB Model-Train Data

```
[[210  97]
 [ 76 678]]
              precision    recall  f1-score   support

           0       0.73      0.68      0.71       307
           1       0.87      0.90      0.89       754

    accuracy                           0.84      1061
   macro avg       0.80      0.79      0.80      1061
weighted avg       0.83      0.84      0.84      1061
```

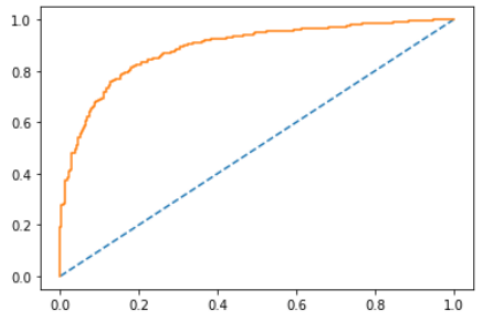
```
ax=sns.heatmap(metrics.confusion_matrix(y_train,y_train_predict),annot=True,fmt='d')
```



AUC-ROC train data NB model

```
AUC of NB Model with GridSearch is 0.888
```

```
[<matplotlib.lines.Line2D at 0x2674814e400>]
```



NB Model With Tuning on Test Data

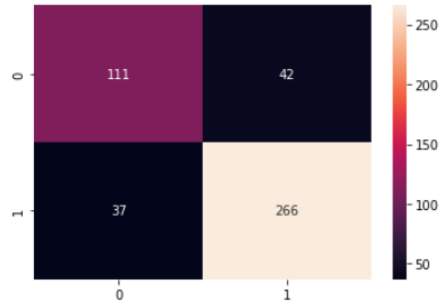
```
y_test_predict=NB_model_grid.predict(X_test)
NB_model_grid_score=NB_model.score(X_test, y_test)
NB_model_grid_score
```

```
0.8223684210526315
```

Confusion matrix and classification report

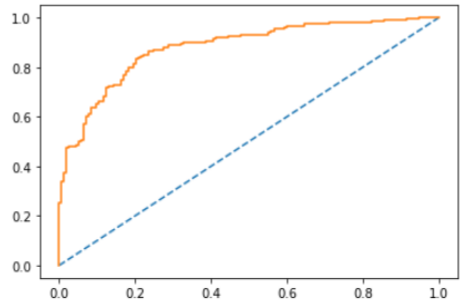
```
[[111  42]
 [ 37 266]]
              precision    recall  f1-score   support

           0       0.75      0.73      0.74       153
           1       0.86      0.88      0.87       303

    accuracy                           0.83       456
   macro avg       0.81      0.80      0.80       456
weighted avg       0.83      0.83      0.83       456
```

```
mx=sns.heatmap(metrics.confusion_matrix(y_test,y_test_predict),annot=True,fmt='d')
```



AUC-ROC test data for NB model

```
AUC of NB Model with GridSearch is 0.879

[<matplotlib.lines.Line2D at 0x2674824a520>]
```



We know that, FPR tells us what proportion of the negative class got incorrectly classified by theclassifier.Here, we have higher TNR and a lower FPR which is desirable to classify the negative class. Here, both Type I Error (False Positives) and Type II Error (False Negatives) are low indicating highSensitivity/Recall, Precision,Specificity and F1 Score. Accuracy of the model is more than 70%, which can be considered as a good accuracy score. Train and Test data scores are mostly in line and the overall performance of model looks good. Hence, it can be inferred that overall this model can be considered as a good model.

After fine tuning the model we can see that model has given mostly the same performance with a very slight improvement in few parameters. Hence, we can say that fine tuning this particular model does not make much of a difference the model performance.

Logistic Regression Model with Tuning- Grid Search

Before using GridSearchCV, listing important parameters below:

estimator: In this we have to pass the models or functions on which we want to use GridSearchCVparam_grid: Dictionary or list of parameters of models or function in which GridSearchCV have to selectthe best.

Scoring: It is used as a evaluating metric for the model performance to decide the best hyperparameters, ifnot especified then it uses estimator score.

solver: string ('liblinear' by default) that decides what solver to use for fitting the model. Other options are'newton-cg', 'lbfgs', 'sag', and 'saga'.Here we are using newton-cg as it adaptively controls the accuracyof the solution without loss of the rapid convergence properties.

max_iter: Defines the maximum number of iterations by the solver during model fitting.Here, we are using10000.

penalty: It imposes a penalty to the logistic model for having too many variables. This results in shrinkingthe coefficients of the less contributive variables toward zero.

verbose: Non-negative integer (0 by default) that defines the verbosity.

n_jobs: controls the number of cores on which the package will attempt to run in parallel.

cv: cross validation generator or an iterable, in this case, there is a 5-fold cross-validation.

scoring: choosing scoring F1 since it computes the Harmonic Mean between Recall and Precision, it tellsus whether both Type I and Type II error is low or high on an average.

Logistic Regression Model- Train Data

Ensemble Machine Learning:

It is a machine learning paradigm where multiple models (often called “weak learners”) are trained to solve thesame problem and combined to get better results. The main hypothesis is that when weak models are correctlycombined we can obtain more accurate and/or robust models.

The three most popular methods for combining the predictions from different models are:

1. Bagging: Building multiple models (typically of the same type) from different subsamples of the trainingdataset.

2. Boosting: Building multiple models (typically of the same type) each of which learns to fix the predictionerrors of a prior model in the sequence of models.

3. Voting: Building multiple models (typically of differing types) and simple statistics (like calculating themean) are used to combine predictions.

Here, we will use the techniques bagging and boosting.

Bagging Idea of Bagging:

To fit several independent models and “average” their predictions in order to obtain a model with a lowervariance. However, in practice, it requires too much data to fit fully independent models . So, we rely on thegood “approximate

properties” of bootstrap samples (representativity and independence) to fit models that arealmost independent.

Note:A Bagging classifier is an ensemble meta-estimator that fits base classifiers each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction. When samples are drawn with replacement, then the method is known as Bagging

Bootstrap Aggregation (or Bagging for short), is a simple and very powerful ensemble method.It is a generalprocedure that can be used to reduce the variance for those algorithm that have high variance. An algorithmthat has high variance are decision trees, like classification and regression trees (CART).Bagging is theapplication of the Bootstrap procedure to a high-variance machine learning algorithm, typically decision trees.

Hence, we will build the Bagging model using Decison Tree as the base estimator and than fit the model.

We know that, Decision trees are sensitive to the specific data on which they are trained. If the training data ischanged (e.g. a tree is trained on a subset of the training data) the resulting decision tree can be quite differentand in turn the predictions can be quite different.Bagging of the

CART algorithm would work as follows:

1. Create many (e.g. 100) random sub-samples of our dataset with replacement.

2. Train a CART model on each sample.

3. Given a new dataset, calculate the average prediction from each model

base_estimator: The base estimator to fit on random subsets of the dataset.Here, we are using DecisionTree Classifier to improve accuracy and reduce variance.

n_estimatorsint:The number of base estimators in the ensemble.HEre we are taking 100.

random_state: Controls the random resampling of the original dataset (sample wise and feature wise).

Bagging Classifier

```
cart=DecisionTreeClassifier()
Bagging_model=BaggingClassifier(base_estimator=cart,n_estimators=100, random_state=1)
Bagging_model.fit(X_train,y_train)
```

```
C:\Users\User\anaconda3\lib\site-packages\sklearn\utils\validation.py:63: DataConversionWarning: A column-vector y was passed w
hen a 1d array was expected. Please change the shape of y to (n_samples, ), for example using ravel().
  return f(*args, **kwargs)
```

```
BaggingClassifier(base_estimator=DecisionTreeClassifier(), n_estimators=100,
                  random_state=1)
```

Bagging model score on train data

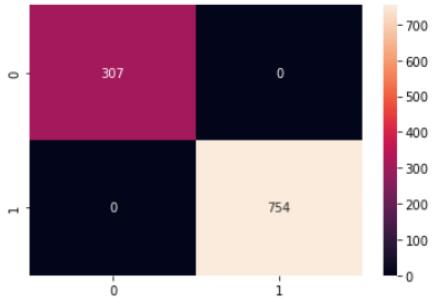
```
y_train_predict=Bagging_model.predict(X_train)
Bagging_model_score=Bagging_model.score(X_train,y_train)
Bagging_model_score
```

```
1.0
```

Confusion matrix and classification report

```
[[307   0]
 [  0 754]]
              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
```

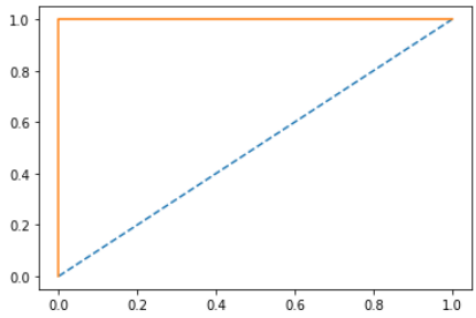
```
ax=sns.heatmap(metrics.confusion_matrix(y_train,y_train_predict),annot=True, fmt='d')
```



AUC-ROC curve

```
AUC: 1.000
```

```
[<matplotlib.lines.Line2D at 0x267484e6520>]
```



Bagging model score for testing data

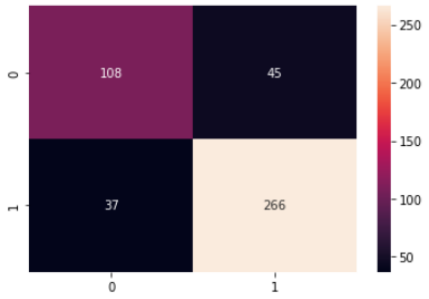
```
y_test_predict=Bagging_model.predict(X_test)
Bagging_model_score=Bagging_model.score(X_test,y_test)
Bagging_model_score
```

```
0.8201754385964912
```

Confusion matrix and classification test report

```
[[108  45]
 [ 37 266]]
              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
```

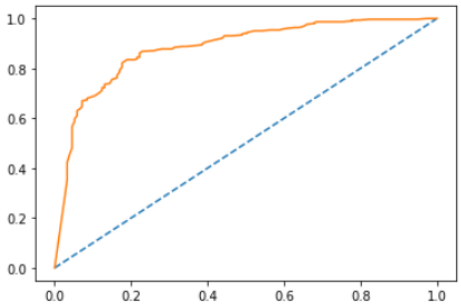
```
mx=sns.heatmap(metrics.confusion_matrix(y_test,y_test_predict),annot=True,fmt='d')
```



AUC-ROC curve for Bagging model-test data

```
AUC: 0.881
```

```
[<matplotlib.lines.Line2D at 0x267485e0d00>]
```



Inference of Bagging Model:

Using the confusion matrix, the True Positive, False Positive, False Negative, and True Negative values can beextracted which will aid in the calculation of the accuracy score, precision score, recall score, and f1score.Listing below model performance metrics before fine tuning the model:

Train Data:

True Positive:307

False Positive:0

False Negative:0

True Negative:754

AUC: 100%

Accuracy: 100%

Precision: 100%

f1-Score: 100%

Recall:100%

Test Data:

True Positive:108

False Positive:37

False Negative:45

True Negative:266

AUC: 88.1%

Accuracy: 82%

Precision: 86%

f1-Score: 87%

Recall:88%

Clearly,our model has better performance on the training set than on the test set, it is likely that model has overfitted.Hence, it might be a big red flag as our model has 100% accuracy on the training set but only82% accuracy on the test set.Generally bagging is used to avoid problems of overfitting but in this modelmay be while sampling with replacements some observations got repeated in each subset.Hence, ourmodel is overfitting.

We know that, FPR tells us what proportion of the negative class got incorrectly classified by theclassifier.Here, we have higher TNR and a lower FPR which is desirable to classify the negative class.

Here, both Type I Error (False Positives) and Type II Error (False Negatives) are low for Test Data indicatinghigh Sensitivity/Recall, Precision,Specificity and F1 Score.

We will now try to build the model using Boosting.While bagging and boosting are both ensemble methods,they approach the problem from opposite directions.Bagging uses complex base models and tries to "smoothout" their predictions, while boosting uses simple base models and tries to "boost" their aggregate complexity.

Boosting:

In Boosting,Base estimators are built sequentially and one tries to reduce the bias of the combined estimator.The idea is to combine several weak models to produce a powerful

ensemble.It makes the boosting algorithmsprone to overfitting. Examples: AdaBoost, Gradient Tree Boosting**.**

Boosting is a sequential process, where each subsequent model attempts to correct the errors of the previousmodel Boosting is focused on reducing the bias. It makes the boosting algorithms prone to overfitting.Tochoose different distribution for each round we use following steps:

Step 1: The base learner takes all the distributions and assign equal weight or attention to each observation.

Step 2: If there is any prediction error caused by first base learning algorithm, then we pay higher attention toobservations having prediction error. Then, we apply the next base learning algorithm.

Step 3: Iterate Step 2 till the limit of base learning algorithm is reached or higher accuracy is achieved.

There are three types of Boosting Algorithms which are as follows:

1. AdaBoost (Adaptive Boosting) algorithm.

2. Gradient Boosting algorithm.

3. XG Boost algorithm.

Here, we will use Adaboost and Gradient Boosting algorithm.

Below are the key parameters for tuning:

n_estimators: It controls the number of weak learners.

learning_rate:Controls the contribution of weak learners in the final combination. There is a trade-offbetween learning_rate and n_estimators.

base_estimators: It helps to specify different ML algorithm.

AdaBoostClassifier

```
ADB_model=AdaBoostClassifier(n_estimators=100,random_state=1)
ADB_model.fit(X_train,y_train)
```

```
C:\Users\User\anaconda3\lib\site-packages\sklearn\utils\validation.py:63: DataConversionWarning: A column-vector y was passed w
hen a 1d array was expected. Please change the shape of y to (n_samples, ), for example using ravel().
  return f(*args, **kwargs)
```

```
AdaBoostClassifier(n_estimators=100, random_state=1)
```

Model score with AdaBoostClassifier for training

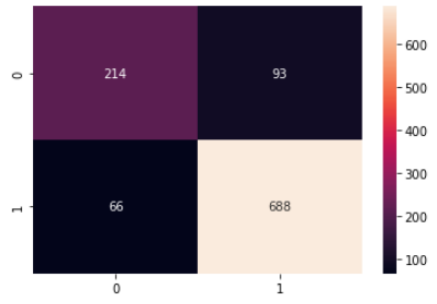
```
y_train_predict=ADB_model.predict(X_train)
ADB_model_score=ADB_model.score(X_train,y_train)
ADB_model_score
```

```
0.8501413760603205
```

Confusion matrix and Classification report

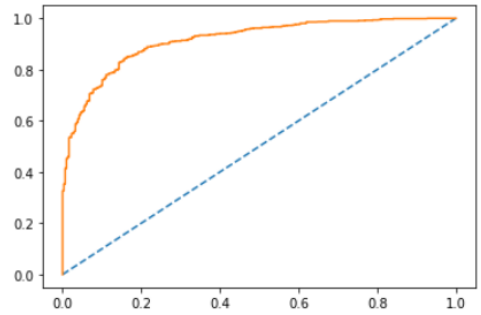
```
[[214  93]
 [ 66 688]]
              precision    recall  f1-score   support

           0       0.76      0.70      0.73       307
           1       0.88      0.91      0.90       754

    accuracy                           0.85      1061
   macro avg       0.82      0.80      0.81      1061
weighted avg       0.85      0.85      0.85      1061
```

```
mx=sns.heatmap(metrics.confusion_matrix(y_train,y_train_predict),annot=True,fmt='d')
```



AUC_ROC Curve of Train Data_ADA Boosting:

```
AUC: 0.915

[<matplotlib.lines.Line2D at 0x267486fe370>]
```



Model score with AdaBoostClassifier for testing

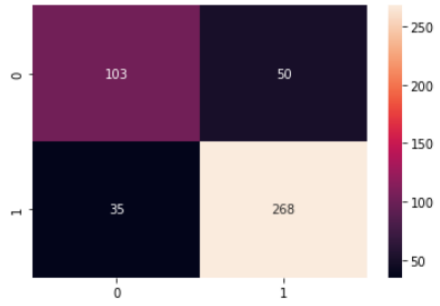
```
y_test_predict=ADB_model.predict(X_test)
ADB_model_score=ADB_model.score(X_test,y_test)
ADB_model_score
```

```
0.8135964912280702
```

Confusion matrix and classification report

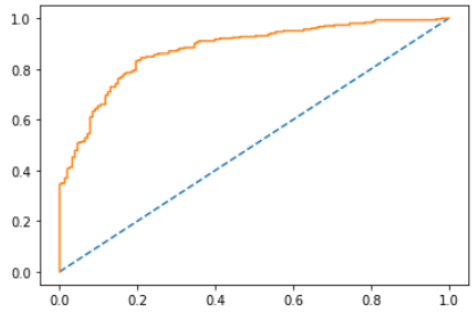
```
[[103  50]
 [ 35 268]]
              precision    recall  f1-score   support

           0       0.75      0.67      0.71       153
           1       0.84      0.88      0.86       303

    accuracy                           0.81       456
   macro avg       0.79      0.78      0.79       456
weighted avg       0.81      0.81      0.81       456
```

```
mx=sns.heatmap(metrics.confusion_matrix(y_test,y_test_predict),annot=True, fmt='d')
```



AUC_ROC Curve of Test Data_ADA Boosting:

```
AUC: 0.877
[<matplotlib.lines.Line2D at 0x267487f5c40>]
```



Inference of ADA Boosting Model:

Using the confusion matrix, the True Positive, False Positive, False Negative, and True Negative values can beextracted which will aid in the calculation of the accuracy score, precision score, recall score, and f1score.Listing below model performance metrics before fine tuning the model:

Train Data:

True Positive:214

False Positive:66

False Negative:93

True Negative:688

AUC: 91.5%

Accuracy: 85%

Precision: 88%

f1-Score: 90%

Recall:91%

Test Data:

True Positive:103

False Positive:35

False Negative:50

True Negative:268

AUC: 87.7%

Accuracy: 81%

Precision: 84%

f1-Score: 86%

Recall:88%

Clearly,our model has better performance on the training set than on the test set.

We know that, FPR tells us what proportion of the negative class got incorrectly classified by theclassifier.Here, we have higher TNR and a lower FPR which is desirable to classify the negative class.

Here, both Type I Error (False Positives) and Type II Error (False Negatives) are low for indicating highSensitivity/Recall, Precision,Specificity and F1 Score.

F1-score, Recall,Precision and AUC are better for train data.

Gradient Boosting:

The principle idea behind this algorithm is to construct new base learners which can be maximally correlated with negative gradient of the loss function, associated with the whole ensemble.

As gradient boosting is same as we done in AdaBoosting

Question 1.8

Based on these predictions, what are the insights?

Answer:

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

2)Labour party is performing better than Conservative from huge margin.

3)Female voters turn out is greater than the male voters.

4)Those who have better national economic conditions are preferring to vote for Labour party.

5)Persons having higher Eurosceptic sentiments conservative party are preferring to vote for Conservative party.

6)Those who have higher political knowledge have voted for Conservative party

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

Problem 2

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

1.President Franklin D. Roosevelt in 1941

2.President John F. Kennedy in 1961

3.President Richard Nixon in 1973

(Hint: use .words(), .raw(), .sent() for extracting counts)

Importing Important Libraries

```
import numpy as np
import pandas as pd
import re
import nltk
import matplotlib.pyplot as plt
%matplotlib inline
import string
import matplotlib
```

Checking the versions of various libraries

```
print('Numpy version:',np.__version__)
print('Pandas version:',pd.__version__)
print('Regular Expression version:',re.__version__)
print('Natural Language Tool Kit version:',nltk.__version__)
print('Matplotlib version:',matplotlib.__version__)
```

```
Numpy version: 1.20.3
Pandas version: 1.3.4
Regular Expression version: 2.2.1
Natural Language Tool Kit version: 3.6.5
Matplotlib version: 3.4.3
```

President Franklin D.Roosevelt in 1941:

For President Franklin D. Roosevelt in 1941

```
inaugural.raw('1941-Roosevelt.txt')
```

```
'On each national day of inauguration since 1789, the people have renewed their sense of dedication to the United States.\n\nIn
Washington\'s day the task of the people was to create and weld together a nation.\n\nIn Lincoln\'s day the task of the people
was to preserve that Nation from disruption from within.\n\nIn this day the task of the people is to save that Nation and its i
nstitutions from disruption from without.\n\nTo us there has come a time, in the midst of swift happenings, to pause for a mome
nt and take stock -- to recall what our place in history has been, and to rediscover what we are and what we may be. If we do n
ot, we risk the real peril of inaction.\n\nLives of nations are determined not by the count of years, but by the lifetime of th
e human spirit. The life of a man is three-score years and ten: a little more, a little less. The life of a nation is the fulln
ess of the measure of its will to live.\n\nThere are men who doubt this. There are men who believe that democracy, as a form of
Government and a frame of life, is limited or measured by a kind of mystical and artificial fate that, for some unexplained rea
son, tyranny and slavery have become the surging wave of the future -- and that freedom is an ebbing tide.\n\nBut we Americans
know that this is not true.\n\nEight years ago, when the life of this Republic seemed frozen by a fatalistic terror, we proved
that this is not true. We were in the midst of shock -- but we acted. We acted quickly, boldly, decisively.\n\nThese later year
s have been living years -- fruitful years for the people of this democracy. For they have brought to us greater security and,
I hope, a better understanding that life\'s ideals are to be measured in other than material things.\n\nMost vital to our prese
nt and our future is this experience of a democracy which successfully survived crisis at home; put away many evil things; buil
t new structures on enduring lines; and, through it all, maintained the fact of its democracy.\n\nFor action has been taken wit
hin the three-way framework of the Constitution of the United States. The coordinate branches of the Government continue freely
to function. The Bill of Rights remains inviolate. The freedom of elections is wholly maintained. Prophets of the downfall of A
merican democracy have seen their dire predictions come to naught.\n\nDemocracy is not dying.\n\nWe know it because we have see
n it revive--and grow.\n\nWe know it cannot die -- because it is built on the unhampered initiative of individual men and women
joined together in a common enterprise -- an enterprise undertaken and carried through by the free expression of a free majorit
y.\n\nWe know it because democracy alone, of all forms of government, enlists the full force of men\'s enlightened will.\n\nWe
know it because democracy alone has constructed an unlimited civilization capable of infinite progress in the improvement of hu
man life.\n\nWe know it because, if we look below the surface, we sense it still spreading on every continent -- for it is the
most humane, the most advanced, and in the end the most unconquerable of all forms of human society.\n\nA nation, like a perso
n, has a body--a body that must be fed and clothed and housed, invigorated and rested, in a manner that measures up to the obje
ctives of our time.\n\nA nation, like a person, has a mind -- a mind that must be kept informed and alert, that must know itsel
f, that understands the hopes and the needs of its neighbors -- all the other nations that live within the narrowing circle of
the world.\n\nAnd a nation, like a person, has something deeper, something more permanent, something larger than the sum of all
its parts. It is that something which matters most to its future -- which calls forth the most sacred guarding of its presen
t.\n\nIt is a thing for which we find it difficult -- even impossible -- to hit upon a single, simple word.\n\nAnd yet we all u
nderstand what it is -- the spirit -- the faith of America. It is the product of centuries. It was born in the multitudes of th
ose who came from many lands -- some of high degree, but mostly plain people, who sought here, early and late, to find freedom
more freely.\n\nThe democratic aspiration is no mere recent phase in human history. It is human history. It permeated the ancie
nt life of early peoples. It blazed anew in the middle ages. It was written in Magna Charta.\n\nIn the Americas its impact has
been irresistible. America has been the New World in all tongues, to all peoples, not because this continent was a new-found la
nd, but because all those who came here believed they could create upon this continent a new life -- a life that should be new
in freedom.\n\nIts vitality was written into our own Mayflower Compact, into the Declaration of Independence, into the Constitu
in freedom.\n\nIts vitality was written into our own Mayflower Compact, into the Declaration of Independence, into the Constitu
tion of the United States, into the Gettysburg Address.\n\nThose who first came here to carry out the longings of their spirit,
and the millions who followed, and the stock that sprang from them -- all have moved forward constantly and consistently toward
an ideal which in itself has gained stature and clarity with each generation.\n\nThe hopes of the Republic cannot forever toler
ate either undeserved poverty or self-serving wealth.\n\nWe know that we still have far to go; that we must more greatly build
the security and the opportunity and the knowledge of every citizen, in the measure justified by the resources and the capacity
of the land.\n\nBut it is not enough to achieve these purposes alone. It is not enough to clothe and feed the body of this Nati
on, and instruct and inform its mind. For there is also the spirit. And of the three, the greatest is the spirit.\n\nWithout th
e body and the mind, as all men know, the Nation could not live.\n\nBut if the spirit of America were killed, even though the N
ation\'s body and mind, constricted in an alien world, lived on, the America we know would have perished.\n\nThat spirit -- tha
t faith -- speaks to us in our daily lives in ways often unnoticed, because they seem so obvious. It speaks to us here in the C
apital of the Nation. It speaks to us through the processes of governing in the sovereignties of 48 States. It speaks to us in
our counties, in our cities, in our towns, and in our villages. It speaks to us from the other nations of the hemisphere, and f
rom those across the seas -- the enslaved, as well as the free. Sometimes we fail to hear or heed these voices of freedom becau
se to us the privilege of our freedom is such an old, old story.\n\nThe destiny of America was proclaimed in words of prophecy
spoken by our first President in his first inaugural in 1789 -- words almost directed, it would seem, to this year of 1941: "Th
e preservation of the sacred fire of liberty and the destiny of the republican model of government are justly considered deepl
y, finally, staked on the experiment intrusted to the hands of the American people."\n\nIf we lose that sacred fire--if we let
it be smothered with doubt and fear -- then we shall reject the destiny which Washington strove so valiantly and so triumphantl
y to establish. The preservation of the spirit and faith of the Nation does, and will, furnish the highest justification for ev
ery sacrifice that we may make in the cause of national defense.\n\nIn the face of great perils never before encountered, our s
trong purpose is to protect and to perpetuate the integrity of democracy.\n\nFor this we muster the spirit of America, and the
faith of America.\n\nWe do not retreat. We are not content to stand still. As Americans, we go forward, in the service of our c
ountry, by the will of God.\n'
```

Checking the Type()

```
type(df_Roosevelt)
```

```
str
```

Splitting the text data at '\n' and creating a corpus of documents and again checking the type ()

```
split_df_Roosevelt= df_Roosevelt.split('\n\n')
```

```
type(split_df_Roosevelt)
```

```
list
```

Creating the Dataframe and Renaming the feature column

```
df_text_Roosevelt= pd.DataFrame(split_df_Roosevelt)
```

```
df_text_Roosevelt.columns=['speech']
```

Head and Tail of Data

```
df_text_Roosevelt.head()
```

	speech
0	On each national day of inauguration since 178...
1	In Washington's day the task of the people was...
2	In Lincoln's day the task of the people was to...
3	In this day the task of the people is to save ...
4	To us there has come a time, in the midst of s...

```
df_text_Roosevelt.tail()
```

	speech
33	The destiny of America was proclaimed in words...
34	If we lose that sacred fire--if we let it be s...
35	In the face of great perils never before encou...
36	For this we muster the spirit of America, and ...
37	We do not retreat. We are not content to stand...

Shape of Data

```
df_text_Roosevelt.shape
```

```
(38, 1)
```

The numbers of rows of Dataframe in 1941- Roosevelt is 38

The numbers of columns of Dtaframe in 1941- Roosevelt is 1

Numbers of Words

	speech	word_count
0	On each national day of inauguration since 178...	20
1	In Washington's day the task of the people was...	16
2	In Lincoln's day the task of the people was to...	17
3	In this day the task of the people is to save ...	20
4	To us there has come a time, in the midst of s...	53

Length of all words in text of 1941-Roosevelt is 1536.

Number of Words

	speech	word_count
0	On each national day of inauguration since 178...	20
1	In Washington's day the task of the people was...	16
2	In Lincoln's day the task of the people was to...	17
3	In this day the task of the people is to save ...	20
4	To us there has come a time, in the midst of s...	53

Number of Characters

	speech	char_count
0	On each national day of inauguration since 178...	120
1	In Washington's day the task of the people was...	84
2	In Lincoln's day the task of the people was to...	96
3	In this day the task of the people is to save ...	108
4	To us there has come a time, in the midst of s...	248

Number of Sentence

```
inaugural.sents('1941-Roosevelt.txt')
```

```
[['On', 'each', 'national', 'day', 'of', 'inauguration', 'since', '1789', ',', 'the', 'people', 'have', 'renewed', 'their', 'se
nse', 'of', 'dedication', 'to', 'the', 'United', 'States', '.'], ['In', 'Washington', "'", 's', 'day', 'the', 'task', 'of', 'th
e', 'people', 'was', 'to', 'create', 'and', 'weld', 'together', 'a', 'nation', '.'], ...]
```

Length of all sentence in text '1941-Roosevelt' is 68

Average Word

	speech	avg_word
0	On each national day of inauguration since 178...	5.050000
1	In Washington's day the task of the people was...	4.312500
2	In Lincoln's day the task of the people was to...	4.705882
3	In this day the task of the people is to save ...	4.450000
4	To us there has come a time, in the midst of s...	3.698113

President John F. Kennedy in 1961

Numbers of character ,words and Sentences for document John F. Kennedy

```
inaugural.raw('1961-Kennedy.txt')
```

```
'Vice President Johnson, Mr. Speaker, Mr. Chief Justice, President Eisenhower, Vice President Nixon, President Truman, reverend
clergy, fellow citizens, we observe today not a victory of party, but a celebration of freedom -- symbolizing an end, as well a
s a beginning -- signifying renewal, as well as change. For I have sworn I before you and Almighty God the same solemn oath our
forebears l prescribed nearly a century and three quarters ago.\n\nThe world is very different now. For man holds in his mortal
hands the power to abolish all forms of human poverty and all forms of human life. And yet the same revolutionary beliefs for w
hich our forebears fought are still at issue around the globe -- the belief that the rights of man come not from the generosity
of the state, but from the hand of God.\n\nWe dare not forget today that we are the heirs of that first revolution. Let the wor
d go forth from this time and place, to friend and foe alike, that the torch has been passed to a new generation of Americans -
- born in this century, tempered by war, disciplined by a hard and bitter peace, proud of our ancient heritage -- and unwilling
to witness or permit the slow undoing of those human rights to which this Nation has always been committed, and to which we are
committed today at home and around the world.\n\nLet every nation know, whether it wishes us well or ill, that we shall pay any
price, bear any burden, meet any hardship, support any friend, oppose any foe, in order to assure the survival and the success
of liberty.\n\nThis much we pledge -- and more.\n\nTo those old allies whose cultural and spiritual origins we share, we pledge
the loyalty of faithful friends. United, there is little we cannot do in a host of cooperative ventures. Divided, there is litt
le we can do -- for we dare not meet a powerful challenge at odds and split asunder.\n\nTo those new States whom we welcome to
the ranks of the free, we pledge our word that one form of colonial control shall not have passed away merely to be replaced by
a far more iron tyranny. We shall not always expect to find them supporting our view. But we shall always hope to find them str
ongly supporting their own freedom -- and to remember that, in the past, those who foolishly sought power by riding the back of
the tiger ended up inside.\n\nTo those peoples in the huts and villages across the globe struggling to break the bonds of mass
misery, we pledge our best efforts to help them help themselves, for whatever period is required -- not because the Communists
may be doing it, not because we seek their votes, but because it is right. If a free society cannot help the many who are poor,
it cannot save the few who are rich.\n\nTo our sister republics south of our border, we offer a special pledge -- to convert ou
r good words into good deeds -- in a new alliance for progress -- to assist free men and free governments in casting off the ch
ains of poverty. But this peaceful revolution of hope cannot become the prey of hostile powers. Let all our neighbors know that
we shall join with them to oppose aggression or subversion anywhere in the Americas. And let every other power know that this H
emisphere intends to remain the master of its own house.\n\nTo that world assembly of sovereign states, the United Nations, our
last best hope in an age where the instruments of war have far outpaced the instruments of peace, we renew our pledge of suppor
t--to prevent it from becoming merely a forum for invective -- to strengthen its shield of the new and the weak -- and to enlar
ge the area in which its writ may run.\n\nFinally, to those nations who would make themselves our adversary, we offer not a ple
dge but a request: that both sides begin anew the quest for peace, before the dark powers of destruction unleashed by science e
ngulf all humanity in planned or accidental self-destruction.\n\nWe dare not tempt them with weakness. For only when our arms a
re sufficient beyond doubt can we be certain beyond doubt that they will never be employed.\n\nBut neither can two great and po
werful groups of nations take comfort from our present course -- both sides overburdened by the cost of modern weapons, both ri
ghtly alarmed by the steady spread of the deadly atom, yet both racing to alter that uncertain balance of terror that stays the
hand of mankind\'s final war.\n\nSo let us begin anew -- remembering on both sides that civility is not a sign of weakness, and
sincerity is always subject to proof. Let us never negotiate out of fear. But let us never fear to negotiate.\n\nLet both sides
explore what problems unite us instead of belaboring those problems which divide us.\n\nLet both sides, for the first time, for
mulate serious and precise proposals for the inspection and control of arms -- and bring the absolute power to destroy other na
tions under the absolute control of all nations.\n\nLet both sides seek to invoke the wonders of science instead of its terror
s. Together let us explore the stars, conquer the deserts, eradicate disease, tap the ocean depths, and encourage the arts and
commerce.\n\nLet both sides unite to heed in all corners of the earth the command of Isaiah -- to "undo the heavy burdens ... a
nd to let the oppressed go free."\n\nAnd if a beachhead of cooperation may push back the jungle of suspicion, let both sides jo
in in creating a new endeavor, not a new balance of power, but a new world of law, where the strong are just and the weak secur
e and the peace preserved.\n\nAll this will not be finished in the first 100 days. Nor will it be finished in the first 1,000 d
ays, nor in the life of this Administration, nor even perhaps in our lifetime on this planet. But let us begin.\n\nIn your hand
s, my fellow citizens, more than in mine, will rest the final success or failure of our course. Since this country was founded,
each generation of Americans has been summoned to give testimony to its national loyalty. The graves of young Americans who ans
wered the call to service surround the globe.\n\nNow the trumpet summons us again -- not as a call to bear arms, though arms we
need; not as a call to battle, though embattled we are -- but a call to bear the burden of a long twilight struggle, year in an
d year out, "rejoicing in hope, patient in tribulation" -- a struggle against the common enemies of man: tyranny, poverty, dise
ase, and war itself.\n\nCan we forge against these enemies a grand and global alliance, North and South, East and West, that ca
n assure a more fruitful life for all mankind? Will you join in that historic effort?\n\nIn the long history of the world, only
a few generations have been granted the role of defending freedom in its hour of maximum danger. I do not shrink from this resp
onsibility -- I welcome it. I do not believe that any of us would exchange places with any other people or any other generatio
n. The energy, the faith, the devotion which we bring to this endeavor will light our country and all who serve it -- and the g
low from that fire can truly light the world.\n\nAnd so, my fellow Americans: ask not what your country can do for you -- ask w
hat you can do for your country.\n\nMy fellow citizens of the world: ask not what America will do for you, but what together we
can do for the freedom of man.\n\nFinally, whether you are citizens of America or citizens of the world, ask of us the same hig
h standards of strength and sacrifice which we ask of you. With a good conscience our only sure reward, with history the final
judge of our deeds, let us go forth to lead the land we love, asking His blessing and His help, but knowing that here on earth
God\'s work must truly be our own.\n'
```

Checking the Type()

```
type(df_Kennedy)
```

```
str
```

Splitting the text data at '\n' and creating a corpus of documents and again checking the type ()

```
split_df_Kennedy= df_Kennedy.split('\n\n')
```

```
type(split_df_Kennedy)
```

```
list
```

Creating the Dataframe and Renaming the feature column

```
df_text_Kennedy= pd.DataFrame(split_df_Kennedy)
```

```
df_text_Kennedy.columns=['speech']
```

Head and Tail of Data

```
df_text_Kennedy.head(
)
```

	speech
0	Vice President Johnson, Mr. Speaker, Mr. Chief...
1	The world is very different now. For man holds...
2	We dare not forget today that we are the heirs...
3	Let every nation know, whether it wishes us we...
4	This much we pledge -- and more.

```
df_text_Kennedy.tail()
```

	speech
22	Can we forge against these enemies a grand and...
23	In the long history of the world, only a few g...
24	And so, my fellow Americans: ask not what your...
25	My fellow citizens of the world: ask not what ...
26	Finally, whether you are citizens of America o...

Shape of Data

```
df_text_Kennedy.shape
```

```
(27, 1)
```

The numbers of rows of Dataframe in 1961- Kennedy is 27

The numbers of columns of Dtaframe in 1961- Kennedy is 1

Numbers of Words

	speech	word_count
0	Vice President Johnson, Mr. Speaker, Mr. Chief...	73
1	The world is very different now. For man holds...	68
2	We dare not forget today that we are the heirs...	96
3	Let every nation know, whether it wishes us we...	40
4	This much we pledge -- and more.	7

Length of all words in text of 1961-Kennedy is 1546.

Number of Words

	speech	word_count
0	Vice President Johnson, Mr. Speaker, Mr. Chief...	73
1	The world is very different now. For man holds...	68
2	We dare not forget today that we are the heirs...	96
3	Let every nation know, whether it wishes us we...	40
4	This much we pledge -- and more.	7

Number of Characters

	speech	char_count
0	Vice President Johnson, Mr. Speaker, Mr. Chief...	445
1	The world is very different now. For man holds...	355
2	We dare not forget today that we are the heirs...	512
3	Let every nation know, whether it wishes us we...	217
4	This much we pledge -- and more.	32

Number of Sentence

```
inaugural.sents('1961-Kennedy.txt')
```

```
[['Vice', 'President', 'Johnson', ',', 'Mr', '.', 'Speaker', ',', 'Mr', '.', 'Chief', 'Justice', ',', 'President', 'Eisenhowe
r', ',', 'Vice', 'President', 'Nixon', ',', 'President', 'Truman', ',', 'reverend', 'clergy', ',', 'fellow', 'citizens', ',',
'we', 'observe', 'today', 'not', 'a', 'victory', 'of', 'party', ',', 'but', 'a', 'celebration', 'of', 'freedom', '--', 'symboli
zing', 'an', 'end', ',', 'as', 'well', 'as', 'a', 'beginning', '--', 'signifying', 'renewal', ',', 'as', 'well', 'as', 'chang
e', '.'], ['For', 'I', 'have', 'sworn', 'I', 'before', 'you', 'and', 'Almighty', 'God', 'the', 'same', 'solemn', 'oath', 'our',
'forebears', 'l', 'prescribed', 'nearly', 'a', 'century', 'and', 'three', 'quarters', 'ago', '.'], ...]
```

Length of all sentence in text ‘1961-Kennedy’ is 52

Average Word

	speech	avg_word
0	Vice President Johnson, Mr. Speaker, Mr. Chief...	5.109589
1	The world is very different now. For man holds...	4.235294
2	We dare not forget today that we are the heirs...	4.343750
3	Let every nation know, whether it wishes us we...	4.450000
4	This much we pledge -- and more.	3.714286

President Richard Nixon in 1973

Numbers of character ,words and Sentences for document Richard Nixon

```
inaugural.raw('1973-Nixon.txt')
```

```
'Mr. Vice President, Mr. Speaker, Mr. Chief Justice, Senator Cook, Mrs. Eisenhower, and my fellow citizens of this great and go
od country we share together:\n\nWhen we met here four years ago, America was bleak in spirit, depressed by the prospect of see
mingly endless war abroad and of destructive conflict at home.\n\nAs we meet here today, we stand on the threshold of a new era
of peace in the world.\n\nThe central question before us is: How shall we use that peace? Let us resolve that this era we are a
bout to enter will not be what other postwar periods have so often been: a time of retreat and isolation that leads to stagnati
on at home and invites new danger abroad.\n\nLet us resolve that this will be what it can become: a time of great responsibilit
ies greatly borne, in which we renew the spirit and the promise of America as we enter our third century as a nation.\n\nThis p
ast year saw far-reaching results from our new policies for peace. By continuing to revitalize our traditional friendships, and
by our missions to Peking and to Moscow, we were able to establish the base for a new and more durable pattern of relationships
among the nations of the world. Because of America\'s bold initiatives, 1972 will be long remembered as the year of the greates
t progress since the end of World War II toward a lasting peace in the world.\n\nThe peace we seek in the world is not the flim
sy peace which is merely an interlude between wars, but a peace which can endure for generations to come.\n\nIt is important th
at we understand both the necessity and the limitations of America\'s role in maintaining that peace.\n\nUnless we in America w
ork to preserve the peace, there will be no peace.\n\nUnless we in America work to preserve freedom, there will be no freedo
m.\n\nBut let us clearly understand the new nature of America\'s role, as a result of the new policies we have adopted over the
se past four years.\n\nWe shall respect our treaty commitments.\n\nWe shall support vigorously the principle that no country ha
s the right to impose its will or rule on another by force.\n\nWe shall continue, in this era of negotiation, to work for the l
imitation of nuclear arms, and to reduce the danger of confrontation between the great powers.\n\nWe shall do our share in defe
nding peace and freedom in the world. But we shall expect others to do their share.\n\nThe time has passed when America will ma
ke every other nation\'s conflict our own, or make every other nation\'s future our responsibility, or presume to tell the peop
le of other nations how to manage their own affairs.\n\nJust as we respect the right of each nation to determine its own futur
e, we also recognize the responsibility of each nation to secure its own future.\n\nJust as America\'s role is indispensable in
preserving the world\'s peace, so is each nation\'s role indispensable in preserving its own peace.\n\nTogether with the rest o
f the world, let us resolve to move forward from the beginnings we have made. Let us continue to bring down the walls of hostil
ity which have divided the world for too long, and to build in their place bridges of understanding -- so that despite profound
differences between systems of government, the people of the world can be friends.\n\nLet us build a structure of peace in the
world in which the weak are as safe as the strong -- in which each respects the right of the other to live by a different syste
m -- in which those who would influence others will do so by the strength of their ideas, and not by the force of their arms.\n
\nLet us accept that high responsibility not as a burden, but gladly -- gladly because the chance to build such a peace is the
noblest endeavor in which a nation can engage; gladly, also, because only if we act greatly in meeting our responsibilities abr
oad will we remain a great Nation, and only if we remain a great Nation will we act greatly in meeting our challenges at hom
e.\n\nWe have the chance today to do more than ever before in our history to make life better in America -- to ensure better ed

ucation, better health, better housing, better transportation, a cleaner environment -- to restore respect for law, to make our
communities more livable -- and to insure the God-given right of every American to full and equal opportunity.\n\nBecause the r
ange of our needs is so great -- because the reach of our opportunities is so great -- let us be bold in our determination to m
eet those needs in new ways.\n\nJust as building a structure of peace abroad has required turning away from old policies that f
ailed, so building a new era of progress at home requires turning away from old policies that have failed.\n\nAbroad, the shift
from old policies to new has not been a retreat from our responsibilities, but a better way to peace.\n\nAnd at home, the shift
from old policies to new will not be a retreat from our responsibilities, but a better way to progress.\n\nAbroad and at home,
the key to those new responsibilities lies in the placing and the division of responsibility. We have lived too long with the c
onsequences of attempting to gather all power and responsibility in Washington.\n\nAbroad and at home, the time has come to tur
n away from the condescending policies of paternalism -- of "Washington knows best."\n\nA person can be expected to act respons
ibly only if he has responsibility. This is human nature. So let us encourage individuals at home and nations abroad to do more
for themselves, to decide more for themselves. Let us locate responsibility in more places. Let us measure what we will do for
others by what they will do for themselves.\n\nThat is why today I offer no promise of a purely governmental solution for every
problem. We have lived too long with that false promise. In trusting too much in government, we have asked of it more than it c
an deliver. This leads only to inflated expectations, to reduced individual effort, and to a disappointment and frustration tha
t erode confidence both in what government can do and in what people can do.\n\nGovernment must learn to take less from people
so that people an do more for themselves.\n\nLet us remember that America was built not by government, but by people -- not by
welfare, but by work -- not by shirking responsibility, but by seeking responsibility.\n\nIn our own lives, let each of us ask
-- not just what will government do for me, but what can I do for myself?\n\nIn the challenges we face together, let each of us
ask -- not just how can government help, but how can I help?\n\nYour National Government has a great and vital role to play. An
d I pledge to you that where this Government should act, we will act boldly and we will lead boldly. But just as important is t
he role that each and every one of us must play, as an individual and as a member of his own community.\n\nFrom this day forwar
d, let each of us make a solemn commitment in his own heart: to bear his responsibility, to do his part, to live his ideals --
so that together, we can see the dawn of a new age of progress for America, and together, as we celebrate our 200th anniversary
as a nation, we can do so proud in the fulfillment of our promise to ourselves and to the world.\n\nAs America\'s longest and m
ost difficult war comes to an end, let us again learn to debate our differences with civility and decency. And let each of us r
each out for that one precious quality government cannot provide -- a new level of respect for the rights and feelings of one a
nother, a new level of respect for the individual human dignity which is the cherished birthright of every American.\n\nAbove a
ll else, the time has come for us to renew our faith in ourselves and in America.\n\nIn recent years, that faith has been chall
enged.\n\nOur children have been taught to be ashamed of their country, ashamed of their parents, ashamed of America\'s record
at home and of its role in the world.\n\nAt every turn, we have been beset by those who find everything wrong with America and
little that is right. But I am confident that this will not be the judgment of history on these remarkable times in which we ar
e privileged to live.\n\nAmerica\'s record in this century has been unparalleled in the world\'s history for its responsibilit
y, for its generosity, for its creativity and for its progress.\n\nLet us be proud that our system has produced and provided mo

y, for its generosity, for its creativity and for its progress.\n\nLet us be proud that our system has produced and provided mo
re freedom and more abundance, more widely shared, than any other system in the history of the world.\n\nLet us be proud that i
n each of the four wars in which we have been engaged in this century, including the one we are now bringing to an end, we have
fought not for our selfish advantage, but to help others resist aggression.\n\nLet us be proud that by our bold, new initiative
s, and by our steadfastness for peace with honor, we have made a break-through toward creating in the world what the world has
not known before -- a structure of peace that can last, not merely for our time, but for generations to come.\n\nWe are embarki
ng here today on an era that presents challenges great as those any nation, or any generation, has ever faced.\n\nWe shall answ
er to God, to history, and to our conscience for the way in which we use these years.\n\nAs I stand in this place, so hallowed
by history, I think of others who have stood here before me. I think of the dreams they had for America, and I think of how eac
h recognized that he needed help far beyond himself in order to make those dreams come true.\n\nToday, I ask your prayers that
in the years ahead I may have God\'s help in making decisions that are right for America, and I pray for your help so that toge
ther we may be worthy of our challenge.\n\nLet us pledge together to make these next four years the best four years in America
\'s history, so that on its 200th birthday America will be as young and as vital as when it began, and as bright a beacon of ho
pe for all the world.\n\nLet us go forward from here confident in hope, strong in our faith in one another, sustained by our fa
ith in God who created us, and striving always to serve His purpose.\n'
```

Checking the Type()

```
type(df_Nixon)
```

```
str
```

Splitting the text data at '\n' and creating a corpus of documents and again checking the type ()

```
split_df_Nixon= df_Nixon.split('\n\n')
```

```
type(split_df_Nixon)
```

```
list
```

Creating the Dataframe and Renaming the feature column

```
df_text_Nixon= pd.DataFrame(split_df_Nixon)

df_text_Nixon.columns=['speech']
```

Head and Tail of Data

```
df_text_Nixon.head()
```

	speech
0	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...
1	When we met here four years ago, America was b...
2	As we meet here today, we stand on the thresho...
3	The central question before us is: How shall w...
4	Let us resolve that this will be what it can b...

```
df_text_Nixon.tail()
```

	speech
46	We shall answer to God, to history, and to our...
47	As I stand in this place, so hallowed by histo...
48	Today, I ask your prayers that in the years ah...
49	Let us pledge together to make these next four...
50	Let us go forward from here confident in hope,...

Shape of Data

```
df_text_Nixon.shape
```

```
(51, 1)
```

The numbers of rows of Dataframe in 1973- Nixon is 51

The numbers of columns of Dtaframe in 1973- Nixon is 1

Numbers of Words

	speech	word_count
0	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	25
1	When we met here four years ago, America was b...	27
2	As we meet here today, we stand on the thresho...	19
3	The central question before us is: How shall w...	51
4	Let us resolve that this will be what it can b...	38

Length of all words in text of 1973-Nixon is 2028.

Number of Words

	speech	word_count
0	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	25
1	When we met here four years ago, America was b...	27
2	As we meet here today, we stand on the thresho...	19
3	The central question before us is: How shall w...	51
4	Let us resolve that this will be what it can b...	38

Number of Characters

	speech	char_count
0	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	155
1	When we met here four years ago, America was b...	156
2	As we meet here today, we stand on the thresho...	84
3	The central question before us is: How shall w...	269
4	Let us resolve that this will be what it can b...	199

Number of Sentence

```
inaugural.sents('1973-Nixon.txt')
```

```
[['Mr', '.', 'Vice', 'President', ',', 'Mr', '.', 'Speaker', ',', 'Mr', '.', 'Chief', 'Justice', ',', 'Senator', 'Cook', ',',
'Mrs', '.', 'Eisenhower', ',', 'and', 'my', 'fellow', 'citizens', 'of', 'this', 'great', 'and', 'good', 'country', 'we', 'shar
e', 'together', ':'], ['When', 'we', 'met', 'here', 'four', 'years', 'ago', ',', 'America', 'was', 'bleak', 'in', 'spirit',
',', 'depressed', 'by', 'the', 'prospect', 'of', 'seemingly', 'endless', 'war', 'abroad', 'and', 'of', 'destructive', 'conflic
t', 'at', 'home', '.'], ...]
```

Length of all sentence in text '1973-Nixon' is 69

Average Word

	speech	avg_word
0	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	5.240000
1	When we met here four years ago, America was b...	4.814815
2	As we meet here today, we stand on the thresho...	3.473684
3	The central question before us is: How shall w...	4.294118
4	Let us resolve that this will be what it can b...	4.263158

Question 2.2

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

Answer:

```
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\User\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

```
True
```

For President Franklin D. Roosevelt in 1941:

- 1st we are spliting each row of the dataframe into words.
- 2nd we are joining all the above words with a space between them.
- 3rd we are making a Series out of it.
- 4th we we are extracting each word one by one and storing it in the variable all_words.

Words Frequency

```
all_words_freq_Roosevelt
```

```
FreqDist({'--': 22, 'know': 9, 'us': 8, 'life': 6, 'people': 5, 'nation': 5, 'human': 5, 'years': 5, 'freedom': 5, 'democracy':
5, ...})
```

Stop words

```
print(stopwords_Roosevelt)
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'y
ourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself',
'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those',
'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'a
n', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'b
etween', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'of
f', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both',
'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very',
's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'ar
en', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "have
n't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "should
n't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't", '!', '"', '#', '$', '%', '&', "'", '(', ')',
'*', '+', ',', '-', '.', '/', ':', ';', '<', '=', '>', '?', '@', '[', '\\', ']', '^', '_', '`', '{', '|', '}', '~']
```

These are the stop words which we have to remove.

Word Features

```
print(word_features_Roosevelt)
```

```
['--', 'know', 'us', 'life', 'people', 'nation', 'human', 'years', 'freedom', 'democracy', 'spirit', 'speaks', 'day', 'men', 'n
ew', 'nation,', 'body', 'must', 'something', 'faith', 'america', 'united', 'states.', 'task', 'nations', 'spirit.', 'governmen
t', 'future', 'democracy.', 'every', 'continent', 'like', 'person,', 'sacred', 'came', 'first', 'destiny', 'national', 'sense',
'create', 'together', 'nation.', 'disruption', 'come', 'midst', 'stock', 'may', 'lives', 'little', 'measure', 'live.', 'doubt',
'measured', 'true.', 'republic', 'security', 'and,', 'many', 'built', 'within', 'constitution', 'american', 'seen', 'cannot',
'enterprise', 'free', 'forms', 'still', 'mind', 'hopes', 'find', 'even', 'upon', 'america.', 'early', 'history.', 'written', 'c
ould', 'enough', 'mind,', 'would', 'words', 'preservation', 'inauguration', 'since', '1789,', 'renewed', 'dedication', "washing
ton's", 'weld', "lincoln's", 'preserve', 'within.', 'save', 'institutions', 'without.', 'time,', 'swift', 'happenings,', 'paus
e', 'moment', 'take', 'recall', 'place', 'history', 'been,', 'rediscover', 'be.', 'not,', 'risk', 'real', 'peril', 'inaction.',
'determined', 'count', 'years,', 'lifetime', 'man', 'three-score', 'ten:', 'more,', 'less.', 'fullness', 'this.', 'believe', 'd
emocracy,', 'form', 'frame', 'life,', 'limited', 'kind', 'mystical', 'artificial', 'fate', 'that,', 'unexplained', 'reason,',
'tyranny', 'slavery', 'become', 'surging', 'wave', 'ebbing', 'tide.', 'americans', 'eight', 'ago,', 'seemed', 'frozen', 'fatali
stic', 'terror,', 'proved', 'shock', 'acted.', 'acted', 'quickly,', 'boldly,', 'decisively.', 'later', 'living', 'fruitful', 'b
rought', 'greater', 'hope,', 'better', 'understanding', "life's", 'ideals', 'material', 'things.', 'vital', 'present', 'experie
nce', 'successfully', 'survived', 'crisis', 'home;', 'put', 'away', 'evil', 'things;', 'structures', 'enduring', 'lines;', 'al
l,', 'maintained', 'fact', 'action', 'taken', 'three-way', 'framework', 'coordinate', 'branches', 'continue', 'freely', 'functi
on.', 'bill', 'rights', 'remains', 'inviolate.', 'elections', 'wholly', 'maintained.', 'prophets', 'downfall', 'dire', 'predict
ions', 'naught.', 'dying.', 'revive--and', 'grow.', 'die', 'unhampered', 'initiative', 'individual', 'women', 'joined', 'commo
n', 'undertaken', 'carried', 'expression', 'majority.', 'alone,', 'government,', 'enlists', 'full', 'force', "men's", 'enlighte
ned', 'will.', 'alone', 'constructed', 'unlimited', 'civilization', 'capable', 'infinite', 'progress', 'improvement', 'life.',
'because,', 'look', 'surface,', 'spreading', 'humane,', 'advanced,', 'end', 'unconquerable', 'society.', 'body--a', 'fed', 'clo
thed', 'housed,', 'invigorated', 'rested,', 'manner', 'measures', 'objectives', 'time.', 'kept', 'informed', 'alert,', 'itsel
f,', 'understands', 'needs', 'neighbors', 'live', 'narrowing', 'circle', 'world.', 'deeper,', 'permanent,', 'larger', 'sum', 'p
arts.', 'matters', 'calls', 'forth', 'guarding', 'present.', 'thing', 'difficult', 'impossible', 'hit', 'single,', 'simple', 'w
ord.', 'yet', 'understand', 'product', 'centuries.', 'born', 'multitudes', 'lands', 'high', 'degree,', 'mostly', 'plain', 'peop
le,', 'sought', 'here,', 'late,', 'freely.', 'democratic', 'aspiration', 'mere', 'recent', 'phase', 'permeated', 'ancient', 'pe
oples.', 'blazed', 'anew', 'middle', 'ages.', 'magna', 'charta.', 'americas', 'impact', 'irresistible.', 'world', 'tongues,',
'peoples,', 'new-found', 'land,', 'believed', 'freedom.', 'vitality', 'mayflower', 'compact,', 'declaration', 'independence,',
'states,', 'gettysburg', 'address.', 'carry', 'longings', 'spirit,', 'millions', 'followed,', 'sprang', 'moved', 'forward', 'co
nstantly', 'consistently', 'toward', 'ideal', 'gained', 'stature', 'clarity', 'generation.', 'forever', 'tolerate', 'either',
'undeserved', 'poverty', 'self-serving', 'wealth.', 'far', 'go;', 'greatly', 'build', 'opportunity', 'knowledge', 'citizen,',
'justified', 'resources', 'capacity', 'land.', 'achieve', 'purposes', 'alone.', 'clothe', 'feed', 'instruct', 'inform', 'min
d.', 'also', 'three,', 'greatest', 'without', 'know,', 'killed,', 'though', "nation's", 'constricted', 'alien', 'world,', 'live
d', 'on,', 'perished.', 'daily', 'ways', 'often', 'unnoticed,', 'seem', 'obvious.', 'capital', 'processes', 'governing', 'sover
eignties', '48', 'counties,', 'cities,', 'towns,', 'villages.', 'hemisphere,', 'across', 'seas', 'enslaved,', 'well', 'free.',
'sometimes', 'fail', 'hear', 'heed', 'voices', 'privilege', 'old,', 'old', 'story.', 'proclaimed', 'prophecy', 'spoken', 'presi
dent', 'inaugural', '1789', 'almost', 'directed,', 'seem,', 'year', '1941:', '"the', 'fire', 'liberty', 'republican', 'model',
'justly', 'considered', 'deeply,', 'finally,', 'staked', 'experiment', 'intrusted', 'hands', 'people."', 'lose', 'fire--if', 'l
et', 'smothered', 'fear', 'shall', 'reject', 'washington', 'strove', 'valiantly', 'triumphantly', 'establish.', 'does,', 'wil
l,', 'furnish', 'highest', 'justification', 'sacrifice', 'make', 'cause', 'defense.', 'face', 'great', 'perils', 'never', 'enco
untered,', 'strong', 'purpose', 'protect', 'perpetuate', 'integrity', 'muster', 'america,', 'retreat.', 'content', 'stand', 'st
ill.', 'americans,', 'go', 'forward,', 'service', 'country,', 'god.']
```

```
#\w: Returns a match where the string contains any word characters (cha
racters from a to Z, digits from 0-9, and the underscore _ character)
#\s: Returns a match where the string contains a white space character.
#[^]: Returns a match for any character EXCEPT what is written after it
.
```

Now we will remove all unwanted character

	speech
0	on each national day of inauguration since 178...
1	in washington's day the task of the people was...
2	in lincoln's day the task of the people was to...
3	in this day the task of the people is to save ...
4	to us there has come a time, in the midst of s...

Word count before stopword removal

```
df_text_Roosevelt['word_count']= df_text_Roosevelt['speech'].apply(lambda x:len(str(x).split()))

df_text_Roosevelt[['speech','word_count']].head()
```

	speech	word_count
0	on each national day of inauguration since 178...	20
1	in washingtons day the task of the people was ...	16
2	in lincolns day the task of the people was to ...	17
3	in this day the task of the people is to save ...	20
4	to us there has come a time in the midst of sw...	52

Removal of Stopwords_Roosevelt

```
from  nltk.corpus import stopwords
stop= stopwords.words('english')
df_text_Roosevelt['speech']= df_text_Roosevelt['speech'].apply(lambda x:" ".join(x for x in x.split() if x not in stop))
df_text_Roosevelt[['speech']].head()
```

	speech
0	national day inauguration since 1789 people re...
1	washingtons day task people create weld togeth...
2	lincolns day task people preserve nation disru...
3	day task people save nation institutions disru...
4	us come time midst swift happenings pause mome...

After removing the stopword checking the word count

	speech	word_count
0	national day inauguration since 1789 people re...	11
1	washingtons day task people create weld togeth...	8
2	lincolns day task people preserve nation disru...	8
3	day task people save nation institutions disru...	8
4	us come time midst swift happenings pause mome...	19

Sample sentence after removal of stopwords_Roosevelt

```
df_text_Roosevelt['speech'][0]
```

```
'national day inauguration since 1789 people renewed sense dedication united states'
```

President John F. Kennedy in 1961:

- 1st we are spliting each row of the dataframe into words.
- 2nd we are joining all the above words with a space between them.
- 3rd we are making a Series out of it.
- 4th we we are extracting each word one by one and storing it in the variable all_words.

Words Frequency

```
all_words_freq_kennedy
```

```
FreqDist({'let': 16, 'us': 12, 'world': 8, 'sides': 8, 'new': 7, 'pledge': 7, 'citizens': 5, 'power': 5, 'shall': 5, 'free': 5,
...})
```

Stop words

```
print(stopwords_kennedy)
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'y
ourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself',
'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those',
'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'a
n', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'b
etween', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'of
f', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both',
'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very',
's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'ar
en', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "have
n't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "should
n't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't", '!', '"', '#', '$', '%', '&', "'", '(', ')',
'*', '+', ',', '-', '.', '/', ':', ';', '<', '=', '>', '?', '@', '[', '\\', ']', '^', '_', '`', '{', '|', '}', '~']
```

These are the stop words which we have to remove.

Word Features

```
print(word_features_kennedy)
```

```
['let', 'us', 'world', 'sides', 'new', 'pledge', 'citizens', 'power', 'shall', 'free', 'nations', 'ask', 'president', 'fellow',
'freedom', 'man', 'first', 'americans', 'war', 'peace', 'always', 'cannot', 'hope', 'help', 'arms', 'country', 'call', 'today',
'well', 'human', 'poverty', 'life', 'globe', 'dare', 'go', 'generation', 'know', 'bear', 'control', 'may', 'good', 'join', 'beg
in', 'never', 'final', 'vice', 'mr', 'god', 'forebears', 'century', 'hands', 'forms', 'yet', 'around', 'rights', 'hand', 'revol
ution', 'word', 'forth', 'time', 'friend', 'foe', 'passed', 'nation', 'committed', 'every', 'whether', 'burden', 'meet', 'oppos
e', 'assure', 'success', 'loyalty', 'united', 'little', 'powerful', 'states', 'welcome', 'merely', 'far', 'tyranny', 'find', 's
upporting', 'back', 'best', 'seek', 'south', 'offer', 'deeds', 'alliance', 'powers', 'instruments', 'weak', 'finally', 'would',
'anew', 'science', 'weakness', 'beyond', 'doubt', 'course', 'balance', 'negotiate', 'fear', 'explore', 'problems', 'unite', 'in
stead', 'bring', 'absolute', 'together', 'disease', 'earth', 'endeavor', 'finished', 'days', 'though', 'long', 'struggle', 'yea
r', 'enemies', 'history', 'light', 'truly', 'america', 'johnson', 'speaker', 'chief', 'justice', 'eisenhower', 'nixon', 'truma
n', 'reverend', 'clergy', 'observe', 'victory', 'party', 'celebration', 'symbolizing', 'end', 'beginning', 'signifying', 'renew
al', 'change', 'sworn', 'almighty', 'solemn', 'oath', '1', 'prescribed', 'nearly', 'three', 'quarters', 'ago', 'different', 'ho
lds', 'mortal', 'abolish', 'revolutionary', 'beliefs', 'fought', 'still', 'issue', 'belief', 'come', 'generosity', 'state', 'fo
rget', 'heirs', 'place', 'alike', 'torch', 'born', 'tempered', 'disciplined', 'hard', 'bitter', 'proud', 'ancient', 'heritage',
'unwilling', 'witness', 'permit', 'slow', 'undoing', 'home', 'wishes', 'ill', 'pay', 'price', 'hardship', 'support', 'order',
'survival', 'liberty', 'much', 'old', 'allies', 'whose', 'cultural', 'spiritual', 'origins', 'share', 'faithful', 'friends', 'h
ost', 'cooperative', 'ventures', 'divided', 'challenge', 'odds', 'split', 'asunder', 'ranks', 'one', 'form', 'colonial', 'awa
y', 'replaced', 'iron', 'expect', 'view', 'strongly', 'remember', 'past', 'foolishly', 'sought', 'riding', 'tiger', 'ended', 'i
nside', 'peoples', 'huts', 'villages', 'across', 'struggling', 'break', 'bonds', 'mass', 'misery', 'efforts', 'whatever', 'peri
od', 'required', 'communists', 'votes', 'right', 'society', 'many', 'poor', 'save', 'rich', 'sister', 'republics', 'border', 's
pecial', 'convert', 'words', 'progress', 'assist', 'men', 'governments', 'casting', 'chains', 'peaceful', 'become', 'prey', 'ho
stile', 'neighbors', 'aggression', 'subversion', 'anywhere', 'americas', 'hemisphere', 'intends', 'remain', 'master', 'house',
'assembly', 'sovereign', 'last', 'age', 'outpaced', 'renew', 'supportto', 'prevent', 'becoming', 'forum', 'invective', 'strengt
hen', 'shield', 'enlarge', 'area', 'writ', 'run', 'make', 'adversary', 'request', 'quest', 'dark', 'destruction', 'unleashed',
'engulf', 'humanity', 'planned', 'accidental', 'selfdestruction', 'tempt', 'sufficient', 'certain', 'employed', 'neither', 'tw
o', 'great', 'groups', 'take', 'comfort', 'present', 'overburdened', 'cost', 'modern', 'weapons', 'rightly', 'alarmed', 'stead
y', 'spread', 'deadly', 'atom', 'racing', 'alter', 'uncertain', 'terror', 'stays', 'mankinds', 'remembering', 'civility', 'sig
n', 'sincerity', 'subject', 'proof', 'belaboring', 'divide', 'formulate', 'serious', 'precise', 'proposals', 'inspection', 'des
troy', 'invoke', 'wonders', 'terrors', 'stars', 'conquer', 'deserts', 'eradicate', 'tap', 'ocean', 'depths', 'encourage', 'art
s', 'commerce', 'heed', 'corners', 'command', 'isaiah', 'undo', 'heavy', 'burdens', 'oppressed', 'beachhead', 'cooperation', 'p
ush', 'jungle', 'suspicion', 'creating', 'law', 'strong', 'secure', 'preserved', '100', '1000', 'administration', 'even', 'perh
aps', 'lifetime', 'planet', 'mine', 'rest', 'failure', 'since', 'founded', 'summoned', 'give', 'testimony', 'national', 'grave
s', 'young', 'answered', 'service', 'surround', 'trumpet', 'summons', 'need', 'battle', 'embattled', 'twilight', 'rejoicing',
'patient', 'tribulation', 'common', 'forge', 'grand', 'global', 'north', 'east', 'west', 'fruitful', 'mankind', 'historic', 'ef
fort', 'generations', 'granted', 'role', 'defending', 'hour', 'maximum', 'danger', 'shrink', 'responsibility', 'believe', 'exch
ange', 'places', 'people', 'energy', 'faith', 'devotion', 'serve', 'glow', 'fire', 'high', 'standards', 'strength', 'sacrific
e', 'conscience', 'sure', 'reward', 'judge', 'lead', 'land', 'love', 'asking', 'blessing', 'knowing', 'gods', 'work', 'must']
```

```
#\w: Returns a match where the string contains any word characters (cha
racters from a to Z, digits from 0-9, and the underscore _ character)
#\s: Returns a match where the string contains a white space character.
#[^]: Returns a match for any character EXCEPT what is written after it
.
```

Now we will remove all unwanted character

	speech
0	vice president johnson mr speaker mr chief jus...
1	the world is very different now for man holds ...
2	we dare not forget today that we are the heirs...
3	let every nation know whether it wishes us wel...
4	this much we pledge and more

Word count before stopword removal

	speech	word_count
0	vice president johnson mr speaker mr chief jus...	71
1	the world is very different now for man holds ...	67
2	we dare not forget today that we are the heirs...	94
3	let every nation know whether it wishes us wel...	40
4	this much we pledge and more	6

Removal of Stopwords_Roosevelt

```
from  nltk.corpus import stopwords
stop= stopwords.words('english')
df_text_Kennedy['speech']= df_text_Kennedy['speech'].apply(lambda x:" ".join(x for x in x.split() if x not in stop))
df_text_Kennedy[['speech']].head()
```

	speech
0	vice president johnson mr speaker mr chief jus...
1	world different man holds mortal hands power a...
2	dare forget today heirs first revolution let w...
3	let every nation know whether wishes us well i...
4	much pledge

After removing the stopword checking the word count

	speech	word_count
0	vice president johnson mr speaker mr chief jus...	46
1	world different man holds mortal hands power a...	31
2	dare forget today heirs first revolution let w...	46
3	let every nation know whether wishes us well i...	25
4	much pledge	2

Sample sentence after removal of stopwords_Roosevelt

```
df_text_Kennedy['speech'][0]
```

```
'vice president johnson mr speaker mr chief justice president eisenhower vice president nixon president truman reverend clergy fellow citizens observe today victory party celebration freedom symbolizing end well beginning signifying renewal well change s worn almighty god solemn oath forebears l prescribed nearly century three quarters ago'
```

President Richard Nixon in 1973:

- 1st we are spliting each row of the dataframe into words.
- 2nd we are joining all the above words with a space between them.

- 3rd we are making a Series out of it.
- 4th we we are extracting each word one by one and storing it in the variable all_words.

Words Frequency

```
(all_words_freq_Nixon)
```

```
FreqDist({'us': 25, 'let': 22, '--': 17, 'new': 15, 'peace': 11, 'great': 9, 'america': 9, 'world.': 8, "america's": 8, 'shal
l': 7, ...})
```

Stop words

```
stopwords_Nixon
```

```
['i',
 'me',
 'my',
 'myself',
 'we',
 'our',
 'ours',
 'ourselves',
 'you',
 "you're",
 "you've",
 "you'll",
 "you'd",
 'your',
 'yours',
 'yourself',
 'yourselves',
 'he',
 'him',
```

These are the stop words which we have to remove.

Word Features

```
word_features_Nixon
```

```
['us',
 'let',
 '--',
 'new',
 'peace',
 'great',
 'america',
 'world.',
 "america's",
 'shall',
 'policies',
 'world',
 'make',
 'every',
 'better',
 'government',
 'abroad',
 'role',
 'people',
```

```
#\w: Returns a match where the string contains any word characters (cha
racters from a to Z, digits from 0-9, and the underscore _ character)
#\s: Returns a match where the string contains a white space character.
#[^]: Returns a match for any character EXCEPT what is written after it
.
```

Now we will remove all unwanted character

	speech
0	mr. vice president, mr. speaker, mr. chief jus...
1	when we met here four years ago, america was b...
2	as we meet here today, we stand on the thresho...
3	the central question before us is: how shall w...
4	let us resolve that this will be what it can b...

Word count before stopword removal

```
df_text_Nixon['word_count']= df_text_Nixon['speech'].apply(lambda x:len(str(x).split()))
df_text_Nixon[['speech','word_count']].head()
```

	speech	word_count
0	mr vice president mr speaker mr chief justice ...	25
1	when we met here four years ago america was bl...	27
2	as we meet here today we stand on the threshol...	19
3	the central question before us is how shall we...	51
4	let us resolve that this will be what it can b...	38

Removal of Stopwords_Roosevelt

```
from  nltk.corpus import stopwords
stop= stopwords.words('english')
df_text_Nixon['speech']= df_text_Nixon['speech'].apply(lambda x:" ".join(x for x in x.split()if x not in stop))
df_text_Nixon[['speech']].head()
```

	speech
0	mr vice president mr speaker mr chief justice ...
1	met four years ago america bleak spirit depres...
2	meet today stand threshold new era peace world
3	central question us shall use peace let us res...
4	let us resolve become time great responsibilit...

After removing the stopword checking the word count

```
df_text_Nixon['word_count']= df_text_Nixon['speech'].apply(lambda x:len(str(x).split()))
df_text_Nixon[['speech','word_count']].head()
```

	speech	word_count
0	mr vice president mr speaker mr chief justice ...	19
1	met four years ago america bleak spirit depres...	16
2	meet today stand threshold new era peace world	8
3	central question us shall use peace let us res...	24
4	let us resolve become time great responsibilit...	17

Sample sentence after removal of stopwords_Roosevelt

```
df_text_Nixon['speech'][0]
```

```
'mr vice president mr speaker mr chief justice senator cook mrs eisenhower fellow citizens great good country share together'
```

Question 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:

For President Franklin D. Roosevelt in 1941:

Removal of Punctuation:

```
#\w: Returns a match where the string contains any word characters (cha
racters from a to Z, digits from 0-9, and the underscore _ character)
#\s: Returns a match where the string contains a white space character.
#[^]: Returns a match for any character EXCEPT what is written after it
```

	speech
0	national day inauguration since 1789 people re...
1	washingtons day task people create weld togeth...
2	lincolns day task people preserve nation disru...
3	day task people save nation institutions disru...
4	us come time midst swift happenings pause mome...

Common Words Removal:

1.We will create a list of 10 frequently occurring words and then decide if we need to remove it or retain it.

2.Reason is that this file has speech related to President Roosevelt. Hence, there is no point in keeping the word like Roosevelt, unless we are focussing on speeches from other President

```
Freq_Roosevelt= pd.Series(' '.join(df_text_Roosevelt['speech']).split()).value_counts()[:10]
Freq_Roosevelt
```

```
nation       11
know         10
democracy     9
spirit        9
life          8
us            8
people        7
america       7
years         6
freedom       6
dtype: int64
```

Since the top 10 frequently used words look important according to a President's Inaugral speech.
Hence, we are not removing any words from here.

Rare Words Removal :

This is done as association of these less occurring words with the existing words could be a noise

```
Freq_Roosevelt= pd.Series(' '.join(df_text_Roosevelt['speech']).split()).value_counts()[-10:]
Freq_Roosevelt
```

```
joined        1
force         1
full          1
enlists       1
majority      1
expression    1
carried       1
undertaken    1
common        1
god           1
dtype: int64
```

As it is difficult to make out if these words will have association in text analytics or not, these words are kept in the dataset.

Stemming

- Stemming is the process of reducing inflection in words to their root forms such as mapping a group ofwords to the same stem even if the stem itself is not a valid word in the Language.
- Stem (root) is the part of the word to which you add inflectional (changing/deriving) affixes such as (-ed,-ize, -s,-de,mis).
- Stems are created by removing the suffixes or prefixes used with a word. Hence, stemming a word or sentence may result in words that are not actual words.

```
from  nltk.stem import PorterStemmer
stem_Kennedy= PorterStemmer()
df_text_Kennedy['speech'].apply(lambda x:" ".join([stem_Kennedy.stem(word) for word in x.split()]))
```

```
0     vice presid johnson mr speaker mr chief justic...
1     world differ man hold mortal hand power abolis...
2     dare forget today heir first revolut let word ...
3     let everi nation know whether wish us well ill...
4                                          much pledg
5     old alli whose cultur spiritu origin share ple...
6     new state welcom rank free pledg word one form...
7     peopl hut villag across globe struggl break bo...
8     sister republ south border offer special pledg...
9     world assembl sovereign state unit nation last...
10    final nation would make adversari offer pledg ...
11    dare tempt weak arm suffici beyond doubt certa...
12    neither two great power group nation take comf...
13    let us begin anew rememb side civil sign weak ...
14    let side explor problem unit us instead belabo...
15    let side first time formul seriou precis propo...
16    let side seek invok wonder scienc instead terr...
17    let side unit heed corner earth command isaiah...
18    beachhead cooper may push back jungl suspicion...
19    finish first 100 day finish first 1000 day lif...
20    hand fellow citizen mine rest final success fa...
21    trumpet summon us call bear arm though arm nee...
22    forg enemi grand global allianc north south ea...
23    long histori world gener grant role defend fre...
24              fellow american ask countri ask countri
25    fellow citizen world ask america togeth freedo...
26    final whether citizen america citizen world as...
Name: speech, dtype: object
```

Sample sentence after BASIC pre processing

```
df_text_Roosevelt['speech'][0]
```

```
'national day inauguration since 1789 people renewed sense dedication united states'
```

Most Common word

```
# most common word
Freq_Roosevelt= pd.Series(' '.join(df_text_Roosevelt['speech']).split()).value_counts()[:3]
Freq_Roosevelt
```

```
nation       11
know         10
democracy     9
dtype: int64
```

For President John F. Kennedy in 1961:

Removal of Punctuation:

```
#\w: Returns a match where the string contains any word characters (cha
racters from a to Z, digits from 0-9, and the underscore _ character)
#\s: Returns a match where the string contains a white space character.
#[^]: Returns a match for any character EXCEPT what is written after it
```

```
df_text_Kennedy[['speech']].head()
```

	speech
0	vice president johnson mr speaker mr chief jus...
1	world different man holds mortal hands power a...
2	dare forget today heirs first revolution let w...
3	let every nation know whether wishes us well i...
4	much pledge

Common Words Removal:

Since amongst the top 10 frequently used words, words 'let' and 'shall' do not look so important according to a President's Inaugral speech. Hence, we will remove these two words to proceed further

```
Freq_Kennedy= pd.Series(' '.join(df_text_Kennedy['speech']).split()).value_counts()[:10]
Freq_Kennedy
```

```
let         16
us          12
sides        8
world        8
pledge       7
new          7
ask          5
citizens     5
nations      5
free         5
dtype: int64
```

Rare Words Removal :
This is done as association of these less occurring words with the existing words could be a noise

```
Freq_Kennedy= pd.Series(' '.join(df_text_Kennedy['speech']).split()).value_counts()[-10:]
Freq_Kennedy
```

```
misery        1
society       1
right         1
votes         1
communists    1
required      1
period        1
whatever      1
efforts       1
must          1
dtype: int64
```

As it is difficult to make out if these words will have association in text analytics or not, these words are kept in the dataset.

Stemming

- Stemming is the process of reducing inflection in words to their root forms such as mapping a group ofwords to the same stem even if the stem itself is not a valid word in the Language.
- Stem (root) is the part of the word to which you add inflectional (changing/deriving) affixes such as (-ed,-ize, -s,-de,mis).
- Stems are created by removing the suffixes or prefixes used with a word. Hence, stemming a word or sentence may result in words that are not actual words.

```
from  nltk.stem import PorterStemmer
stem_Kennedy= PorterStemmer()
df_text_Kennedy['speech'].apply(lambda x:" ".join([stem_Kennedy.stem(word) for word in x.split()]))
```

```
0     vice presid johnson mr speaker mr chief justic...
1     world differ man hold mortal hand power abolis...
2     dare forget today heir first revolut let word ...
3     let everi nation know whether wish us well ill...
4                                          much pledg
5     old alli whose cultur spiritu origin share ple...
6     new state welcom rank free pledg word one form...
7     peopl hut villag across globe struggl break bo...
8     sister republ south border offer special pledg...
9     world assembl sovereign state unit nation last...
10    final nation would make adversari offer pledg ...
11    dare tempt weak arm suffici beyond doubt certa...
12    neither two great power group nation take comf...
13    let us begin anew rememb side civil sign weak ...
14    let side explor problem unit us instead belabo...
15    let side first time formul seriou precis propo...
16    let side seek invok wonder scienc instead terr...
17    let side unit heed corner earth command isaiah...
18    beachhead cooper may push back jungl suspicion...
19    finish first 100 day finish first 1000 day lif...
20    hand fellow citizen mine rest final success fa...
21    trumpet summon us call bear arm though arm nee...
22    forg enemi grand global allianc north south ea...
23    long histori world gener grant role defend fre...
24              fellow american ask countri ask countri
25    fellow citizen world ask america togeth freedo...
26    final whether citizen america citizen world as...
Name: speech, dtype: object
```

Sample sentence after BASIC pre processing

```
df_text_Kennedy['speech'][0]
```

```
'vice president johnson mr speaker mr chief justice president eisenhower vice president nixon president truman reverend clergy fellow citizens observe today victory party celebration freedom symbolizing end well beginning signifying renewal well change s worn almighty god solemn oath forebears l prescribed nearly century three quarters ago'
```

Most Common word

```
# most common word
Freq_Kennedy= pd.Series(' '.join(df_text_Kennedy['speech']).split()).value_counts()[:3]
Freq_Kennedy
```

```
let      16
us       12
sides     8
dtype: int64
```

For President Richard Nixon in 1973:

Removal of Punctuation:

```
#\w: Returns a match where the string contains any word characters (cha
racters from a to Z, digits from 0-9, and the underscore _ character)
#\s: Returns a match where the string contains a white space character.
#[^]: Returns a match for any character EXCEPT what is written after it
```

```
df_text_Nixon[['speech']].head()
```

	speech
0	mr vice president mr speaker mr chief justice ...
1	met four years ago america bleak spirit depres...
2	meet today stand threshold new era peace world
3	central question us shall use peace let us res...
4	let us resolve become time great responsibilit...

Common Words Removal:

Since amongst the top 10 frequently used words, words 'let' and 'shall' do not look so important according to a President's Inaugral speech. Hence, we will remove these two words to proceed further

```
Freq_Nixon= pd.Series(' '.join(df_text_Nixon['speech']).split()).value_counts()[:10]
Freq_Nixon
```

```
us                26
let               22
peace             19
world             16
new               15
america           13
responsibility    11
government        10
great              9
home               9
dtype: int64
```

Rare Words Removal :

This is done as association of these less occurring words with the existing words could be a noise

```
Freq_Nixon= pd.Series(' '.join(df_text_Nixon['speech']).split()).value_counts()[-10:]
Freq_Nixon
```

```
strength     1
influence    1
would        1
different    1
respects     1
safe         1
weak         1
friends      1
systems      1
purpose      1
dtype: int64
```

As it is difficult to make out if these words will have association in text analytics or not, these words are kept in the dataset.

Stemming

- Stemming is the process of reducing inflection in words to their root forms such as mapping a group ofwords to the same stem even if the stem itself is not a valid word in the Language.
- Stem (root) is the part of the word to which you add inflectional (changing/deriving) affixes such as (-ed,-ize, -s,-de,mis).
- Stems are created by removing the suffixes or prefixes used with a word. Hence, stemming a word or sentence may result in words that are not actual words.

```
from  nltk.stem import PorterStemmer
stem_Nixon= PorterStemmer()
df_text_Nixon['speech'].apply(lambda x:" ".join([stem_Nixon.stem(word) for word in x.split()]))
```

```
0     mr vice presid mr speaker mr chief justic sena...
1     met four year ago america bleak spirit depress...
2         meet today stand threshold new era peac world
3     central question us shall use peac let us reso...
4     let us resolv becom time great respons greatli...
5     past year saw farreach result new polici peac ...
6     peac seek world flimsi peac mere interlud war ...
7     import understand necess limit america role ma...
8                 unless america work preserv peac peac
9           unless america work preserv freedom freedom
10    let us clearli understand new natur america ro...
11                          shall respect treati commit
12    shall support vigor principl countri right imp...
13    shall continu era negoti work limit nuclear ar...
14    shall share defend peac freedom world shall ex...
15    time pass america make everi nation conflict m...
16    respect right nation determin futur also recog...
17    america role indispens preserv world peac nati...
18    togeth rest world let us resolv move forward b...
19    let us build structur peac world weak safe str...
20    let us accept high respons burden gladli gladl...
21    chanc today ever histori make life better amer...
22    rang need great reach opportun great let us bo...
23    build structur peac abroad requir turn away ol...
24    abroad shift old polici new retreat respons be...
25    home shift old polici new retreat respons bett...
26    abroad home key new respons lie place divis re...
27    abroad home time come turn away condescend pol...
28    person expect act respons respons human natur ...
29    today offer promis pure government solut everi...
30              govern must learn take less peopl peopl
31    let us rememb america built govern peopl welfa...
32                               live let us ask govern
33     challeng face togeth let us ask govern help help
34    nation govern great vital role play pledg gove...
35    day forward let us make solemn commit heart be...
36    america longest difficult war come end let us ...
37                  els time come us renew faith america
38                           recent year faith challeng
39    children taught asham countri asham parent ash...
40    everi turn beset find everyth wrong america li...
41    america record centuri unparallel world histor...
42    let us proud system produc provid freedom abun...
43    let us proud four war engag centuri includ one...
44    let us proud bold new initi steadfast peac hon...
45    embark today era present challeng great nation...
46      shall answer god histori conscienc way use year
47    stand place hallow histori think other stood t...
48    today ask prayer year ahead may god help make ...
49    let us pledg togeth make next four year best f...
50    let us go forward confid hope strong faith one...
Name: speech, dtype: object
```

Sample sentence after BASIC pre processin

```
df_text_Nixon['speech'][0]
```

```
'mr vice president mr speaker mr chief justice senator cook mrs eisenhower fellow citizens great good country share together'
```

Most Common word

```
# most common word
Freq_Nixon= pd.Series(' '.join(df_text_Nixon['speech']).split()).value_counts()[:3]
Freq_Nixon
```

```
us       26
let      22
peace    19
dtype: int64
```

Question 2.4

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

Answer:

Word Cloud Roosevelt

Removing stopwords from the corpus

```
corpus_Roosevelt= df_text_Roosevelt['speech'].apply(lambda x:' '.join([z for z in x.split() if z not in stop_words]))
corpus_Roosevelt
```

```
0     national day inauguration since 1789 people re...
1     washingtons day task people create weld togeth...
2     lincolns day task people preserve nation disru...
3     day task people save nation institutions disru...
4     us come time midst swift happenings pause mome...
5     lives nations determined count years lifetime ...
6     men doubt men believe democracy form governmen...
7                                 americans know true
8     eight years ago life republic seemed frozen fa...
9     later years living years fruitful years people...
10    vital present future experience democracy succ...
11    action taken within threeway framework constit...
12                                  democracy dying
13                           know seen reviveand grow
14    know cannot die built unhampered initiative in...
15    know democracy alone forms government enlists ...
16    know democracy alone constructed unlimited civ...
17    know look surface sense still spreading every ...
18    nation like person bodya body must fed clothed...
19    nation like person mind mind must kept informe...
20    nation like person something deeper something ...
21    thing find difficult even impossible hit upon ...
22    yet understand spirit faith america product ce...
23    democratic aspiration mere recent phase human ...
24    americas impact irresistible america new world...
25    vitality written mayflower compact declaration...
26    first came carry longings spirit millions foll...
27    hopes republic cannot forever tolerate either ...
28    know still far go must greatly build security ...
29    enough achieve purposes alone enough clothe fe...
30         without body mind men know nation could live
31    spirit america killed even though nations body...
32    spirit faith speaks us daily lives ways often ...
33    destiny america proclaimed words prophecy spok...
34    lose sacred fireif let smothered doubt fear sh...
35    face great perils never encountered strong pur...
36                  muster spirit america faith america
37    retreat content stand still americans go forwa...
Name: speech, dtype: object
```

```
Word_Cloud_Roosevelt= ' '.join(corpus_Roosevelt)
Word_Cloud_Roosevelt
```

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

Plotting Word Cloud

```
from wordcloud import WordCloud,STOPWORDS
```

```
wordcloud = WordCloud(stopwords=STOPWORDS,
                      width = 3000, height = 3000,
                background_color ='white',
                min_font_size = 10, random_state=100).generate(Word_Cloud_Roosevelt)

# plot the WordCloud image
plt.figure(figsize = (8, 8), facecolor = None)
plt.imshow(wordcloud)
plt.axis("off")
plt.xlabel('Word Cloud')
plt.tight_layout(pad = 0)

print("Word Cloud for President Roosevelt after removing stopwords (after cleaning)!!")
plt.show()
```

Word Cloud for President Roosevelt after removing stopwords (after cleaning)!!

Figure 10: wordcloud for Roosevelt

For President Kennedy

Removing stopwords from the corpus

```
stop_words= list(stopwords.words('english'))
```

```
corpus_Kennedy= df_text_Kennedy['speech'].apply(lambda x:' '.join([z for z in x.split() if z not in stop_words]))
corpus_Kennedy
```

```
0     vice president johnson mr speaker mr chief jus...
1     world different man holds mortal hands power a...
2     dare forget today heirs first revolution let w...
3     let every nation know whether wishes us well i...
4                                          much pledge
5     old allies whose cultural spiritual origins sh...
6     new states welcome ranks free pledge word one ...
7     peoples huts villages across globe struggling ...
8     sister republics south border offer special pl...
9     world assembly sovereign states united nations...
10    finally nations would make adversary offer ple...
11    dare tempt weakness arms sufficient beyond dou...
12    neither two great powerful groups nations take...
13    let us begin anew remembering sides civility s...
14    let sides explore problems unite us instead be...
15    let sides first time formulate serious precise...
16    let sides seek invoke wonders science instead ...
17    let sides unite heed corners earth command isa...
18    beachhead cooperation may push back jungle sus...
19    finished first 100 days finished first 1000 da...
20    hands fellow citizens mine rest final success ...
21    trumpet summons us call bear arms though arms ...
22    forge enemies grand global alliance north sout...
23    long history world generations granted role de...
24             fellow americans ask country ask country
25    fellow citizens world ask america together fre...
26    finally whether citizens america citizens worl...
Name: speech, dtype: object
```

```
Word_Cloud_Kennedy= ' '.join(corpus_Kennedy)
Word_Cloud_Kennedy
```

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

Wordcloud plotting

```
wordcloud = WordCloud(stopwords=STOPWORDS,
                      width = 3000, height = 3000,
                background_color ='white',
                min_font_size = 10, random_state=100).generate(Word_Cloud_Kennedy)

# plot the WordCloud image
plt.figure(figsize = (8, 8), facecolor = None)
plt.imshow(wordcloud)
plt.axis("off")
plt.xlabel('Word Cloud')
plt.tight_layout(pad = 0)

print("Word Cloud for President Kennedy after removing stopword (after cleaning)!!")
plt.show()
```

Figure 11: wordcloud for Kennedy

For president Nixon

Removing stopwords from the corpus

```
corpus_Nixon= df_text_Nixon['speech'].apply(lambda x:' '.join([z for z in x.split() if z not in stop_words]))
corpus_Nixon
```

```
0     mr vice president mr speaker mr chief justice ...
1     met four years ago america bleak spirit depres...
2        meet today stand threshold new era peace world
3     central question us shall use peace let us res...
4     let us resolve become time great responsibilit...
5     past year saw farreaching results new policies...
6     peace seek world flimsy peace merely interlude...
7     important understand necessity limitations ame...
8              unless america work preserve peace peace
9          unless america work preserve freedom freedom
10    let us clearly understand new nature americas ...
11                     shall respect treaty commitments
12    shall support vigorously principle country rig...
13    shall continue era negotiation work limitation...
14    shall share defending peace freedom world shal...
15    time passed america make every nations conflic...
16    respect right nation determine future also rec...
17    americas role indispensable preserving worlds ...
18    together rest world let us resolve move forwar...
19    let us build structure peace world weak safe s...
20    let us accept high responsibility burden gladl...
21    chance today ever history make life better ame...
22    range needs great reach opportunities great le...
23    building structure peace abroad required turni...
24    abroad shift old policies new retreat responsi...
25    home shift old policies new retreat responsibi...
26    abroad home key new responsibilities lies plac...
27    abroad home time come turn away condescending ...
28    person expected act responsibly responsibility...
29    today offer promise purely governmental soluti...
30        government must learn take less people people
31    let us remember america built government peopl...
31    let us remember america built government peopl...
32                          lives let us ask government
33    challenges face together let us ask government...
34    national government great vital role play pled...
35    day forward let us make solemn commitment hear...
36    americas longest difficult war comes end let u...
37                else time come us renew faith america
38                        recent years faith challenged
39    children taught ashamed country ashamed parent...
40    every turn beset find everything wrong america...
41    americas record century unparalleled worlds hi...
42    let us proud system produced provided freedom ...
43    let us proud four wars engaged century includi...
44    let us proud bold new initiatives steadfastnes...
45    embarking today era presents challenges great ...
46    shall answer god history conscience way use years
47    stand place hallowed history think others stoo...
48    today ask prayers years ahead may gods help ma...
49    let us pledge together make next four years be...
50    let us go forward confident hope strong faith ...
Name: speech, dtype: object
```

```
Word_Cloud_Nixon= ' '.join(corpus_Nixon)
Word_Cloud_Nixon
```

```
'mr vice president mr speaker mr chief justice senator cook mrs eisenhower fellow citizens great good country share together me
t four years ago america bleak spirit depressed prospect seemingly endless war abroad destructive conflict home meet today stan
d threshold new era peace world central question us shall use peace let us resolve era enter postwar periods often time retreat
isolation leads stagnation home invites new danger abroad let us resolve become time great responsibilities greatly borne renew
spirit promise america enter third century nation past year saw farreaching results new policies peace continuing revitalize tr
aditional friendships missions peking moscow able establish base new durable pattern relationships among nations world americas
bold initiatives 1972 long remembered year greatest progress since end world war ii toward lasting peace world peace seek world
flimsy peace merely interlude wars peace endure generations come important understand necessity limitations americas role maint
aining peace unless america work preserve peace peace unless america work preserve freedom freedom let us clearly understand ne
w nature americas role result new policies adopted past four years shall respect treaty commitments shall support vigorously pr
inciple country right impose rule another force shall continue era negotiation work limitation nuclear arms reduce danger confr
ontation great powers shall share defending peace freedom world shall expect others share time passed america make every nation
s conflict make every nations future responsibility presume tell people nations manage affairs respect right nation determine f
uture also recognize responsibility nation secure future americas role indispensable preserving worlds peace nations role indis
pensable preserving peace together rest world let us resolve move forward beginnings made let us continue bring walls hostility
divided world long build place bridges understanding despite profound differences systems government people world friends let u
s build structure peace world weak safe strong respects right live different system would influence others strength ideas force
arms let us accept high responsibility burden gladly gladly chance build peace noblest endeavor nation engage gladly also act g
reatly meeting responsibilities abroad remain great nation remain great nation act greatly meeting challenges home chance today
ever history make life better america ensure better education better health better housing better transportation cleaner enviro
nment restore respect law make communities livable insure godgiven right every american full equal opportunity range needs grea
t reach opportunities great let us bold determination meet needs new ways building structure peace abroad required turning away
old policies failed building new era progress home requires turning away old policies failed abroad shift old policies new retr
eat responsibilities better way peace home shift old policies new retreat responsibilities better way progress abroad home key
new responsibilities lies placing division responsibility lived long consequences attempting gather power responsibility washin
gton abroad home time come turn away condescending policies paternalism washington knows best person expected act responsibly r
esponsibility human nature let us encourage individuals home nations abroad decide let us locate responsibility places let us m
easure others today offer promise purely governmental solution every problem lived long false promise trusting much government
asked deliver leads inflated expectations reduced individual effort disappointment frustration erode confidence government peop
le government must learn take less people people let us remember america built government people welfare work shirking responsi
bility seeking responsibility lives let us ask government challenges face together let us ask government help help national gov
ernment great vital role play pledge government act act boldly lead boldly important role every one us must play individual mem
ber community day forward let us make solemn commitment heart bear responsibility part live ideals together see dawn new age pr
ogress america together celebrate 200th anniversary nation proud fulfillment promise world americas longest difficult war comes
end let us learn debate differences civility decency let us reach one precious quality government cannot provide new level resp
ect rights feelings one another new level respect individual human dignity cherished birthright every american else time come u
s renew faith america recent years faith challenged children taught ashamed country ashamed parents ashamed americas record hom
e role world every turn beset find everything wrong america little right confident judgment history remarkable times privileged
live americas record century unparalleled worlds history responsibility generosity creativity progress let us proud system prod
uced provided freedom abundance widely shared system history world let us proud four wars engaged century including one bringin
g end fought selfish advantage help others resist aggression let us proud bold new initiatives steadfastness peace honor made b
reakthrough toward creating world world known structure peace last merely time generations come embarking today era presents ch
allenges great nation generation ever faced shall answer god history conscience way use years stand place hallowed history thin
k others stood think dreams america think recognized needed help far beyond order make dreams come true today ask prayers years
ahead may gods help making decisions right america pray help together may worthy challenge let us pledge together make next fou
r years best four years americas history 200th birthday america young vital began bright beacon hope world let us go forward co
nfident hope strong faith one another sustained faith god created us striving always serve purpose'
```

.

Word Cloud for President Nixon after removing stopword (after cleaning)!!

Figure 12: wordcloud for Nixon

The END