PROJECT - MACHINE LEARNING

JAYA PREETHI R M

TABLE OF CONTENTS

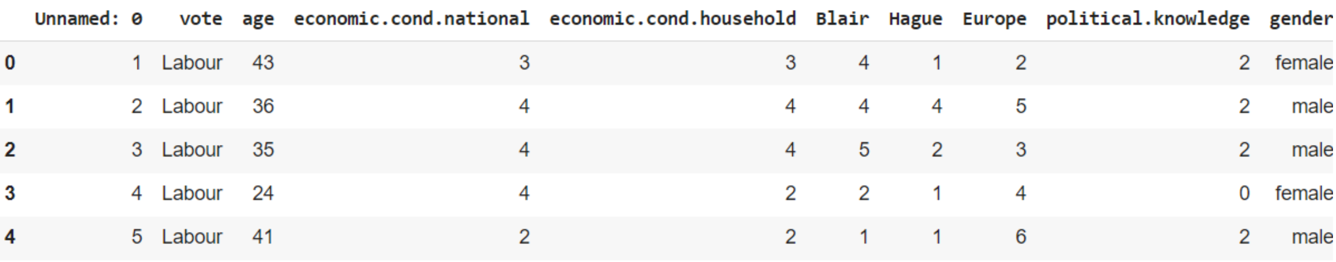
TABLES & DIAGRAMS

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.

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.

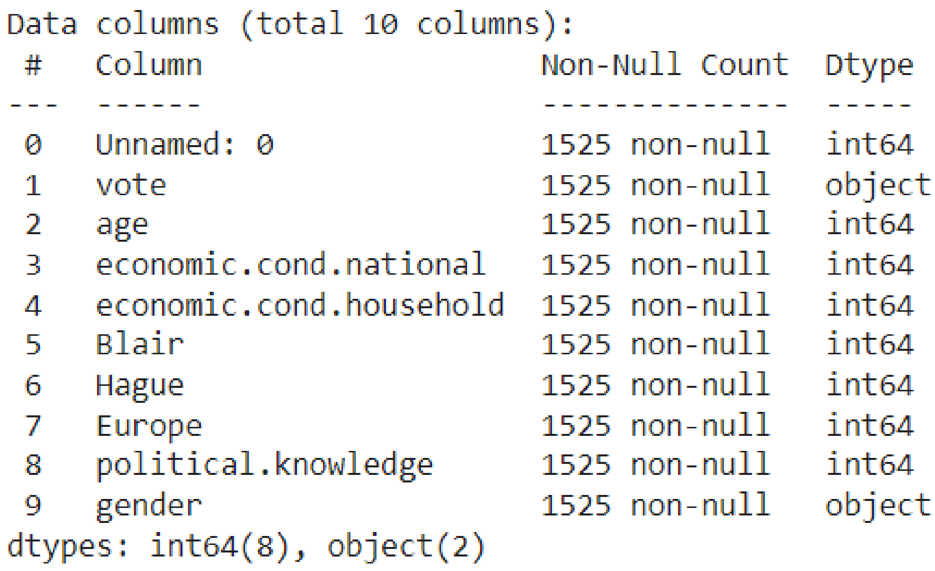
HEAD



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

The dataset has 10 variables or 10 columns. Most of the variables are ordinal or categorical in nature, except age which is continuous in nature. There is also a serial number column which must be removed, since it is redundant.

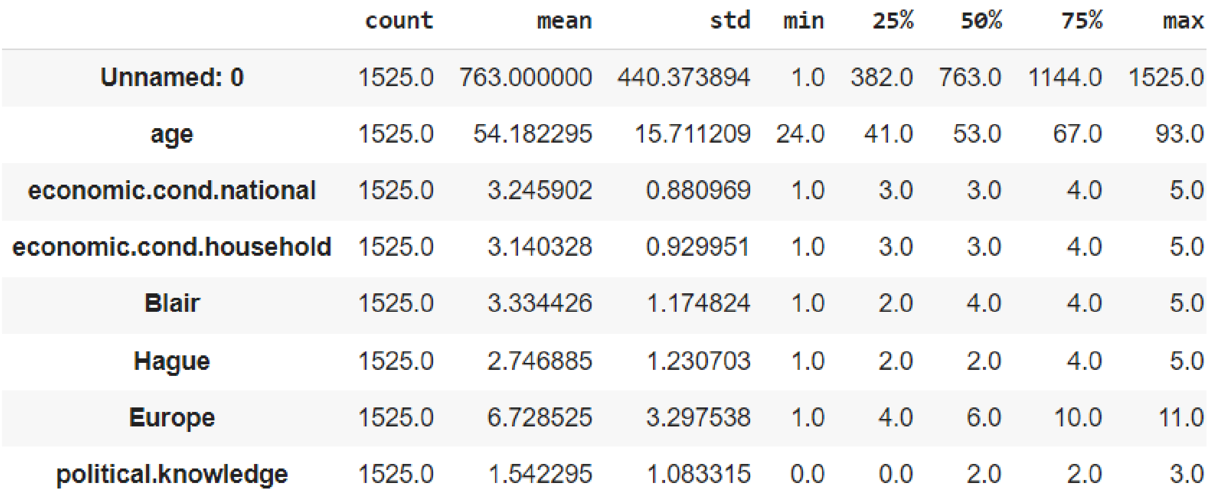
INFO & DATA TYPES



```
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)
```

There are 1525 entries or 1525 rows in total. We can also say that there are no null variables in the dataset, since the total number of entries are the same all throughout the info table. There are 8 variables of integer data type and 2 variables of object data type.

SUMMARY



	count	mean	std	min	25%	50%	75%	max
Unnamed: 0	1525.0	763.000000	440.373894	1.0	382.0	763.0	1144.0	1525.0
age	1525.0	54.182295	15.711209	24.0	41.0	53.0	67.0	93.0
economic.cond.national	1525.0	3.245902	0.880969	1.0	3.0	3.0	4.0	5.0
economic.cond.household	1525.0	3.140328	0.929951	1.0	3.0	3.0	4.0	5.0
Blair	1525.0	3.334426	1.174824	1.0	2.0	4.0	4.0	5.0
Hague	1525.0	2.746885	1.230703	1.0	2.0	2.0	4.0	5.0
Europe	1525.0	6.728525	3.297538	1.0	4.0	6.0	10.0	11.0
political.knowledge	1525.0	1.542295	1.083315	0.0	0.0	2.0	2.0	3.0

Most ordinal variables range between 1 to 5 in terms of the respondent's views on the topic. With variable political knowledge having the range of 0 to 3 and Europe variable ranging from 1 to 11.

NULL VARIABLE

```
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
```

There are no null variables in this dataset.

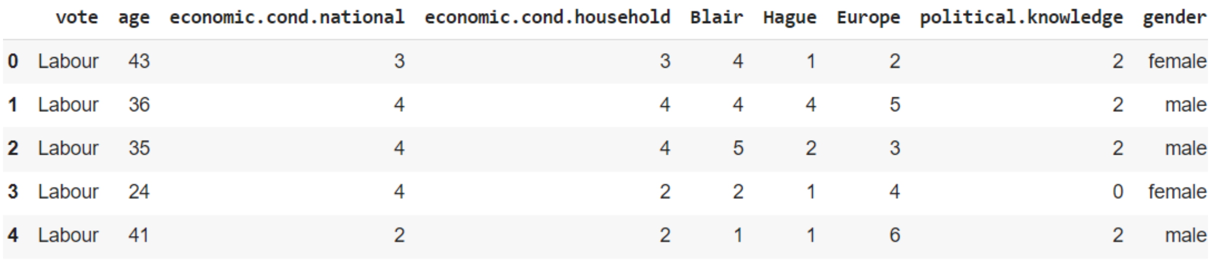
DUPLICATES

```
False 1525
```

There are no duplicated rows in this dataset.

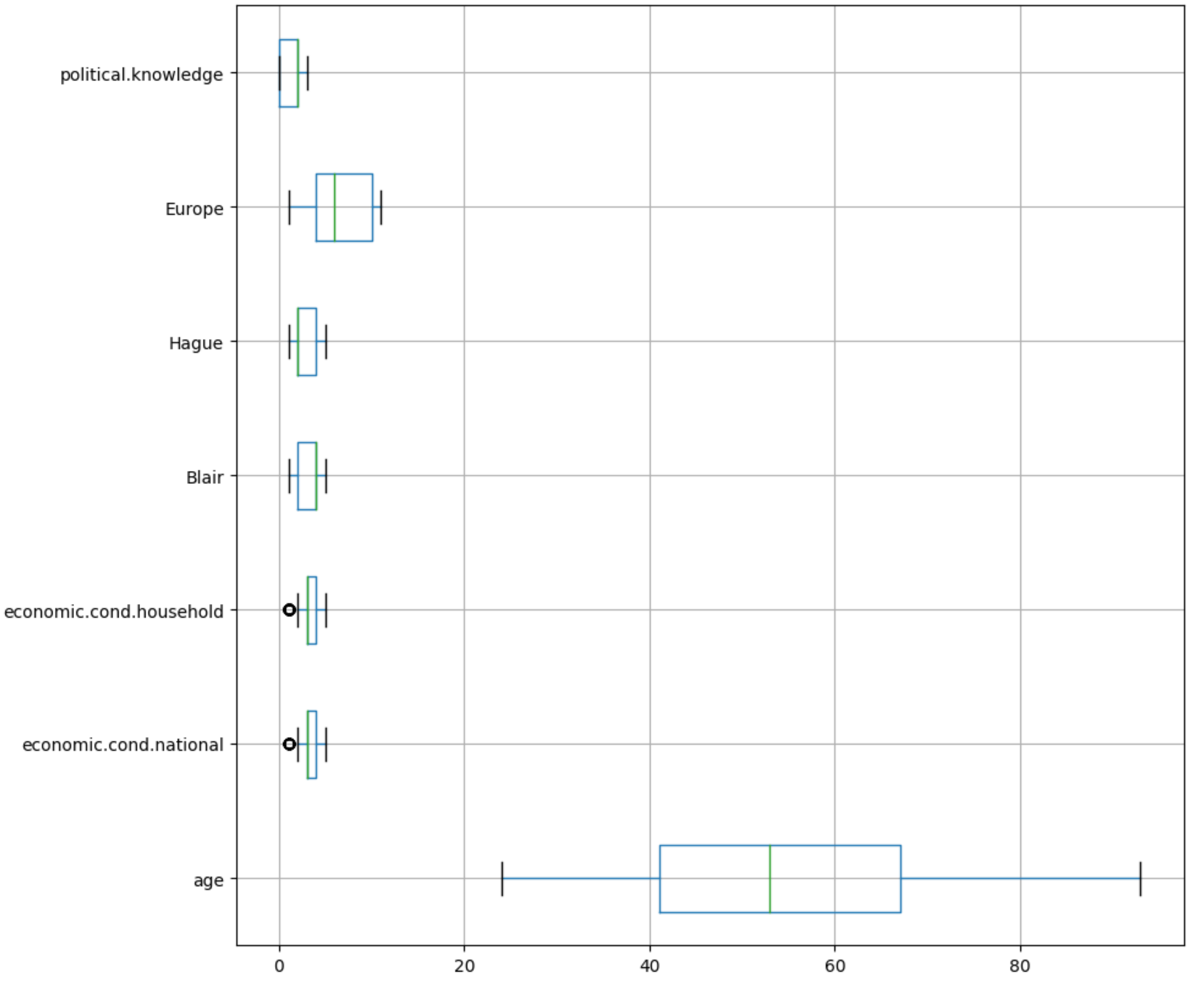
DROPPING SR. NUMBER VARIABLE

Before checking the skewness of the variables, I dropped the serial number variable - 'Unnamed : 0'. Following the sample of the dataset after this process.



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

SKEWNESS



Most ordinal variables are normally distributed. Political knowledge variable is positively skewed. Similarly, age is slightly positively skewed. Whereas, the Europe variable is negatively skewed. Boxplot for the age variable lies in a different range than the other variables because of its continuous nature

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.

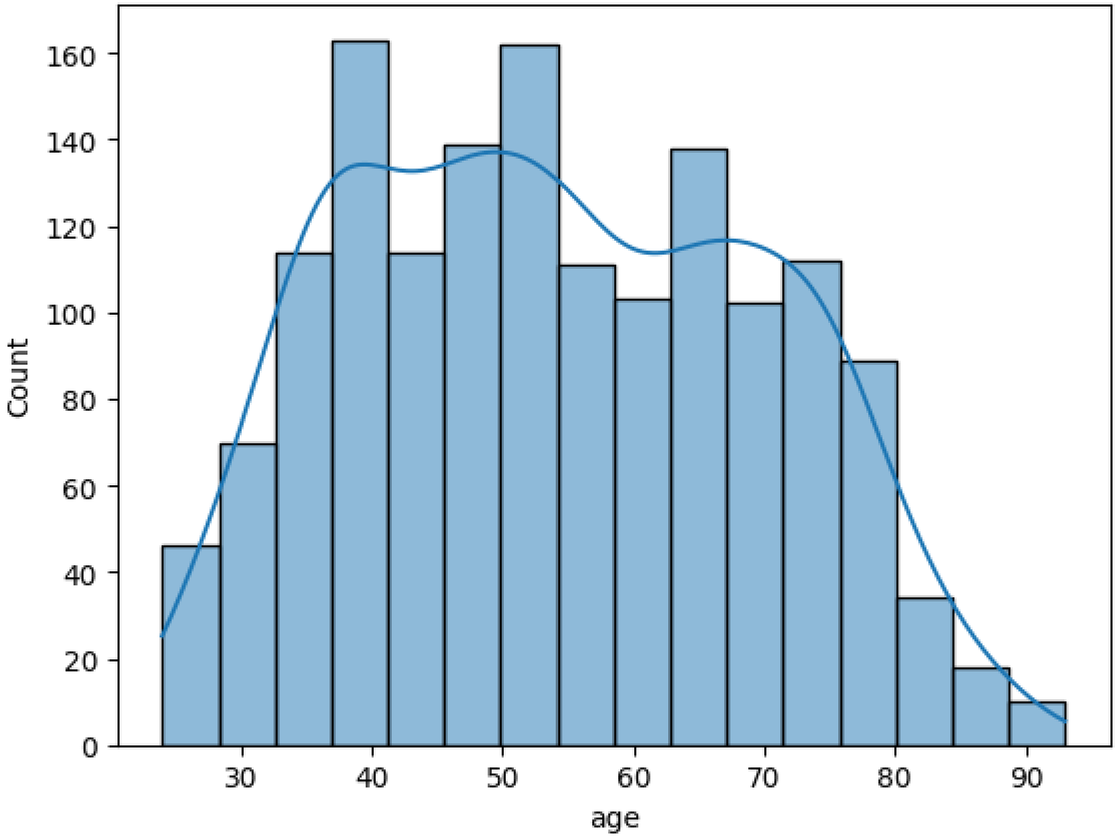
There is no restriction on how the learner wishes to implement this but the code should be able to represent the correct output and inferences should be logical and correct.

OUTLIERS

From the above given box plot diagram, there are some outliers in the economic.cond.household variable and economic.cond.national variable. I have decided not to treat these outliers since these variables are ordinal in nature and treating the outliers might affect the prediction accuracy in the model building.

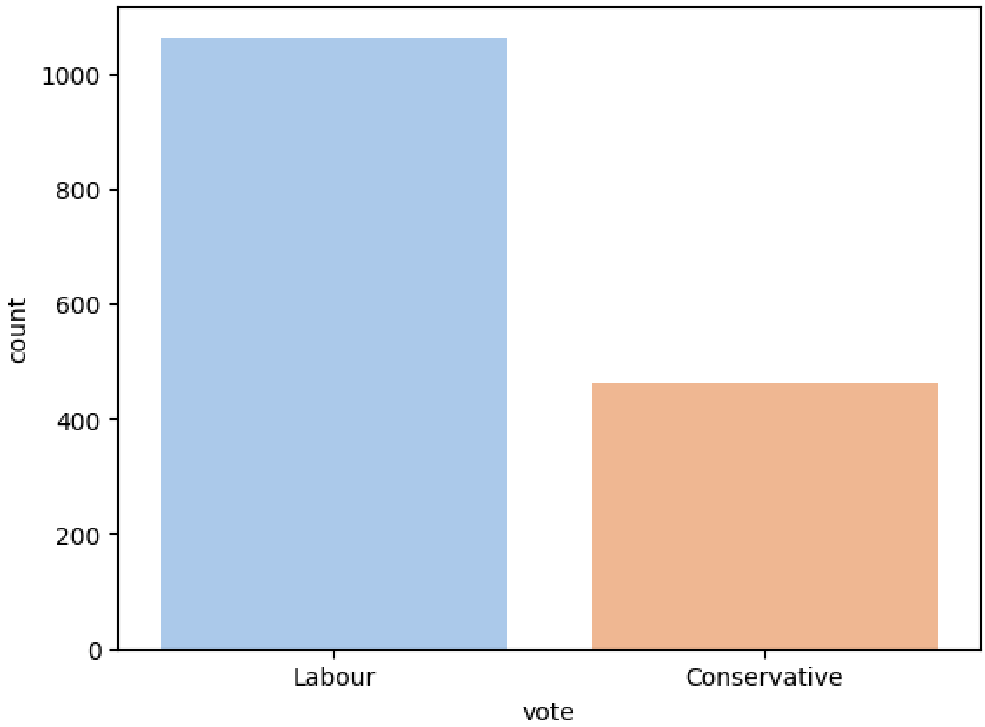
DESCRIPTIVE ANALYSIS

AGE



The distribution of the data must be significant on the left of the graph, i.e., the lower end of the scale. Thus, there are more respondents of the age less than 60. The lowest aged respondent is around 25 years of age. TThe highest aged respondent is around 95 years of age.

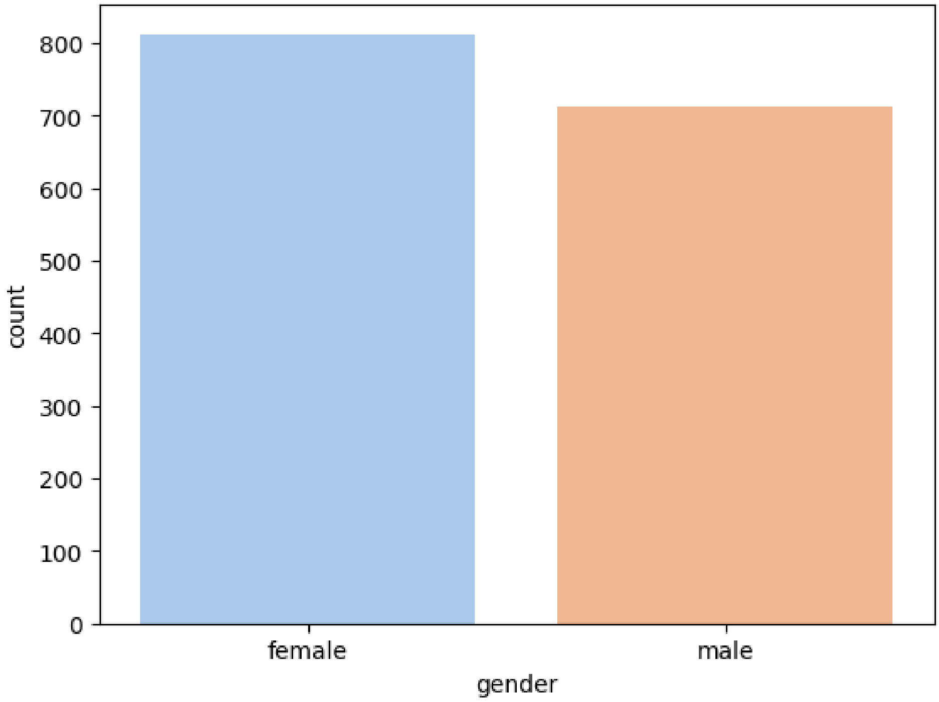
VOTE



```
Total number of people who voted for Labour party :1063
Total number of people who voted for Conservative party :462
```

This variable accounts the total number of people who voted for the labour party and the conservative party each. There are significantly higher numbers of individuals who voted for the Labour party than the conservative party.

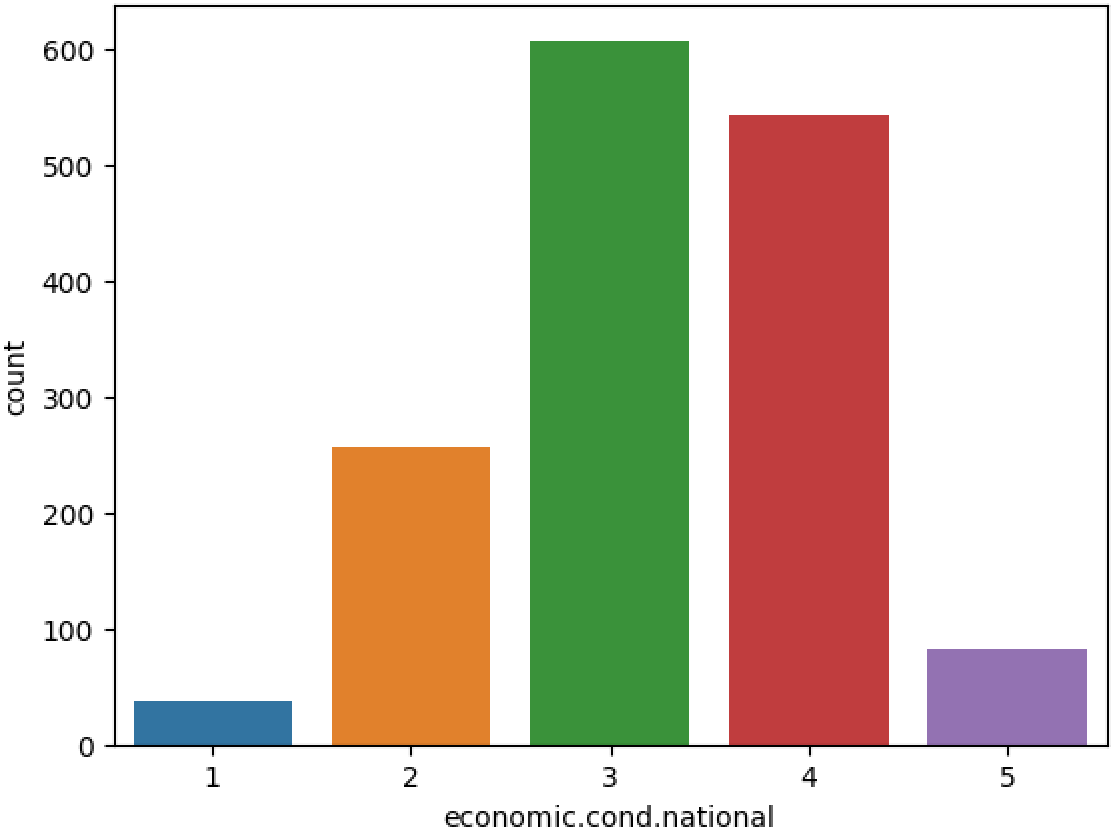
GENDER



```
There are 812 female respondents in total.
There are 713 male respondents in total.
```

There are more female respondents than male respondents in this dataset.

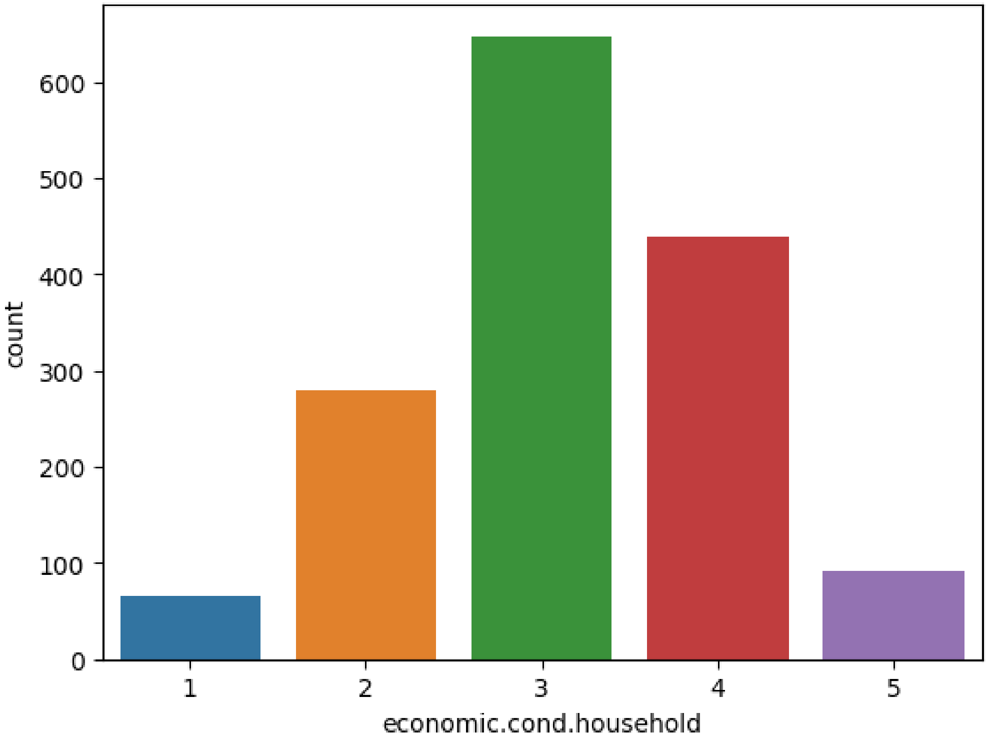
ECONOMIC.COND.NATIONAL



```
SCORE: NO. OF RESPONDENTS: 1:37; 2:257; 3:607; 4:542; 5:82
```

This variable shows us the respondent's assessment of the current national economic condition on the scale of 1 to 5, with 1 being the lowest score and 5 being the highest score. Most consumers have felt pretty neutral on this accord with most recurring scores being 3. There is another set of significant population who feel that the current economic condition can be described with a score of 4 on 5 followed by 3 on 5.

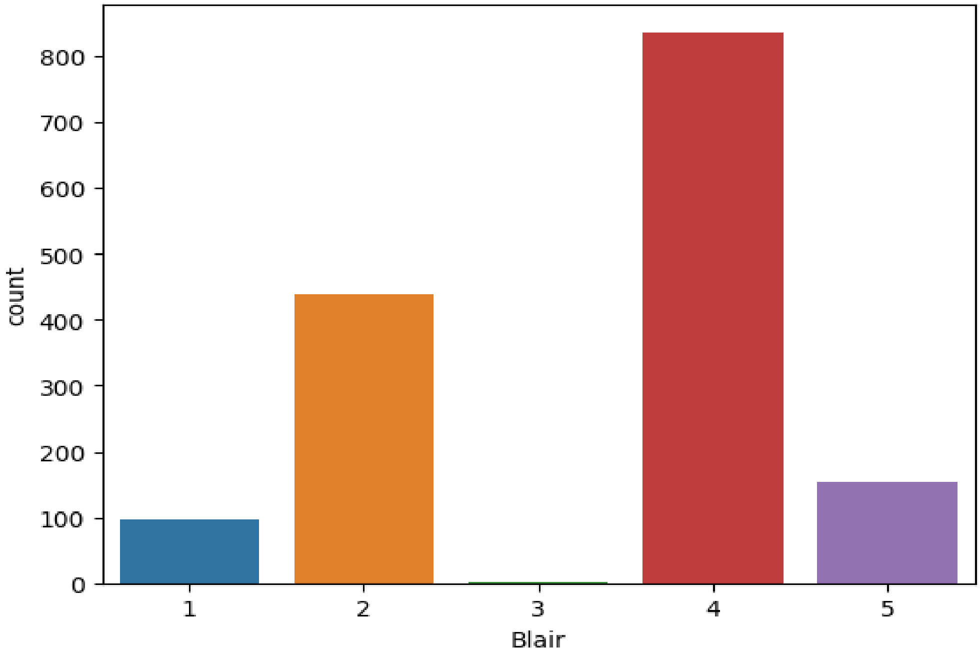
ECONOMIC CONDITION HOUSEHOLD



`SCORE: NO. OF RESPONDENTS: 1:65; 2:280; 3:648; 4:440; 5:92`

This variable shows us the respondent's assessment score on the current household economic conditions on the scale of 1 to 5, with 1 being the lowest score and 5 being the highest score. Most consumers have felt pretty neutral on this accord with most recurring scores being 3. There is another set of significant population who feel that the current economic condition can be described with a score of 4 on 5 followed by 3 on 5.

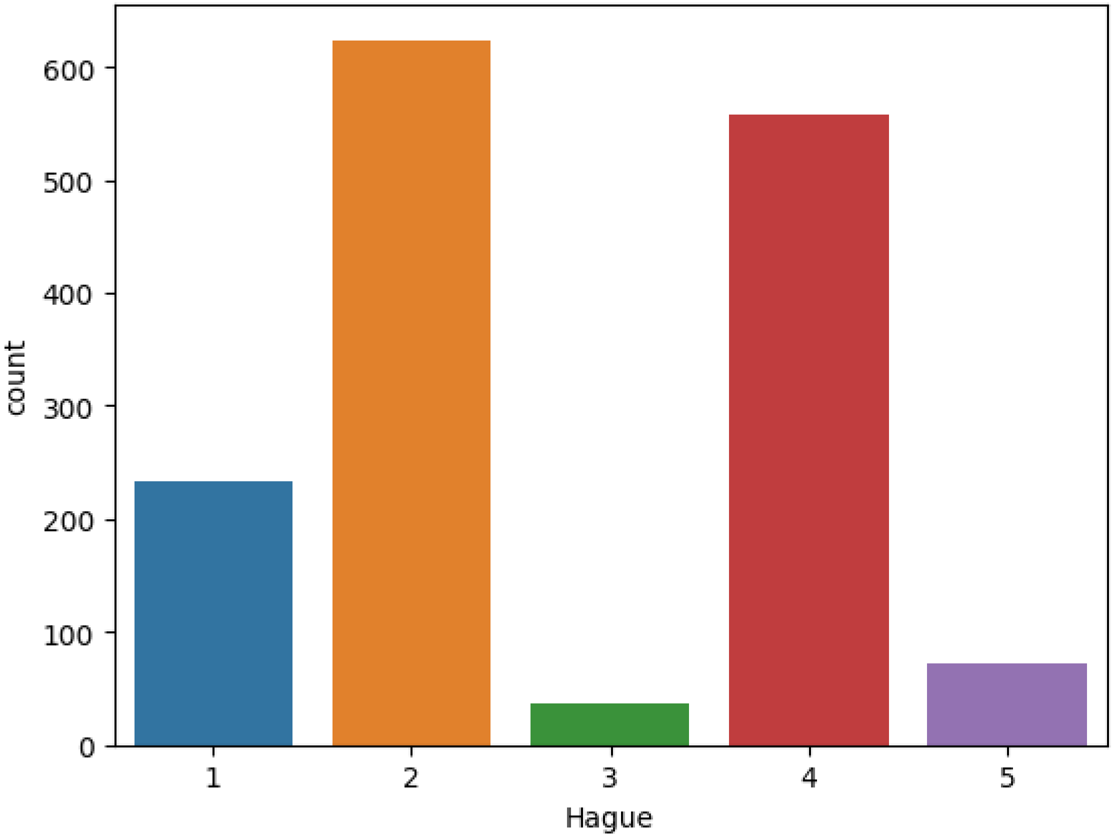
BLAIR



`SCORE: NO. OF RESPONDENTS: 1: 97; 2:438; 3:1; 4:836; 5:153`

This variable gives us the respondent's assessment score on the Labour Party contestant: Tony Blair. Most respondents gave Blair a score of 4 on 5 followed by a score of 2 on 5. Thus we can say that most consumers had a positive assessment of Blair.

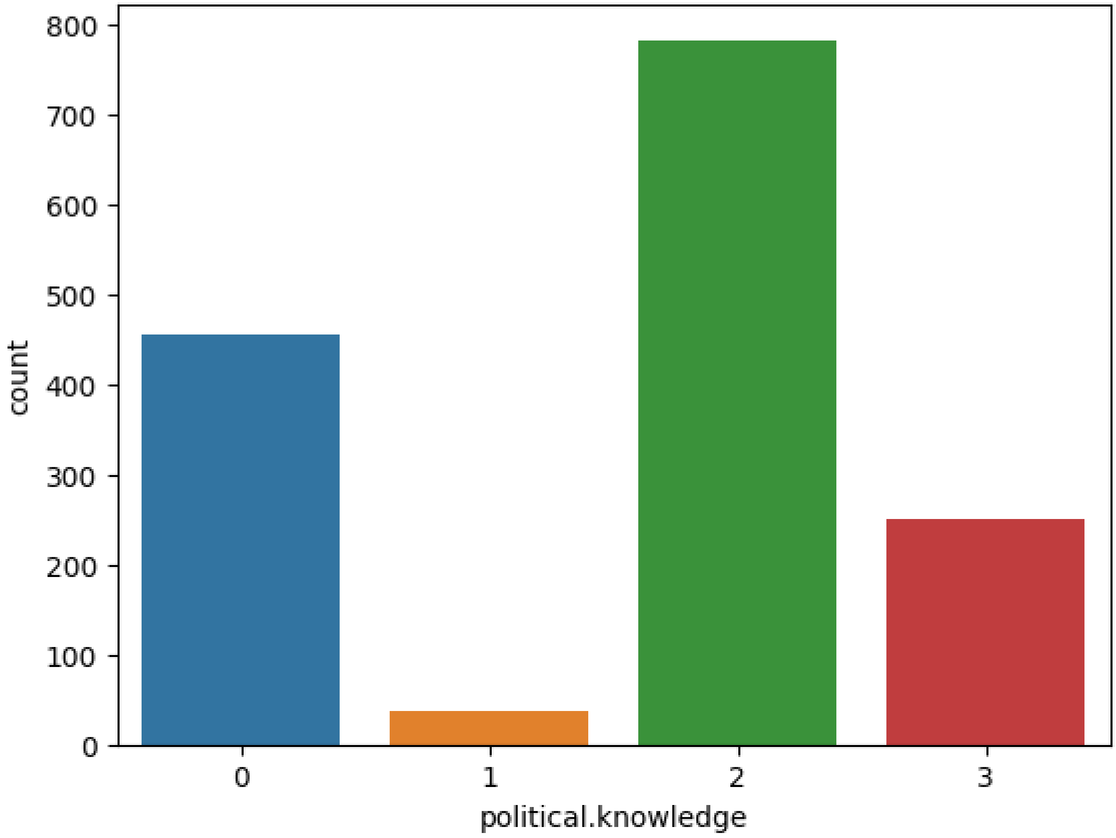
HAGUE



```
SCORE: NO. OF RESPONDENTS: 1:233; 2:624; 3:37; 4:558; 5:73
```

This variable gives us the respondent’s assessment score on the Conservative Party contestant: Hague. Most respondents gave Hague a score of 2 on 5 followed by a score of 4 on 5. Thus we can say that most consumers had a relatively negative assessment of Blair.

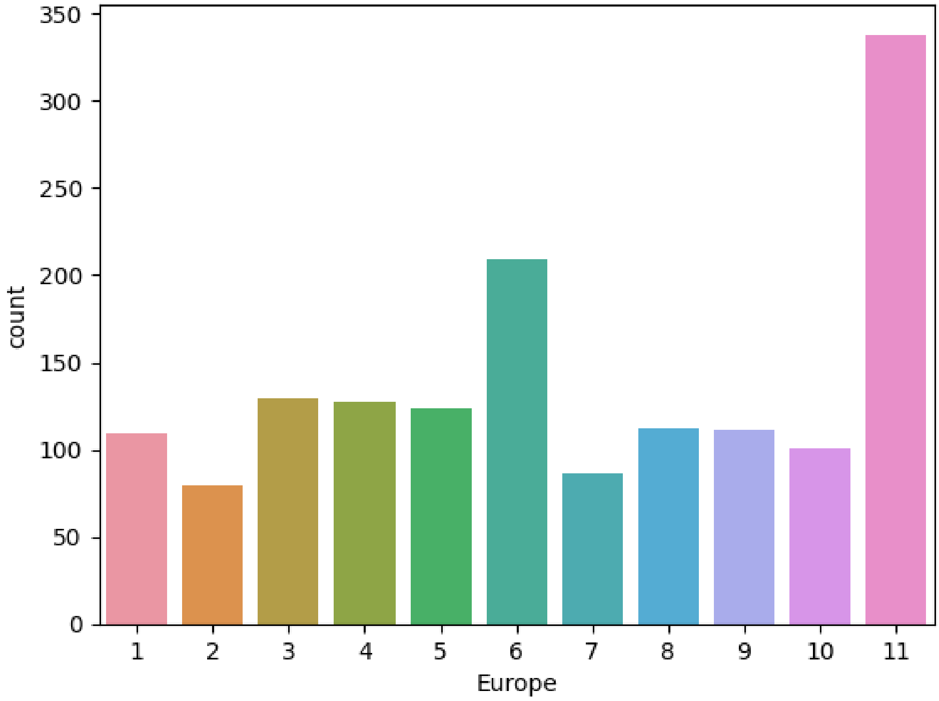
POLITICAL KNOWLEDGE



```
SCORE: TOTAL RESPONDENTS:  0:455; 1:38; 2:782; 3:250
```

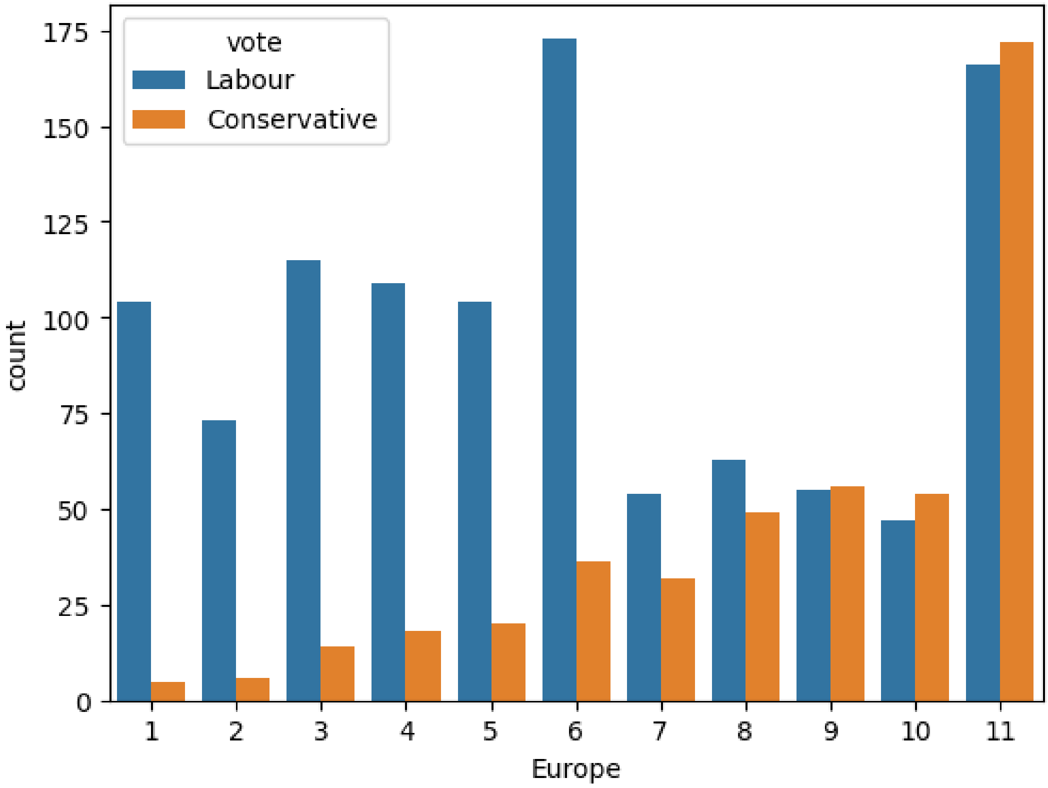
This variable measures the level of knowledge of the respondent’s on the parties position on Europe integration on a scale of 0 to 3. Most respondents are aware of the party's position on Europe integration since the highest number of respondents chose 2 on 3. The second most popular choice was 0 on 3, there are a significant number among the respondents who are unaware of the party's position on Europe integration.

EUROPE



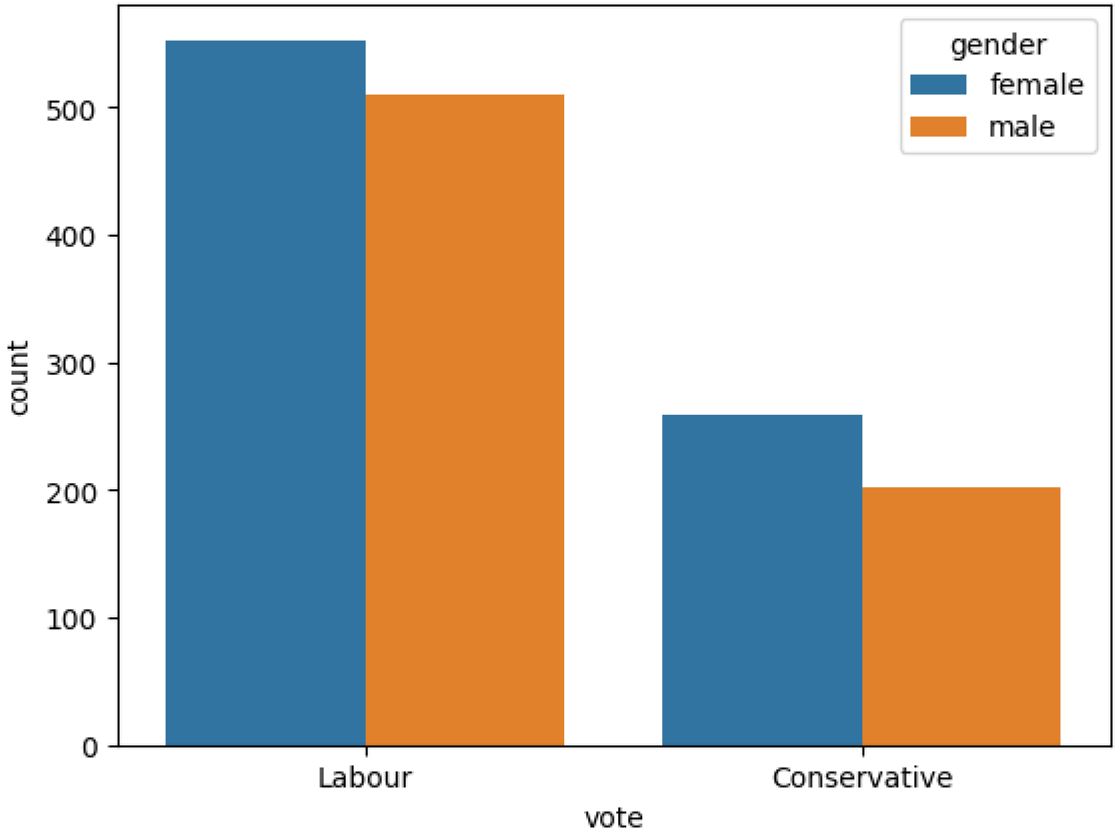
This variable measures the respondent’s attitude towards European integration with a scale of 1 to 11, where 11 represents ‘Eurosceptic’ sentiment. Most respondents chose 11 on 11, Thus, respondents exhibit highly Eurosceptic sentiment.

EUROPE - VOTE



The respondents who voted for the Labour party tend to have a varied level of response for European integration. With a significant number of Labour party members choosing 6 on 11 and 11 on 11. On the other hand, the majority of respondents who voted for conservative party chose 11 on 11, exhibiting high eurosceptic sentiment.

VOTE - GENDER



The gender ratio in between Labour party voters and Conservative party voters is pretty similar, with Conservative party voters being lesser in number than the Labour party voters.

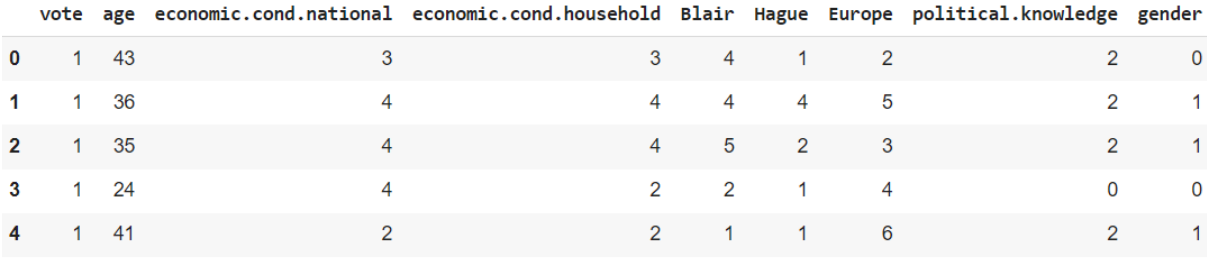
1.3) Encode the data (having string values) for Modelling. Is Scaling necessary here or not?(2 pts), Data Split: Split the data into train and test (70:30) (2 pts). The learner is expected to check and comment about the difference in scale of different features on the bases of appropriate measure for 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.

DATA SCALING

I would prefer not to scale the data. Since the independent variables are predominantly ordinal or categorical variables, scaling the data would result in changing the nature of the dataset. Thus, affecting the performance of the model including predictability and accuracy of the model.

CONVERT DATA TYPE

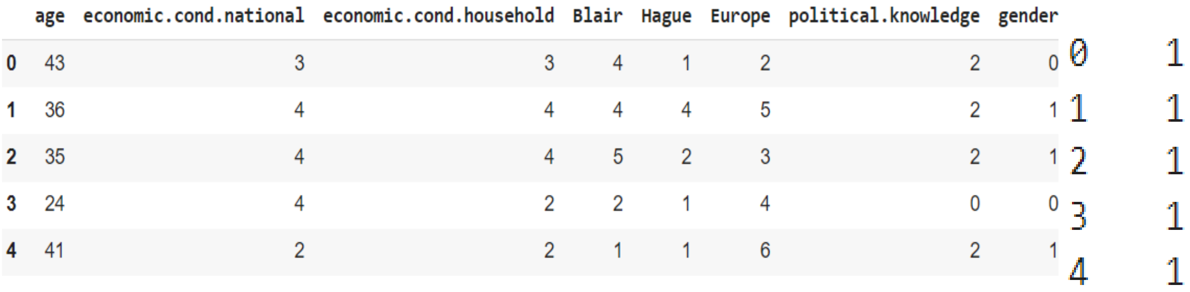
Vote and Gender variables are of object type types. Thus, they were converted into dummy variables by using the get_dummies command. Following sample is the result of that.



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

DATA SPLIT

Further data was split into X and Y dataset, with X being independent variables and Y being dependent variables.



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

TRAIN-TEST SPLIT

The data set was split into train and test data using a test ratio of 70:30. The random state hyper parameter is used to control the shuffling process while creating the test dataset. This hyper parameter is set at 1, since this might reproduce the same split each time the code is run. Stratify is a hyper parameter used to split the dataset into two subsets based on a categorical variable. In this case the variable vote has been used for stratification.

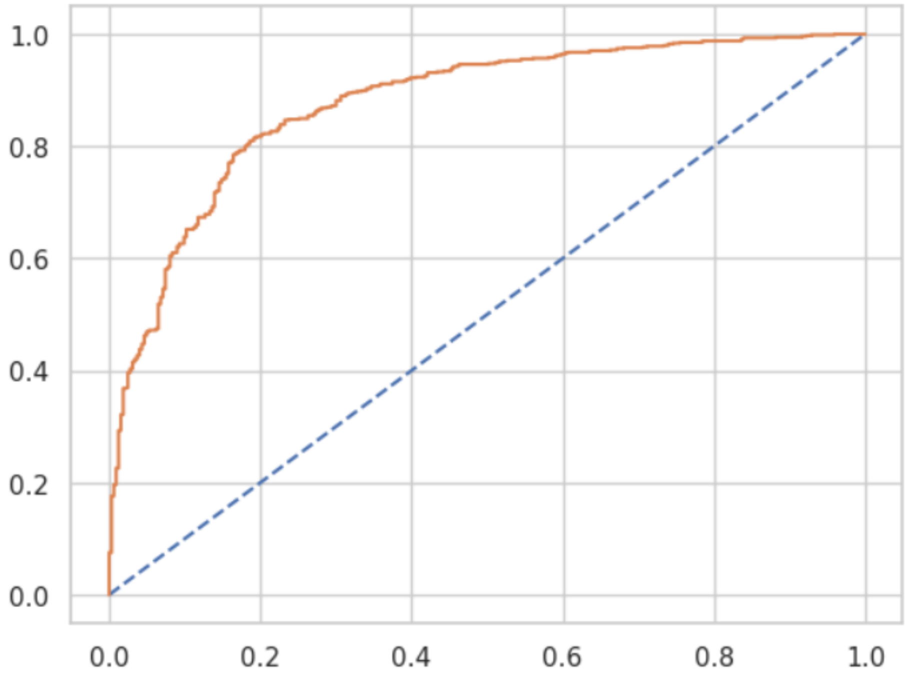
1.4) Apply Logistic Regression and LDA (Linear Discriminant Analysis) (2 pts). Interpret the inferences of both model s (2 pts). Successful implementation of each model. Logical reason should be shared if any custom changes are made to the parameters while building the model. Calculate Train and Test Accuracies for each model. Comment on the validness of models (over fitting or under fitting)

LOGISTIC REGRESSION

A Logistic Regression tool from sklearn library was used to create a logistic regression model. The solver hyperparameter helps in reducing the optimization problems of the target variable by performing approximate minimisation. Here, the solver was set with newton-cg, which uses Hessian matrix to improve model optimization. This setting can make the code run slow for larger dataset, since our dataset is relatively smaller it was not an issue. Max_iter hyperparameter controls the total times k_means runs from the starting point. The iteration was set at 10,000 in this case. The penalty hyperparameter helps in regularisation of the model or shrinking the coefficients of less contributive variables. Here, it was set at ‘none’. N_jobs controls the parallelism of all processes in the available processor. It was set at 2.

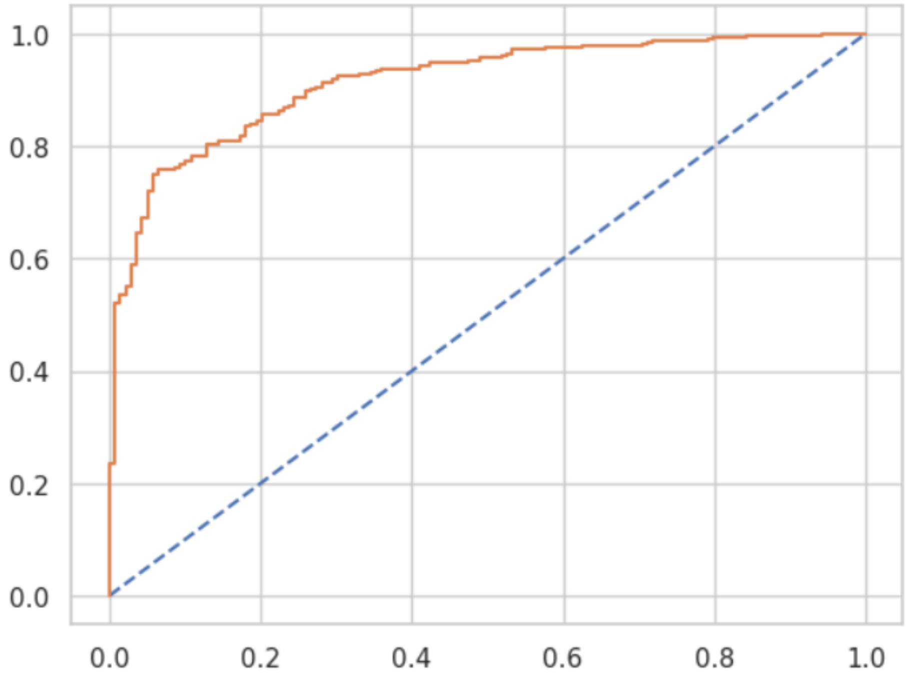
TRAIN DATA ACCURACY
MODEL SCORE: 0.83
AUC SCORE:

```
AUC: 0.877
```



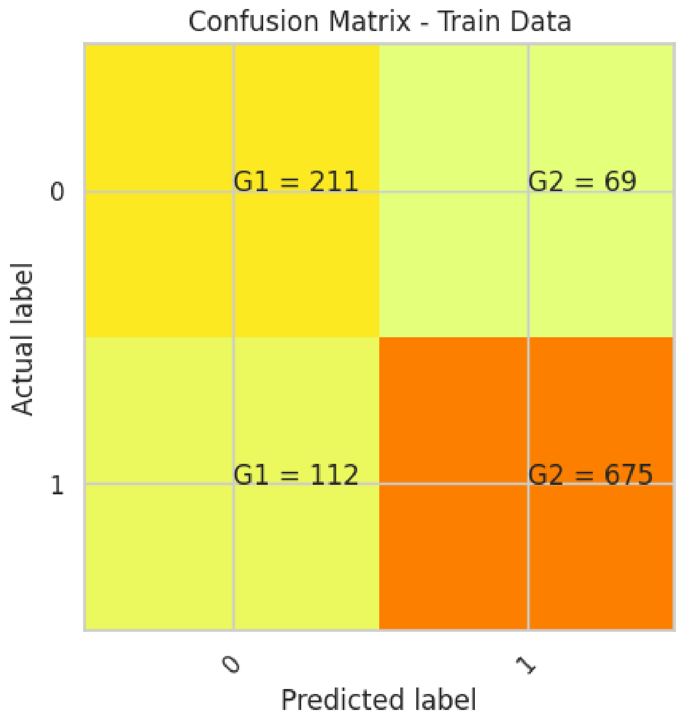
TEST DATA ACCURACY
MODEL SCORE: 0.85
AUC SCORE:

```
AUC: 0.877
```

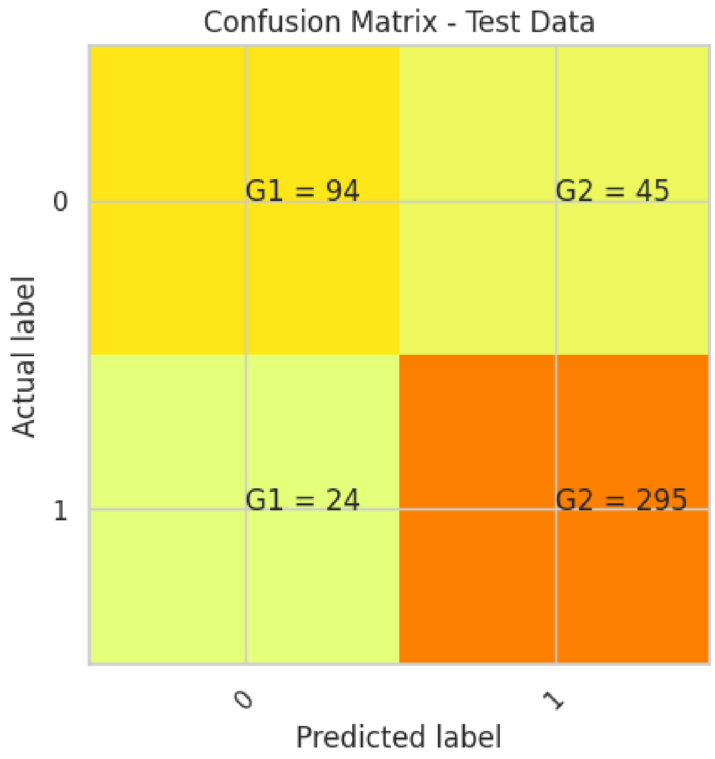


Both train data and test data have the same AUC score and the AUC curve are similar. However, train data has the model score of 83% and test data has model score of 85%. Thus, there is a minimal case of overfitting and the test data performs better than the train data.

CONFUSION MATRIX
TRAIN DATA



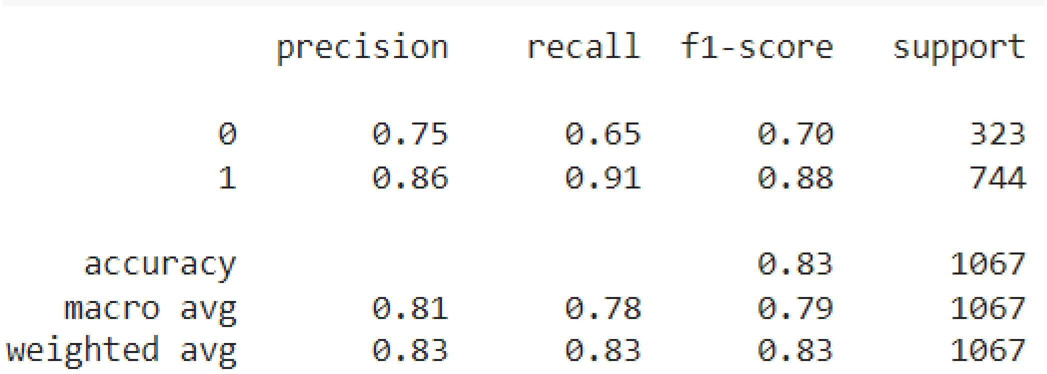
TEST DATA



Both train data and test data are good at predicting true negatives followed by true positives. However,overall test data predictions are mostly lesser than the train data.

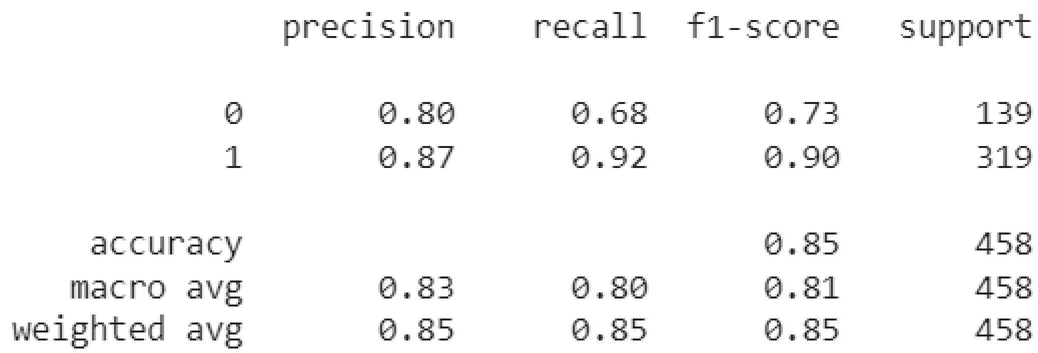
CLASSIFICATION REPORT

TRAIN DATA



```
              precision    recall  f1-score   support

           0       0.75      0.65      0.70       323
           1       0.86      0.91      0.88       744

    accuracy                           0.83      1067
   macro avg       0.81      0.78      0.79      1067
weighted avg       0.83      0.83      0.83      1067
```

TEST DATA



```
              precision    recall  f1-score   support

           0       0.80      0.68      0.73       139
           1       0.87      0.92      0.90       319

    accuracy                           0.85       458
   macro avg       0.83      0.80      0.81       458
weighted avg       0.85      0.85      0.85       458
```

Test data has slightly better precision score and f1 score than train data.

There is minor overfitting with the test data, but it does not hamper the performance of the test data in any way.

LDA

One of the mandatory preprocessing to be done to create a LDA model is scaling the independent variables. As mentioned before, Independent variables are mostly ordinal or categorical. Thus, scaling them would affect the predictability of the model. I decided to create two models, once scaled and another that is not scaled, to compare the model and choose the best model.

Here are the results:

LDA FORMULA:

SCALED MODEL

```
Y= 1.458 - 0.373*age + 0.371*economic.cond.national +
0.067*economic.cond.household + 0.899*Blair - 1.185*Hague -
0.761*Europe - 0.540*political.knowledge - 0.034*gender
```

UNSCALED MODEL

```
Y= 8.924 - 0.0633*age + 0.728*economic.cond.national +
0.184*economic.cond.household + 1.922*Blair - 2.234*Hague -
0.526*Europe - 1.289*political.knowledge - 0.043*gender
```

According to the scaled model, Blair, Hague and Europe are the most important variables in predicting the dependent variables. According to the unscaled model, economic.cond.national, Blair, Hague and Political knowledge are the most important variables in predicting dependent variables.

CONFUSION MATRIX

SCALED DATA

```
[315, 147]
[106, 957]
```

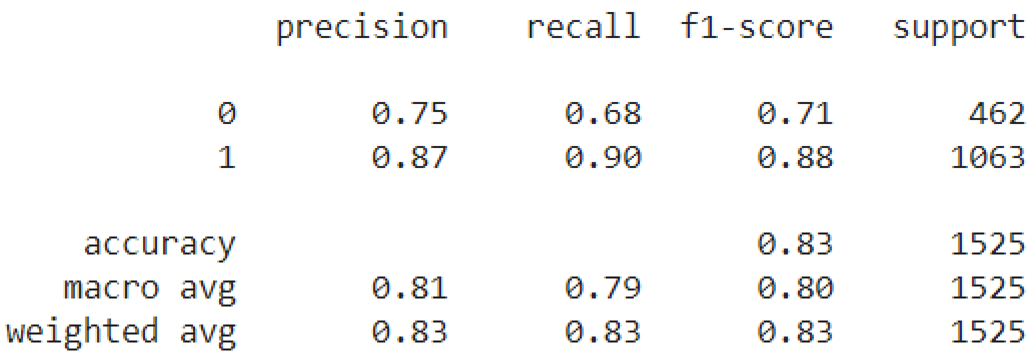
UNSCALED MODEL

```
[ 418, 3]
[ 16, 1088]
```

Scaled data has high truly negative predictions followed by true positives. Same is the case with unscaled data. However, scaled data has more false predictions than unscaled data.

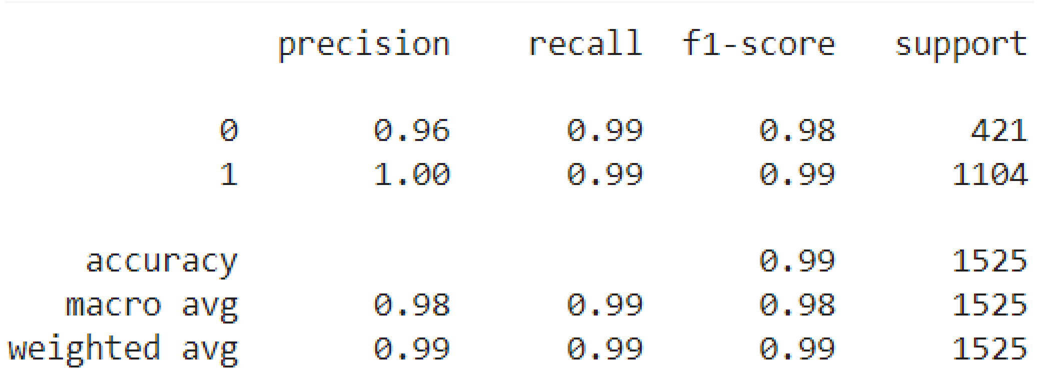
CLASSIFICATION REPORT

SCALED DATA



```
              precision    recall  f1-score   support

           0       0.75      0.68      0.71       462
           1       0.87      0.90      0.88      1063

    accuracy                           0.83      1525
   macro avg       0.81      0.79      0.80      1525
weighted avg       0.83      0.83      0.83      1525
```

UNSCALED DATA



```
              precision    recall  f1-score   support

           0       0.96      0.99      0.98       421
           1       1.00      0.99      0.99      1104

    accuracy                           0.99      1525
   macro avg       0.98      0.99      0.98      1525
weighted avg       0.99      0.99      0.99      1525
```

Unscaled data has overall better precision and recall score. The accuracy of the unscaled data is 99% and scaled data is 83%. Therefore, we can certainly say that **unscaled data has performed better than the scaled data.**

1.5) Apply KNN Model and Naïve Bayes Model (2pts). Interpret the inferences of each model (2 pts). Successful implementation of each model. Logical reasons 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-neighbours classifier command was downloaded from the sklearn library. The tuning parameter of the model, n_neighbours was set at the default setting of 5. The model also had weight at distance, i.e. to set weight to the points based on how distant it is from the main point. The random state was set at 1, so whenever the code is implemented it will give us the same split.

MODEL SCORE

TRAIN MODEL: `0.999`

TEST MODEL: `0.777`

There is some **drastic case of underfitting with the test model.**

CONFUSION MATRIX

TEST MODEL

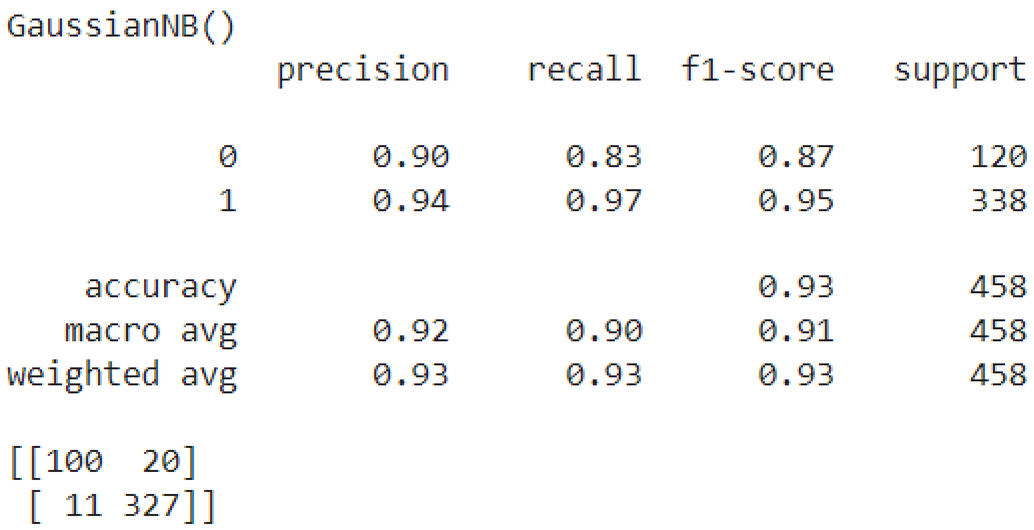
```
[107 13]
[ 23 315]
```

The test model had predicted the true negatives the most followed by true positives.

NAIVE BAYES MODEL

Gaussian_NB was downloaded from sklearn library to create the naive Bayes model. The seed function was set at 7 to reproduce the same random numbers at multiple runs of the code.

CLASSIFICATION TABLE AND THE CONFUSION MATRIX



```
GaussianNB()
              precision    recall  f1-score   support

           0       0.90      0.83      0.87       120
           1       0.94      0.97      0.95       338

    accuracy                           0.93       458
   macro avg       0.92      0.90      0.91       458
weighted avg       0.93      0.93      0.93       458

[[100  20]
 [ 11 327]]
```

The test model has high precision and recall scores. However, the recall score is better than the precision score. The model also has an overall accuracy score of 93%. As evidenced by the recall score, The model predicted better true negatives than the true positives.

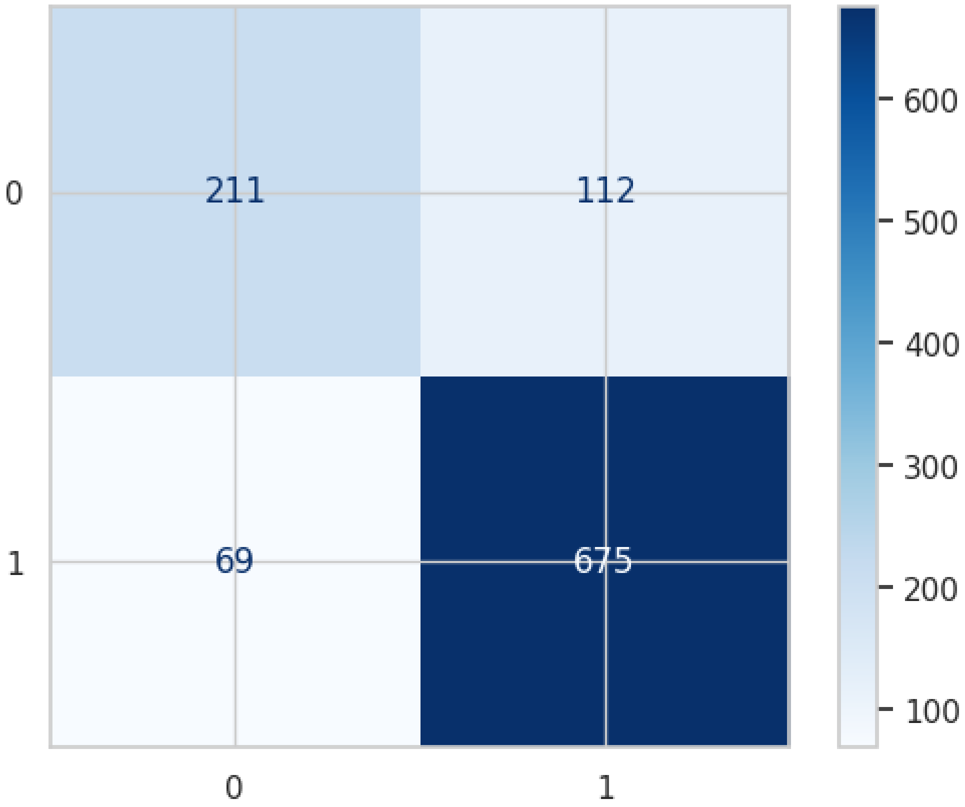
1.6) Model Tuning (4 pts) , Bagging (1.5 pts) and Boosting (1.5 pts). Apply grid search on each model (include all models) and make models on best_params. Compare and comment on performances of all. Comment on feature importance if applicable. Successful implementation of both algorithms along with inferences and comments on the model performances.

GRIDSEARCHCV - LOGISTIC REGRESSION MODEL

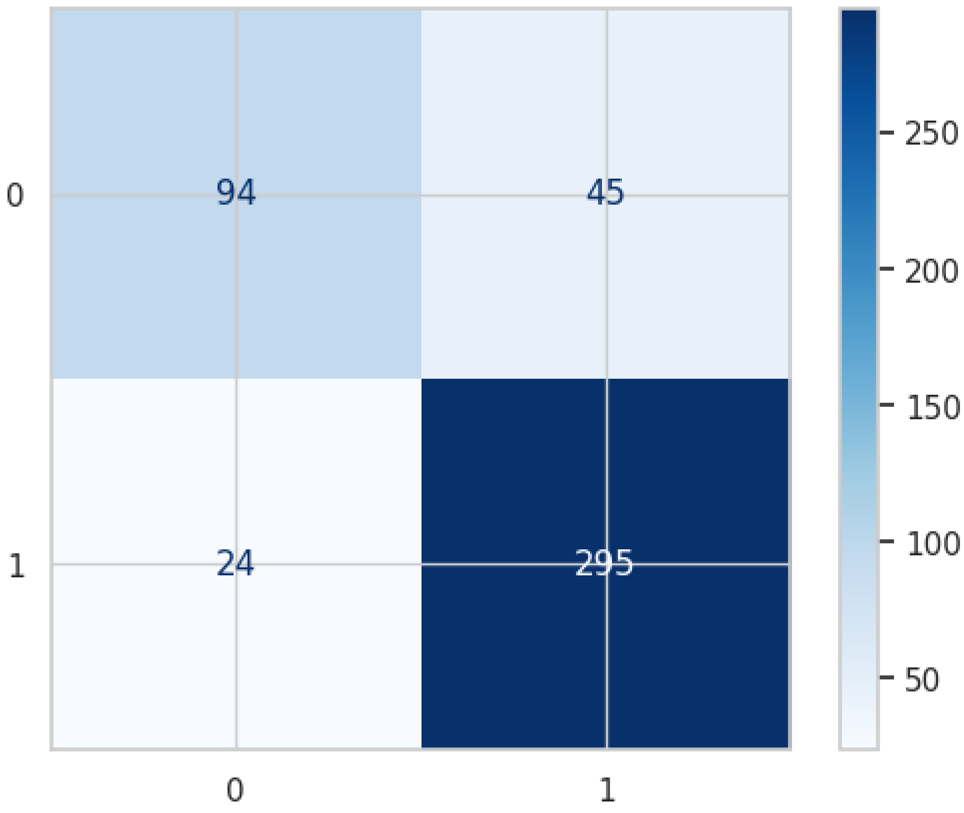
GridsearchCV was downloaded from sklearn to find the best parameters. While fitting the model, the cv parameter was set at 3, n_jobs were set at -1 to reduce the total number of functions running simultaneously for faster output. Best parameters according to Gridsearch were `(max_iter=10000, n_jobs=2, solver='sag', tol=1e-05, verbose=True).` Following are the results after running the model on best estimators:

CONFUSION MATRIX

TRAIN MODEL

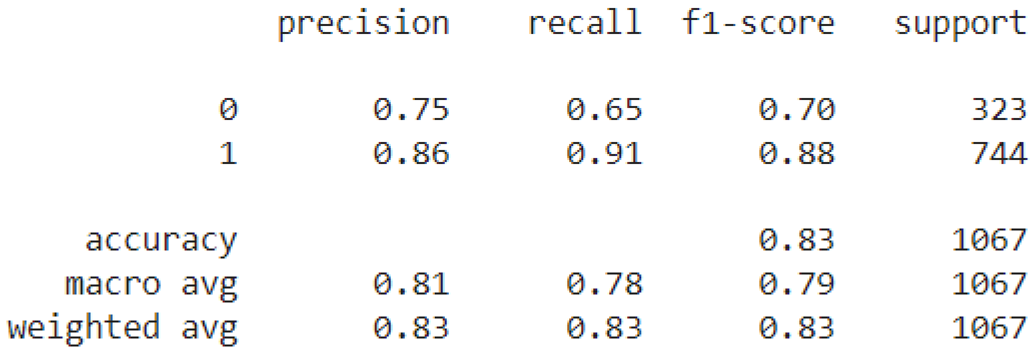


TEST MODEL



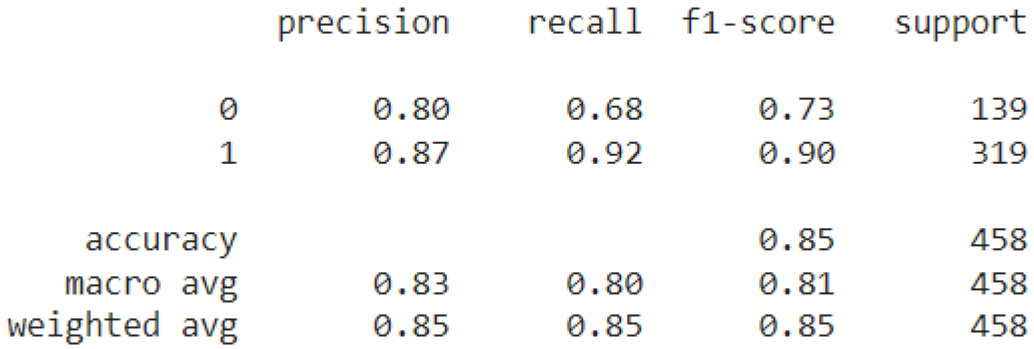
Both models are good at predicting true negatives followed by true positives.

CLASSIFICATION REPORT
TRAIN MODEL



```
              precision    recall  f1-score   support

           0       0.75      0.65      0.70       323
           1       0.86      0.91      0.88       744

    accuracy                           0.83      1067
   macro avg       0.81      0.78      0.79      1067
weighted avg       0.83      0.83      0.83      1067
```

TEST MODEL



```
              precision    recall  f1-score   support

           0       0.80      0.68      0.73       139
           1       0.87      0.92      0.90       319

    accuracy                           0.85       458
   macro avg       0.83      0.80      0.81       458
weighted avg       0.85      0.85      0.85       458
```

The test model slightly performed better than the train model, including precision, recall and overall accuracy score.

GRIDSEARCHCV - KNN MODEL
The hyper parameter CV was set at 10 and scoring was set at accuracy. The best parameters according to the gridsearch CV are `{'n_neighbors': 13}.`
The best estimators according to the gridsearchCV are `(n_neighbors=13, weights='distance')`
Following are the accuracy scores as the result of the models built on best parameters.

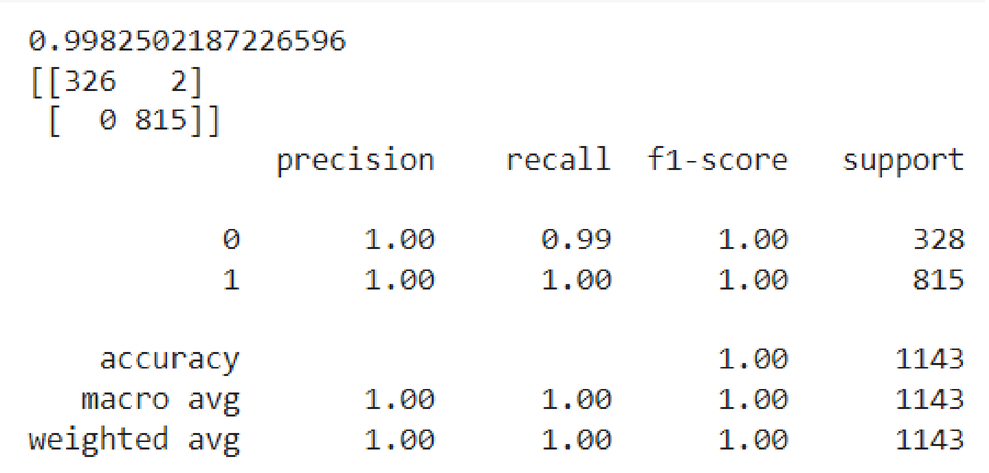
```
Accuracy for our training dataset with tuning is : 81.36%
Accuracy for our testing dataset with tuning is : 100.00%
```

There seems to be a significant overfitting in the test model compared to the train model.

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.

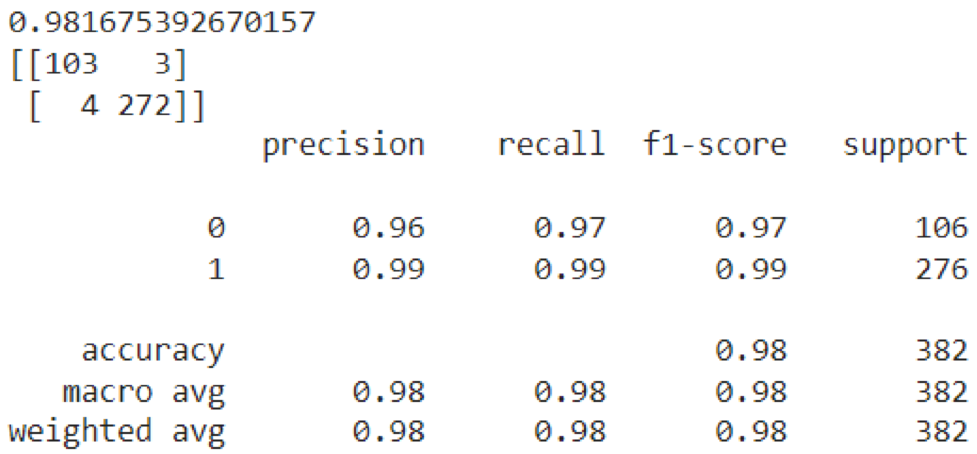
BOOSTING
ADABOOSTING
The adaboosting model is a type of boosting model. In this particular case it was built on n_estimators at 100.

MODEL SCORE, CONFUSION MATRIX AND CLASSIFICATION REPORT
TRAIN MODEL



```
0.9982502187226596
[[326   2]
 [  0 815]]
              precision    recall  f1-score   support

           0       1.00      0.99      1.00       328
           1       1.00      1.00      1.00       815

    accuracy                           1.00      1143
   macro avg       1.00      1.00      1.00      1143
weighted avg       1.00      1.00      1.00      1143
```

TEST MODEL

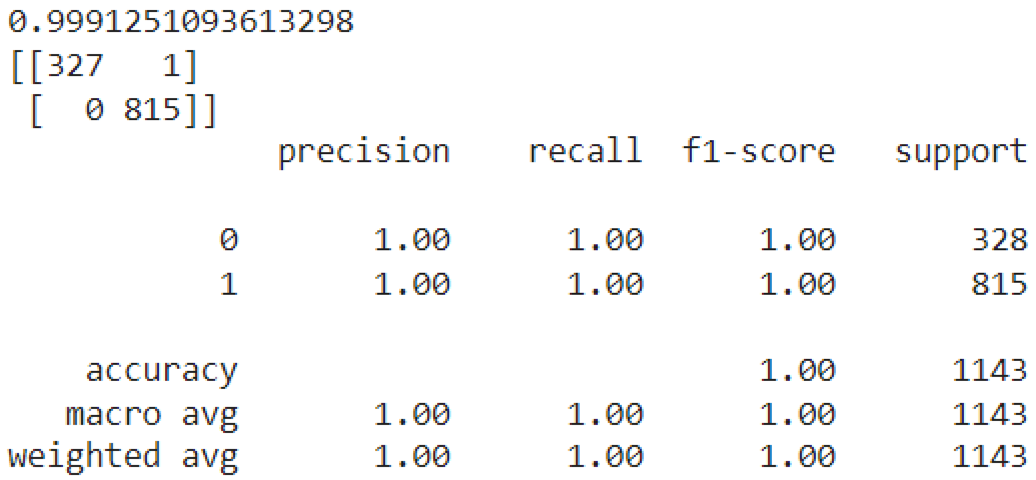


```
0.981675392670157
[[103   3]
 [  4 272]]
              precision    recall  f1-score   support

           0       0.96      0.97      0.97       106
           1       0.99      0.99      0.99       276

    accuracy                           0.98       382
   macro avg       0.98      0.98      0.98       382
weighted avg       0.98      0.98      0.98       382
```

Based on the overall model score, there seems to be slight underfitting in the test model. But based on the rule of thumb, it is not a major issue.
Train model, when compared to the test model in terms of ratio, is better at predicting true positives and true negatives.
There are only slight differences in the test model predictability compared to the train model.
Thus, we can say that the test model has solid predictions in comparison to the train model.

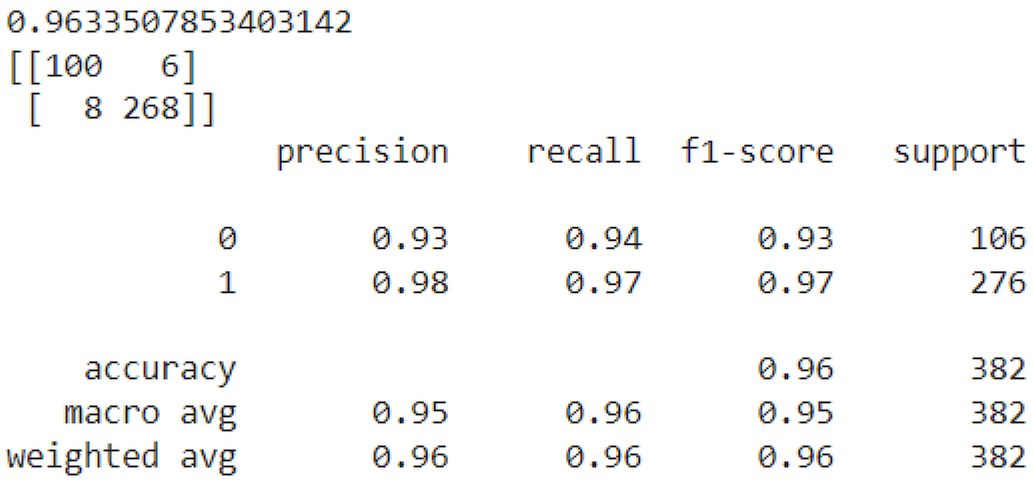
GRADIENT BOOSTING
Gradient boosting is another type of model boosting classifier. Here are the results of Gradient boosting model:

MODEL SCORE, CONFUSION MATRIX AND CLASSIFICATION REPORT
TRAIN MODEL



```
0.9991251093613298
[[327   1]
 [  0 815]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       328
           1       1.00      1.00      1.00       815

    accuracy                           1.00      1143
   macro avg       1.00      1.00      1.00      1143
weighted avg       1.00      1.00      1.00      1143
```

TEST MODEL



```
0.9633507853403142
[[100   6]
 [  8 268]]
              precision    recall  f1-score   support

           0       0.93      0.94      0.93       106
           1       0.98      0.97      0.97       276

    accuracy                           0.96       382
   macro avg       0.95      0.96      0.95       382
weighted avg       0.96      0.96      0.96       382
```

Similar to adaboosting model, there is **slight underfitting in the gradient boosting test model.** The test model has slightly lesser prediction score and recall score as evidenced by the confusion matrix where ratio-wise, the train model has better true predictions than test model. Even the overall accuracy score is higher with the train model than the test model. Similar to adaboosting, this test model is sufficient in giving good predictions of the dependent variable.

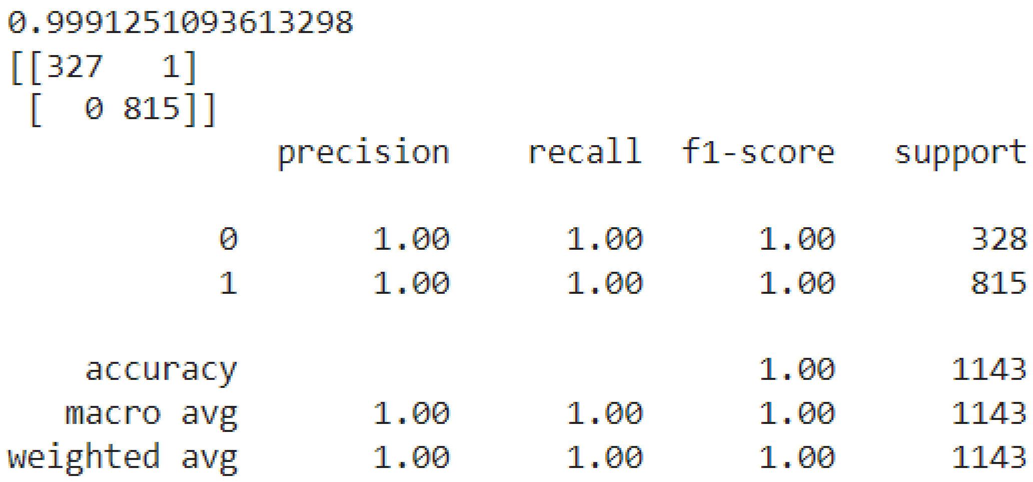
GRIDSEARCHCV - GRADIENT BOOSTING

With gridsearch the best parameters were found to be `'learning_rate': 0.05, 'n_estimators': 250.`

Following are the result of the tuned model:

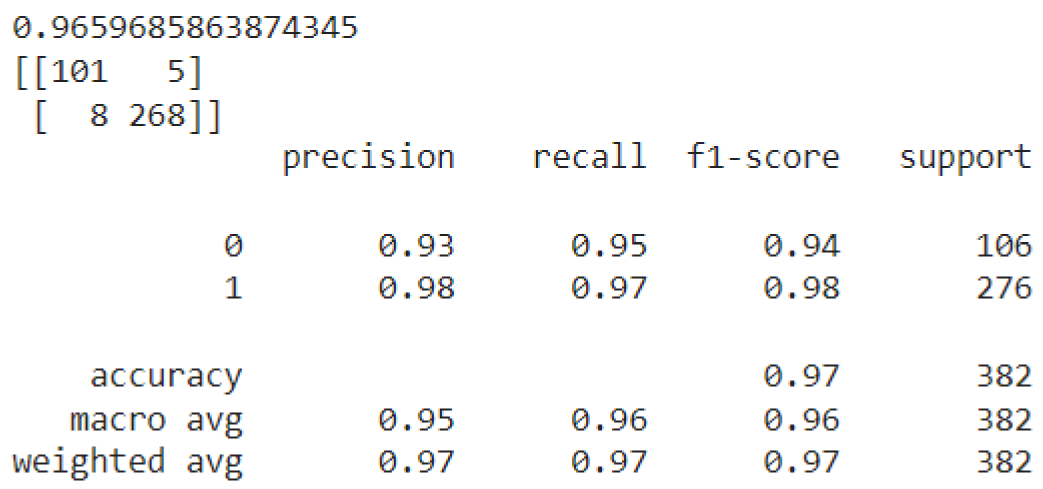
MODEL SCORE, CONFUSION MATRIX AND CLASSIFICATION REPORT

TRAIN MODEL



```
0.9991251093613298
[[327   1]
 [  0 815]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       328
           1       1.00      1.00      1.00       815

    accuracy                           1.00      1143
   macro avg       1.00      1.00      1.00      1143
weighted avg       1.00      1.00      1.00      1143
```

TEST MODEL

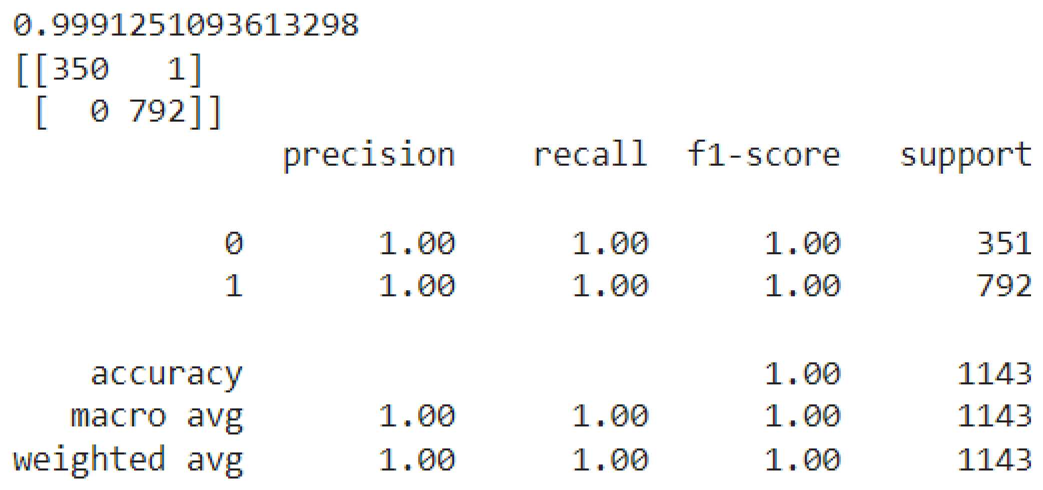


```
0.9659685863874345
[[101   5]
 [  8 268]]
              precision    recall  f1-score   support

           0       0.93      0.95      0.94       106
           1       0.98      0.97      0.98       276

    accuracy                           0.97       382
   macro avg       0.95      0.96      0.96       382
weighted avg       0.97      0.97      0.97       382
```

Similar to the previous model, there is **slight underfitting in the tuned gradient boosting test model.** The test model has slightly lesser prediction score and recall score as evidenced by the confusion matrix where ratio-wise, the train model has better true predictions than test model. Even the overall accuracy score is higher with the train model than the test model. Similar to the previous model, **this test model is sufficient in giving good predictions of the dependent variable.**

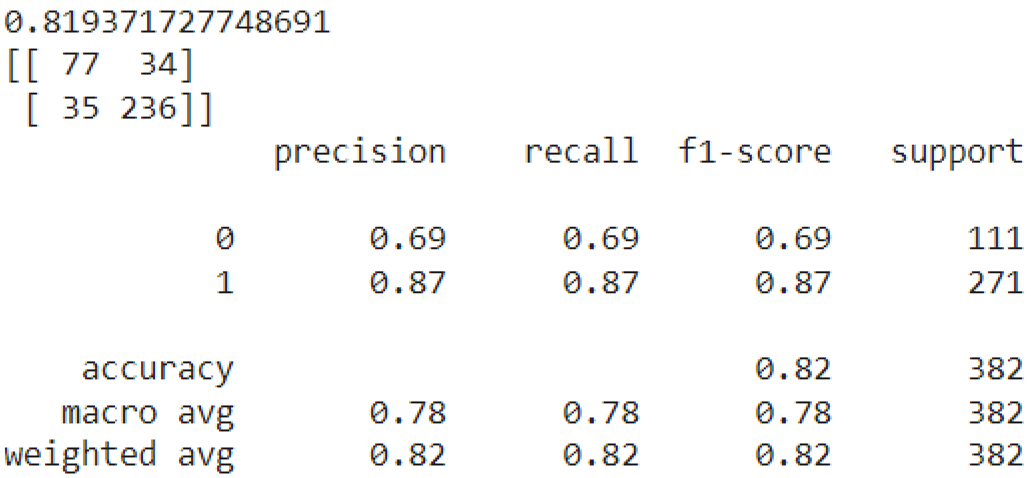
BAGGING CLASSIFIER
A bagging model was created with base set at decision tree classifier and n_estimators set at 100. Following are the results:

MODEL ACCURACY, CONFUSION MATRIX AND CLASSIFICATION REPORT
TRAIN DATA



```
0.9991251093613298
[[350   1]
 [  0 792]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       351
           1       1.00      1.00      1.00       792

    accuracy                           1.00      1143
   macro avg       1.00      1.00      1.00      1143
weighted avg       1.00      1.00      1.00      1143
```

TEST DATA



```
0.819371727748691
[[ 77  34]
 [ 35 236]]
              precision    recall  f1-score   support

           0       0.69      0.69      0.69       111
           1       0.87      0.87      0.87       271

    accuracy                           0.82       382
   macro avg       0.78      0.78      0.78       382
weighted avg       0.82      0.82      0.82       382
```

The test data has a significantly lower model accuracy score. While the train data has perfect precision, recall and accuracy scores, test data scores are much far behind. Similarly, train data is much better at predicting true values than the test model. Therefore,**we can certainly say that there is a case of underfitting.**

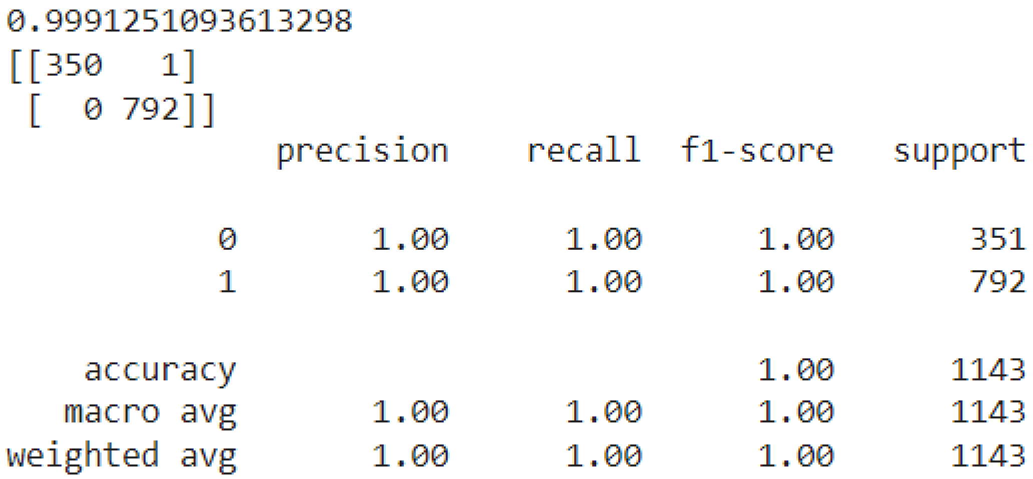
GRIDSEARCHCV - BAGGING CLASSIFIER

After using GridsearchCV on the bagging classifier model, the best estimators were `n_estimators=50, n_jobs=100, random_state=1, verbose=True.`

Following are the results:

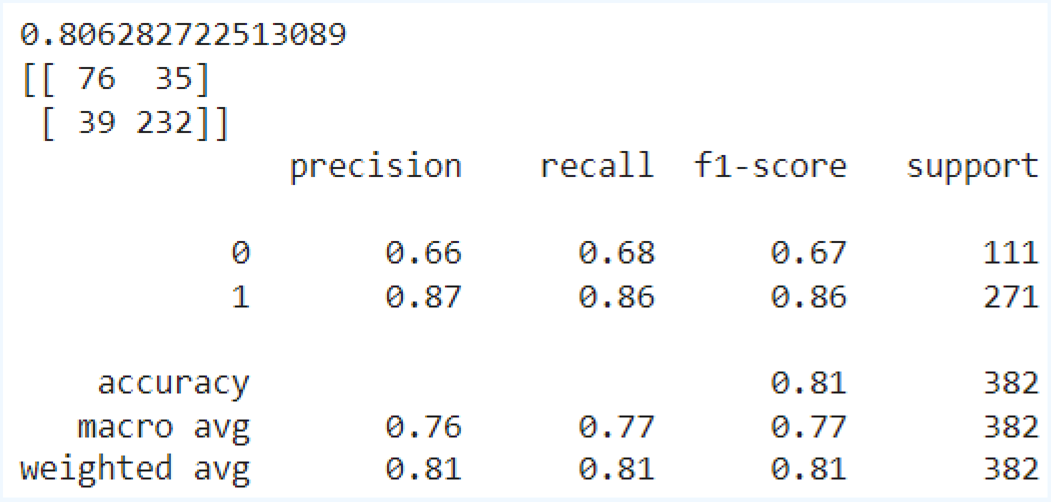
MODEL ACCURACY, CONFUSION MATRIX AND CLASSIFICATION REPORT

TRAIN DATA



```
0.9991251093613298
[[350   1]
 [  0 792]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       351
           1       1.00      1.00      1.00       792

    accuracy                           1.00      1143
   macro avg       1.00      1.00      1.00      1143
weighted avg       1.00      1.00      1.00      1143
```

TEST DATA



```
0.806282722513089
[[ 76  35]
 [ 39 232]]
              precision    recall  f1-score   support

           0       0.66      0.68      0.67       111
           1       0.87      0.86      0.86       271

    accuracy                           0.81       382
   macro avg       0.76      0.77      0.77       382
weighted avg       0.81      0.81      0.81       382
```

After model tuning, the test model has performed worse than the test model before model tuning. Since, **overall model accuracy and predictability of the test model has slightly reduced.**

MODEL COMPARISONS AND THE BEST MODEL

As the first step of narrowing down the best models, let us find the models that have good test models, i.e. The test models have good accuracy and predictability, where the test models are closest to the train model and have minimal underfitting or overfitting issues. They are:

- LOGISTIC REGRESSION / LOGISTIC REGRESSION (GRIDSEARCH-CV)
- LDA - UNSCALED
- NAIVE BAYES MODEL
- ADABOOSTING
- GRADIENT BOOSTING / GRADIENT BOOSTING (GRIDSEARCH-CV)

Another inference to be noted is that all these test models had great recall value. Since, Recall refers to the percentage of total relevant results correctly classified by the algorithm, all these models are good at predicting the dependent variable.

However, I believe that the **LDA-Unscaled** model is the best model to predict the dependent variable. This is because, given the business question, the LDA model has better interpretability and great results to implement business strategies.

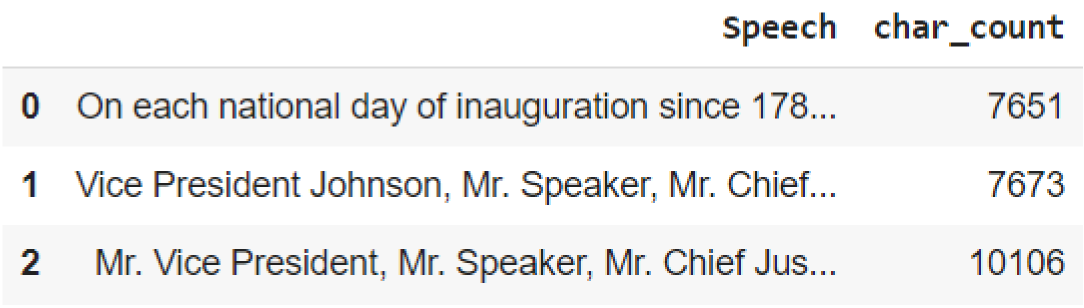
1.8) Based on your analysis and working on the business problem, detail out appropriate insights and recommendations to help the management solve the business objective. There should be at least 3-4 Recommendations and insights in total. Recommendations should be easily understandable and business specific, students should not give any technical suggestions. Full marks should only be allotted if the recommendations are correct and business specific.

- The more ‘Eurosceptic’ a person’s sentiment is, the more likely they might vote for the conservative party.
- It is more likely for an individual to vote for the Labour party if they have a high accord with the current economic conditions.
- The voter’s positive opinion of a party representative has high correlation with that person voting for that party.

- Age and Gender are not always a great indicator of who they will vote for.

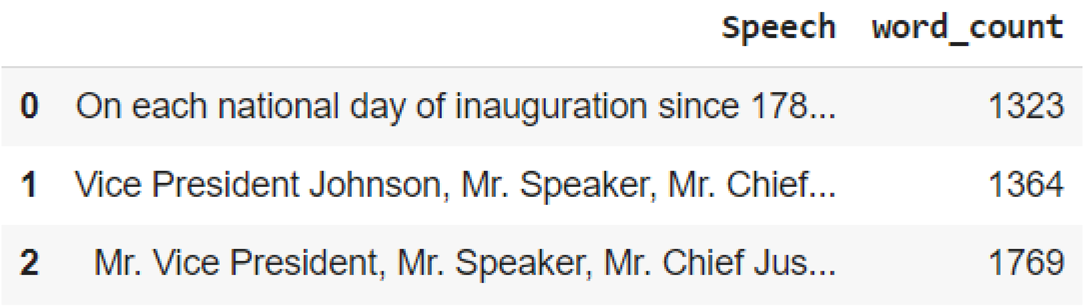
2.1) Find the number of characters, words and sentences for the mentioned documents. (Hint: use .words(), .raw(), .sent() for extracting counts)

NUMBER OF CHARACTERS



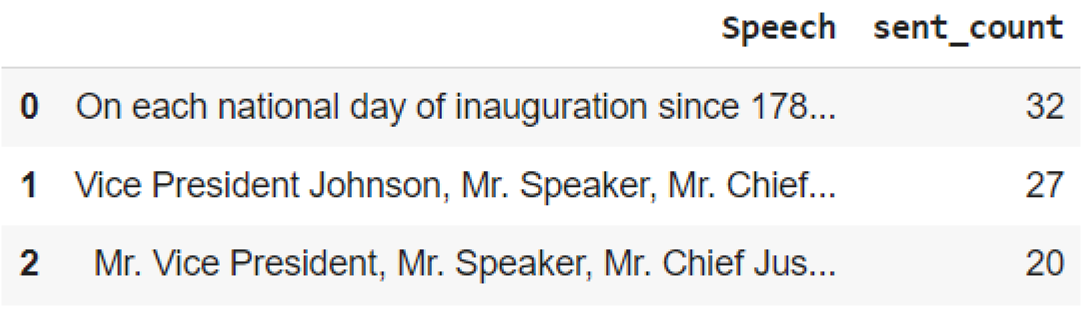
	Speech	char_count
0	On each national day of inauguration since 178...	7651
1	Vice President Johnson, Mr. Speaker, Mr. Chief...	7673
2	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	10106

NUMBERS OF WORDS



	Speech	word_count
0	On each national day of inauguration since 178...	1323
1	Vice President Johnson, Mr. Speaker, Mr. Chief...	1364
2	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	1769

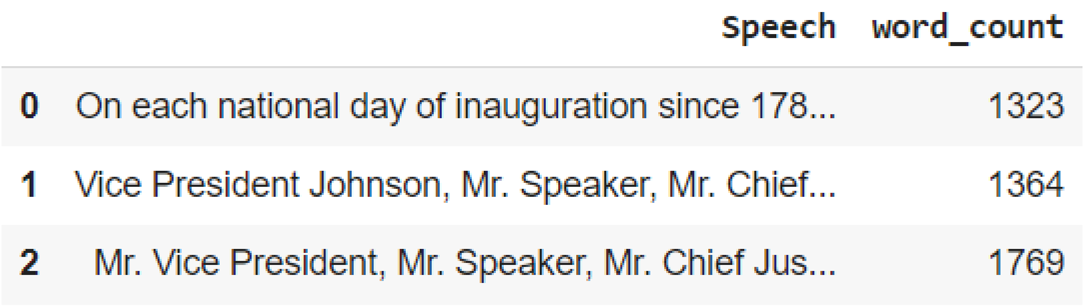
NUMBER OF SENTENCES



	Speech	sent_count
0	On each national day of inauguration since 178...	32
1	Vice President Johnson, Mr. Speaker, Mr. Chief...	27
2	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	20

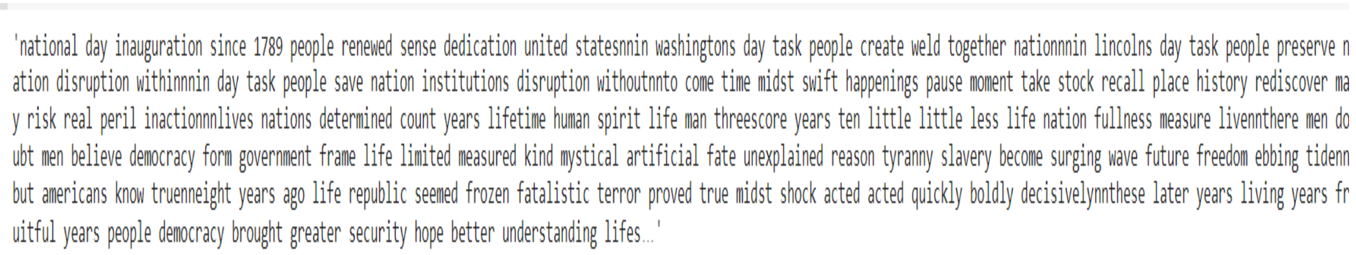
2.2) Remove all the stopwords from the three speeches. Show the word count before and after the removal of stopwords. Show a sample sentence after the removal of stopwords.

WORD COUNT BEFORE THE REMOVAL OF STOPWORDS



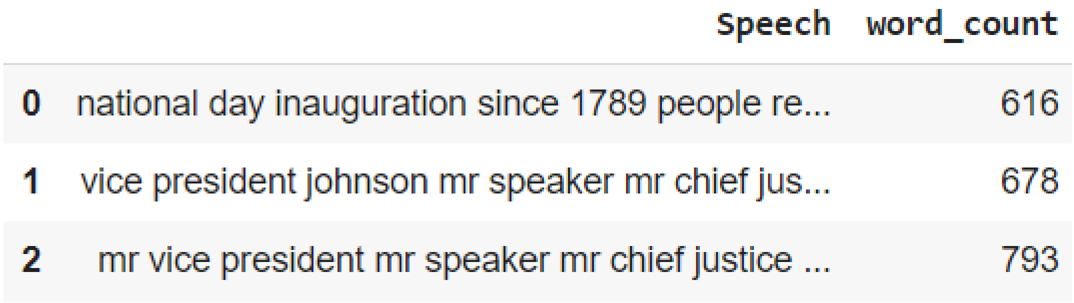
	Speech	word_count
0	On each national day of inauguration since 178...	1323
1	Vice President Johnson, Mr. Speaker, Mr. Chief...	1364
2	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	1769

SAMPLE SENTENCES BEFORE REMOVAL OF STOPWORDS



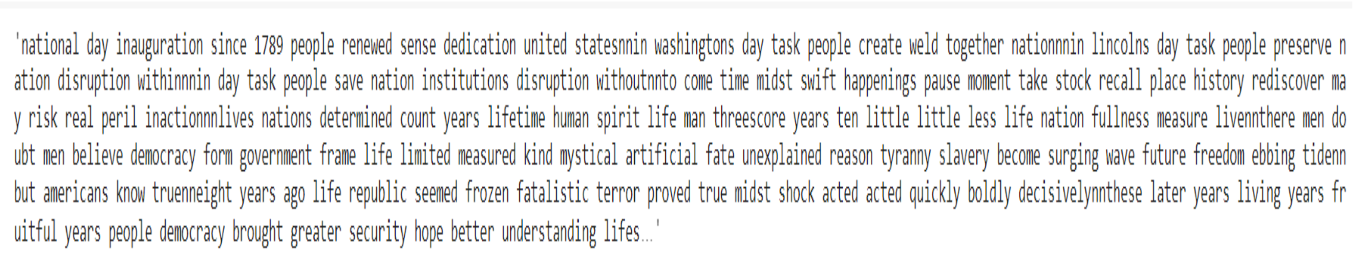
```
'national day inauguration since 1789 people renewed sense dedication united statesnnin washingtons day task people create weld together nationnnin lincolns day task people preserve n
ation disruption withinnnin day task people save nation institutions disruption withoutnnto come time midst swift happenings pause moment take stock recall place history rediscover ma
y risk real peril inactionnnlives nations determined count years lifetime human spirit life man threescore years ten little little less life nation fullness measure livennthere men do
ubt men believe democracy form government frame life limited measured kind mystical artificial fate unexplained reason tyranny slavery become surging wave future freedom ebbing tidenn
but americans know truenneight years ago life republic seemed frozen fatalistic terror proved true midst shock acted acted quickly boldly decisivelynnthese later years living years fr
uitful years people democracy brought greater security hope better understanding lifes...'
```

WORD COUNT AFTER THE REMOVAL OF STOPWORDS



	Speech	word_count
0	national day inauguration since 1789 people re...	616
1	vice president johnson mr speaker mr chief jus...	678
2	mr vice president mr speaker mr chief justice ...	793

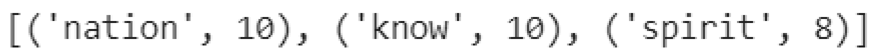
SAMPLE SENTENCES AFTER REMOVAL OF STOPWORDS



```
'national day inauguration since 1789 people renewed sense dedication united statesnnin washingtons day task people create weld together nationnnin lincolns day task people preserve n
ation disruption withinnnin day task people save nation institutions disruption withoutnnto come time midst swift happenings pause moment take stock recall place history rediscover ma
y risk real peril inactionnnlives nations determined count years lifetime human spirit life man threescore years ten little little less life nation fullness measure livennthere men do
ubt men believe democracy form government frame life limited measured kind mystical artificial fate unexplained reason tyranny slavery become surging wave future freedom ebbing tidenn
but americans know truenneight years ago life republic seemed frozen fatalistic terror proved true midst shock acted acted quickly boldly decisivelynnthese later years living years fr
uitful years people democracy brought greater security hope better understanding lifes...'
```

2.3) Which word occurs the most number of times in his inaugural address for each president? Mention the top three words. (after removing the stopwords)

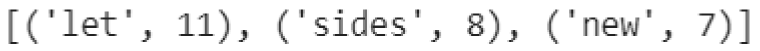
THE MOST RECURRENT WORDS IN ROOSEVELT'S SPEECH



```
[('nation', 10), ('know', 10), ('spirit', 8)]
```

'Nation', 'know' and 'Spirit' are the most recurrent words in his speech.

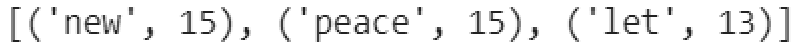
THE MOST RECURRENT WORDS IN KENNEDY’S SPEECH



```
[('let', 11), ('sides', 8), ('new', 7)]
```

‘Let’, ‘sides’ and ‘new’ are the most recurrent words in his speech.

THE MOST RECURRENT WORDS IN NIXON’S SPEECH

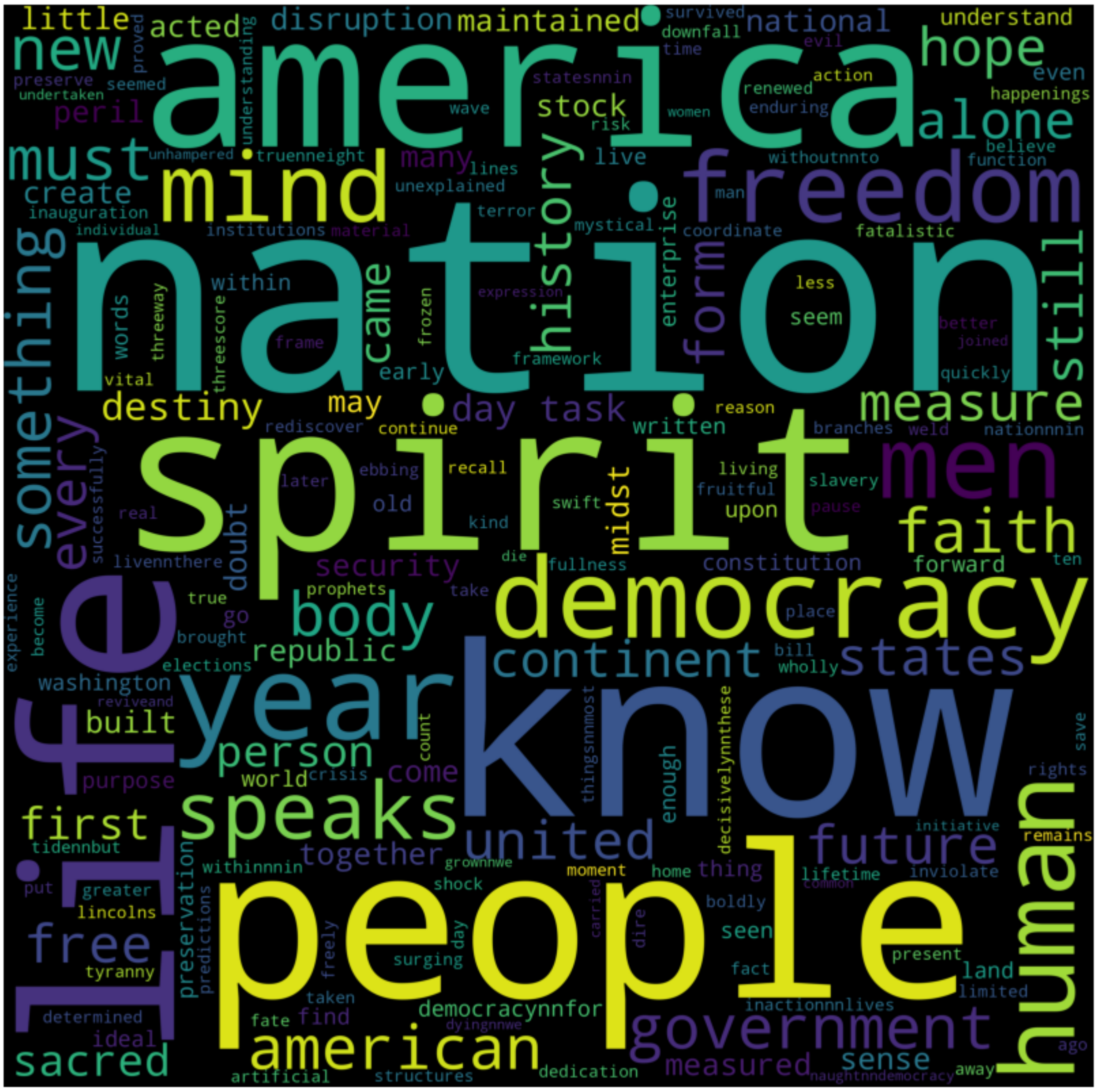


```
[('new', 15), ('peace', 15), ('let', 13)]
```

‘New’, ‘peace’ and ‘let’ are the most recurrent words in his speech.

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

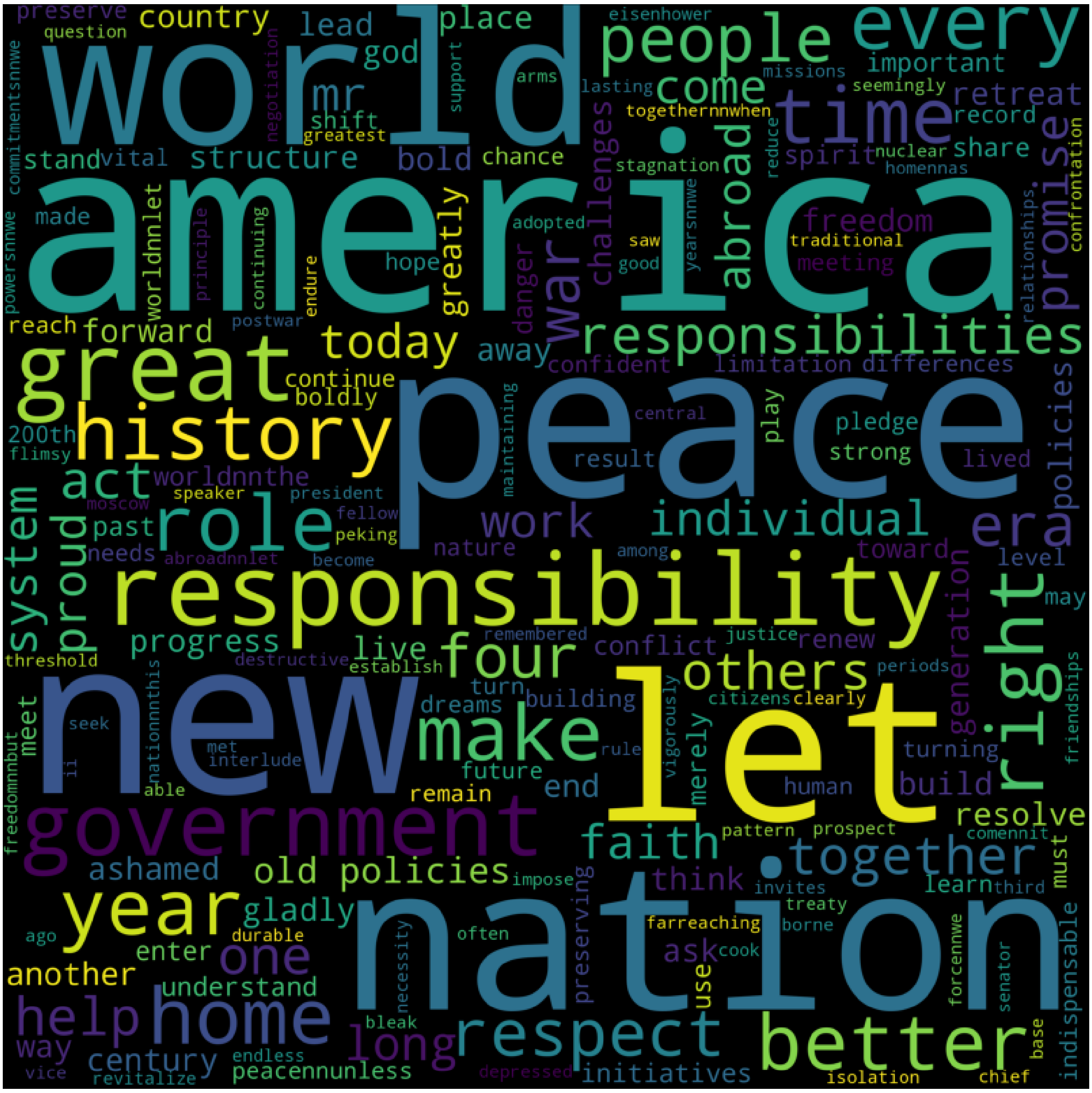
WORD CLOUD FOR ROOSEVELT SPEECH



WORD CLOUD FOR KENNEDY SPEECH

let
sides
power
pledge
citizens
nation
new
world
ask
president
freedom
fellow
first
generation
free
form
always
forebears
united
wishes
join
balance
dare
assure
finished
well
burden
human
bear
hand
call
control
loyalty
observe
tempered
eisenhower
faithful
host
undoing
alike
heirs
support
price
still
symbolizing
weak
mankind
quarters
absolute
proud
south
heritage
meet
celebration
man
long
godnnwe
good
final
science
welcome
revolutionary
supporting
enemies
word
right
victory
prescribed
change
disease
know
hope
come
origins
party
old
back
seek
reverend
share
truly
place
slow
weakness
fear
hardship
allies
different
together
hard
permit
nixon
americans
never
yet
people
foe
year
merely
best
endeavor
instead
abolish
forget
offer
passed
struggle
explore
doubt
disciplined
bitter
beginning
though
peace
mortal
agonnthe
course
beyond
truman
pay
today
holds
worldnnlet
committed
fought
born
belief
chief
globe
whether
around
anew
little
poverty
revolution
oppose
success
generosity
home
torch
terror
begin
renewal
god
solemn
justice
help
go
every
spiritual
find
history
order
cultural
war
tyranny
unite
forth
sworn
survival
johnson
earth
nearly
problems
far
bring
days
century
life
ancient
three
deeds
mr
end
almighty
libertynnthis
whose
much
unwilling
powerful
country
light
time
signifying
may
arms
ill
oath
speaker
friend
state
clergy
america
witness
instruments
alliance
morennto
issue
vice

WORD CLOUD FOR NIXON SPEECH


preserve
country
lead
place
eisenhower
question
world
people
every
commitmentsnnwe
god
support
arms
missions
important
negotiation
mr
lasting
come
seemingly
retreat
shift
togethernnwhen
time
record
greatest
reduce
confrontation
stand
vital
structure
bold
chance
spirit
nuclear
share
stagnation
homennas
promise
america
made
worldnnlet
principle
continuing
endure
greatly
adopted
challenges
saw
yearsnnwe
abroad
freedom
traditional
good
meeting
hope
war
danger
relationships
powersnnwe
postwar
reach
forward
today
away
responsibilities
great
continue
confident
limitation
differences
boldly
peace
policies
central
play
pledge
200th
maintaining
history
flimsy
result
strong
lived
act
worldnnthe
moscow
speaker
president
fellow
role
past
system
proud
peking
work
individual
era
needs
abroadnnlet
become
nature
among
toward
level
responsibility
may
remembered
generation
justice
progress
renew
threshold
destructive
live
establish
four
conflict
periods
right
nationnnthis
others
new
turn
dreams
building
vigorously
citizens
clearly
friendships
meet
seek
make
freedomnnbut
ii
met
interlude
rule
merely
turning
let
future
able
remain
end
human
build
government
resolve
pattern
prospect
comennit
faith
together
ashamed
old policies
impose
think
invites
must
learn
third
year
gladly
preserving
treaty
durable
ago
often
farreaching
borne
enter
one
necessity
ask
cook
forcennwe
senator
nation
another
use
understand
indispensable
help
home
bleak
way
respect
better
long
base
century
endless
vice
revitalize
peacennunless
depressed
initiatives
isolation
chief