BUSINESS REPORT OF

MACHINE LEARNING

BY

BALAVIGNESH S

13/10/2023

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.

Data Dictionary:

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.**

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.

Read the dataset.

	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

Information of the dataset

```
<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   float64
 1   economic.cond.national  1517 non-null   int64
 2   economic.cond.household 1517 non-null   int64
 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(1), int64(6), uint8(2)
memory usage: 97.8 KB
```

FINDING ANY MISSING VALUE SOF THE DATASET.

```
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
```

Unnamed 0 has no meaning in the column so we can drop it.

After dropping we can see that there 8 duplicated are to be removed

We can see there are duplicated 8 rows so we need to remove

After removing the duplicated we can see the shape of the data.

The row are 1517 and column is 9.

Describe the data set

	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

Age minimum is 24 and average age is 53 and max age is 93 in the data set.

Economic condition national and household are rated by the people is average and most is 3 and maximum is 5 in the data set.

Blair support by the people is very positive more than 75% shows that blair has best leader.

Hague support by the people is very average show the poor leadership.

Europe integration are support to average 6 that vary from 11. it show only 40% are interested.

Political knowledge most of the most have much knowledge as 2 to 3 is been average

Observation:

There are Totally 10 column and 1524 rows with 0 null values for each feature

Vote and Gender are Object and the rest are all integer Data type.

Unnamed 0 feature to be removed because no meaning insight can be driven.

There are 8 duplicated in the Row are to be removed.

After removing it it become 1517 rows and 9 column

Based on the Describe statistics, skewness of the variables:

1. age: The distribution of age appears to be positively skewed, as the mean (54.24) is greater than the median (50th percentile or 53.0), and the 75th percentile (67.0) is greater than the mean. This indicates a potential concentration of younger individuals on the left side of the distribution.

2. economic.cond.national: The data for national economic conditions appears to be approximately symmetric, as the mean (3.25) is close to the median (50th percentile or 3.0).

3. economic.cond.household: Similar to national economic conditions, household economic conditions seem to be symmetric, with the mean (3.14) close to the median (50th percentile or 3.0).

4. Blair: The distribution for the Blair variable seems to be positively skewed, with the mean (3.34) greater than the median (50th percentile or 4.0).

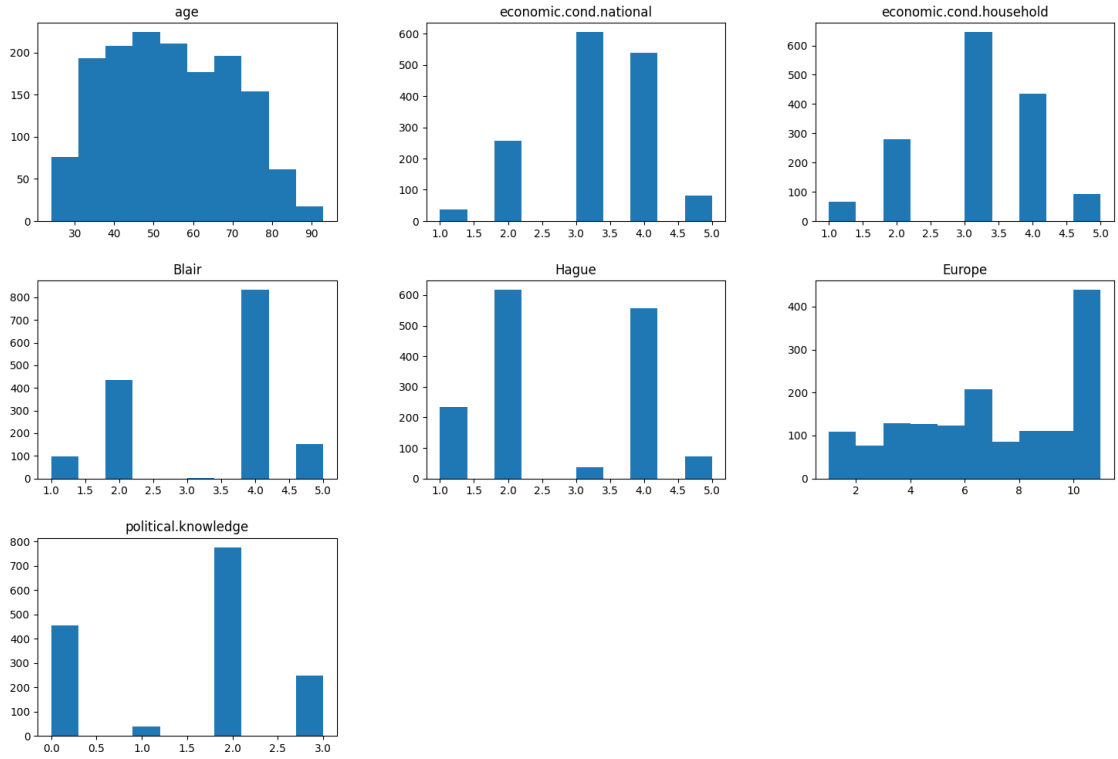
5. Hague: The distribution for the Hague variable also seems to be negative skewed, with the mean (2.75) less than the median (50th percentile or 2.0).

6. Europe: The distribution for the Europe variable may be slightly positively skewed, with the mean (6.74) greater than the median (50th percentile or 6.0).

7. political.knowledge: The data for political knowledge appears to be approximately symmetric, as the mean (1.54) is close to the median (50th percentile or 2.0).

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. 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.

UNIVARITE ANALYSIS:



Sure, here are short one-line notes for each variable's skewness observation:

1. age: The distribution is positively skewed, suggesting a concentration of younger individuals on the left side.

2. **economic.cond.national:** The data distribution appears approximately symmetric around the median.

3. **economic.cond.household :** The data distribution is symmetric around the median, reflecting consistent household economic conditions.

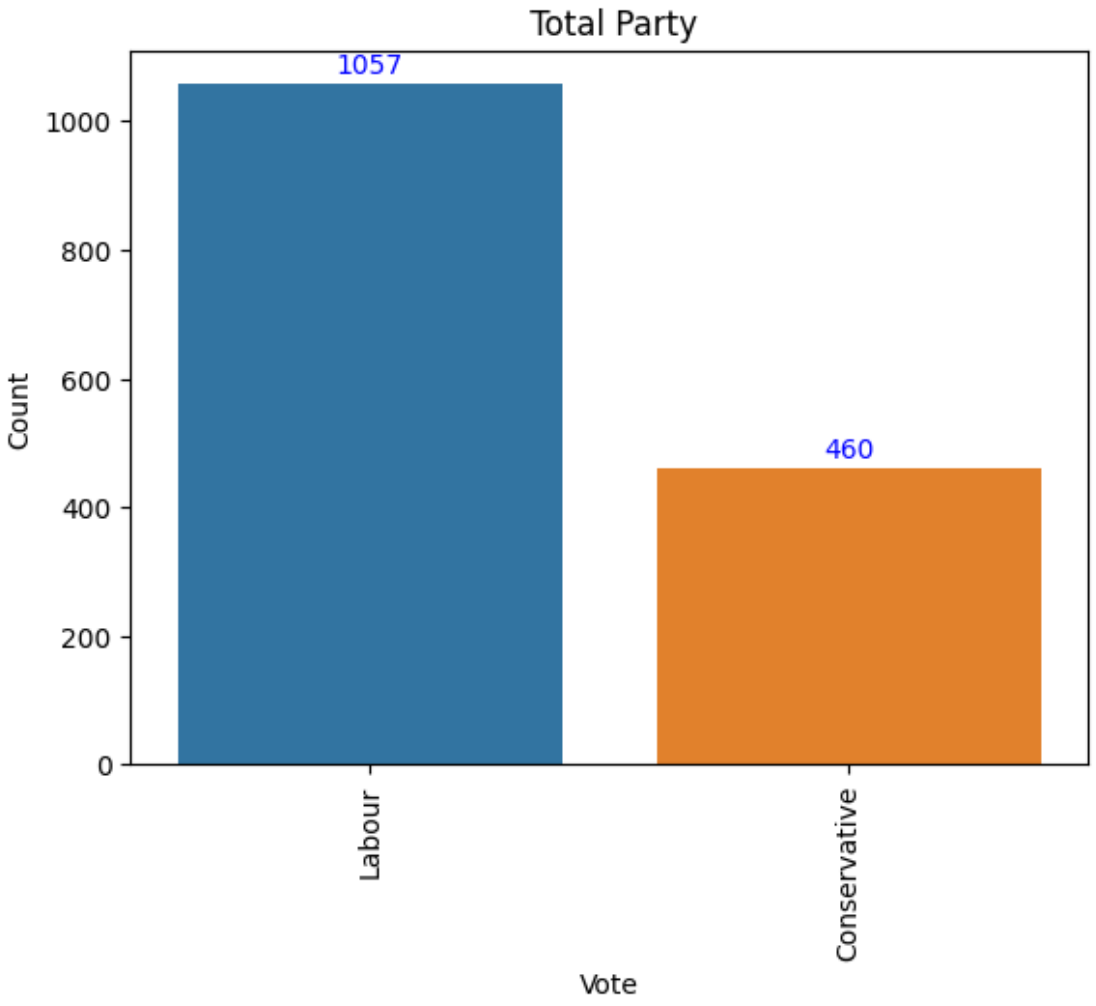
4. **Blair :** The distribution is positively skewed, with some higher values pulling the mean to the right.

5. Hague: The distribution is negatively skewed, indicating lower values that pull the mean to the left.

6. Europe: The distribution is slightly positively skewed, with some larger values extending the right tail.

7. political.knowledge: The data distribution appears approximately symmetric, centered around the median..

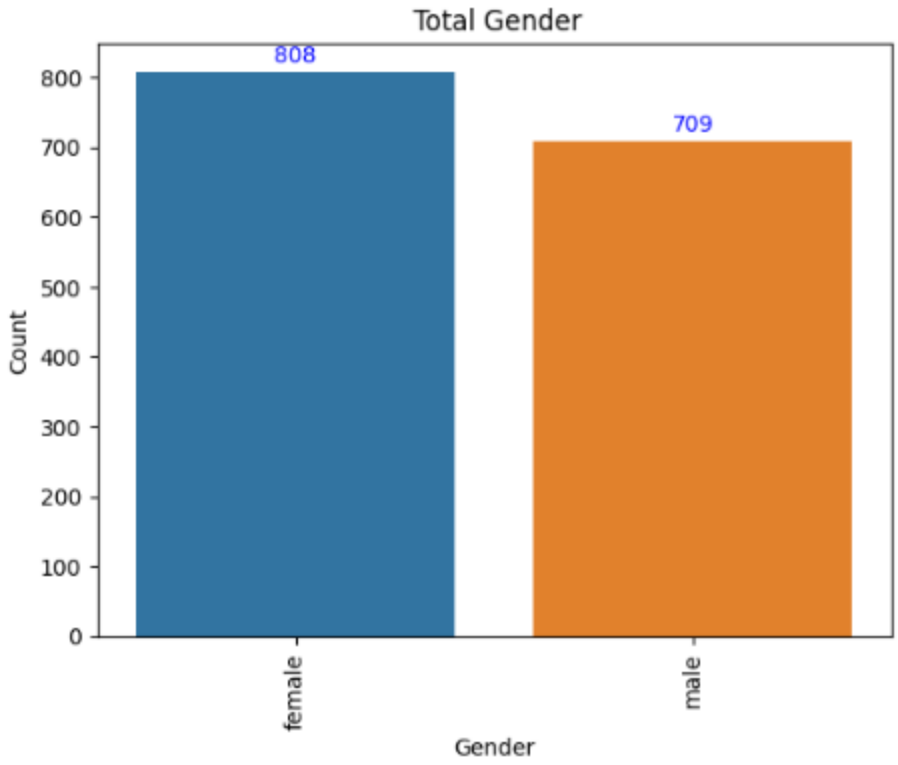
1) What is the distribution of respondents among different political parties?



The Labour are 1057

The conservative are 460

2) What is the gender distribution of the respondents in the dataset?

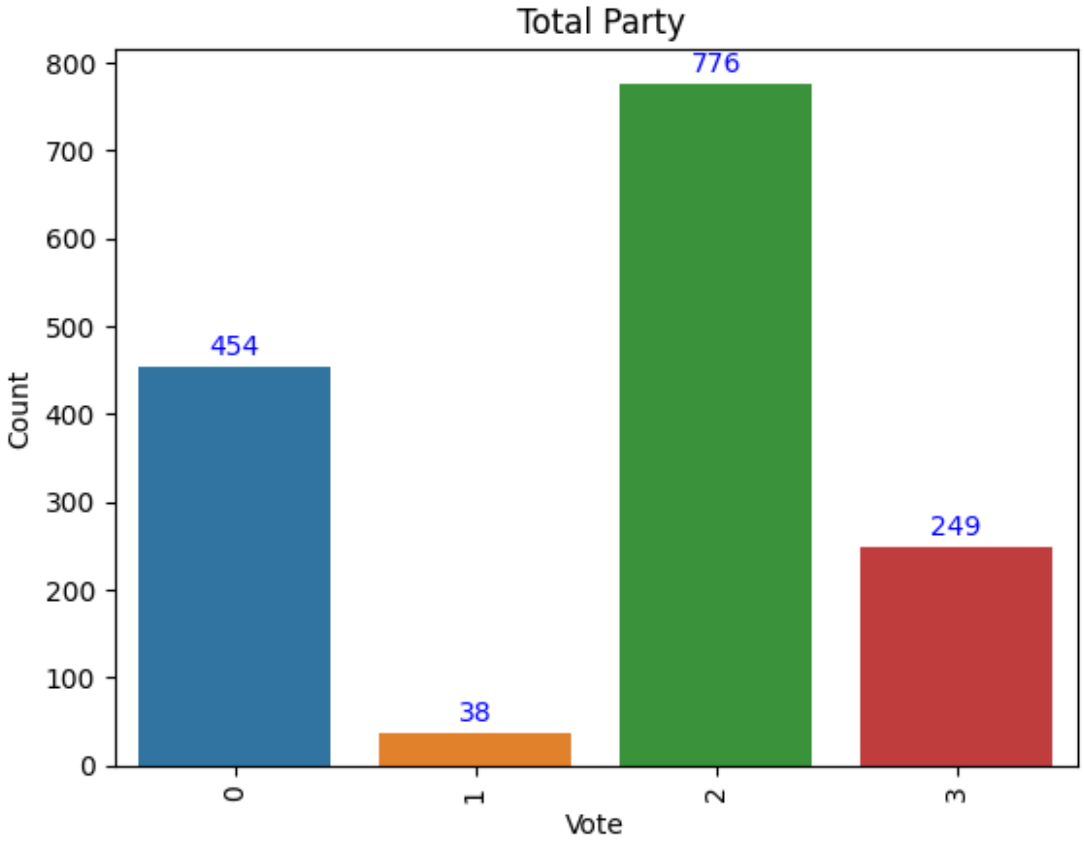


The Female are 808

The Male are 709

Female are more compared to men

3)What is the distribution of respondents' knowledge levels of parties' positions on European integration?



The value of 0 indicate that 454 has no knowledge or awareness of the positions that political parties hold on European integration.

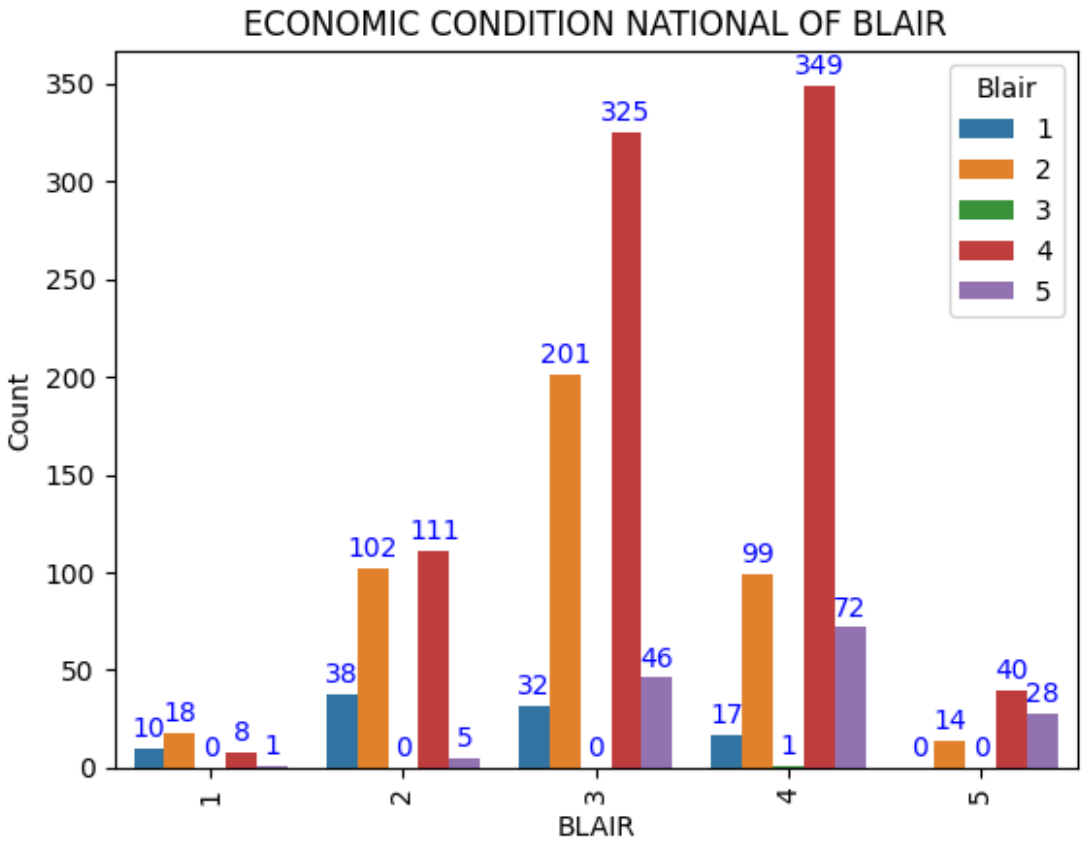
The value of 1 indicate that 38 have limited knowledge of the political parties hold on European integration but they don't have much knowledge.

The value of 2 indicate that 776 have Moderate knowledge the political parties hold on European integration

The value of 3 indicate that 249 have High knowledge of the political parties hold on European integration

BIVARIATE ANALYSIS:

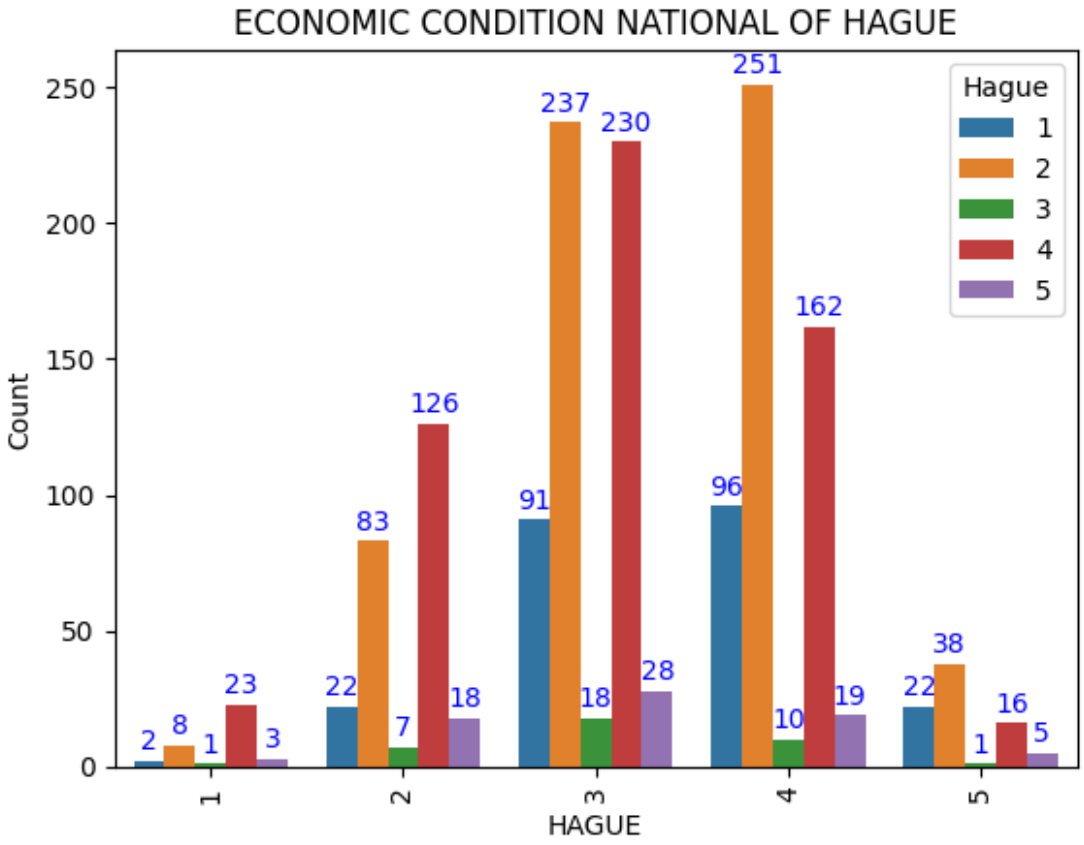
4) How does the economic condition (national) vary across different levels of the factor related to Blair?



WE CAN CLEARLY SAY ECONOMIC CONDITION NATIONAL OF BLAIR HAS GREATE VALUE IN 4 IT INDICATE THAT ECONOMIC CONDITION IS IN POSITIVE

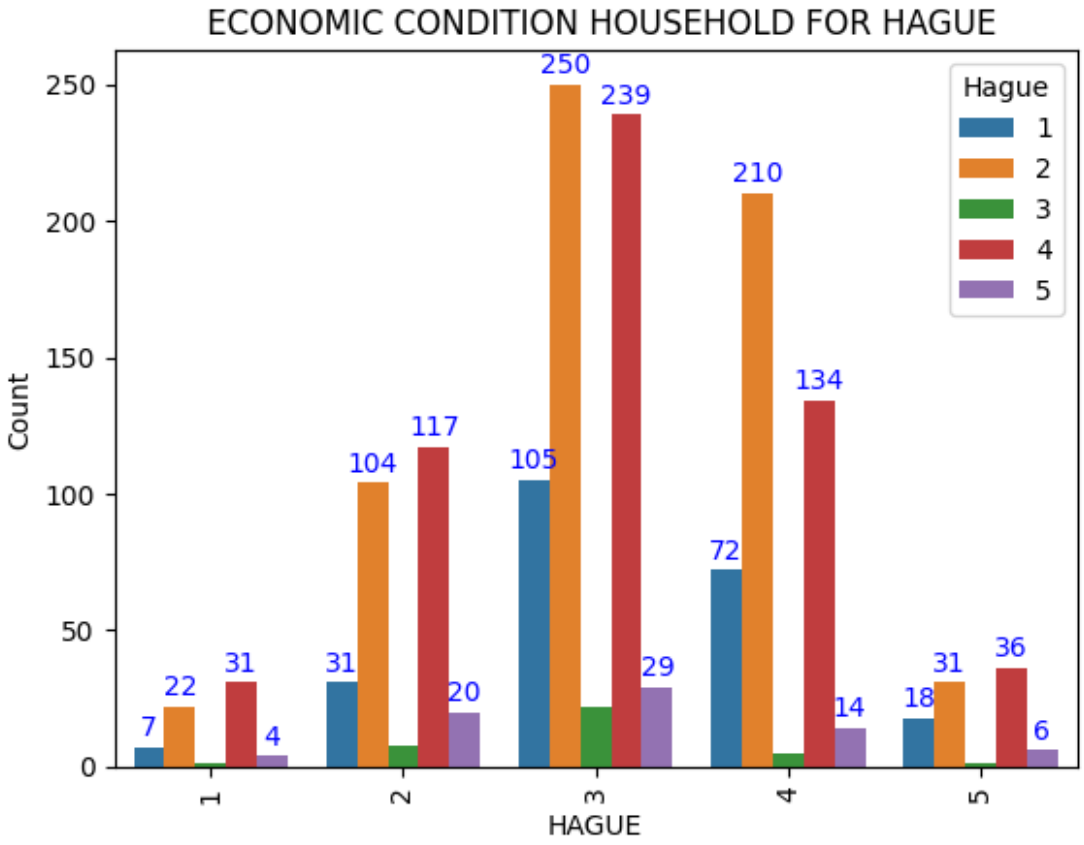
IN THE LEADERSHIP OF BLAIR

5) How does the economic condition (national) vary across different levels of the factor related to Hague?



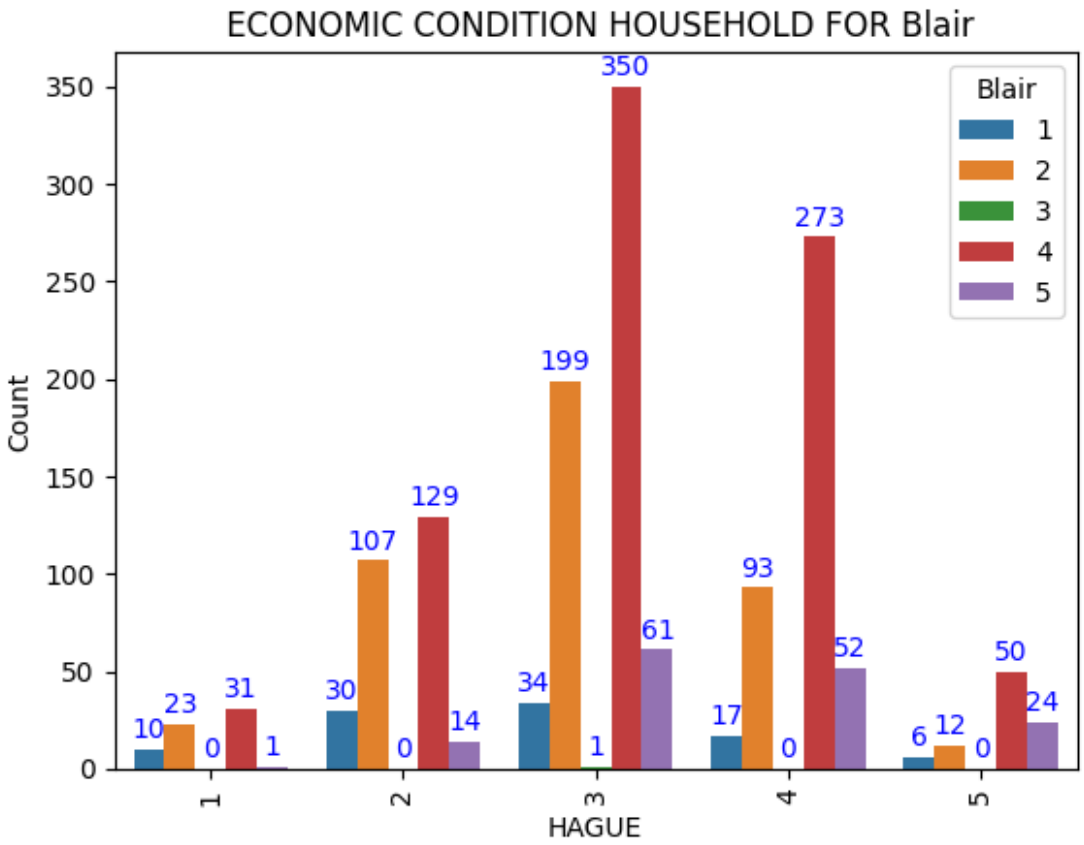
THE ECONOMIC CONDITION NATION IS MIXED WITH AVERAGE AND POSITIVE WITH THE LEADERSHIP UNDER HAGUE

6) Which level of the factor related to HAGUE appears to have the highest proportion of respondents with positive household economic perceptions?



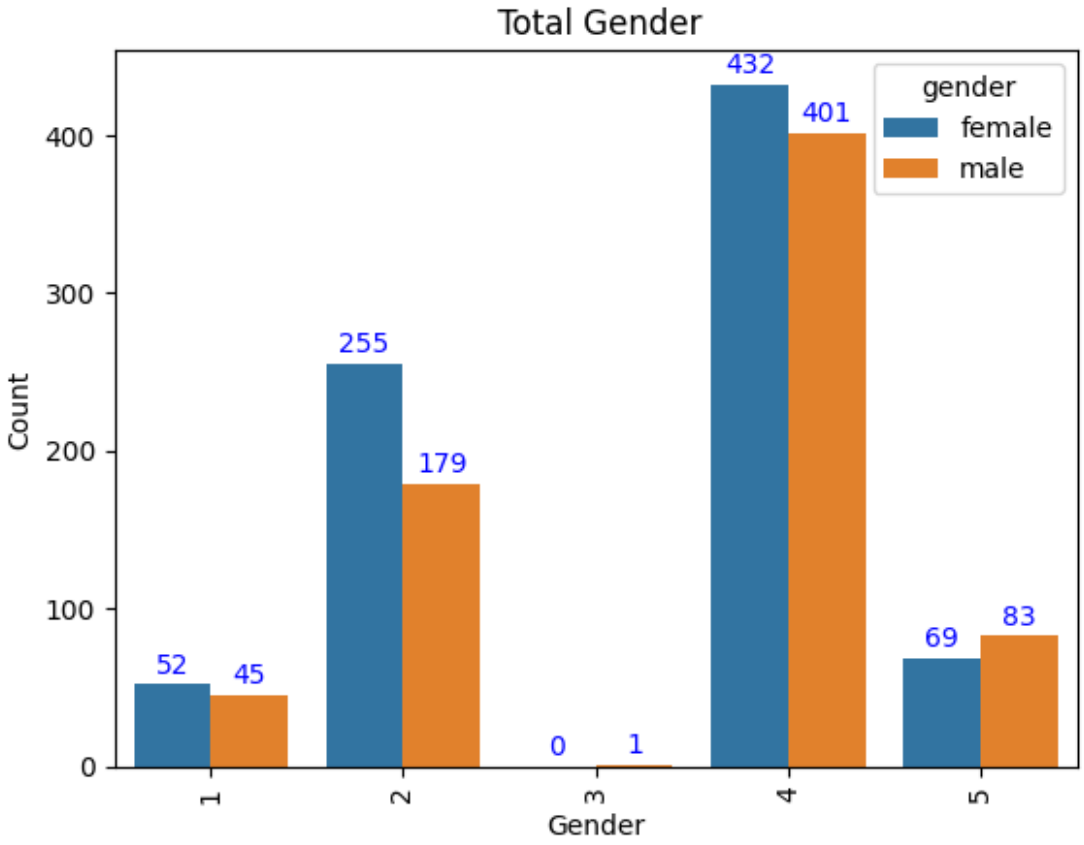
THE ECONOMIC CONDITION HOUSEHOLD ALSO HAVE THE MIXED OF AVERAGE AND POSITIVE RATING UNDER THE LEADERSHIP OF HAGUE

7) Which level of the factor related to BLAIR appears to have the highest proportion of respondents with positive household economic perceptions?



THE ECONOMIC CONDITION HOUSEHOLD ALSO HAVE THE MIXED OF AVERAGE AND POSITIVE BUT HAS THE HIGHEST RATING COMPARED WITH HAGUE , UNDER THE LEADERSHIP OF BLAIR ECONOMIC CONDITION IS VERY GOOD

8) How does the distribution of respondents across different levels of "Blair" vary by gender?



The value of 1 indicate that Female of 52 and Male of 45 says that negative or low leadership of blair labour leader.

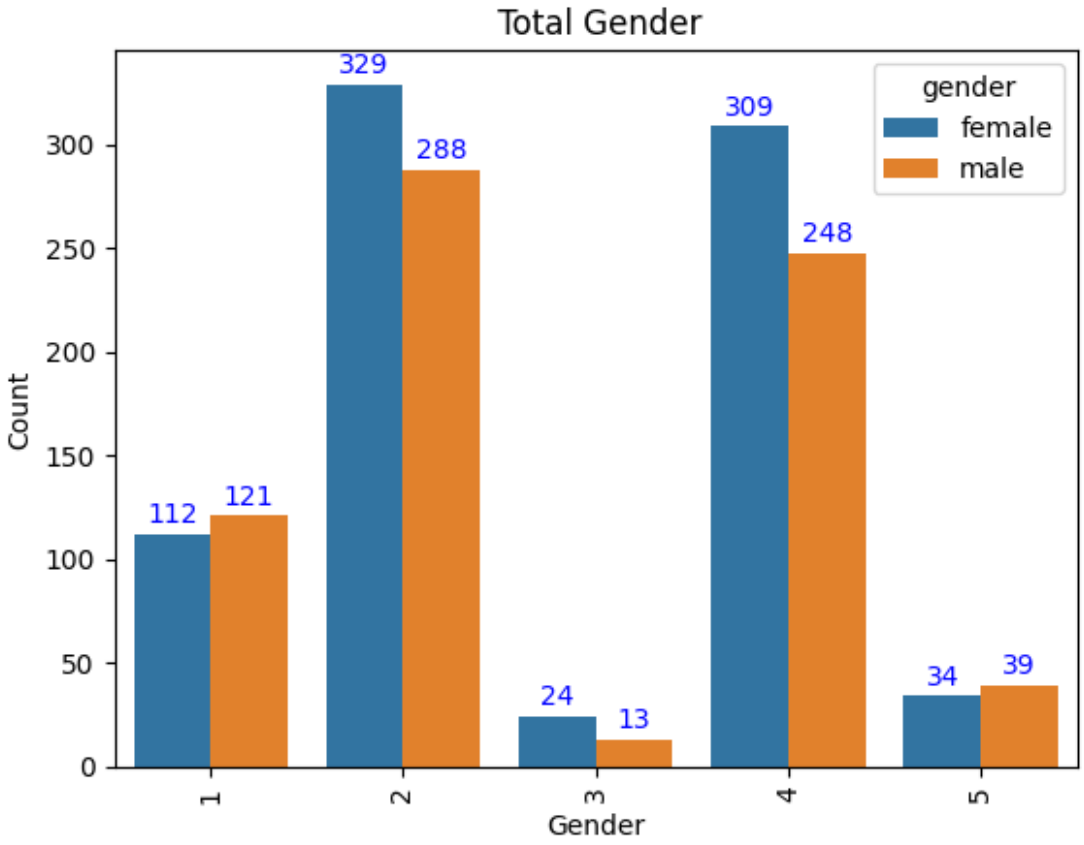
The value of 2 indicate that Female of 255 and Male of 179 says that somewhat negative leadership of blair labour leader

The value of 3 indicate that Female of 0 and Male of 1 says that Moderate leadership of blair labour leader its show nor strong or nor poor leadership.

The value of 4 indicate that Female of 432 and Male of 401 says that Some what Positive leadership of blair labour leader .Its indicate that the leader achievements.

The value of 5 indicate that Female of 69 and Male of 83 says that of Strong positive blair labour leader increases the employment, Income facility and increases quality of life.

9) How does the distribution of respondents across different levels of HAGUE vary by gender?



The value of 1 indicate that Female of 112 and Male of 121 says that negative or low leadership of Hague conservative leader.

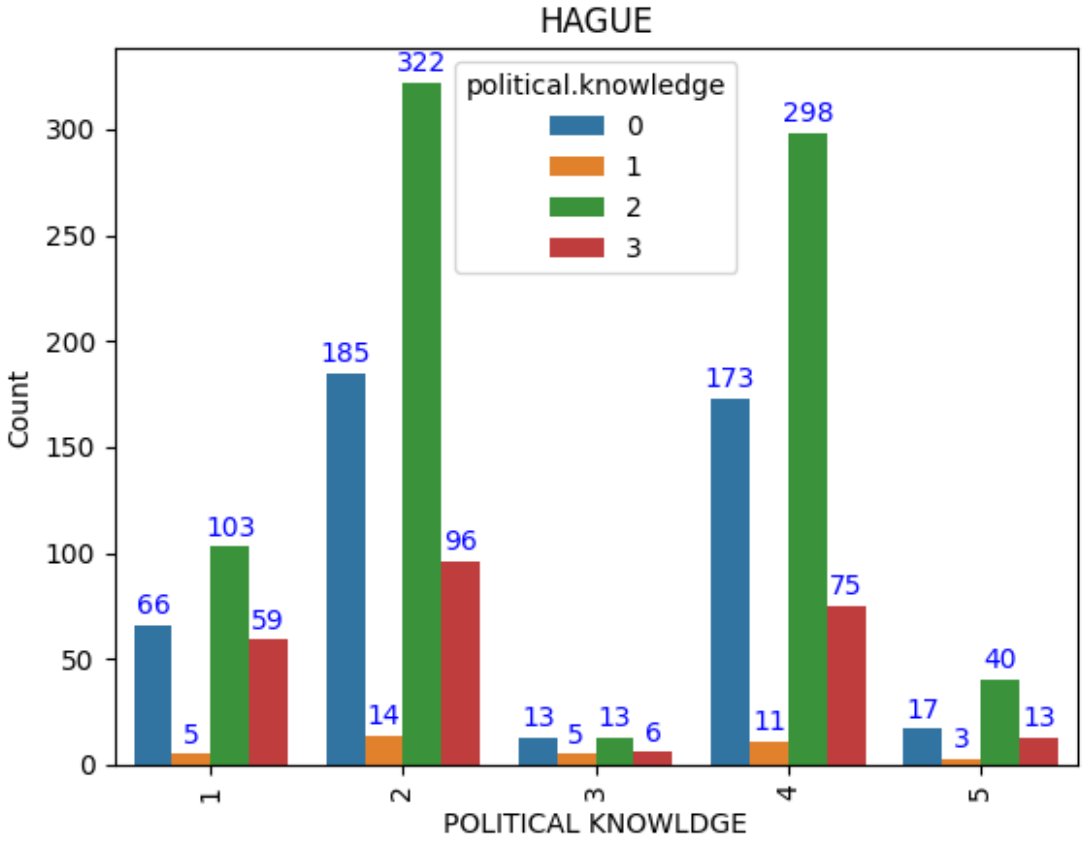
The value of 2 indicate that Female of 329 and Male of 288 says that somewhat negative leadership of Hague conservative leader.

The value of 3 indicate that Female of 24 and Male of 13 says that Moderate leadership of Hague conservative leader. its show nor strong or nor poor leadership.

The value of 4 indicate that Female of 309 and Male of 248 says that Some what Postive leadership of Hague conservative leader. Its indicate that the leader achievements.

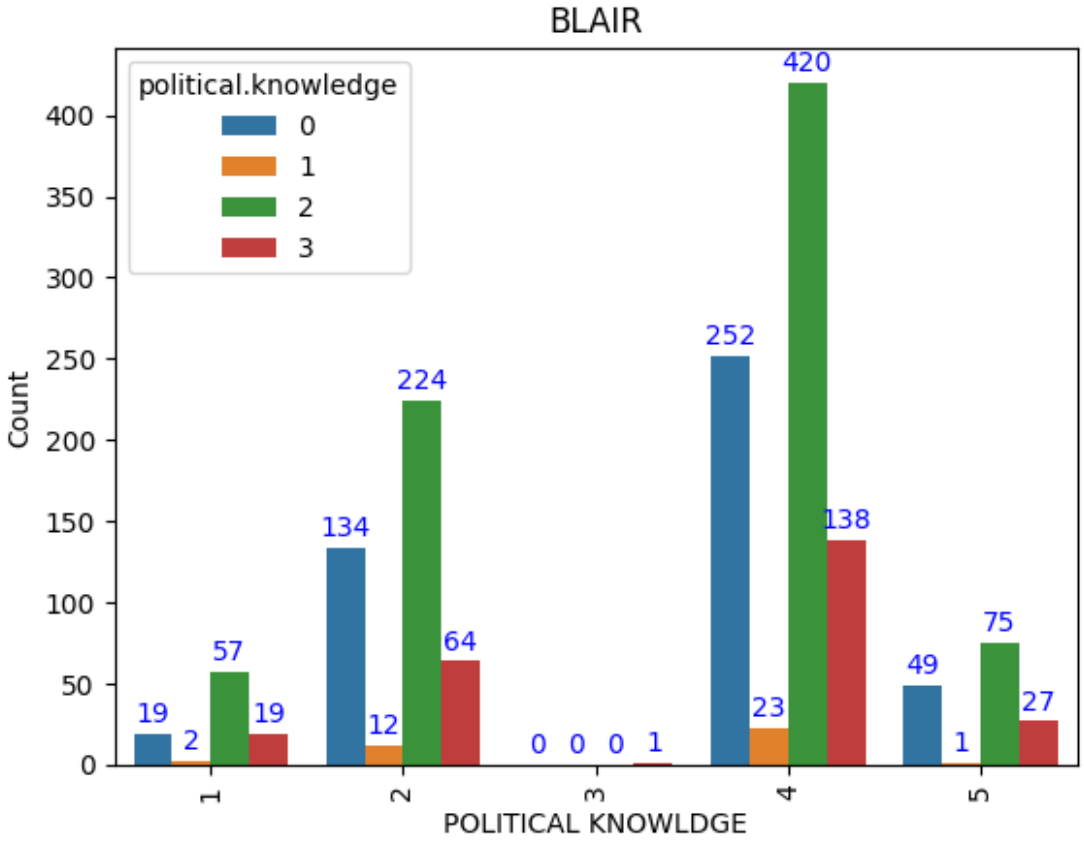
The value of 5 indicate that Female of 34 and Male of 39 says that of Strong positive Hague conservative leader the employment, income facility and increases quality of life.

10) How does the distribution of respondents' preferences for Hague vary across different levels of political knowledge?



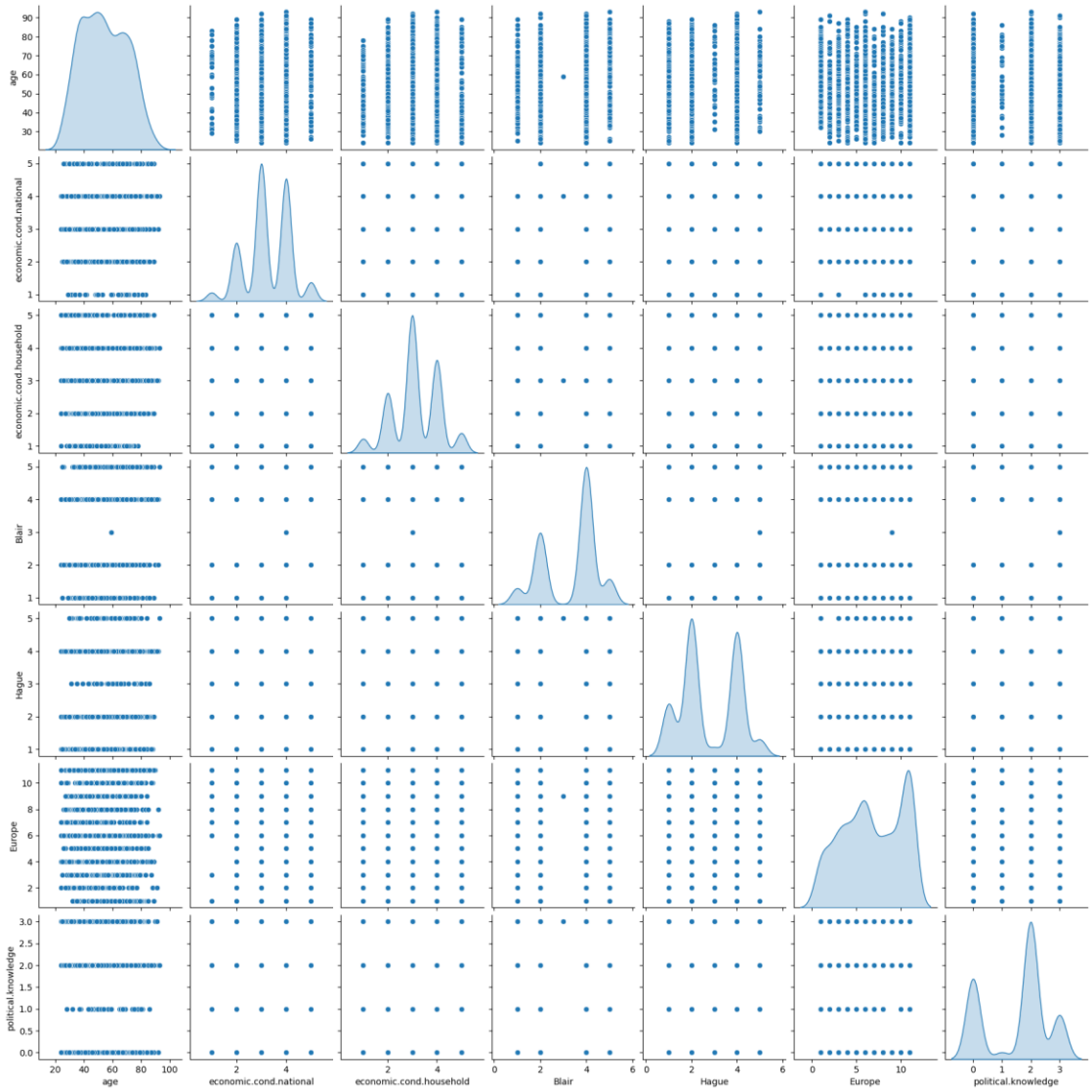
THE POLITICAL KNOWLEGHE OF HAGUE ITS LOW AND AVERAGE.

11) How does the distribution of respondents' preferences for Blair vary across different levels of political knowledge?



THE POLITICAL KNOWLEGHE OF BLAIR HAS VERY GOOD.

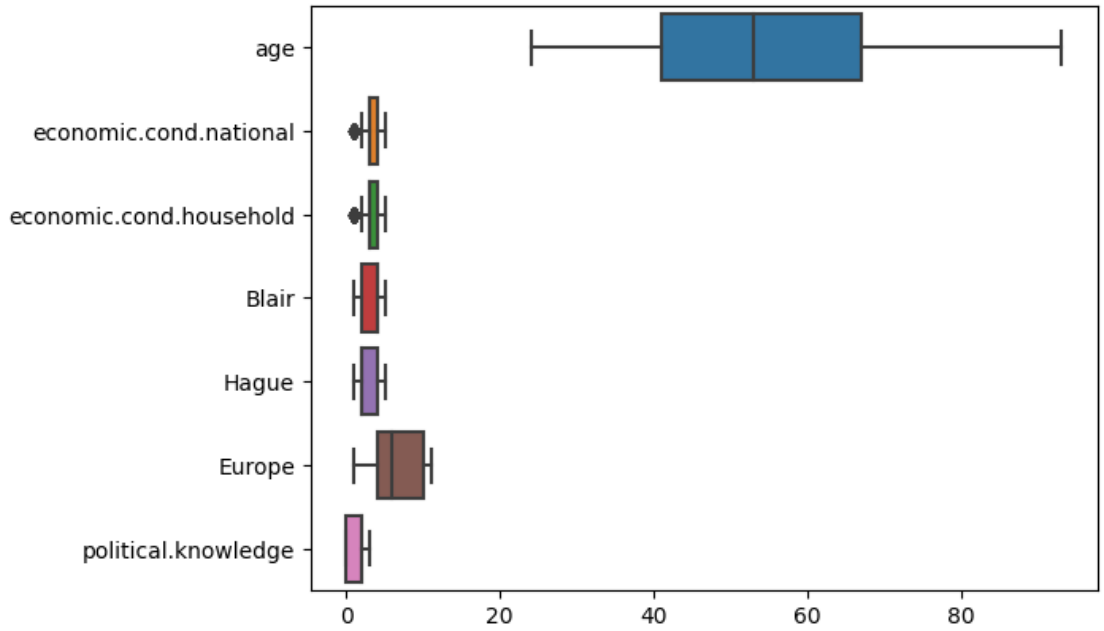
MULTIVARIATE



The pair plot shows that there is no relationship with each others.

We can national and house hold has the left tail which is mean by there can be outliers.

FINDING OUTLIERS:



There are Outlier in National and Household.

Will check the proportion Whether need to be removed are not

Outlier proportions:

	economic.cond.national	economic.cond.household
Proportion	0.02439	0.042848

Outlier proportion for economic.cond.national is acceptable.
Outlier proportion for economic.cond.household is acceptable.

we have used IQR method to check the proportion.

The national and household outlier won't impact the insights we can continue without removing the outlier

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.

Scaling is essential in machine learning to ensure that all input features have a consistent influence on the model, preventing attributes with larger scales from dominating the learning process. Proper scaling enhances convergence, accelerates training, and improves model accuracy, enabling reliable and unbiased predictions in real-world scenarios.

As many machine language learning models cannot work with string values we will encode the categorical variables and convert their data types to integer type. From the data type we have two categorical type variable so we can need to encode these 2 variable with the one hot encoding.

We are converting the categorical feature into binary for supply into algorithm

female 812

male 713

0 - FEMALE

1- MALE

We are converting the categorical feature into binary for supply into algorithm

Labour 1063

Conservative 462

0 - LABOUR

1- CONSERVATIVE

AFTER CREATING ONE HOT ENCODINGG FOR VOTE AND GENDER CHECKING FOR DATA TYPE.

```
<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   int64
 2   economic.cond.household 1517 non-null   int64
 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: int64(7), uint8(2)
memory usage: 97.8 KB
```

Age is the only column has to be scaled because the data is continue .

and the rest of the column are ordinal not necessary to be scaled.

	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	vote_Labour	gender_male
0	-0.716161	3	3	4	1	2	2	1	0
1	-1.162118	4	4	4	4	5	2	1	1
2	-1.225827	4	4	5	2	3	2	1	1
3	-1.926617	4	2	2	1	4	0	1	0
4	-0.843577	2	2	1	1	6	2	1	1

We can only age has been scaled and the rest of the feature are same.

Counter({1: 1057, 0: 460})

We can consider dependent variable as vote labour

X consist of independent variable

Y consist of Dependent variable

We can split the model into X and Y to supply data point into algorithm.

We are splitting the model into 70 : 30

AFTER SPLITING IN X ND Y SHAPE OF THE DEPENDENT AND INDPENDENT ARE:

The X Train set is independent : (1061, 8)

The X Test set is independent : (456, 8)

The y Test set is dependent: (456,)

The y Train set is dependent: (1061,)

CHECKING FOR DATA IS IMBALANCE OR BALANCED

```
Counter({1: 1057, 0: 460})
```

1 shows the 1057 and 0 shows the 460 its represent the data in not imbalanced data.

more than 70% in the 1 can be imbalanced data so we can consider this as the balanced data.

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 should be shared if any custom changes are made to the parameters while building the model. Calculate Train and Test Accuracies for each model. Comment on the validness of models (over fitting or under fitting)

Logistic Regression and Hyperparameters:

Logistic Regression predicts class probabilities in business tasks like churn prediction and fraud detection.

Hyperparameters like regularization strength (C), penalty type, solver, max iterations, and class weights control model performance and overfitting.

GridSearchCV:

GridSearchCV optimizes model hyperparameters by systematically testing combinations from a predefined grid. It enhances accuracy and generalization fo r business tasks, ensuring the model performs well on new data.

```
the accuracy score of test is : 0.8201754385964912
the accuracy score of train is : 0.8444863336475024
```

THE ACCURACY SCORE OF TEST DATA IN LOGISTC REGRESION IS 82%

THE ACCURACY SCORE OF TRAIN DATA IN LOGISTC REGRESION IS 84%

GRID SEARCH FOR LOGISTIC REGRESSION

```
Accuracy on Test Set: 0.8201754385964912
Accuracy on Train Set: 0.8416588124410933
```

THE BEST HYPERPARAMETERS FOR LOGISTIC REGRESSION IS:

C : 1

PENALTY : 11

SOLVER : liblinear

after using hyperparameters we check the accuracy score for test and train model

THE ACCURACY SCORE OF TEST DATA IN LOGISTC REGRESION IS 82%

THE ACCURACY SCORE OF TRAIN DATA IN LOGISTC REGRESION IS 84%

THE LOGISTIC REGRESSION PERFORM SAME ACCURACY WITH OR WITHOUT ANY HYPER PARAMETER .

THE ACCURACY SCORE OF TEST DATA IN LOGISTC REGRESION IS 82%

THE ACCURACY SCORE OF TRAIN DATA IN LOGISTC REGRESION IS 84%

The model does well in both and ther is no underfit or overfit data in the accuracy

LDA

LDA is a dimensionality reduction technique and classifier useful in business scenarios like customer segmentation. Its main hyperparameter is the number of components or classes used for projection. Adjusting this parameter influences the balance between dimension reduction and classification accuracy.

The Hyperparameter are n components , solver.

The Accuracy Score of Test is : 0.8245614035087719
The Accuracy Score of Train is : 0.8388312912346843

THE ACCURACY SCORE OF TEST DATA IN LINEAR DISCRIMINANT ANALYSIS IS 82%

THE ACCURACY SCORE OF TRAIN DATA IN LINEAR DISCRIMINANT ANALYSIS IS 84%

LDA GRID SEARCH

Accuracy on Test Set: 0.8245614035087719
the accuracy score of train is : 0.8388312912346843

The best Hyperparameter for lda are:

n_components : 1

solver : svd

After using the hyper parameter accuracy score are:

THE ACCURACY SCORE OF TEST DATA IN LINEAR DISCRIMINANT ANALYSIS IS 82%

THE ACCURACY SCORE OF TRAIN DATA IN LINEAR DISCRIMINANT ANALYSIS IS 84%

THE LINEAR DISCRIMINANT ANALYSIS PERFORM SAME ACCURACY WITH OR WITHOUT ANY HYPER PARAMETER .

THE ACCURACY SCORE OF TEST DATA IN LINEAR DISCRIMINANT ANALYSIS IS 82%

THE ACCURACY SCORE OF TRAIN DATA IN LINEAR DISCRIMINANT ANALYSISIS 84%

The model does well in both and there is no under fit or over fit data in the accuracy

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 should be shared if any custom changes are made to the parameters while building the model. Calculate Train and Test Accuracies for each model. Comment on the validness of models (over fitting or under fitting)

KNN MODEL

K-Nearest Neighbors is a versatile classification and regression algorithm with applications in customer profiling and recommendation systems. Key hyperparameters include:

1. **n_neighbors**: Number of neighbors to consider during prediction. Influences model sensitivity to local variations.

2. **weights**: Specifies how neighbors' contributions are weighted (uniform or distance-based).

3. algorithm: Algorithm used to compute neighbors (ball_tree, kd_tree, brute, or auto).

4. **p**: Power parameter for the Minkowski distance metric (1 for Manhattan, 2 for Euclidean).

5. **metric**: Distance metric to measure similarity between instances.

THE ACCURACY SCORE OF TEST DATA IN KNN IS 82%

THE ACCURACY SCORE OF TRAIN DATA IN KNN IS 100%

IF THE ACCURACY OF TRAIN DATA SCORE VARY THE RANGE MORE THAN 100% WE CAN CONSIDER IT AS AN OVERFIT MODEL.

HERE WE CAN SEE THAT TEST DATA IS 82 % BUT THE TRAIN DATA IS 100 % THESE MODEL IS CONSIDER AS THE OVERFIT MODEL.

GRID SEARCH KNN

Best hyperparameters: 'metric': 'manhattan', 'n_neighbors': 9, 'weights': 'uniform'
Test accuracy: 0.8092105263157895

The best hyperparameter for the model are:

metric : mahanttan

n_neighbors : 9

weighths: uniform

```
the accuracy score of test is : 0.8092105263157895
the accuracy score of train is : 0.8576814326107446
```

AFTER APPLYING THE GRID SEARCH CV FOR KNN MODE

THE ACCURACY SCORE TEST DATA FOR THE KNN MODEL IS 81%

THE ACCURACY SCORE TRAIN DATA FOR THE KNN MODEL IS 86%

FROM 100% TO IT BECOME 86% NOW THERE IS NO OVER FIT IN THE MODEL.

NAVIEE

The Naive Bayes Gaussian model is a probabilistic algorithm frequently used for text classification and spam filtering in business contexts. While it's relatively simple, it has a smoothing hyperparameter that can be important:

alpha: Smoothing parameter for handling zero probabilities. Adjusting alpha can impact model sensitivity to unseen data.

the accuracy score of test is : 0.8157894736842105
the accuracy score of train is : 0.8407163053722903

THE ACCURACY SCORE OF TEST DATA IN NAVIE IS 82%

THE ACCURACY SCORE OF TRAIN DATA IN NAVIE IS 84%

NAVIEE GRID SEARCH

Fitting 5 folds for each of 100 candidates, totaling 500 fits

Test accuracy: 0.8157894736842105
Train accuracy: 0.8416588124410933

The best Hyperparameter for Gaussian navie bayes are:

var_smoothing : 0.01

After using the hyper parameter accuracy score are:

THE ACCURACY SCORE OF TEST DATA IN Gaussian navie bayes IS 82%

THE ACCURACY SCORE OF TRAIN DATA IN Gaussian navie bayes IS 84%

THE Gaussian navie bayes PERFORM SAME ACCURACY WITH OR WITHOUT ANY HYPER PARAMETER .

THE ACCURACY SCORE OF TEST DATA IN Gaussian navie bayes IS 82%

THE ACCURACY SCORE OF TRAIN DATA IN Gaussian navie bayes 84%

The model does well in both and ther is no underfit or overfit data in the accuracy

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. 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

BAGGING CLASSIFIER

The Bagging Classifier is an ensemble technique that combines multiple base models to improve prediction accuracy and reduce variance. In a business context, it can be employed for customer churn prediction and fraud detection. Key hyper parameters to consider include:

1. **n_estimators:** Number of base estimators (models) to include in the ensemble. More estimators can improve stability but may increase computation time.

2. **max_samples**: Proportion of training data to use for each base estimator. Balances diversity and stability of the ensemble.

3. **max_features**: Number of features to consider for each base estimator. Controls feature diversity and model correlation.

4. **bootstrap**: Whether to use bootstrap sampling for creating subsets of data for each base estimator.

```
Bagging Accuracy test: 0.8092105263157895
Bagging Accuracy train : 0.88124410933082
```

THE ACCURACY SCORE OF TEST DATA IN BBAGGING IS 81%

THE ACCURACY SCORE OF TRAIN DATA IN BAGGING IS 88%

The model is not overfit or underfit but its near to be an overfit.

BAGGING CLASSIFIER FOR GRID SEARCH CV

The test accuracy is 0.8114035087719298
The train accuracy is 0.8718190386427899

The best Hyperparameter for BAGGING CLASSIFIER are:

BSE ESTIMATOR_MAX DEPTH: 5

N_ESTIMATORS : 10

After using the hyper parameter accuracy score are:

THE ACCURACY SCORE OF TEST DATA IN BAGGING CLASSIFIER IS 81%

THE ACCURACY SCORE OF TRAIN DATA IN BAGGING CLASSIFIER IS 87%

RANDOM FOREST

The Random Forest Classifier is a powerful ensemble method widely used in business applications such as customer segmentation and recommendation systems. It combines multiple decision trees for improved prediction accuracy. Important hyperparameters include:

1. **n_estimators:** Number of trees in the ensemble. More trees enhance performance but may increase computation time.

2. **max_depth:** Maximum depth of individual decision trees. Controls model complexity and potential overfitting.

3. **min_samples_split**: Minimum number of samples required to split an internal node. Influences tree depth and generalization.

4. **min_samples_leaf:** Minimum number of samples required in a leaf node. Affects tree depth and overfitting.

5. **max_features:** Number of features considered for the best split. Controls feature diversity and model correlation.

6. **bootstrap:** Whether to use bootstrap sampling for creating subsets of data for each tree.

the accuracy score of test is : 0.8026315789473685
the accuracy score of train is : 0.9990574929311969

THE ACCURACY SCORE OF TEST DATA IN RANDOM FOREST IS 81%

THE ACCURACY SCORE OF TRAIN DATA IN RANDOM FOREST IS 87%

RANDOM FOREST GRID SEARCH CV

THE BEST HYPERPARAMETER FOR RANDOM FOREST IS :

MAX_DEPTH : 10

N_ESTIMATORS : 5

Bagging Accuracy test: 0.8026315789473685
Bagging Accuracy train : 0.8755890669180019

THE ACCURACY SCORE FOR TEST DATA AFTER GRID SEARCH CV FOR RANDOM FOREST IS : 80%

THE ACCURACY SCORE FOR TRAIN DATA AFTER GRID SEARCH CV FOR RANDOM FOREST IS: 88%

AND THERE IS NO UNDERFIT OR OVRFIT.

GRAIDENT BOOSTING CLASSIFIER

The Gradient Boosting Classifier is a powerful ensemble method used for business tasks like customer churn prediction and fraud detection. It combines weak learners sequentially, building a strong predictive model. Key hyperparameters to consider include:

1. **n_estimators**: Number of boosting stages (weak learners) to include in the ensemble. Increasing estimators can lead to better performance but longer training time.

2.**learning_rate:** Controls the contribution of each weak learner to the ensemble. A lower learning rate may require more estimators but can improve robustness.

3. **max_depth**: Maximum depth of individual weak learners (typically shallow trees). Controls model complexity and potential overfitting.

4. **min_samples_split**:Minimum number of samples required to split an internal node in the weak learners.

5. **min_samples_leaf:** Minimum number of samples required in a leaf node in the weak learners.

6. **subsample**: Proportion of training data to use for each weak learner. Controls model diversity.

The accuracy of test : 0.8114035087719298
The accuracy of train: 0.8765315739868049

THE ACCURACY SCORE OF TEST IN GRAIDENT BOOSTING CLASSIFIER IS 81%

THE ACCURACY SCORE OF TRAIN IN GRAIDENT BOOSTING CLASSIFIER IS 88%

GRID SEARCH CV FOR GRADIENT BOOSTING CLASSFIER

```
The accuracy of test : 0.8114035087719298
The accuracy of train: 0.8765315739868049
```

THE BEST HYPER PARATMETER FOR GRADIENT BOOSTING CLASSFIER IS :

LEARNING RATE : 0.1

MAX DEPTH : 3

N ESTIMATE IS: 50

THE ACCURACY SCORE FOR GRID SERACH CV OF TEST IS :82%

THE ACCURACY SCORE FOR GRID SERACH CV OF TEST IS :88%

AND THE BOTH WITH AND WITHOUT HYPER PARAMTER GIVE THE SAME ACCURACY SCORE.

AND THERE IS NO OVERFIT AND UNDERFIT FOR THE MODEL.

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.

Confusion matrix for Test set of Logistic regression.

```
[ 88,  53],
[ 29, 286]
```

True Negative (TN) :88 The number of instances that were actually negative (belonging to the negative class) and were correctly predicted as negative by the classifier.

False Positive (FP): 53 The number of instances that were actually negative but were incorrectly predicted as positive by the classifier.

False Negative (FN): 29 The number of instances that were actually positive (belonging to the positive class) but were incorrectly predicted as negative by the classifier.

True Positive (TP): 286 The number of instances that were actually positive and were correctly predicted as positive by the classifier.

Confusion matrix of Logistic regression For Train set :

```
[214, 105],
[ 63, 679]
```

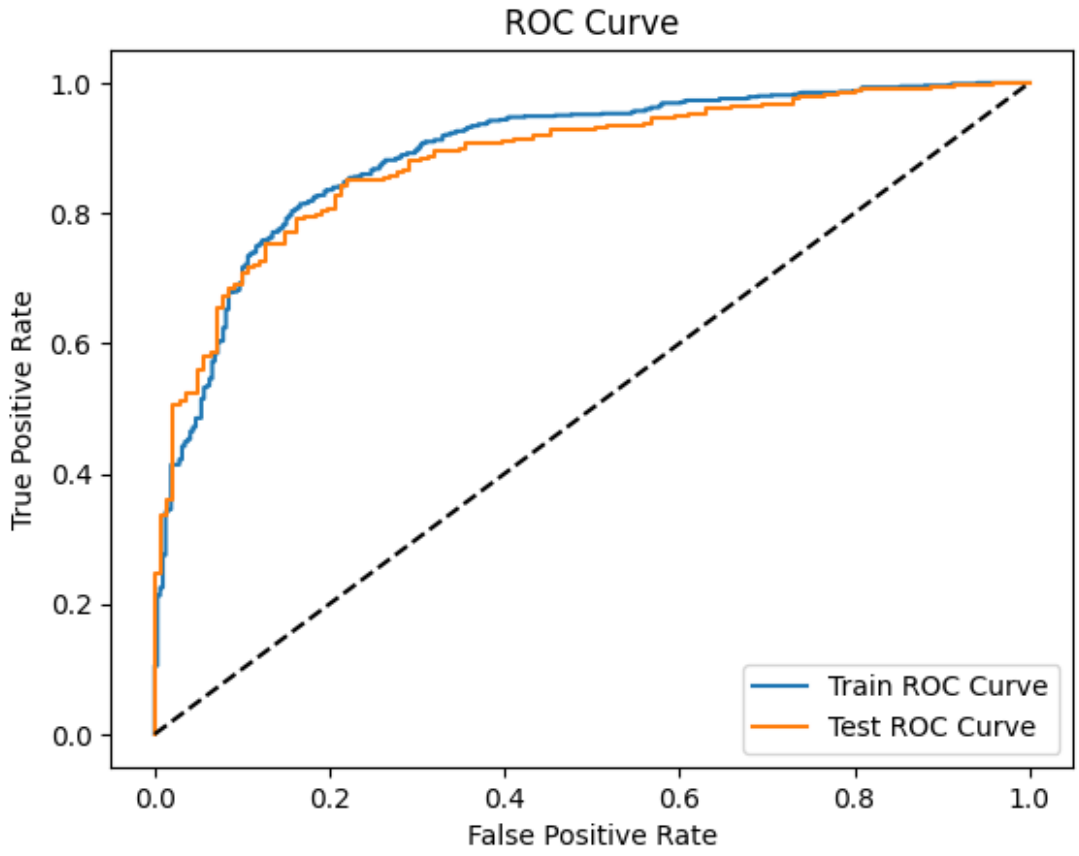
Classification report for Test set of Logistic regression:

	precision	recall	f1-score	support
0	0.75	0.62	0.68	141
1	0.84	0.91	0.87	315
accuracy			0.82	456
macro avg	0.80	0.77	0.78	456
weighted avg	0.82	0.82	0.82	456

Classification report for Train set of logistic regression :

	precision	recall	f1-score	support
0	0.77	0.67	0.72	319
1	0.87	0.92	0.89	742
accuracy			0.84	1061
macro avg	0.82	0.79	0.80	1061
weighted avg	0.84	0.84	0.84	1061

LOGISTIC REGRESSION FOR TRAIN AND TEST ROC CURVE :




```
LOGISTIC REGRESSION Train AUC is: 0.89230369500376
LOGISTIC REGRESSION Test AUC is : 0.8827423167848699
```

Train AUC is: 0.89230369500376

This means that when the Logistic Regression model is evaluated on the training da ta, the AUC score is approximately 0.892. A higher AUC indicates that the model is better at differentiating between the two classes, and its predictions have a higher probability of being correct

Test AUC is: 0.8827423167848699

This means that when the Logistic Regression model is evaluated on the test data (unseen data), the AUC score is approximately 0.883. A similar interpretation applies here – a higher AUC suggests better model performance in terms of classification accuracy

LDA (LINEAR DISCRIMINANT ANALYSIS).

LDA CONFUSION MATRIX FOR TEST SET :

```
[91,  50],
[ 30, 285]
```

Interpretation of the values in the given confusion matrix:

True Negative (TN) : 91 There are 91 instances that were actually negative (belonging to the negative class) and were correctly predicted as negative by the classifier.

False Positive (FP): 50 There are 50 instances that were actually negative but were incorrectly predicted as positive by the classifier.

False Negative (FN): 30 There are 30 instances that were actually positive (belonging to the positive class) but were incorrectly predicted as negative by the classifier.

True Positive (TP) : 285 There are 285 instances that were actually positive and were correctly predicted as positive by the classifier.

In summary, the confusion matrix provides insights into how well the binary classification model is performing. It shows the counts of instances that were classified correctly and incorrectly for both the positive and negative classes.

Confusion Matrix for train dataset of lda:

```
[219, 100],
[ 71, 671]
```

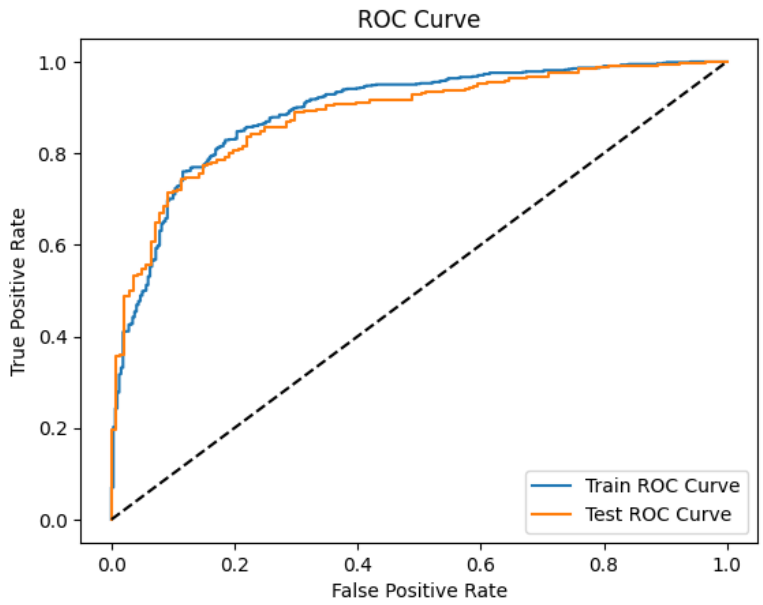
Classification report for Test set of Lda:

	precision	recall	f1-score	support
0	0.75	0.65	0.69	141
1	0.85	0.90	0.88	315
accuracy			0.82	456
macro avg	0.80	0.78	0.79	456
weighted avg	0.82	0.82	0.82	456

Classification report for Train set of Lda:

	precision	recall	f1-score	support
0	0.76	0.69	0.72	319
1	0.87	0.90	0.89	742
accuracy			0.84	1061
macro avg	0.81	0.80	0.80	1061
weighted avg	0.84	0.84	0.84	1061

ROC CURVE FOR TRAIN AND TEST OF LDA:




```
lda Train AUC is: 0.8919868355457163
lda Test AUC is : 0.8820443543847799
```

The LDA model exhibits consistent performance on both training and test data with Train AUC of approximately 0.892 and Test AUC of around 0.882. This suggests that the model effectively discriminates between classes and generalizes well to new data, demonstrating its reliability for classification tasks.

Navie baye Gaussian:

Confusion matrix of Test set Gaussian:

```
[ 91,  50],
[ 34, 281]
```

Interpretation of the values in the given confusion matrix:

True Negative (TN): 91 There are 91 instances that were actually negative (belonging to the negative class) and were correctly predicted as negative by the classifier.

False Positive (FP): 50 There are 50 instances that were actually negative but were incorrectly predicted as positive by the classifier.

False Negative (FN): 34 There are 34 instances that were actually positive (belonging to the positive class) but were incorrectly predicted as negative by the classifier.

True Positive (TP): 281 There are 281 instances that were actually positive and were correctly predicted as positive by the classifier.

In summary, the confusion matrix provides insights into the model's performance, indicating the accuracy of class predictions and highlighting instances that were classified correctly and incorrectly for both the positive and negative classes.

Confusion matrix for Train set of naviee

[229, 90],
[79, 663]

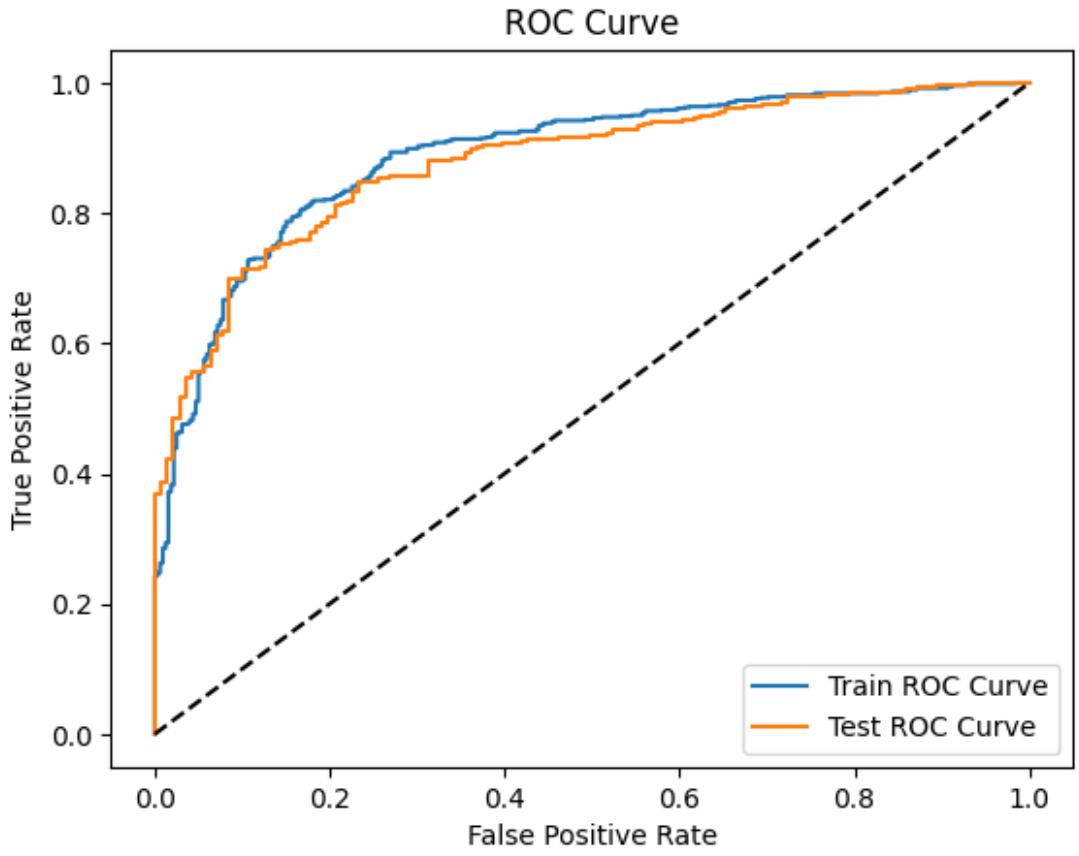
Classification report for Test set of Gaussian:

	precision	recall	f1-score	support
0	0.73	0.65	0.68	141
1	0.85	0.89	0.87	315
accuracy			0.82	456
macro avg	0.79	0.77	0.78	456
weighted avg	0.81	0.82	0.81	456

Classification report for Train set of Gaussian:

	precision	recall	f1-score	support
0	0.74	0.72	0.73	319
1	0.88	0.89	0.89	742
accuracy			0.84	1061
macro avg	0.81	0.81	0.81	1061
weighted avg	0.84	0.84	0.84	1061

ROC CURVE OF TEST AND TRAIN DATA SET OF GAUSSIAN:



NAVIE Train AUC is: 0.8884126608589848
NAVIE Test AUC is : 0.8786220871327254

The Naive Bayes model demonstrates consistent AUC performance across training and test datasets, with a Train AUC of approximately 0.888 and Test AUC of around 0.879. This suggests the model's ability to effectively discriminate between classes and maintain strong generalization, indicating its reliability for classification tasks.

KNN:

CONFUSION MATRIX KNN OF TEST SET:

```
[ 91,  50],
[ 37, 278]
```

Interpretation of the values in the given confusion matrix:

True Negative (TN): 91 There are 91 instances that were actually negative (belonging to the negative class) and were correctly predicted as negative by the classifier.

False Positive (FP): 50 There are 50 instances that were actually negative but were incorrectly predicted as positive by the classifier.

False Negative (FN): 37 There are 37 instances that were actually positive (belonging to the positive class) but were incorrectly predicted as negative by the classifier.

True Positive (TP): 278 There are 278 instances that were actually positive and were correctly predicted as positive by the classifier.

In summary, the confusion matrix provides insights into the model's performance, indicating the accuracy of class predictions and highlighting instances that were classified correctly and incorrectly for both the positive and negative classes.

CONFUSION MATRIX OF TRAIN SET KNN

```
[236,  83],
[ 68, 674]
```

Classification report for Test set of knn:

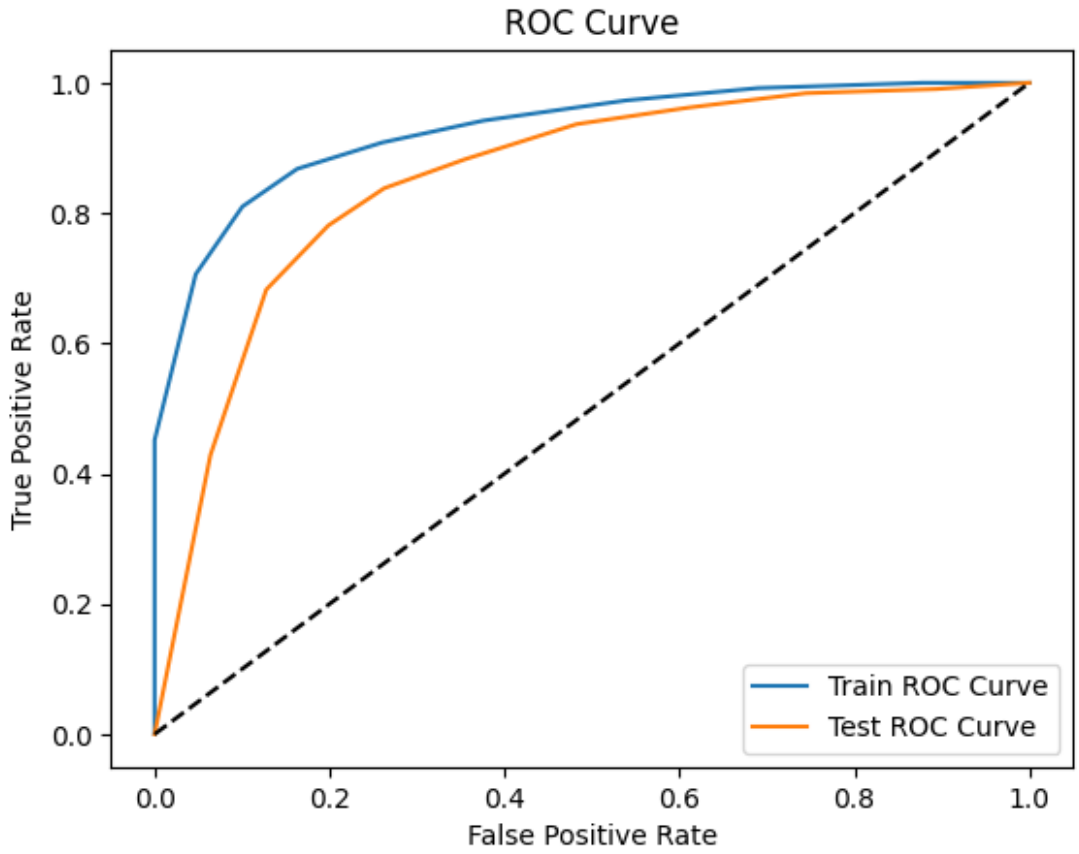
	precision	recall	f1-score	support
0	0.71	0.65	0.68	141
1	0.85	0.88	0.86	315
accuracy			0.81	456
macro avg	0.78	0.76	0.77	456
weighted avg	0.81	0.81	0.81	456

Classification report for Train set of knn:

	precision	recall	f1-score	support
0	0.78	0.74	0.76	319
1	0.89	0.91	0.90	742
accuracy			0.86	1061
macro avg	0.83	0.82	0.83	1061

```
weighted avg       0.86      0.86      0.86      1061
```

ROC CURVE OF TEST AND TRAIN DATA SET OF KNN:




```
KNN Train AUC is: 0.9273441262706064
KNN Test AUC is : 0.8533490937746258
```

The KNN model exhibits a higher AUC score of approximately 0.927 on the training data, indicating strong discrimination ability. However, its Test AUC of around 0.853 suggests some potential performance drop on unseen data, implying a need for further evaluatio n or tuning to enhance generalization.

BAGGING:

CONFUSION MATRIX BAGGING OF TEST SET

```
[ 87,  54],
[ 32, 283]
```

Interpretation of the values in the given confusion matrix:

True Negative (TN): 87 There are 87 instances that were actually negative (belonging to the negative class) and were correctly predicted as negative by the classifier.

False Positive (FP): 54 There are 54 instances that were actually negative but were incorrectly predicted as positive by the classifier.

False Negative (FN): 32 There are 32 instances that were actually positive (belonging to the positive class) but were incorrectly predicted as negative by the classifier.

True Positive (TP): 283 There are 283 instances that were actually positive and were correctly predicted as positive by the classifier.

In summary, the confusion matrix provides insights into the model's performance, indicating the accuracy of class predictions and highlighting instances that were classified correctly and incorrectly for both the positive and negative classes.

CONFUSION MATRIX BAGGING OF TRAIN SET

```
[232,  87],
[ 49, 693]
```

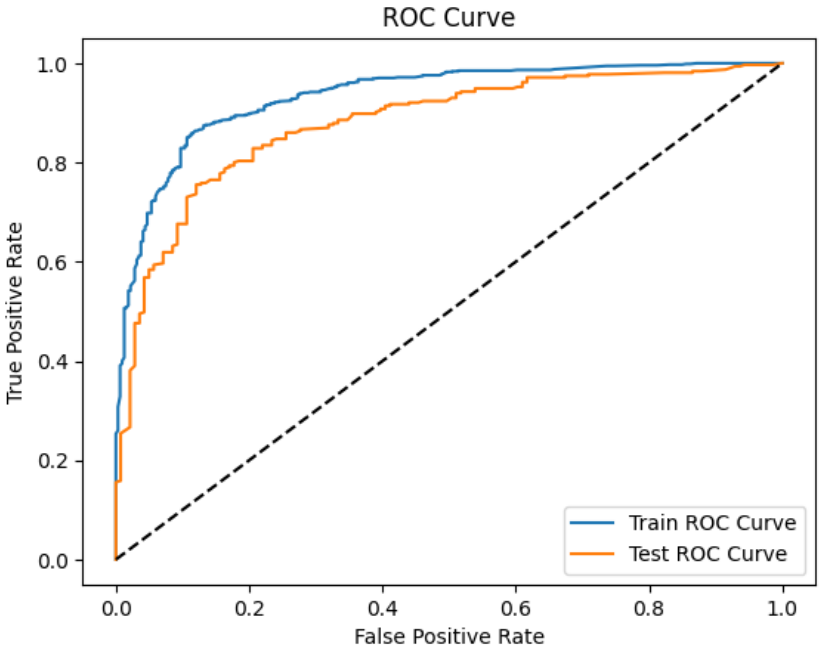
Classification report for Test set of BAGGING:

	precision	recall	f1-score	support
0	0.80	0.66	0.72	147
1	0.85	0.92	0.89	311
accuracy			0.84	458
macro avg	0.83	0.79	0.80	458
weighted avg	0.84	0.84	0.83	458

Classification report for TRAIN set of BAGGING:

	precision	recall	f1-score	support
0	0.83	0.73	0.77	319
1	0.89	0.93	0.91	742
accuracy			0.87	1061
macro avg	0.86	0.83	0.84	1061
weighted avg	0.87	0.87	0.87	1061

ROC CURVE OF TEST AND TRAIN DATA SET OF Bagging:




```
Bagging Train AUC is: 0.9343234839331132
Bagging Test AUC is : 0.8788697512101769
```

The Bagging ensemble model shows impressive Train AUC performance of about 0.934, highlighting strong discrimination on training data. While its Test AUC of approximately 0.879 suggests good generalization, further evaluation may be warranted to ensure consistent model effectiveness on new, unseen data.

RANDOM FOREST

CONFUSION RANDOM FOREST OF TEST SET

```
[ 79,  62],
[ 28, 287]
```

Interpretation of the values in the given confusion matrix:

True Negative (TN): 79 There are 79 instances that were actually negative (belonging to the negative class) and were correctly predicted as negative by the classifier.

False Positive (FP): 62 There are 62 instances that were actually negative but were incorrectly predicted as positive by the classifier.

False Negative (FN): 28 There are 28 instances that were actually positive (belonging to the positive class) but were incorrectly predicted as negative by the classifier.

True Positive (TP): 287 There are 287 instances that were actually positive and were correctly predicted as positive by the classifier.

In summary, the confusion matrix provides insights into the model's performance, indicating the accuracy of class predictions and highlighting instances that were classified correctly and incorrectly for both the positive and negative classes **.**

CONFUSION RANDOM FOREST OF TRAIN SET :

[228, 91],

[41, 701]

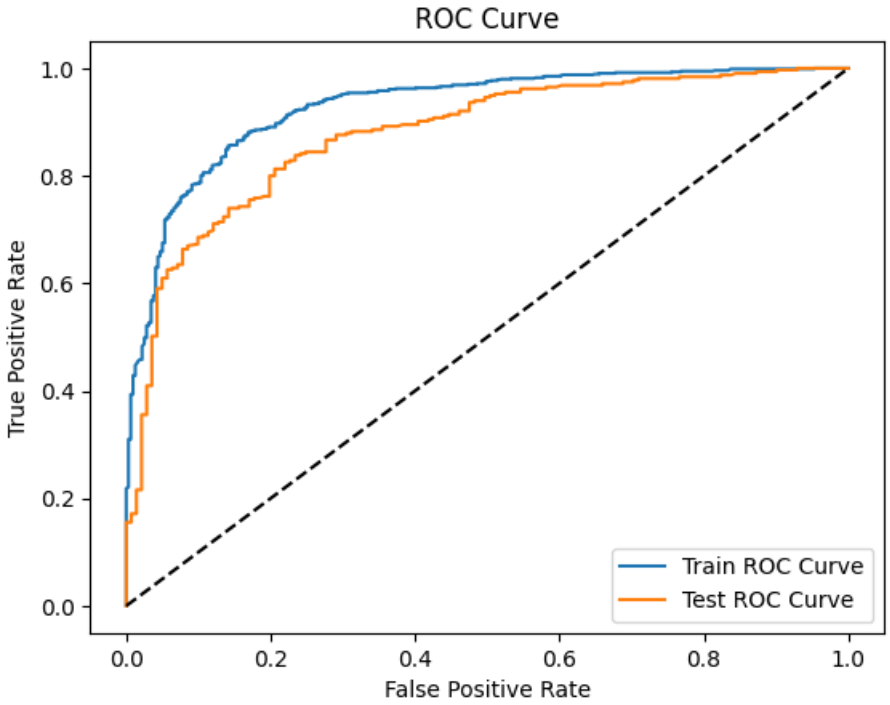
Classification report for Test set of random forest :

	precision	recall	f1-score	support
0	0.74	0.56	0.64	141
1	0.82	0.91	0.86	315
accuracy			0.80	456
macro avg	0.78	0.74	0.75	456
weighted avg	0.80	0.80	0.79	456

Classification report for TRAIN set of random forest :

	precision	recall	f1-score	support
0	0.85	0.71	0.78	319
1	0.89	0.94	0.91	742
accuracy			0.88	1061
macro avg	0.87	0.83	0.84	1061
weighted avg	0.87	0.88	0.87	1061

ROC CURVE OF TEST AND TRAIN DATA SET OF Random Forest :




```
Random Forest Train AUC is: 0.9281151509518459
Random Forest Test AUC is : 0.8769222109647641
```

The Random Forest model demonstrates robust Train AUC performance of around 0.928, showcasing effective class discrimination on the training data. Its Test AUC of approximately 0.877 indicates promising generalization ability, suggesting reliable performance on unseen data as well.

Gradient Boosting:

CONFUSION Gradient Boosting OF TRAIN SET :

```
[ 90,  51],
[ 35, 280]
```

Interpretation of the values in the given confusion matrix:

True Negative (TN): 90 There are 90 instances that were actually negative (belonging to the negative class) and were correctly predicted as negative by the classifier.

False Positive (FP): 51 There are 51 instances that were actually negative but were incorrectly predicted as positive by the classifier.

False Negative (FN): 35 There are 35 instances that were actually positive (belonging to the positive class) but were incorrectly predicted as negative by the classifier.

True Positive (TP): 280 There are 280 instances that were actually positive and were correctly predicted as positive by the classifier.

In summary, the confusion matrix provides insights into the model's perfo rmance, indicating the accuracy of class predictions and highlighting instances that were classified correctly and incorrectly for both the positive and negative classes **.**

CONFUSION Gradient Boosting OF TRAIN SET :

```
[238,  81],
[ 50, 692]
```

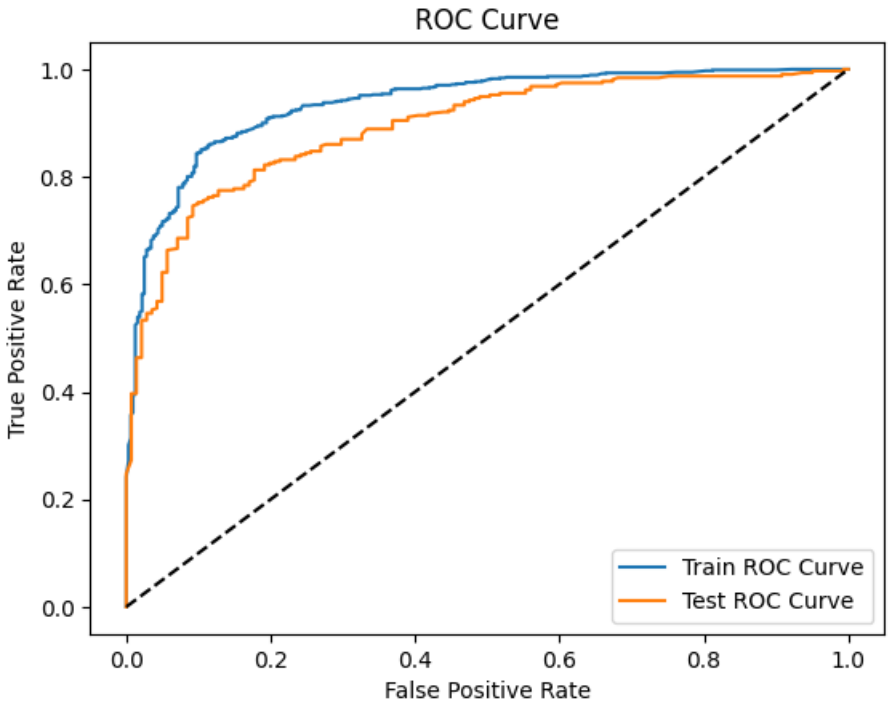
Classification report for Test set of Gradient Boosting:

	precision	recall	f1-score	support
0	0.72	0.64	0.68	141
1	0.85	0.89	0.87	315
accuracy			0.81	456
macro avg	0.78	0.76	0.77	456
weighted avg	0.81	0.81	0.81	456

Classification report for Train set of Gradient Boosting:

	precision	recall	f1-score	support
0	0.83	0.75	0.78	319
1	0.90	0.93	0.91	742
accuracy			0.88	1061
macro avg	0.86	0.84	0.85	1061
weighted avg	0.87	0.88	0.87	1061

ROC CURVE OF TEST AND TRAIN DATA SET OF Gradient Boosting:



Gradient Boosting Train AUC is: 0.9362246406813747
Gradient Boosting Test AUC is : 0.893121693121693

The Gradient Boosting model achieves strong Train AUC of about 0.936, showcasing effective class separation on training data. Its Test AUC of approximately 0.893 demonstrates excellent generalization, indicating high predictive performance on new, unseen data.

Final model we can consider as:

Based on the evaluation metrics you've provided for each model, it appears that Logistic Regression the top-performing models for dataset.

Logistic Regression:

Test Accuracy: 0.82

Test F1-Score (weighted average): 0.82

Pros: Logistic Regression shows balanced performance in terms of accuracy and F1-score. It has relatively high precision and recall values for both classes. It's a simple and interpretable model.

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.

Insights:

1. The logistic regression model indicates higher precision, recall, and F1-score for Party 1 (Conservative), implying stronger predicted voter support.

2. The model's balanced metrics suggest stable performance in predicting Party 1, indicating reliability in forecasting voter preferences.

Recommendations:

1. Allocate campaign resources strategically to engage Party 1 supporters based on the model's predictions for optimized voter outreach.

2. Craft targeted campaign messages that resonate with the Party 1 demographic, leveraging the model's insights for impactful communication.

Problem 2:

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

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

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

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

The Number of **Characters** For 1941- Roosevelt: 7571

The Number of **words** for 1941- Roosevelt: 1536

The Number of **Sentences** For 1941- Roosevelt : 68

- **President Richard Nixon in 1973**

The Number of **Characters** For 1973-Nixon.txt: 9991

The Number of **words** for 1973-Nixon.txt : 2028

The Number of **Sentences** For 1973-Nixon.txt : 69

- **President John F. Kennedy in 1961**

The Number of **Characters** For 1961-Kennedy.txt: 7618

The Number of **words** for 1961-Kennedy.txt : 1546

The Number of **Sentences** For 1961-Kennedy.txt : 52

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. & 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)

Installing Stopwords its is used to find the punctuation, special character and words which below mentioned.

```
 ['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you',
"you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself',
'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', 'an',
'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of',
'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through',
'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down',
'in', 'out', 'on', 'off', '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', 'aren', "aren't", 'couldn', "couldn't", 'didn',
"didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven',
"haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't",
'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn',
"wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't", '!', '"',
'#', '$', '%', '&', "'", '(', ')', '*', '+', ',', '-', '.', '/', ':', ';',
'<', '=', '>', '?', '@', '[', '\\', ']', '^', '_', '`', '{', '|', '}', '~']
```

President Franklin D. Roosevelt in 1941

After removed stop word in President Franklin D.Roosevelt in 1941

```
‘national', 'day', 'inauguration', 'since', '1789', 'people', 'renewed',
'sense', 'dedication', 'united', 'states', 'washington', 'day', 'task',
'people', 'create', 'weld', 'together', 'nation', 'lincoln', 'day', 'task',
'people', 'preserve', 'nation', 'disruption', 'within', 'day', 'task',
'people', 'save', 'nation', 'institutions', 'disruption', 'without', 'us',
'come', 'time', 'midst', 'swift', 'happenings', 'pause', 'moment', 'take',
'stock', 'recall', 'place', 'history', 'rediscover', 'may', 'risk', 'real',
'peril', 'inaction', 'lives', 'nations', 'determined', 'count', 'years',
'lifetime', 'human', 'spirit', 'life', 'man', 'three-score', 'years', 'ten',
'little', 'little', 'less', 'life', 'nation', 'fullness', 'measure', 'live',
'men', 'doubt', 'men', 'believe', 'democracy', 'form', 'government', 'frame',
'life', 'limited', 'measured', 'kind', 'mystical', 'artificial', 'fate',
'unexplained', 'reason', 'tyranny', 'slavery', 'become', 'surging', 'wave',
'future', 'freedom', 'ebbing', 'tide', 'americans', 'know', 'true', 'eight',
'years', 'ago', 'life', 'republic', 'seemed', 'frozen', 'fatalistic',
'terror', 'proved', 'true', 'midst', 'shock', 'acted', 'acted', 'quickly',
'boldly', 'decisively', 'later', 'years', 'living', 'years', 'fruitful',
'years', 'people', 'democracy', 'brought', 'us', 'greater', 'security',
```

'hope', 'better', 'understanding', 'life', 'ideals', 'measured', 'material',
'things', 'vital', 'present', 'future', 'experience', 'democracy',
'successfully', 'survived', 'crisis', 'home', 'put', 'away', 'many', 'evil',
'things', 'built', 'new', 'structures', 'enduring', 'lines', 'maintained',
'fact', 'democracy', 'action', 'taken', 'within', 'three-way', 'framework',
'constitution', 'united', 'states', 'coordinate', 'branches', 'government',
'continue', 'freely', 'function', 'bill', 'rights', 'remains', 'inviolate',
'freedom', 'elections', 'wholly', 'maintained', 'prophets', 'downfall',
'american', 'democracy', 'seen', 'dire', 'predictions', 'come', 'naught',
'democracy', 'dying', 'know', 'seen', 'revive', 'grow', 'know', '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', 'men', 'enlightened',
'know', 'democracy', 'alone', 'constructed', 'unlimited', 'civilization',
'capable', 'infinite', 'progress', 'improvement', 'human', 'life', 'know',
'look', 'surface', 'sense', 'still', 'spreading', 'every', 'continent',
'humane', 'advanced', 'end', 'unconquerable', 'forms', 'human', 'society',
'nation', 'like', 'person', 'body', 'body', 'must', 'fed', 'clothed',
'housed', 'invigorated', 'rested', 'manner', 'measures', 'objectives',
'time', 'nation', 'like', 'person', 'mind', 'mind', 'must', 'kept',
'informed', 'alert', 'must', 'know', 'understands', 'hopes', 'needs',
'neighbors', 'nations', 'live', 'within', 'narrowing', 'circle', 'world',
'nation', 'like', 'person', 'something', 'deeper', 'something', 'permanent',
'something', 'larger', 'sum', 'parts', 'something', 'matters', 'future',
'calls', 'forth', 'sacred', 'guarding', 'present', 'thing', 'find',
'difficult', 'even', 'impossible', 'hit', 'upon', 'single', 'simple', 'word',
'yet', 'understand', 'spirit', 'faith', 'america', 'product', 'centuries',
'born', 'multitudes', 'came', 'many', 'lands', 'high', 'degree', 'mostly',
'plain', 'people', 'sought', 'early', 'late', 'find', 'freedom', 'freely',
'democratic', 'aspiration', 'mere', 'recent', 'phase', 'human', 'history',
'human', 'history', 'permeated', 'ancient', 'life', 'early', 'peoples',
'blazed', 'anew', 'middle', 'ages', 'written', 'magna', 'charta', 'americas',
'impact', 'irresistible', 'america', 'new', 'world', 'tongues', 'peoples',
'continent', 'new-found', 'land', 'came', 'believed', 'could', 'create',
'upon', 'continent', 'new', 'life', 'life', 'new', 'freedom', 'vitality',
'written', 'mayflower', 'compact', 'declaration', 'independence',
'constitution', 'united', 'states', 'gettysburg', 'address', 'first', 'came',
'carry', 'longings', 'spirit', 'millions', 'followed', 'stock', 'sprang',
'moved', 'forward', 'constantly', 'consistently', 'toward', 'ideal',
'gained', 'stature', 'clarity', 'generation', 'hopes', 'republic', 'forever',
'tolerate', 'either', 'undeserved', 'poverty', 'self-serving', 'wealth',
'know', 'still', 'far', 'go', 'must', 'greatly', 'build', 'security',
'opportunity', 'knowledge', 'every', 'citizen', 'measure', 'justified',
'resources', 'capacity', 'land', 'enough', 'achieve', 'purposes', 'alone',
'enough', 'clothe', 'feed', 'body', 'nation', 'instruct', 'inform', 'mind',
'also', 'spirit', 'three', 'greatest', 'spirit', 'without', 'body', 'mind',
'men', 'know', 'nation', 'could', 'live', 'spirit', 'america', 'killed',
'even', 'though', 'nation', 'body', 'mind', 'constricted', 'alien', 'world',
'lived', 'america', 'know', 'would', 'perished', 'spirit', 'faith', 'speaks',

'us', 'daily', 'lives', 'ways', 'often', 'unnoticed', 'seem', 'obvious',
'speaks', 'us', 'capital', 'nation', 'speaks', 'us', 'processes',
'governing', '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',
'first', 'president', 'first', 'inaugural', '1789', 'words', 'almost',
'directed', 'would', 'seem', 'year', '1941', 'preservation', 'sacred',
'fire', 'liberty', 'destiny', 'republican', 'model', 'government', 'justly',
'considered', 'deeply', 'finally', 'staked', 'experiment', 'intrusted',
'hands', 'american', 'people', 'lose', 'sacred', 'fire', 'let', 'smothered',
'doubt', 'fear', 'shall', 'reject', 'destiny', 'washington', 'strove',
'valiantly', 'triumphantly', 'establish', 'preservation', 'spirit', 'faith',
'nation', 'furnish', 'highest', 'justification', 'every', 'sacrifice', 'may',
'make', 'cause', 'national', 'defense', 'face', 'great', 'perils', 'never',
'encountered', 'strong', 'purpose', 'protect', 'perpetuate', 'integrity',
'democracy', 'muster', 'spirit', 'america', 'faith', 'america', 'retreat',
'content', 'stand', 'still', 'americans', 'go', 'forward', 'service',
'country', 'god']

The Number of before remove stopword 1941- Roosevelt: 7571

The words after removed stopword: 625

The most common words are:

nation: 12

know: 10

spirit: 9

President John F. Kennedy in 1961

After removed stop word in President John F. Kennedy in 1961

'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', 'sworn', 'almighty',
'god', 'solemn', 'oath', 'forebears', 'l', 'prescribed', 'nearly', 'century',
'three', 'quarters', 'ago', 'world', 'different', 'man', 'holds', 'mortal',
'hands', 'power', 'abolish', 'forms', 'human', 'poverty', 'forms', 'human',
'life', 'yet', 'revolutionary', 'beliefs', 'forebears', 'fought', 'still',
'issue', 'around', 'globe', 'belief', 'rights', 'man', 'come', 'generosity',
'state', 'hand', 'god', 'dare', 'forget', 'today', 'heirs', 'first',
'revolution', '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',
'whether', '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', 'host', 'cooperative',
'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', 'across', 'globe', 'struggling', 'break', 'bonds',
'mass', 'misery', 'pledge', 'best', 'efforts', 'help', 'help', 'whatever',
'period', 'required', 'communists', 'may', 'seek', 'votes', 'right', 'free',
'society', 'help', 'many', 'poor', 'save', 'rich', 'sister', 'republics',
'south', 'border', 'offer', 'special', 'pledge', 'convert', 'good', 'words',
'good', 'deeds', 'new', 'alliance', 'progress', 'assist', 'free', 'men',
'free', 'governments', 'casting', 'chains', 'poverty', 'peaceful',
'revolution', 'hope', 'become', 'prey', 'hostile', 'powers', 'let',
'neighbors', 'know', 'shall', 'join', 'oppose', 'aggression', 'subversion',
'anywhere', 'americas', 'let', 'every', 'power', 'know', 'hemisphere',
'intends', 'remain', 'master', 'house', 'world', 'assembly', 'sovereign',
'states', 'united', 'nations', 'last', 'best', 'hope', 'age', 'instruments',
'war', 'far', 'outpaced', 'instruments', 'peace', 'renew', 'pledge',
'support', 'prevent', 'becoming', 'merely', 'forum', 'invective',
'strengthen', 'shield', 'new', 'weak', 'enlarge', 'area', 'writ', 'may',

'run', 'finally', 'nations', 'would', 'make', 'adversary', 'offer', 'pledge',
'request', 'sides', 'begin', 'anew', 'quest', 'peace', 'dark', 'powers',
'destruction', 'unleashed', 'science', 'engulf', 'humanity', 'planned',
'accidental', 'self-destruction', 'dare', 'tempt', 'weakness', 'arms',
'sufficient', 'beyond', 'doubt', 'certain', 'beyond', 'doubt', 'never',
'employed', 'neither', 'two', 'great', 'powerful', 'groups', 'nations',
'take', 'comfort', 'present', 'course', 'sides', 'overburdened', 'cost',
'modern', 'weapons', 'rightly', 'alarmed', 'steady', 'spread', 'deadly',
'atom', 'yet', 'racing', 'alter', 'uncertain', 'balance', 'terror', 'stays',
'hand', 'mankind', 'final', 'war', 'let', 'us', 'begin', 'anew',
'remembering', 'sides', 'civility', 'sign', 'weakness', 'sincerity',
'always', 'subject', 'proof', 'let', 'us', 'never', 'negotiate', 'fear',
'let', 'us', 'never', 'fear', 'negotiate', 'let', 'sides', 'explore',
'problems', 'unite', 'us', 'instead', 'belaboring', 'problems', 'divide',
'us', 'let', 'sides', 'first', 'time', 'formulate', 'serious', 'precise',
'proposals', 'inspection', 'control', 'arms', 'bring', 'absolute', 'power',
'destroy', 'nations', 'absolute', 'control', 'nations', 'let', 'sides',
'seek', 'invoke', 'wonders', 'science', 'instead', 'terrors', 'together',
'let', 'us', 'explore', 'stars', 'conquer', 'deserts', 'eradicate',
'disease', 'tap', 'ocean', 'depths', 'encourage', 'arts', 'commerce', 'let',
'sides', 'unite', 'heed', 'corners', 'earth', 'command', 'isaiah', 'undo',
'heavy', 'burdens', 'let', 'oppressed', 'go', 'free', 'beachhead',
'cooperation', 'may', 'push', 'back', 'jungle', 'suspicion', 'let', 'sides',
'join', 'creating', 'new', 'endeavor', 'new', 'balance', 'power', 'new',
'world', 'law', 'strong', 'weak', 'secure', 'peace', 'preserved', 'finished',
'first', '100', 'days', 'finished', 'first', '1,000', 'days', 'life',
'administration', 'even', 'perhaps', 'lifetime', 'planet', 'let', 'us',
'begin', 'hands', 'fellow', 'citizens', 'mine', 'rest', 'final', 'success',
'failure', 'course', 'since', 'country', 'founded', 'generation',
'americans', 'summoned', 'give', 'testimony', 'national', 'loyalty',
'graves', 'young', 'americans', 'answered', 'call', 'service', 'surround',
'globe', 'trumpet', 'summons', 'us', 'call', 'bear', 'arms', 'though',
'arms', 'need', 'call', 'battle', 'though', 'embattled', 'call', 'bear',
'burden', 'long', 'twilight', 'struggle', 'year', 'year', 'rejoicing',
'hope', 'patient', 'tribulation', 'struggle', 'common', 'enemies', 'man',
'tyranny', 'poverty', 'disease', 'war', 'forge', 'enemies', 'grand',
'global', 'alliance', 'north', 'south', 'east', 'west', 'assure', 'fruitful',
'life', 'mankind', 'join', 'historic', 'effort', 'long', 'history', 'world',
'generations', 'granted', 'role', 'defending', 'freedom', 'hour', 'maximum',
'danger', 'shrink', 'responsibility', 'welcome', 'believe', 'us', 'would',
'exchange', 'places', 'people', 'generation', 'energy', 'faith', 'devotion',
'bring', 'endeavor', 'light', 'country', 'serve', 'glow', 'fire', 'truly',
'light', 'world', 'fellow', 'americans', 'ask', 'country', 'ask', 'country',
'fellow', 'citizens', 'world', 'ask', 'america', 'together', 'freedom',
'man', 'finally', 'whether', 'citizens', 'america', 'citizens', 'world',
'ask', 'us', 'high', 'standards', 'strength', 'sacrifice', 'ask', 'good',
'conscience', 'sure', 'reward', 'history', 'final', 'judge', 'deeds', 'let',
'us', 'go', 'forth', 'lead', 'land', 'love', 'asking', 'blessing', 'help',
'knowing', 'earth', 'god', 'work', 'must', 'truly'

The Number of words before stopward For 1961-Kennedy.txt : 7618

The words after removed stopword: 689

The most common words are:

let: 16

us: 12

world: 8

President Richard Nixon in 1973

AFTER REMOVING STOPWORD FROM PRESIDENT RICAHARD NIXON IN 1973

'mr.', 'vice', 'president', 'mr.', 'speaker', 'mr.', 'chief', 'justice',
'senator', 'cook', 'mrs.', 'eisenhower', 'fellow', 'citizens', 'great',
'good', 'country', 'share', 'together', 'met', 'four', 'years', 'ago',
'america', 'bleak', 'spirit', 'depressed', 'prospect', 'seemingly',
'endless', 'war', 'abroad', 'destructive', 'conflict', 'home', 'meet',
'today', 'stand', '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', 'far-reaching', 'results', 'new',
'policies', 'peace', 'continuing', 'revitalize', 'traditional',
'friendships', 'missions', 'peking', 'moscow', 'able', 'establish', 'base',
'new', 'durable', 'pattern', 'relationships', 'among', 'nations', 'world',
'america', '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', 'america', 'role',
'maintaining', 'peace', 'unless', 'america', 'work', 'preserve', 'peace',
'peace', 'unless', 'america', 'work', 'preserve', 'freedom', 'freedom',
'let', 'us', 'clearly', 'understand', 'new', 'nature', 'america', 'role',
'result', 'new', 'policies', 'adopted', 'past', 'four', 'years', 'shall',
'respect', 'treaty', 'commitments', 'shall', 'support', 'vigorously',
'principle', 'country', 'right', 'impose', 'rule', 'another', 'force',
'shall', 'continue', 'era', 'negotiation', 'work', 'limitation', 'nuclear',
'arms', 'reduce', 'danger', 'confrontation', 'great', 'powers', 'shall',

'share', 'defending', 'peace', 'freedom', 'world', 'shall', 'expect',
'others', 'share', 'time', 'passed', 'america', 'make', 'every', 'nation',
'conflict', 'make', 'every', 'nation', 'future', 'responsibility', 'presume',
'tell', 'people', 'nations', 'manage', 'affairs', 'respect', 'right',
'nation', 'determine', 'future', 'also', 'recognize', 'responsibility',
'nation', 'secure', 'future', 'america', 'role', 'indispensable',
'preserving', 'world', 'peace', 'nation', 'role', 'indispensable',
'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', 'us', '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', 'greatly',
'meeting', 'responsibilities', 'abroad', 'remain', 'great', 'nation',
'remain', 'great', 'nation', 'act', 'greatly', 'meeting', 'challenges',
'home', 'chance', 'today', 'ever', 'history', 'make', 'life', 'better',
'america', 'ensure', 'better', 'education', 'better', 'health', 'better',
'housing', 'better', 'transportation', 'cleaner', 'environment', 'restore',
'respect', 'law', 'make', 'communities', 'livable', 'insure', 'god-given',
'right', 'every', 'american', 'full', 'equal', 'opportunity', 'range',
'needs', 'great', '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', 'retreat', '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',
'washington', 'abroad', 'home', 'time', 'come', 'turn', 'away',
'condescending', 'policies', 'paternalism', 'washington', 'knows', 'best',
'person', 'expected', 'act', 'responsibly', 'responsibility', 'human',
'nature', 'let', 'us', 'encourage', 'individuals', 'home', 'nations',
'abroad', 'decide', 'let', 'us', 'locate', 'responsibility', 'places', 'let',
'us', 'measure', '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',
'people', 'government', 'must', 'learn', 'take', 'less', 'people', 'people',
'let', 'us', 'remember', 'america', 'built', 'government', 'people',
'welfare', 'work', 'shirking', 'responsibility', 'seeking', 'responsibility',
'lives', 'let', 'us', 'ask', 'government', 'challenges', 'face', 'together',
'let', 'us', 'ask', 'government', 'help', 'help', 'national', 'government',
'great', 'vital', 'role', 'play', 'pledge', 'government', 'act', 'act',

'boldly', 'lead', 'boldly', 'important', 'role', 'every', 'one', 'us',
'must', 'play', 'individual', 'member', 'community', 'day', 'forward', 'let',
'us', 'make', 'solemn', 'commitment', 'heart', 'bear', 'responsibility',
'part', 'live', 'ideals', 'together', 'see', 'dawn', 'new', 'age',
'progress', 'america', 'together', 'celebrate', '200th', 'anniversary',
'nation', 'proud', 'fulfillment', 'promise', 'world', 'america', 'longest',
'difficult', 'war', 'comes', 'end', 'let', 'us', 'learn', 'debate',
'differences', 'civility', 'decency', 'let', 'us', 'reach', 'one',
'precious', 'quality', 'government', 'provide', 'new', 'level', 'respect',
'rights', 'feelings', 'one', 'another', 'new', 'level', 'respect',
'individual', 'human', 'dignity', 'cherished', 'birthright', 'every',
'american', 'else', 'time', 'come', 'us', 'renew', 'faith', 'america',
'recent', 'years', 'faith', 'challenged', 'children', 'taught', 'ashamed',
'country', 'ashamed', 'parents', 'ashamed', 'america', 'record', 'home',
'role', 'world', 'every', 'turn', 'beset', 'find', 'everything', 'wrong',
'america', 'little', 'right', 'confident', 'judgment', 'history',
'remarkable', 'times', 'privileged', 'live', 'america', 'record', 'century',
'unparalleled', 'world', 'history', 'responsibility', 'generosity',
'creativity', 'progress', 'let', 'us', 'proud', 'system', 'produced',
'provided', 'freedom', 'abundance', 'widely', 'shared', 'system', 'history',
'world', 'let', 'us', 'proud', 'four', 'wars', 'engaged', 'century',
'including', 'one', 'bringing', 'end', 'fought', 'selfish', 'advantage',
'help', 'others', 'resist', 'aggression', 'let', 'us', 'proud', 'bold',
'new', 'initiatives', 'steadfastness', 'peace', 'honor', 'made', 'break-
through', 'toward', 'creating', 'world', 'world', 'known', 'structure',
'peace', 'last', 'merely', 'time', 'generations', 'come', 'embarking',
'today', 'era', 'presents', 'challenges', 'great', 'nation', 'generation',
'ever', 'faced', 'shall', 'answer', 'god', 'history', 'conscience', 'way',
'use', 'years', 'stand', 'place', 'hallowed', 'history', 'think', 'others',
'stood', 'think', 'dreams', 'america', 'think', 'recognized', 'needed',
'help', 'far', 'beyond', 'order', 'make', 'dreams', 'come', 'true', 'today',
'ask', 'prayers', 'years', 'ahead', 'may', 'god', 'help', 'making',
'decisions', 'right', 'america', 'pray', 'help', 'together', 'may', 'worthy',
'challenge', 'let', 'us', 'pledge', 'together', 'make', 'next', 'four',
'years', 'best', 'four', 'years', 'america', 'history', '200th', 'birthday',
'america', 'young', 'vital', 'began', 'bright', 'beacon', 'hope', 'world',
'let', 'us', 'go', 'forward', 'confident', 'hope', 'strong', 'faith', 'one',
'another', 'sustained', 'faith', 'god', 'created', 'us', 'striving',
'always', 'serve', 'purpose'

The most common words are:

us: 26

let: 22

america: 21

The Number of words before removing stopward For 1973-Nixon.txt : 9991

Number of words after removing stopwords: 832

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

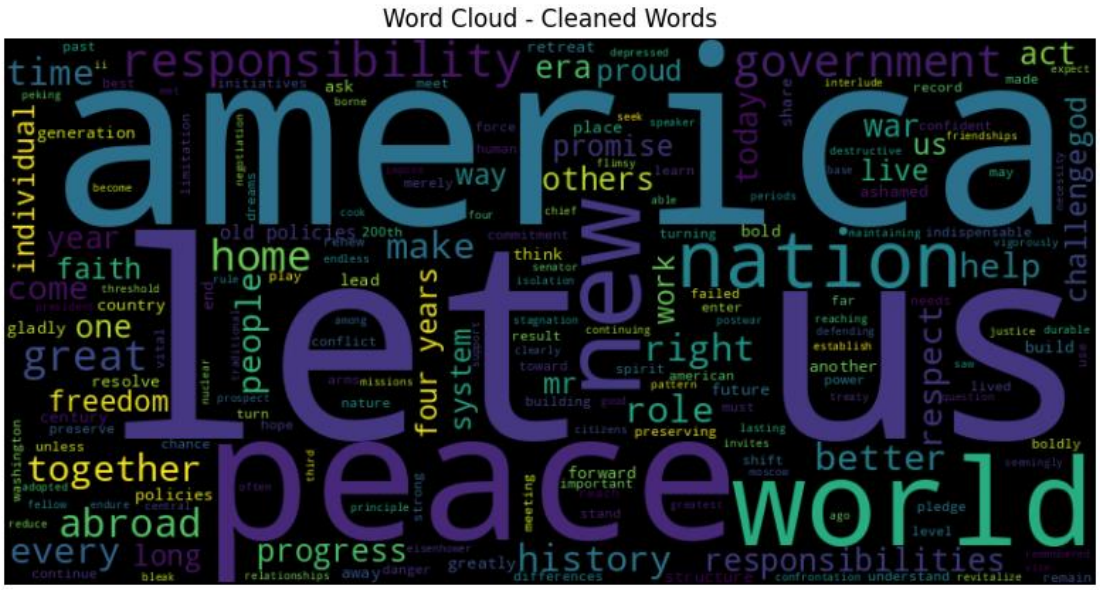
WE ARE IMPORTING THE WORD CLOUD TO CHECK HOW STRONG OR(WEIGHTED WORDS)

THERE ARE 3 WORDS ARE STRONG IN President Franklin D. Roosevelt in 1941 is NATION, PEOPLE, KNOW.

President John F. Kennedy in 1961 Word Cloud:

The Strong words in President John F. Kennedy in 1961 are World, let, nation.

President Richard Nixon in 1973



The strong words in President Richard Nixon in 1973 are America, Let, Us.

INSIGHTS:

IN THE 3 SPEECH Of president we can say nation are repeated in the all among the speech.