MACHINE LEARNING PROJECT

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

Read the dataset.

We can see below the head of the dataset. The unnamed variable will be removed since it does not provide any inference for our modelling.

	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

Below is the tail of the data set. There are total of 1525 rows. There are 9 variables excluding the unnamed.

	Unnamed: 0	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
1520	1521	Conservative	67	5	3	2	4	11	3	male
1521	1522	Conservative	73	2	2	4	4	8	2	male
1522	1523	Labour	37	3	3	5	4	2	2	male
1523	1524	Conservative	61	3	3	1	4	11	2	male
1524	1525	Conservative	74	2	3	2	4	11	0	female

Below we can see the describe function of the data set.

	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
count	1517	1517.000000	1517.000000	1517.000000	1517.000000	1517.000000	1517.000000	1517.000000	1517
unique	2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2
top	Labour	NaN	NaN	NaN	NaN	NaN	NaN	NaN	female
freq	1057	NaN	NaN	NaN	NaN	NaN	NaN	NaN	808
mean	NaN	54.241266	3.245221	3.137772	3.335531	2.749506	6.740277	1.540541	NaN
std	NaN	15.701741	0.881792	0.931069	1.174772	1.232479	3.299043	1.084417	NaN
min	NaN	24.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.000000	NaN
25%	NaN	41.000000	3.000000	3.000000	2.000000	2.000000	4.000000	0.000000	NaN
50%	NaN	53.000000	3.000000	3.000000	4.000000	2.000000	6.000000	2.000000	NaN
75%	NaN	67.000000	4.000000	4.000000	4.000000	4.000000	10.000000	2.000000	NaN
max	NaN	93.000000	5.000000	5.000000	5.000000	5.000000	11.000000	3.000000	NaN

We can see how many unique values are there for the categorical variables. We can see the mean of age is 54. We can see the minimum age of the voter is 24 and maximum age of the voter is 93.

Descriptive statistics:

The average age of the voters is 54 years.

```
df['age'].mean()
```

```
54.18229508196721
```

Out of the total voters' majority of them looks like they might be voting for the Labour party. Around 70% of the voters might be voting for Labour party and 30% will might vote for Conservatives.

```
Labour          1063
Conservative     462
Name: vote, dtype: int64
```

```
Labour          0.697049
Conservative    0.302951
Name: vote, dtype: float64
```

Out of the total voters' female voters seems to outnumber the male voters. 53% of the voters are female candidates and 47% of them are male voters. So female voters are marginally more than the male voters.

```
female    812
male      713
Name: gender, dtype: int64
```

```
female    0.532459
male      0.467541
Name: gender, dtype: float64
```

The average rating for Blair the Labour party candidate is 4. While for Hague the conservatives party candidate is 2. If the values allotted are from best to worst, then assessment for Blair is better than that of Hague's.

```
df.Blair.mode()
```

```
0    4
dtype: int64
```

```
df['Hague'].mode()
```

```
0    2
dtype: int64
```

There are no null values in the dataset.

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

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)

There are no null values present in the data set.
We can see from the describe function also that all the variable has 1525 entries present.

Data types:

We can see that all numerical variables are correctly assigned as int 64 and categorical variables are as object.
Variables such as economic cond national, economic cond household, Blair, Hague, Europe, political knowledge are all categorical variables in nature, but the data type is in int. so in further process they will be converted to object data type.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1525 entries, 0 to 1524
Data columns (total 9 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   vote                    1525 non-null   object
 1   age                     1525 non-null   int64
 2   economic.cond.national  1525 non-null   int64
 3   economic.cond.household 1525 non-null   int64
 4   Blair                   1525 non-null   int64
 5   Hague                   1525 non-null   int64
 6   Europe                  1525 non-null   int64
 7   political.knowledge     1525 non-null   int64
 8   gender                  1525 non-null   object
dtypes: int64(7), object(2)
memory usage: 107.4+ KB
```

It is also observed that there are no other values or strings present in the variables. From the unique value check we can see that.

```
['Labour' 'Conservative']


[43 36 35 24 41 47 57 77 39 70 66 59 51 79 37 38 53 44 60 56 61 55 62 76
 27 52 48 72 42 54 50 46 33 58 64 32 71 28 34 68 67 88 40 78 65 74 82 49
 84 81 45 69 31 63 89 83 29 92 73 75 26 90 25 80 30 86 85 87 93 91]


[3 4 2 1 5]


[3 4 2 1 5]


[4 5 2 1 3]


[1 4 2 5 3]


[ 2  5  3  4  6 11  1  7  9 10  8]


[2 0 3 1]


['female' 'male']
```

The other categorical variables are converted to object for further process.

```
vote                    object
age                      int64
economic.cond.national  object
economic.cond.household object
Blair                   object
Hague                   object
Europe                  object
political.knowledge     object
Gender_male              uint8
dtype: object
```

Duplicate records and shape:

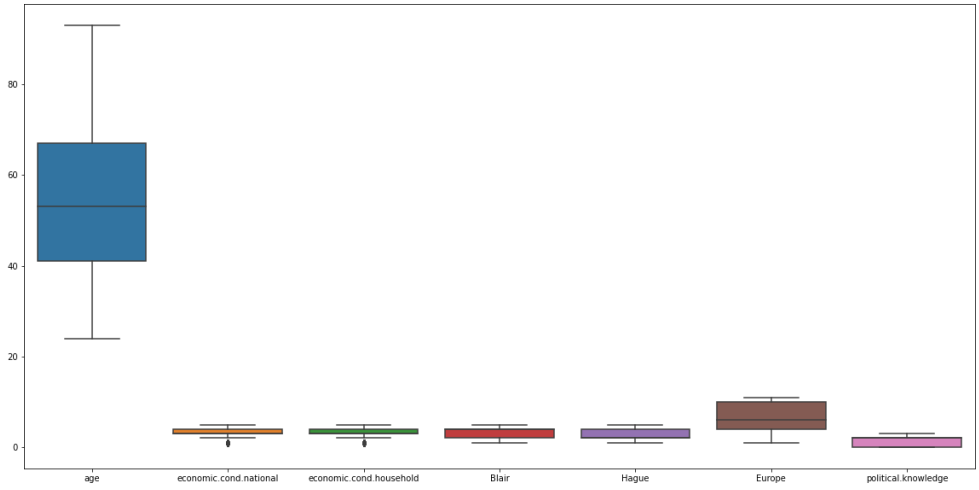
```
Number of duplicate rows = 8
```

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

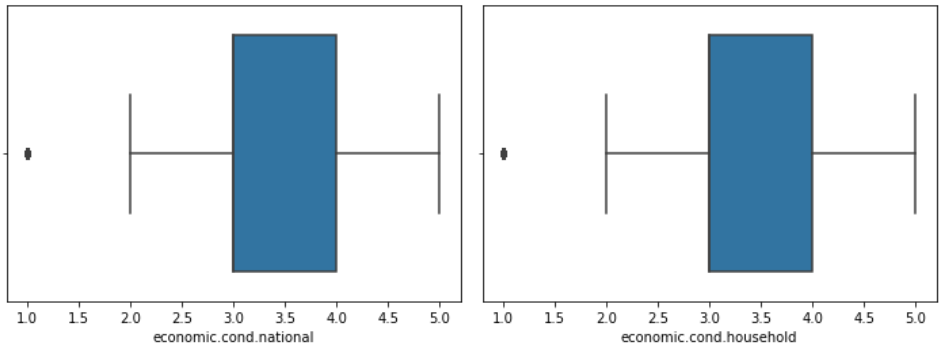
```
Before (1525, 9)
After (1517, 9)
Number of duplicate rows = 0
```

There are 8 records which are duplicates. These records will be removed since they are duplicates. The data set now has 1517 values instead of 1525.

Outliers:

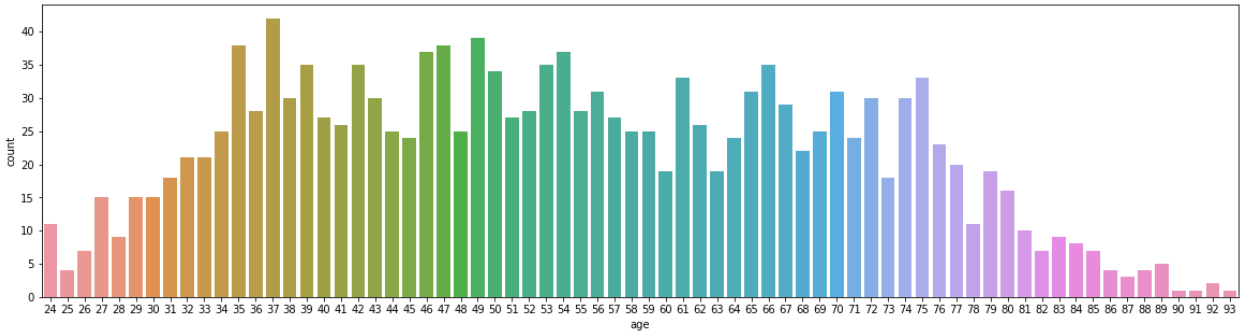


We can see that there are not much of outliers in the variables. We can only see one outlier in economic Cond national and household. And I do not think these can be considered as outliers since they are not continuos and are categorical variables which have meaning. This can be a genuine number also.

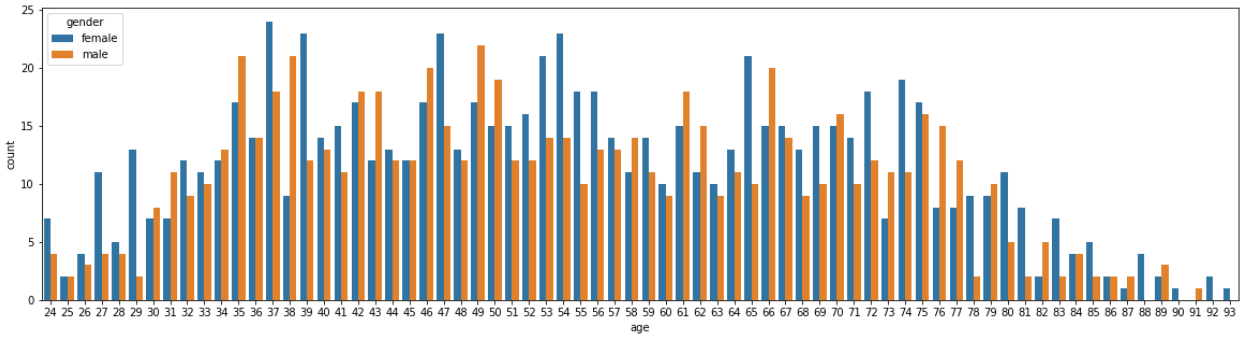


EDA:

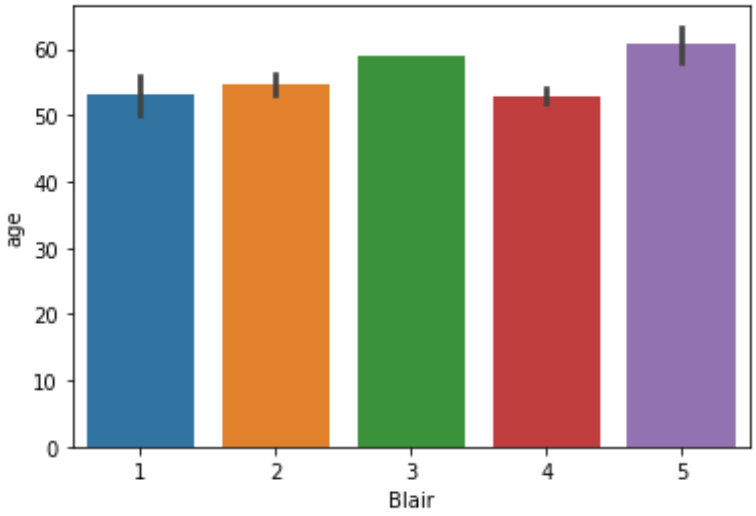
Below we can see the number of people in different age groups. We can see a steep decline in number of voters above the age of 75. Maximum number of voters are around the age of 37 with 50 people in that age group.



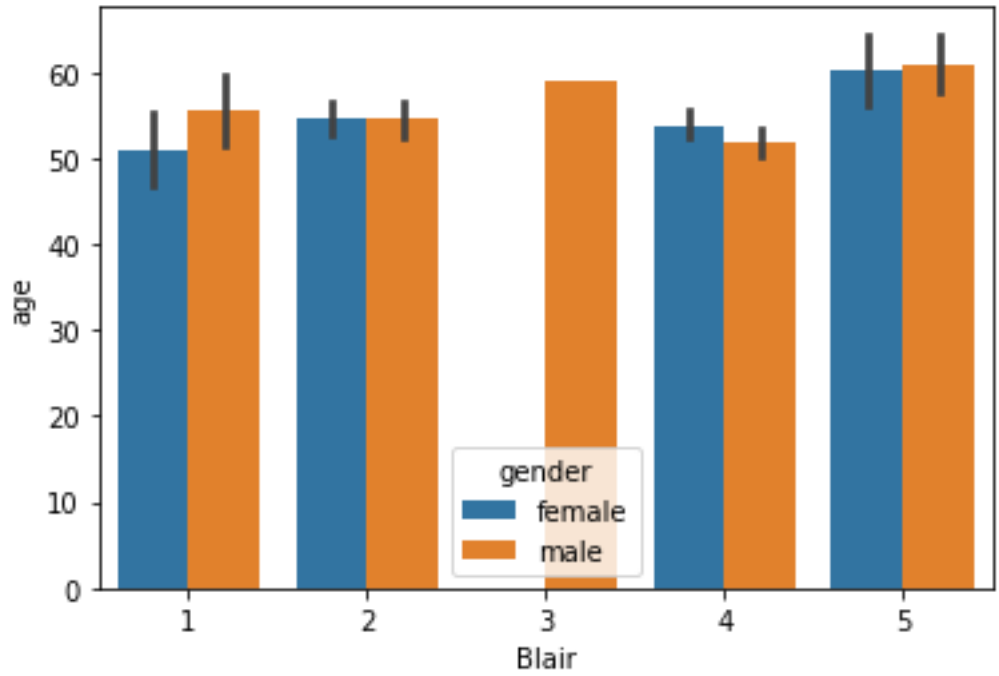
We can see here that in all the ages female voter are more than male voters. Except a few age groups where male voters are high like for age 35,38 male voters are high.



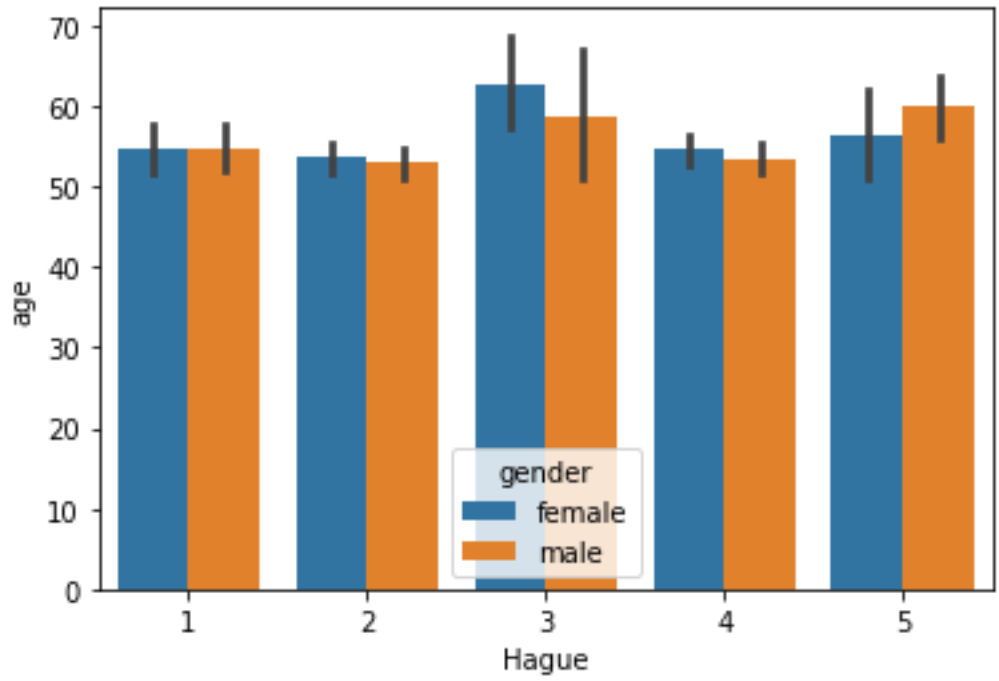
The average assessment for Blair is 3.



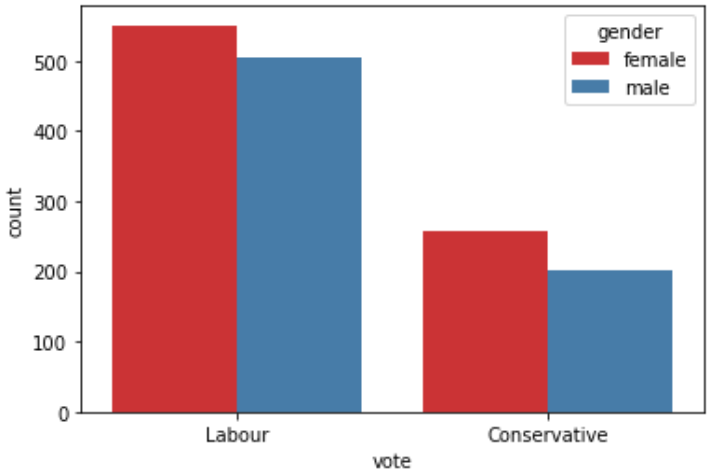
Below show the assessment of Blair and we can see that only male voters have given rating 3 assessment.



However, for Hague the conservative leader females have voted more for assessment of 3. For other ratings it is equally split between male and female.

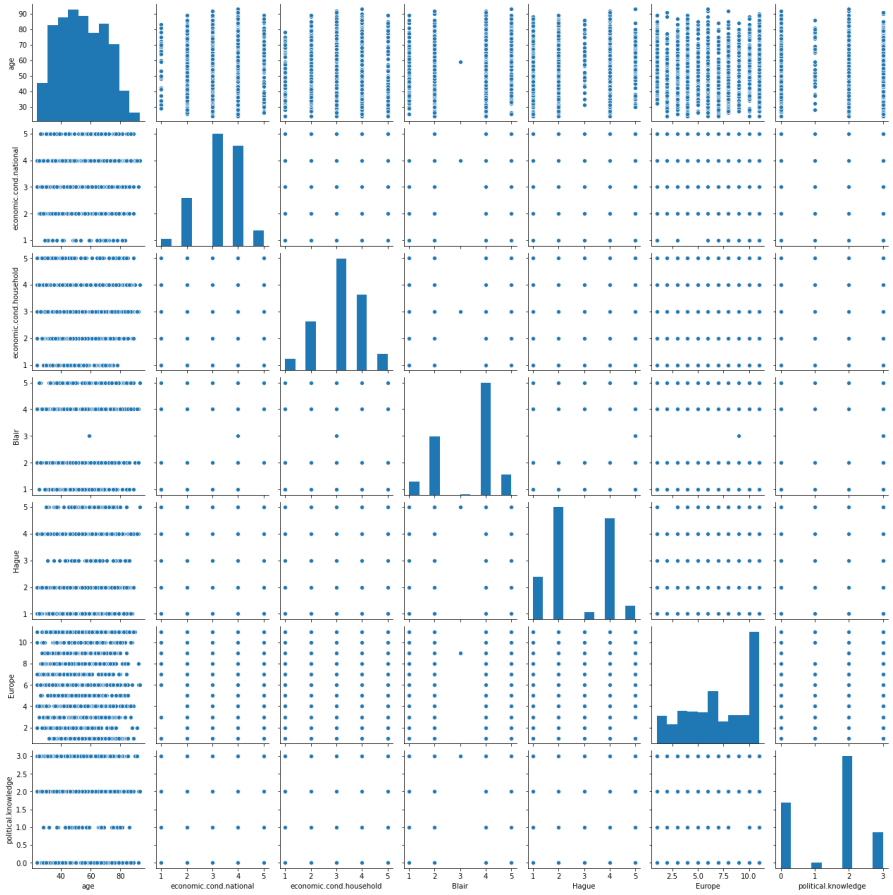


As seen earlier most of the voters might vote for the Labour party and female voters are around 500+ and male voters are around 500.
Conservative party also has more female voters than male.

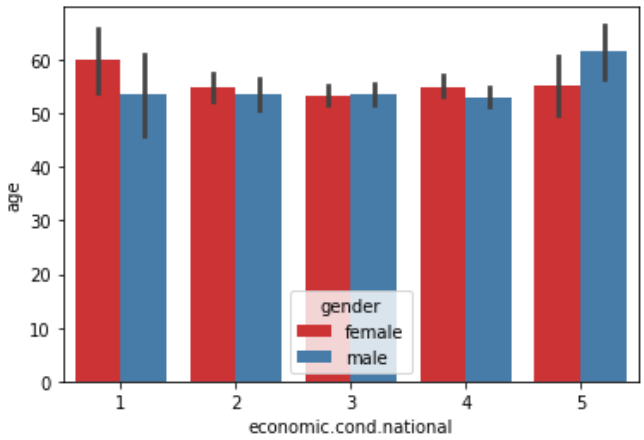


From the co relation heatmap we can see that none of the variable are highly co related here. We can see that Blair and economic cond national are co related.

From the below pair plot we can see that the variables do not have a normal distribution.



Here we can see that Assessment of current national economic conditions male is contributing more to the economic condition at level 5 and female for level 1.



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

The gender columns are converted from categorical to int using one hot encoding. And new variable Gender_male and Gender _female is created. Drop first and gender_female is removed from the list of variables.

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

The dependent variable is also changed from object to numerical of 1 for Labour party and 0 for Conservative party.

Other categorical variables are converted to int using the .codes function

```
feature: vote
[Labour, Conservative]
Categories (2, object): [Conservative, Labour]
[1 0]


feature: economic.cond.national
[3, 4, 2, 1, 5]
Categories (5, object): [1, 2, 3, 4, 5]
[2 3 1 0 4]


feature: economic.cond.household
[3, 4, 2, 1, 5]
Categories (5, object): [1, 2, 3, 4, 5]
[2 3 1 0 4]


feature: Blair
[4, 5, 2, 1, 3]
Categories (5, object): [1, 2, 3, 4, 5]
[3 4 1 0 2]


feature: Hague
[1, 4, 2, 5, 3]
Categories (5, object): [1, 2, 3, 4, 5]
[0 3 1 4 2]



feature: Europe
[2, 5, 3, 4, 6, ..., 1, 7, 9, 10, 8]
Length: 11
Categories (11, object): [1, 10, 11, 2, ..., 6, 7, 8, 9]
[ 3  6  4  5  7  2  0  8 10  1  9]


feature: political.knowledge
[2, 0, 3, 1]
Categories (4, object): [0, 1, 2, 3]
[2 0 3 1]
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1525 entries, 0 to 1524
Data columns (total 9 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   vote                     1525 non-null   object
 1   age                      1525 non-null   int64
 2   economic.cond.national   1525 non-null   int64
 3   economic.cond.household  1525 non-null   int64
 4   Blair                    1525 non-null   int64
 5   Hague                    1525 non-null   int64
 6   Europe                   1525 non-null   int64
 7   political.knowledge      1525 non-null   int64
 8   gender                   1525 non-null   object
dtypes: int64(7), object(2)
memory usage: 107.4+ KB
```

We further split the data into X and y. X will have all the independent variables and y will have all the dependent variables. We split the data into 70% for training and 30% for testing.

Scaling is necessary for the age variables since it is not in same magnitude of the other variables, so we are doing the min max scaling here for only the age variable.

	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	Gender_male
0	0.275362	3	3	4	1	2	2	0
1	0.173913	4	4	4	4	5	2	1
2	0.159420	4	4	5	2	3	2	1
3	0.000000	4	2	2	1	4	0	0
4	0.246377	2	2	1	1	6	2	1

Below is the shape of the dataset after spliting.

```
X_train (1061, 8)
X_test (456, 8)
y_train (1061, 1)
y_test (456, 1)
```

We are scaling the data set here for this because for KNN models it is important for scaling the data set. However, for Logistic regression and LDA scaling is not mandatory. But we will be using the scaled data frame for all the models.

	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	Gender_male
0	0.275362	3	3	4	1	2	2	0
1	0.173913	4	4	4	4	5	2	1
2	0.159420	4	4	5	2	3	2	1
3	0.000000	4	2	2	1	4	0	0
4	0.246377	2	2	1	1	6	2	1

Age, which was in continuos, has been scaled here. Rest since they are ordinal data, we are not scaling them.

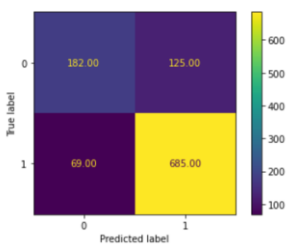
1.4) Apply Logistic Regression and LDA (Linear Discriminant Analysis) (3 pts). Interpret the inferences of both models (2 pts).

Logistic regression:

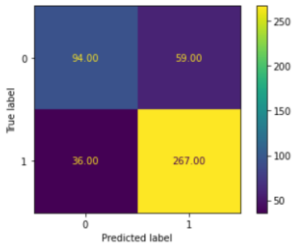
We applied the logistic regression to our data set, and we have got an accuracy of 81% for the training data set and for testing data set also we have got 79%.

Below is the classification and confusion matrix for the training and testing data set. Here we have to see how good the precision is compared to recall because we need to find who all will vote for labour and conservation party and based on the precision only the exit polls will be done.

```
81.71536286522148
[[182 125]
 [ 69 685]]
              precision    recall  f1-score   support

           0       0.73      0.59      0.65       307
           1       0.85      0.91      0.88       754

    accuracy                           0.82      1061
   macro avg       0.79      0.75      0.76      1061
weighted avg       0.81      0.82      0.81      1061
```




```
79.16666666666666
[[ 94  59]
 [ 36 267]]
              precision    recall  f1-score   support

           0       0.72      0.61      0.66       153
           1       0.82      0.88      0.85       303

    accuracy                           0.79       456
   macro avg       0.77      0.75      0.76       456
weighted avg       0.79      0.79      0.79       456
```



0 is the conservative party and 1 is the labour party.

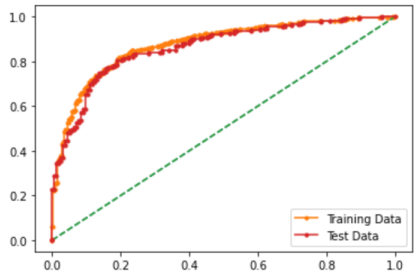
Below is the classification report for the testing data.

The AUC score is

Training: 87

Testing: 86

```
AUC for the Training Data: 0.876
AUC for the Test Data: 0.863
```



We can see that the training data is performing slightly better than the testing data.

LDA:

Accuracy: 82% for training and 80% testing.

Confusion matrix:

AUC:

We can see that both the training and testing are performing similarly.

Logistic and LDA:

Both Logistic and LDA have a similar performance. Both models are performing with an accuracy of 79% and 80%. Even other parameters like the AUC recall precision and F1 are all similar results.

	LOGISTICS TRAIN	LOGISTICS Test	LDA Train	LDA Test	T
Accuracy	0.82	0.79	0.82	0.80	
AUC	0.88	0.86	0.88	0.87	
Recall	0.91	0.88	0.90	0.88	
Precision	0.85	0.82	0.85	0.83	
F1 Score	0.88	0.85	0.88	0.85	

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

The KNN model is applied for the problem and the accuracy is

Training: 85%. **Testing**: 80%.

Below is the classification report for the training data:

```
0.8548539114043355
[[224  83]
 [ 71 683]]
              precision    recall  f1-score   support

           0       0.76      0.73      0.74       307
           1       0.89      0.91      0.90       754

    accuracy                           0.85      1061
   macro avg       0.83      0.82      0.82      1061
weighted avg       0.85      0.85      0.85      1061
```

Classification for testing data:

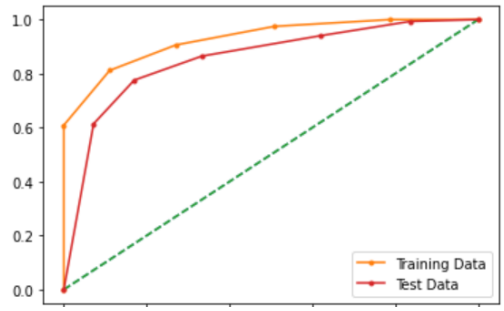
```
0.7982456140350878
[[102  51]
 [ 41 262]]
              precision    recall  f1-score   support

           0       0.71      0.67      0.69       153
           1       0.84      0.86      0.85       303

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

we see here that the precision for the testing data without any parameters for the KNN model is lesser than the training data.

KNN AUC for training and testing:

We can see that the AUC for training us 93% and testing it is 85%. Looking at the curve also we can see a slightly better training score.

```
AUC for the Training Data: 0.928
AUC for the Test Data: 0.855
```



Naïve Bayes:

Training data set:

The accuracy for training data for naïve bayes 82%. with a precision of 71% for the conservative and 85% for the labour party.

```
0.8162111215834119
[[192 115]
 [ 80 674]]
              precision    recall  f1-score   support

           0       0.71      0.63      0.66       307
           1       0.85      0.89      0.87       754

    accuracy                           0.82      1061
   macro avg       0.78      0.76      0.77      1061
weighted avg       0.81      0.82      0.81      1061
```

Testing data:

```
0.793859649122807
[[103  50]
 [ 44 259]]
              precision    recall  f1-score   support

           0       0.70      0.67      0.69       153
           1       0.84      0.85      0.85       303

    accuracy                           0.79       456
   macro avg       0.77      0.76      0.77       456
weighted avg       0.79      0.79      0.79       456
```

The accuracy for the testing data has reduce compared to the training with 79%. Here the precision for labour party is 70 and conservative is 84%.

Both KNN and the naïve bayes model have an accuracy of 79% in the testing data. The precision scores for both are also same. There is no significant difference between the two models. Therefore, both the models perform similar on the data set.

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

I will be using the random forest classifier for model tuning and applying the tuned random forest as the estimator for bagging.

With the use of grid search the different parameter that was entered are run and the best parameters are selected. Below are the different parameter that was entered into the random forest.

```
GridSearchCV(cv=5, estimator=RandomForestClassifier(random_state=1),
             param_grid={'max_depth': [2, 5, 7], 'max_features': [2, 5, 7],
                         'min_samples_leaf': [5, 7],
                         'min_samples_split': [25, 30], 'n_estimators': [250]})
```

The best param function select the best opted parameters for the model and that is applied, and the accuracy is shown. Below are the best parameters from the grid search we entered. The best max depth is 7 from 2 and 5. Similarly above we can see the different parameters that was entered and below the best is selected.

```
{'max_depth': 7,
 'max_features': 5,
 'min_samples_leaf': 5,
 'min_samples_split': 30,
 'n_estimators': 250}
```

This tuned model of random forest is used as the base estimator for bagging.

```
BaggingClassifier(base_estimator=RandomForestClassifier(max_depth=7,
                                                        max_features=5,
                                                        min_samples_leaf=5,
                                                        min_samples_split=30,
                                                        n_estimators=250,
                                                        random_state=1),
                  n_estimators=100, random_state=1)
```

The accuracy for the training data set is 84%.

```
0.8369462770970783
[[195 112]
 [ 61 693]]
              precision    recall  f1-score   support

           0       0.76      0.64      0.69       307
           1       0.86      0.92      0.89       754

    accuracy                           0.84      1061
   macro avg       0.81      0.78      0.79      1061
weighted avg       0.83      0.84      0.83      1061
```

The accuracy for the testing data set is 82%.

```
0.8114035087719298
[[ 99  54]
 [ 32 271]]
              precision    recall  f1-score   support

           0       0.76      0.65      0.70       153
           1       0.83      0.89      0.86       303

    accuracy                           0.81       456
   macro avg       0.79      0.77      0.78       456
weighted avg       0.81      0.81      0.81       456
```

Boosting:

I am using the ada booster and gradient booster here.

And the other parameter such as learning rate and algorithm is also added to tune the ada booster. Learning rate is kept at 1.0 and algorithm as SAMME.R

The accuracy for the training model is 86%

```
0.8576814326107446
[[218  89]
 [ 62 692]]
              precision    recall  f1-score   support

           0       0.78      0.71      0.74       307
           1       0.89      0.92      0.90       754

    accuracy                           0.86      1061
   macro avg       0.83      0.81      0.82      1061
weighted avg       0.85      0.86      0.86      1061
```

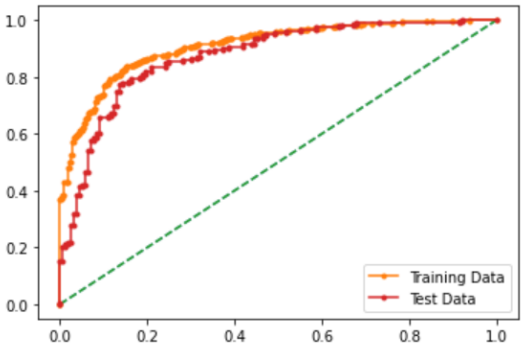
The accuracy for the training model is 81%

```
0.8114035087719298
[[101  52]
 [ 34 269]]
              precision    recall  f1-score   support

           0       0.75      0.66      0.70       153
           1       0.84      0.89      0.86       303

    accuracy                           0.81       456
   macro avg       0.79      0.77      0.78       456
weighted avg       0.81      0.81      0.81       456
```

Here model is not performing as good as the earlier models.

AUC Curve:

```
AUC for the Training Data: 0.910
AUC for the Test Data: 0.875
```



Gradient Boosting:

For this the model is tuned using the param grid. The below are the parameters used to tune the model.

```
GridSearchCV(cv=5, estimator=GradientBoostingClassifier(random_state=1),
             param_grid={'max_depth': [2, 5], 'max_features': [2, 5],
                         'min_samples_leaf': [5, 7],
                         'min_samples_split': [30, 60, 90],
                         'n_estimators': [50, 100, 250], 'tol': [0.0001]})
```

From the entered parameters the best parameters are selected using the best params. Below are the best parameters that is selected for the model.

```
{'max_depth': 2,
 'max_features': 2,
 'min_samples_leaf': 5,
 'min_samples_split': 60,
 'n_estimators': 100,
 'tol': 0.0001}
```

The training data set accuracy is 85%

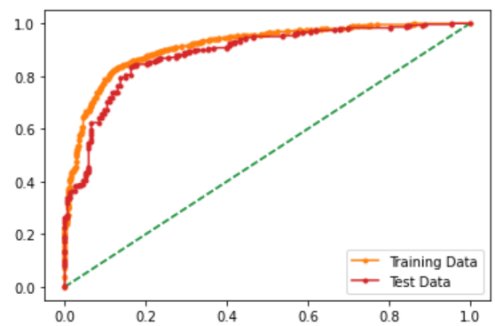
```
0.8539114043355325
[[207 100]
 [ 55 699]]
              precision    recall  f1-score   support

           0       0.79      0.67      0.73       307
           1       0.87      0.93      0.90       754

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

The testing data set accuracy is 82%

```
0.8245614035087719
[[105  48]
 [ 32 271]]
              precision    recall  f1-score   support

           0       0.77      0.69      0.72       153
           1       0.85      0.89      0.87       303

    accuracy                           0.82       456
   macro avg       0.81      0.79      0.80       456
weighted avg       0.82      0.82      0.82       456
```

AUC:

```
AUC for the Training Data: 0.915
AUC for the Test Data: 0.887
```



1.7 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model (4 pts) Final Model - Compare all models on the basis of the performance metrics in a structured tabular manner. Describe on which model is best/optimized (3 pts).

	LOGISTICS TRAIN	LOGISTICS Test	LDA Train	LDA Test	KNN TRAIN	KNN TEST	NAIVE Train	NAIVE TEST	RF BAGGING TRAIN	RF BAGGING TEST	ADA TRAIN	ADA TEST	GRADIENT TRAIN	GRADIENT TEST
Accuracy	0.82	0.79	0.82	0.80	0.85	0.80	0.82	0.79	0.84	0.81	0.86	0.81	0.85	0.82
AUC	0.88	0.86	0.88	0.87	0.93	0.86	0.88	0.86	0.91	0.87	0.92	0.87	0.92	0.89
Recall	0.91	0.88	0.90	0.88	0.91	0.86	0.89	0.85	0.92	0.89	0.92	0.88	0.93	0.88
Precision	0.85	0.82	0.85	0.83	0.89	0.84	0.85	0.84	0.86	0.83	0.89	0.84	0.87	0.84
F1 Score	0.88	0.85	0.88	0.85	0.90	0.85	0.87	0.85	0.89	0.86	0.90	0.86	0.90	0.86

Above is the table with the accuracy, AUC scores for all the models.

From all the models the best test score is for gradient model 82%. Rest all the model are performing very similarly.

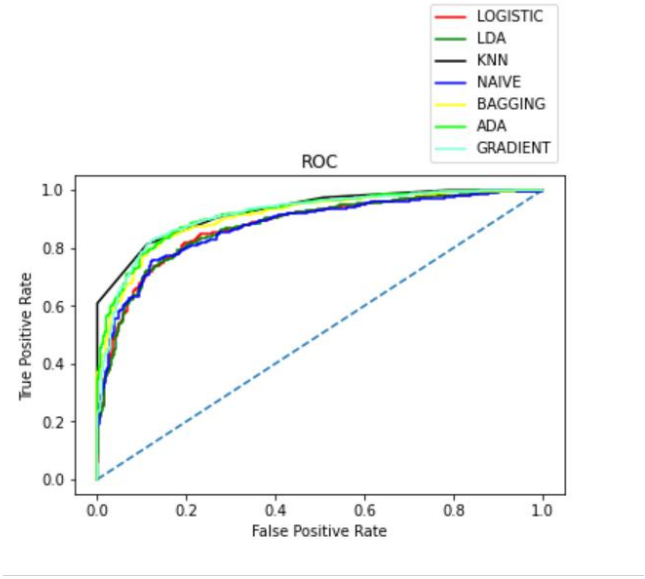
On the training data set the ADA boost and gradient boosting is the best performing with an accuracy of 86% and 85%.

The AUC scores are the best in gradient boosting with 89%.

Overall compared to all the model gradient boosting has done marginally better in accuracy and AUC. However, in recall the bagging is doing best in the test data with 92.

There is not much difference in all the models.

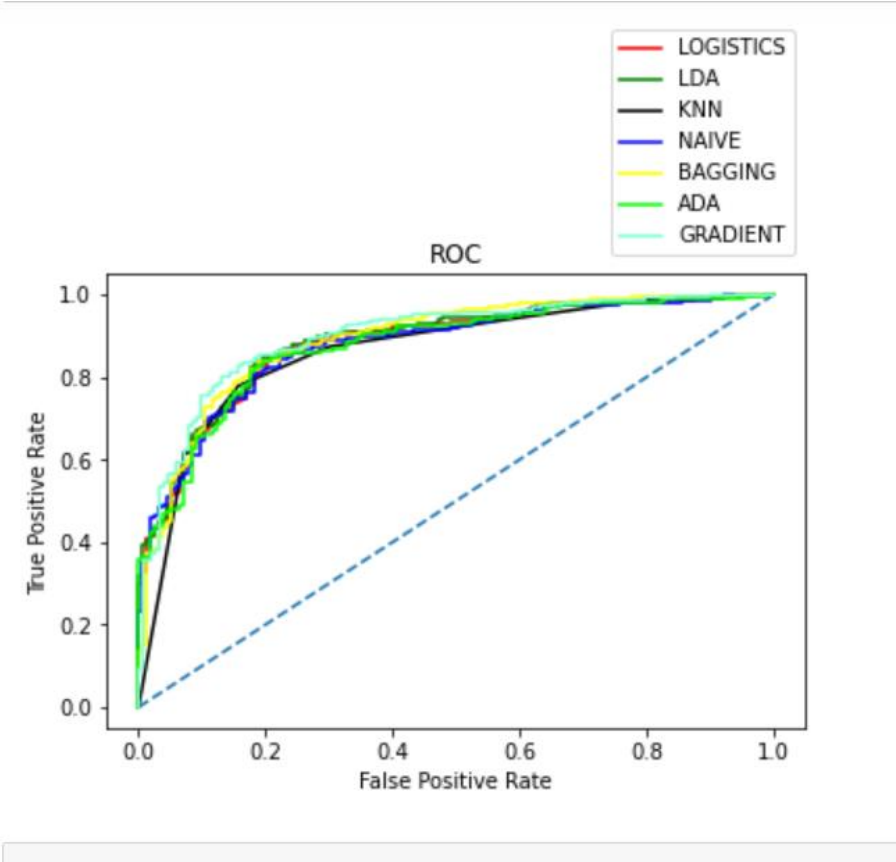
AUC curve Training data:



As discussed, we can see the KNN curve is performing slightly better than other in the training data.

AUC curve testing data:

However, in the testing data all the models are performing very similarly.

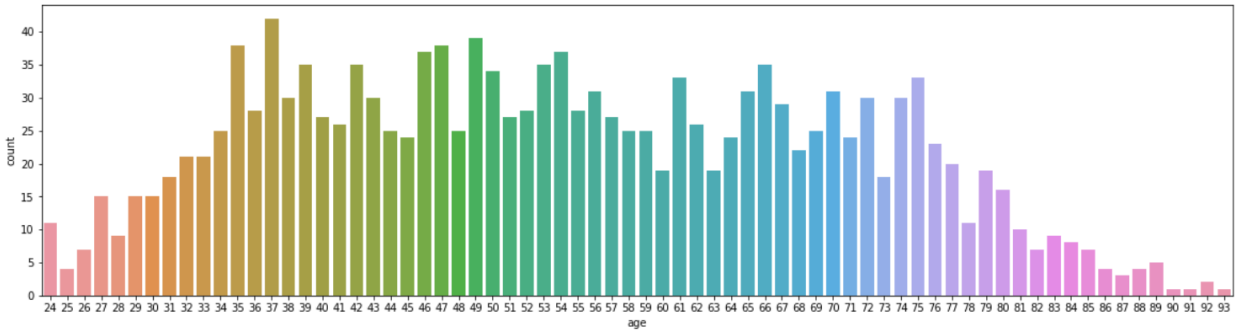


1.7) Based on your analysis and working on the business problem, detail out appropriate insights and recommendations to help the management solve the business objective.

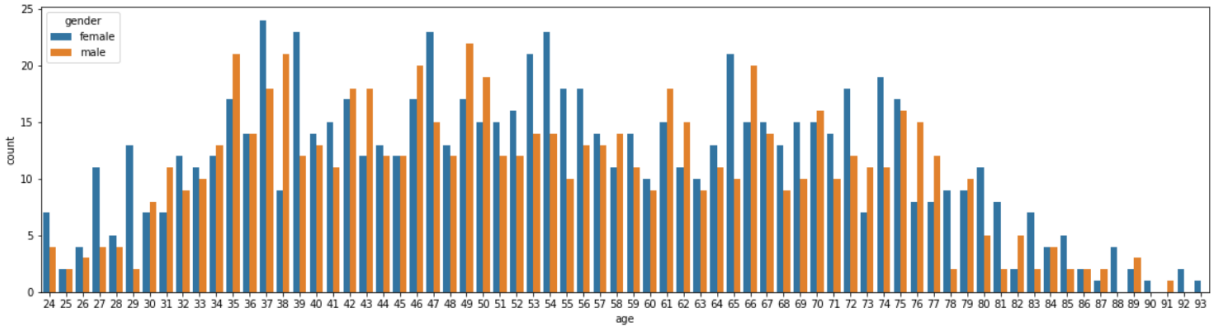
From the random forest classifier that we did for model tuning for boosting we found out the important variables that are contributing more to the model using the feature importance.
The below variable are the most important features for the model. As known the two candidates are the major contributor in deciding whom to vote.

	Imp
Hague	0.357886
Blair	0.267366
Europe	0.104371
economic.cond.national	0.100854
age	0.090914
political.knowledge	0.039541
economic.cond.household	0.031199
Gender_male	0.007868

- Most of the voters are female voters compared to male voters. And the average age of voters is 54. it is possible that the male voters are not tapped, and both the parties can use this as an opportunity to focus some campaign to boost the male voters to vote.
- It is also noticed that young voters are not voting much. We can see that very few voters in the age group between 24 to 32. Some schemes can be organized to get these voters to vote.

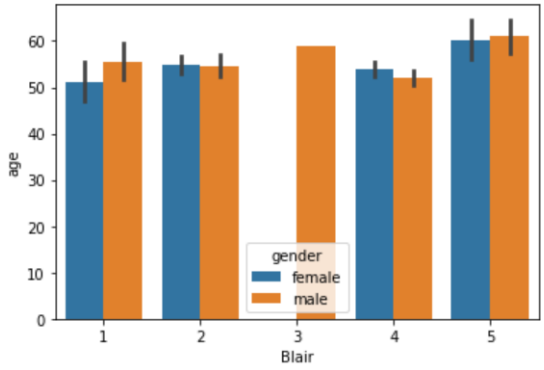


- As seen earlier female voters are voting more in all the age groups. From the young to the old age group.



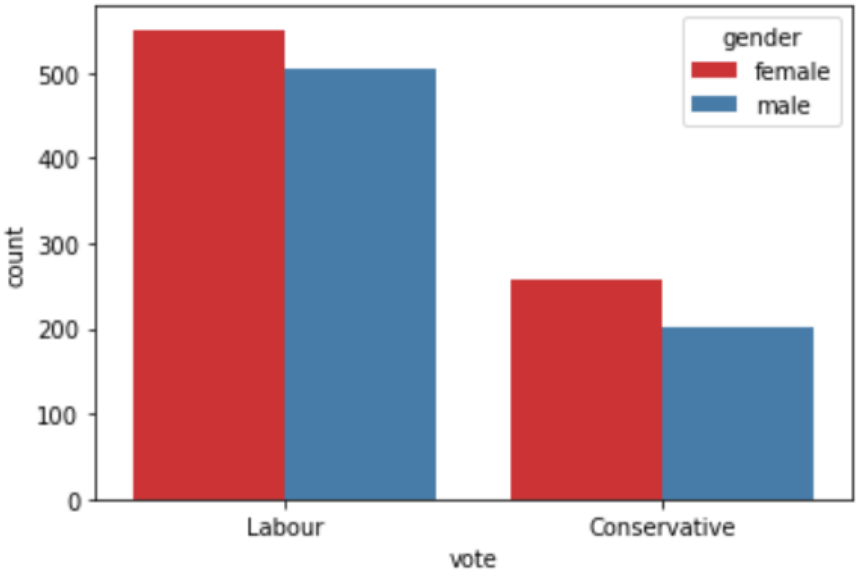
- We can see from the Blair rating that for rating 3 all the votes have been done by the male. These are important votes that can be converted to higher rating. Therefore, the Blair party has to focus on making their male voters or else the 3 rating voters can vote for Hague's side. Similarly, females have given less rating,

so Blair team has to focus more on male voters and bring them to vote for him.

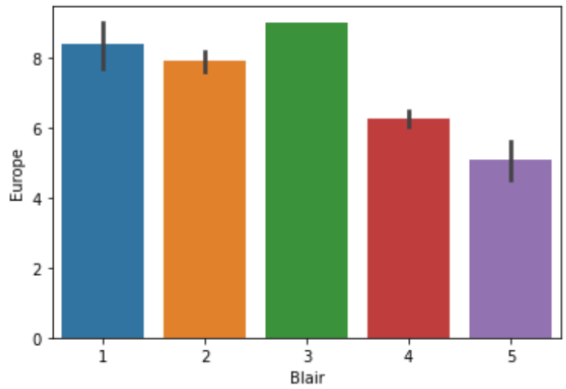


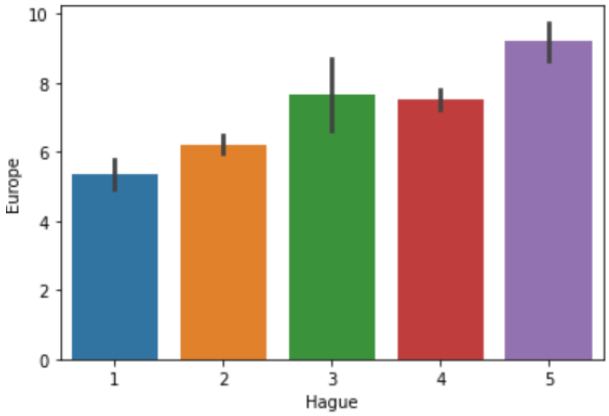
- Most of them look like they will be voting for Blair. Hague chances of winning is very less. Almost 70% of the voters are voting for Blair(Labour party).

```
Labour          0.697049
Conservative    0.302951
Name: vote, dtype: float64
```

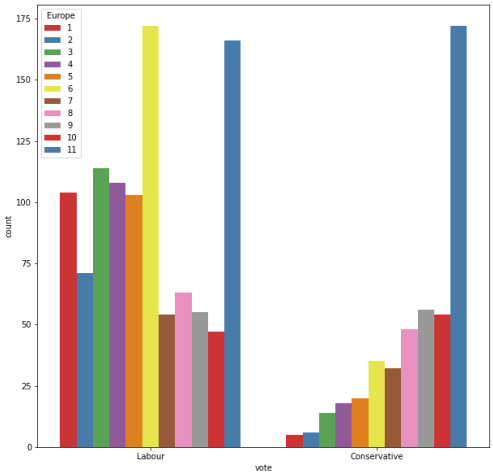


```
<AxesSubplot:xlabel='Blair', ylabel='Europe'>
```





Above the barplot shows that people who gave 3 rating to Blair have more sentiment to Europe sentiment. And people who gave 5 rating to Hague have more sentiment for Europe integration.



Voters of labour party have less sentiment to European integration compared to conservative party. We can see that conservative party have given more rating of 11 which mean they have more sentiment for European integration.

So, the conservative party can catch on this point and attract the European votes because labour party does not have huge sentiment for European integration.

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

```
Total number of charaters in Roosevelt are 7571
Total number of charaters in Kennedy are 7618
Total number of charaters in Nixon are 9991
```

The total number of characters are highest for Nixon with 9991.

Words including all the punctuation and string.

```
Number of words in Roosevelt are 1526
Number of words in Kennedy are 1543
Number of words in Nixon are 2006
```

The above apart from English words even the strings and punctuations are included.

```
Total number of words in Roosevelt without punctuation is 1350
Total number of words in Kennedy without punctuation is 1370
Total number of words in Nixon without punctuation is 1819
```

Above we can see that only the actual word counts for each individual president speeches.

The total number of sentences in each president speech are

```
Total number of sentence in Roosevelt is 68
Total number of sentence in Kennedy is 52
Total number of sentence in Nixon is 68
```

2.2) Remove all the stopwords from the three speeches.

For this process I have removed all the stop words and the punctuation and other unwanted string items. Also, words are changed to lower case so that no two

words of different case are duplicated. Have also done stemming to remove the same meaning words which are ing extension and other extension words.

Below are the speeches after removal of all these.

Roosevelt speech.

```
['on', 'each', 'nation', 'day', 'of', 'inaugur', 'sinc', 'the', 'peopl', 'have', 'renew', 'their', 'sens', 'of', 'dedic', 't
o', 'the', 'unit', 'state', 'in', 'washington', "'s", 'day', 'the', 'task', 'of', 'the', 'peopl', 'was', 'to', 'creat', 'an
d', 'weld', 'togeth', 'a', 'nation', 'in', 'lincoln', "'s", 'day', 'the', 'task', 'of', 'the', 'peopl', 'was', 'to', 'preser
v', 'that', 'nation', 'from', 'disrupt', 'from', 'within', 'in', 'this', 'day', 'the', 'task', 'of', 'the', 'peopl', 'is', 't
o', 'save', 'that', 'nation', 'and', 'it', 'institut', 'from', 'disrupt', 'from', 'without', 'to', 'us', 'there', 'has', 'com
e', 'a', 'time', 'in', 'the', 'midst', 'of', 'swift', 'happen', 'to', 'paus', 'for', 'a', 'moment', 'and', 'take', 'stock',
'to', 'recal', 'what', 'our', 'place', 'in', 'histori', 'has', 'been', 'and', 'to', 'rediscov', 'what', 'we', 'are', 'and',
'what', 'we', 'may', 'be', 'if', 'we', 'do', 'not', 'we', 'risk', 'the', 'real', 'peril', 'of', 'inact', 'live', 'of', 'natio
n', 'are', 'determin', 'not', 'by', 'the', 'count', 'of', 'year', 'but', 'by', 'the', 'lifetim', 'of', 'the', 'human', 'spiri
t', 'the', 'life', 'of', 'a', 'man', 'is', 'three-scor', 'year', 'and', 'ten', 'a', 'littl', 'more', 'a', 'littl', 'less', 't
he', 'life', 'of', 'a', 'nation', 'is', 'the', 'full', 'of', 'the', 'measur', 'of', 'it', 'will', 'to', 'live', 'there', 'ar
e', 'men', 'who', 'doubt', 'this', 'there', 'are', 'men', 'who', 'believ', 'that', 'democraci', 'as', 'a', 'form', 'of', 'gov
ern', 'and', 'a', 'frame', 'of', 'life', 'is', 'limit', 'or', 'measur', 'by', 'a', 'kind', 'of', 'mystic', 'and', 'artifici',
'fate', 'that', 'for', 'some', 'unexplain', 'reason', 'tyranni', 'and', 'slaveri', 'have', 'becom', 'the', 'surg', 'wave', 'o
f', 'the', 'futur', 'and', 'that', 'freedom', 'is', 'an', 'eb', 'tide', 'but', 'we', 'american', 'know', 'that', 'this', 'i
s', 'not', 'true', 'eight', 'year', 'ago', 'when', 'the', 'life', 'of', 'this', 'republ', 'seem', 'frozen', 'by', 'a', 'fatal
ist', 'terror', 'we', 'prove', 'that', 'this', 'is', 'not', 'true', 'we', 'were', 'in', 'the', 'midst', 'of', 'shock', 'but',
'we', 'act', 'we', 'act', 'quick', 'bold', 'decis', 'these', 'later', 'year', 'have', 'been', 'live', 'year', 'fruit', 'yea
r', 'for', 'the', 'peopl', 'of', 'this', 'democraci', 'for', 'they', 'have', 'brought', 'to', 'us', 'greater', 'secur', 'an
d', 'i', 'hope', 'a', 'better', 'understand', 'that', 'life', "'s", 'ideal', 'are', 'to', 'be', 'measur', 'in', 'other', 'tha
```

Kennedy Speech:

```
['vice', 'presid', 'johnson', 'mr.', 'speaker', 'mr.', 'chief', 'justic', 'presid', 'eisenhow', 'vice', 'presid', 'nixon', 'p
resid', 'truman', 'reverend', 'clergi', 'fellow', 'citizen', 'we', 'observ', 'today', 'not', 'a', 'victori', 'of', 'parti',
'but', 'a', 'celebr', 'of', 'freedom', 'symbol', 'an', 'end', 'as', 'well', 'as', 'a', 'begin', 'signifi', 'renew', 'as', 'we
ll', 'as', 'chang', 'for', 'i', 'have', 'sworn', 'i', 'befor', 'you', 'and', 'almighti', 'god', 'the', 'same', 'solemn', 'oat
h', 'our', 'forebear', 'l', 'prescrib', 'near', 'a', 'centuri', 'and', 'three', 'quarter', 'ago', 'the', 'world', 'is', 'ver
i', 'differ', 'now', 'for', 'man', 'hold', 'in', 'his', 'mortal', 'hand', 'the', 'power', 'to', 'abolish', 'all', 'form', 'o
f', 'human', 'poverti', 'and', 'all', 'form', 'of', 'human', 'life', 'and', 'yet', 'the', 'same', 'revolutionari', 'belief',
'for', 'which', 'our', 'forebear', 'fought', 'are', 'still', 'at', 'issu', 'around', 'the', 'globe', 'the', 'belief', 'that',
'the', 'right', 'of', 'man', 'come', 'not', 'from', 'the', 'generos', 'of', 'the', 'state', 'but', 'from', 'the', 'hand', 'o
f', 'god', 'we', 'dare', 'not', 'forget', 'today', 'that', 'we', 'are', 'the', 'heir', 'of', 'that', 'first', 'revolut', 'le
t', 'the', 'word', 'go', 'forth', 'from', 'this', 'time', 'and', 'place', 'to', 'friend', 'and', 'foe', 'alik', 'that', 'th
e', 'torch', 'has', 'been', 'pass', 'to', 'a', 'new', 'generat', 'of', 'american', 'born', 'in', 'this', 'centuri', 'temper',
'by', 'war', 'disciplin', 'by', 'a', 'hard', 'and', 'bitter', 'peac', 'proud', 'of', 'our', 'ancient', 'heritag', 'and', 'unw
il', 'to', 'wit', 'or', 'permit', 'the', 'slow', 'undo', 'of', 'those', 'human', 'right', 'to', 'which', 'this', 'nation', 'h
as', 'alway', 'been', 'commit', 'and', 'to', 'which', 'we', 'are', 'commit', 'today', 'at', 'home', 'and', 'around', 'the',
'world', 'let', 'everi', 'nation', 'know', 'whether', 'it', 'wish', 'us', 'well', 'or', 'ill', 'that', 'we', 'shall', 'pay',
'ani', 'price', 'bear', 'ani', 'burden', 'meet', 'ani', 'hardship', 'support', 'ani', 'friend', 'oppos', 'ani', 'foe', 'in',
'order', 'to', 'assur', 'the', 'surviv', 'and', 'the', 'success', 'of', 'liberti', 'this', 'much', 'we', 'pledg', 'and', 'mor
e', 'to', 'those', 'old', 'alli', 'whose', 'cultur', 'and', 'spiritu', 'origin', 'we', 'share', 'we', 'pledg', 'the', 'loyalt
i', 'of', 'faith', 'friend', 'unit', 'there', 'is', 'littl', 'we', 'can', 'not', 'do', 'in', 'a', 'host', 'of', 'cooper', 've
```

Nixon Speech:

```
['mr.', 'vice', 'presid', 'mr.', 'speaker', 'mr.', 'chief', 'justic', 'senat', 'cook', 'mrs.', 'eisenhow', 'and', 'my', 'fell
ow', 'citizen', 'of', 'this', 'great', 'and', 'good', 'countri', 'we', 'share', 'togeth', 'when', 'we', 'met', 'here', 'fou
r', 'year', 'ago', 'america', 'was', 'bleak', 'in', 'spirit', 'depress', 'by', 'the', 'prospect', 'of', 'seem', 'endless', 'w
ar', 'abroad', 'and', 'of', 'destruct', 'conflict', 'at', 'home', 'as', 'we', 'meet', 'here', 'today', 'we', 'stand', 'on',
'the', 'threshold', 'of', 'a', 'new', 'era', 'of', 'peac', 'in', 'the', 'world', 'the', 'central', 'question', 'befor', 'us',
'is', 'how', 'shall', 'we', 'use', 'that', 'peac', 'let', 'us', 'resolv', 'that', 'this', 'era', 'we', 'are', 'about', 'to',
'enter', 'will', 'not', 'be', 'what', 'other', 'postwar', 'period', 'have', 'so', 'often', 'been', 'a', 'time', 'of', 'retrea
t', 'and', 'isol', 'that', 'lead', 'to', 'stagnat', 'at', 'home', 'and', 'invit', 'new', 'danger', 'abroad', 'let', 'us', 're
solv', 'that', 'this', 'will', 'be', 'what', 'it', 'can', 'becom', 'a', 'time', 'of', 'great', 'respons', 'great', 'born', 'i
n', 'which', 'we', 'renew', 'the', 'spirit', 'and', 'the', 'promis', 'of', 'america', 'as', 'we', 'enter', 'our', 'third', 'c
enturi', 'as', 'a', 'nation', 'this', 'past', 'year', 'saw', 'far-reach', 'result', 'from', 'our', 'new', 'polici', 'for', 'p
eac', 'by', 'continu', 'to', 'revit', 'our', 'tradit', 'friendship', 'and', 'by', 'our', 'mission', 'to', 'peke', 'and', 't
o', 'moscow', 'we', 'were', 'abl', 'to', 'establish', 'the', 'base', 'for', 'a', 'new', 'and', 'more', 'durabl', 'pattern',
'of', 'relationship', 'among', 'the', 'nation', 'of', 'the', 'world', 'becaus', 'of', 'america', "'s", 'bold', 'initi', 'wil
l', 'be', 'long', 'rememb', 'as', 'the', 'year', 'of', 'the', 'greatest', 'progress', 'sinc', 'the', 'end', 'of', 'world', 'w
ar', 'ii', 'toward', 'a', 'last', 'peac', 'in', 'the', 'world', 'the', 'peac', 'we', 'seek', 'in', 'the', 'world', 'is', 'no
t', 'the', 'flimsi', 'peac', 'which', 'is', 'mere', 'an', 'interlud', 'between', 'war', 'but', 'a', 'peac', 'which', 'can',
'endur', 'for', 'generat', 'to', 'come', 'it', 'is', 'import', 'that', 'we', 'understand', 'both', 'the', 'necess', 'and', 't
he', 'limit', 'of', 'america', "'s", 'role', 'in', 'maintain', 'that', 'peac', 'unless', 'we', 'in', 'america', 'work', 'to',
'preserv', 'the', 'peac', 'there', 'will', 'be', 'no', 'peac', 'unless', 'we', 'in', 'america', 'work', 'to', 'preserv', 'fre
```

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 top 3 most occurring words for the president Roosevelt speech are:

Nation which has occurred 17 times.
Know which has occurred 10 times in the document
People which have occurred 9 times.

```
[('nation', 17), ('know', 10), ('peopl', 9)]
```

Kennedy speech most common occurring 3 words are:

Let accruing 16 times
US occurring 12 times
Power occurring 9 times

```
[('let', 16), ('us', 12), ('power', 9)]
```

Nixon speech most common occurring 3 words are:

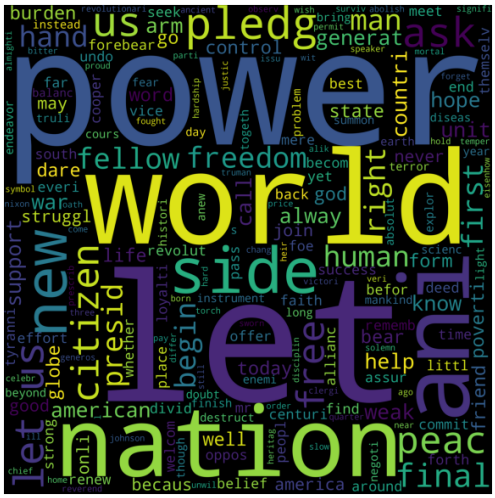
Us 26 times
Let 22 times
America 21 times

```
[('us', 26), ('let', 22), ('america', 21)]
```

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

Roosevelt:

Kennedy word cloud:



Nixon word cloud: