Machine Learning

Project

Pavithra Devi
DSBA
Great Learning

Index

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.	
1.7 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model, classification report (4 pts) Final Model - Compare and comment on all models on the basis of the performance metrics in a structured tabular manner. Describe on which model is best/optimized, After comparison which model suits the best for the problem in hand on the basis of different measures. Comment on the final model.(3 pts)	7
1.8) Based on your analysis and working on the business problem, detail out appropriate insights and recommendations to help the management solve the business objective. There should be at least 3-4 Recommendations and insights in total. Recommendations should be easily understandable and business specific, students should not give any technical suggestions. Full marks should only be allotted if the recommendations are correct and business specific.	5
2.1) Find the number of characters, words and sentences for the mentioned documents. (Hint: use .words(), .raw(), .sent() for extracting counts)	3
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.	3
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)	3
2.4) Plot the word cloud of each of the three speeches. (after removing the stopwords)	3
Quality of Business Report (Please refer to the Evaluation Guidelines for Business report checklist. Marks in this criteria are at the moderator's discretion)	6

Problem 1:

Context:

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

Data Dictionary:

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

Method used:

Logistic Regression:

Logit classifier is a supervised learning method for classification. It establishes relationship between dependent class variable and independent variables using regression. The dependent variable is categorical

LDA(Linear Discriminant Analysis) :-

LDA is used for classifying observations to a class or category based on predictor(independent variable).
LDA creats a model to predict the classes of the new or future observation.

KNN:

The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point.

Naives Bayes Model:

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.
It is mainly used in text classification that includes a high-dimensional training dataset.
Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

Model Tuning:

Model tuning is the experimental process of finding the optimal values of hyperparameters to maximize model performance. Hyperparameters are the set of variables whose values cannot be estimated by the model from the training data. These values control the training process.

Bagging :

Bagging, also known as bootstrap aggregation, is the ensemble learning method that is commonly used to reduce variance within a noisy dataset. In bagging, a random sample of data in a training set is selected with replacement—meaning that the individual data points can be chosen more than once.

Boosting:

Boosting is an ensemble learning method that combines a set of weak learners into a strong learner to minimize training errors. In boosting, a random sample of data is selected, fitted with a model and then trained sequentially—that is, each model tries to compensate for the weaknesses of its predecessor.

1.1 Exploratory Data Analysis:

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

Figure 1

Dataset Size: Dataset contains 1525 rows and 9 columns.

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

Figure 2

Figure 3 provides summary statistics for each of the variables in the dataset. Here's an interpretation of the key information:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
vote	1525	2	Labour	1063	NaN	NaN	NaN	NaN	NaN	NaN	NaN
age	1525.0	NaN	NaN	NaN	54.182295	15.711209	24.0	41.0	53.0	67.0	93.0
economic.cond.national	1525.0	NaN	NaN	NaN	3.245902	0.880969	1.0	3.0	3.0	4.0	5.0
economic.cond.household	1525.0	NaN	NaN	NaN	3.140328	0.929951	1.0	3.0	3.0	4.0	5.0
Blair	1525.0	NaN	NaN	NaN	3.334426	1.174824	1.0	2.0	4.0	4.0	5.0
Hague	1525.0	NaN	NaN	NaN	2.746885	1.230703	1.0	2.0	2.0	4.0	5.0
Europe	1525.0	NaN	NaN	NaN	6.728525	3.297538	1.0	4.0	6.0	10.0	11.0
political.knowledge	1525.0	NaN	NaN	NaN	1.542295	1.083315	0.0	0.0	2.0	2.0	3.0
gender	1525	2	female	812	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figure 3

1. vote (Party Choice):
   - There are 1525 entries in the dataset.
   - There are two unique values: "Labour" and another party (possibly "Conservative" or a different party).
   - "Labour" is the most frequently occurring choice, appearing 1063 times.

2. age (Age of Voters):
   - The age variable contains data for all 1525 voters.
   - The mean age is approximately 54.18 years.
   - The minimum age is 24 years, and the maximum age is 93 years.
   - Age distribution shows some variability (standard deviation of approximately 15.71).

3. economic.cond.national (Assessment of National Economic Conditions):
   - The variable contains data for all 1525 voters.
   - Voters' assessments of national economic conditions range from 1 (poor) to 5 (excellent).
   - The mean assessment is approximately 3.25.

4. economic.cond.household (Assessment of Household Economic Conditions):
   - The variable contains data for all 1525 voters.
   - Voters' assessments of their household economic conditions range from 1 (poor) to 5 (excellent).
   - The mean assessment is approximately 3.14.

5. Blair (Assessment of the Labour Leader):
   - The variable contains data for all 1525 voters.
   - Assessments of the Labour leader range from 1 (poor) to 5 (excellent).
   - The mean assessment is approximately 3.33.

6. Hague (Assessment of the Conservative Leader):
   - The variable contains data for all 1525 voters.
   - Assessments of the Conservative leader range from 1 (poor) to 5 (excellent).
   - The mean assessment is approximately 2.75.

7. Europe (Attitudes Toward European Integration):
   - The variable contains data for all 1525 voters.
   - Attitudes toward European integration are measured on an 11-point scale, with high scores representing 'Eurosceptic' sentiment.
   - The mean score is approximately 6.73, and the standard deviation is about 3.30.

8. political.knowledge (Knowledge of Parties' Positions on European Integration):
- The variable contains data for all 1525 voters.
- Knowledge levels range from 0 to 3.
- The mean knowledge level is approximately 1.54.

9. gender (Gender of Voters):
- There are two unique values: "female" and "male."
- "Female" is the most frequently occurring gender, appearing 812 times.

This summary provides a snapshot of the dataset's characteristics, distributions, and central tendencies for each variable. It can be helpful for understanding the dataset and planning data analysis and modeling.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1525 entries, 1 to 1525
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
```

Figure 4

In the dataset (figure 4), the majority of columns are of integer data type, which is suitable for variables involving numerical measurements such as age, assessments, and knowledge levels. It's noteworthy that there are no missing values (null values) in the dataset, indicating good data quality.

Specifically, the "gender" and "vote" columns are of object data type, which is expected since they contain categorical and text-based information. The "gender" column represents gender categories ("female" and "male"), while the "vote" column indicates party choices and may contain party names.

Number of duplicates: (Figure 5)Although there are eight rows in the dataset that appear to have identical values across multiple columns, it's essential to note that these instances may not necessarily represent duplicated or redundant data. In this context, having similar characteristics, such as age and party choice, among different individuals is entirely plausible.

It's possible that these rows represent different people who happen to share the same age and voting preferences. This underscores the importance of distinguishing between genuine duplicates and instances where individuals possess similar attributes. In this dataset, these eight rows do not appear to be true duplicates but rather unique entries with shared characteristics.

	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
68	Labour	35	4	4	5	2	3	2	male
627	Labour	39	3	4	4	2	5	2	male
871	Labour	38	2	4	2	2	4	3	male
984	Conservative	74	4	3	2	4	8	2	female
1155	Conservative	53	3	4	2	2	6	0	female
1237	Labour	36	3	3	2	2	6	2	female
1245	Labour	29	4	4	4	2	2	2	female
1439	Labour	40	4	3	4	2	2	2	male

Figure 5

Therefore, we are continuing our analysis without dropping the columns.

```
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
```

Figure 6

Having no missing, null values and appropriately assigned data types ensures that the dataset is clean and well-prepared for subsequent data analysis and modeling. Figure 6, shows us that there are no null values in the dataset.

	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
1	0	43	3	3	4	1	2	2	0
2	0	36	4	4	4	4	5	2	1
3	0	35	4	4	5	2	3	2	1
4	0	24	4	2	2	1	4	0	0
5	0	41	2	2	1	1	6	2	1
...	...	...	...	...	...	...	...	...	...
1521	1	67	5	3	2	4	11	3	1
1522	1	73	2	2	4	4	8	2	1
1523	0	37	3	3	5	4	2	2	1
1524	1	61	3	3	1	4	11	2	1
1525	1	74	2	3	2	4	11	0	0

Figure 7

To prepare the data for modeling, **encoding** has been applied to variables that originally contained string values. Specifically, the "gender" variable has been encoded such that "male" is represented as 1, and "female" is represented as 0. Likewise, the "vote" variable has undergone encoding, with "Labour" being assigned a value of 0, and "Conservative" assigned a value of 1.

As a result of this encoding process, all variables in the dataset now have an integer data type. This transformation enables the data to be used effectively in machine learning models that typically require numerical input, ensuring that the dataset is ready for modeling and analysis. [Figure 7 and 8]

```
Data columns (total 9 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   vote                    1525 non-null   int64
 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   int64
```

Figure 8

vote	0.858449
age	0.144621
economic.cond.national	-0.240453
economic.cond.household	-0.149552
Blair	-0.535419
Hague	0.152100
Europe	-0.135947
political.knowledge	-0.426838
gender	0.130239

Figure 9

(Figure 9) The provided values are measures of skewness for each variable in the dataset. Skewness is a statistical measure that indicates the asymmetry of the data's distribution.

1. vote (Party Choice): A skewness value of approximately 0.86 indicates a moderate positive skew. This suggests that there may be a slight tendency for more voters to lean towards one party over the other, resulting in a positively skewed distribution.

2. age (Age of Voters): With a skewness value of around 0.14, the age variable is very close to a symmetric distribution. This indicates that the age distribution is relatively balanced and does not show a significant skew in either direction.

3. economic.cond.national (Assessment of National Economic Conditions): A skewness value of approximately -0.24 suggests a slight negative skew. This could indicate that voters tend to have a more negative assessment of national economic conditions, resulting in a slightly left-skewed distribution.

4. economic.cond.household (Assessment of Household Economic Conditions): Similarly, a skewness value of around -0.15 indicates a slight negative skew. This implies that voters may, on average, assess their household economic conditions slightly more negatively, contributing to a left-skewed distribution.

5. Blair (Assessment of the Labour Leader): With a skewness value of approximately -0.54, the Blair assessment variable exhibits a moderate negative skew. This indicates that, on average, assessments of the Labour leader tend to be lower, resulting in a left-skewed distribution.

6. Hague (Assessment of the Conservative Leader): A skewness value of about 0.15 indicates a slight positive skew. This suggests that assessments of the Conservative leader may be slightly higher on average, contributing to a right-skewed distribution.

7. Europe (Attitudes Toward European Integration): A skewness value of approximately -0.14 suggests a slight negative skew in attitudes toward European integration. This could mean that more voters tend to have slightly more Eurosceptic sentiments, contributing to a left-skewed distribution.

8. political.knowledge (Knowledge of Parties' Positions on European Integration): With a skewness value of around -0.43, the political knowledge variable exhibits a moderate negative skew. This implies that, on average, voters may possess lower levels of knowledge regarding parties' positions on European integration, resulting in a left-skewed distribution.

9. gender (Gender of Voters): A skewness value of about 0.13 indicates a slight positive skew. This suggests that there may be a slightly higher number of male voters in the dataset, contributing to a positively skewed distribution.

In summary, positive skewness indicates a tail extending to the right, while negative skewness indicates a tail extending to the left in the distribution. Understanding skewness is important for data analysis and modeling, as it can impact the choice of statistical techniques and the interpretation of results.

1.2 Data Visualization:

Univariate Analysis:

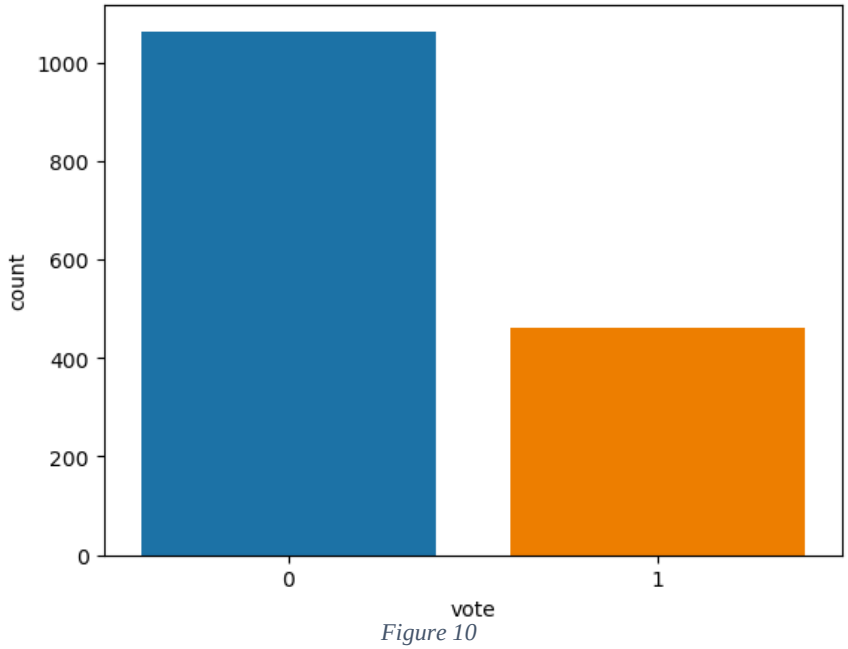


Figure 10

- There are two main parties: "Labour" (represented as 0) and "Conservative" (represented as 1).
- "Labour" is the more common choice, with 1063 occurrences, while "Conservative" appears 462 times.

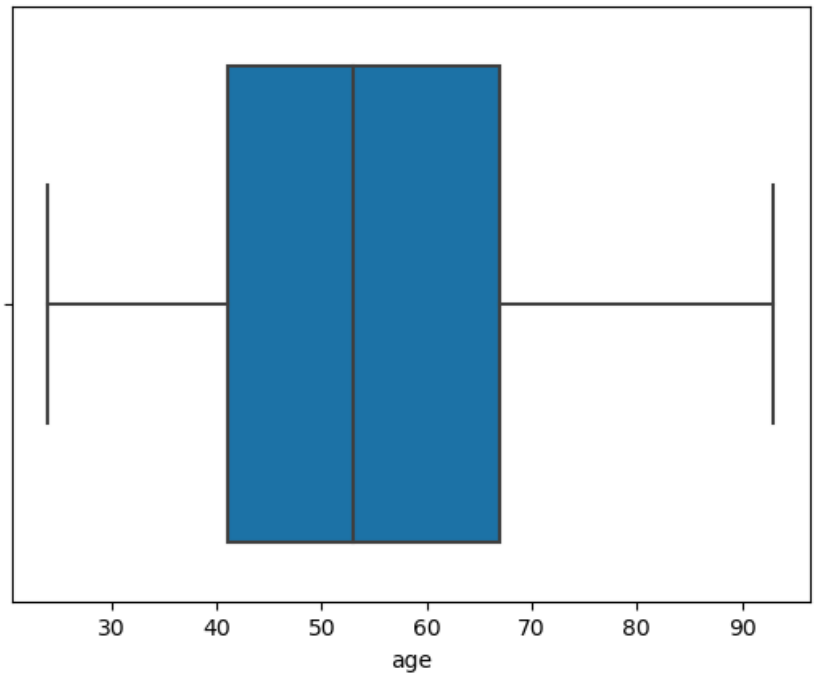


Figure 11

- The mean age is approximately 54 years, with ages ranging from 24 to 93 years.
- The age distribution shows variability with a standard deviation of about 15.71.
- The age observations are grouped into categories with frequencies for each category, spanning from 37 to 93.

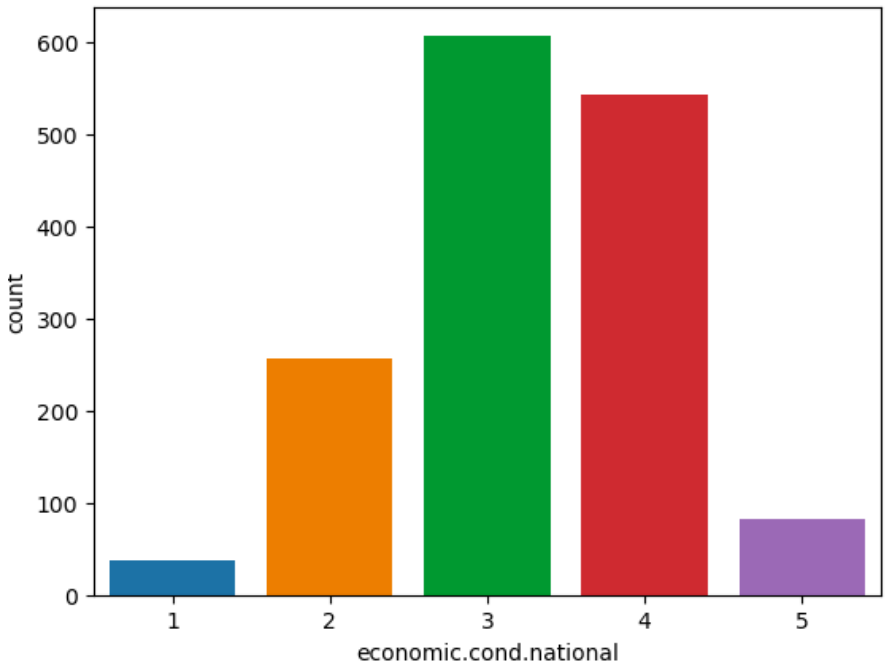


Figure 12

- Voter assessments of national economic conditions are on a scale from 1 (poor) to 5 (excellent).
- The mean assessment is approximately 3.25.
- The distribution of assessments is primarily concentrated around values 3 and 4, with some voters rating conditions lower or higher.

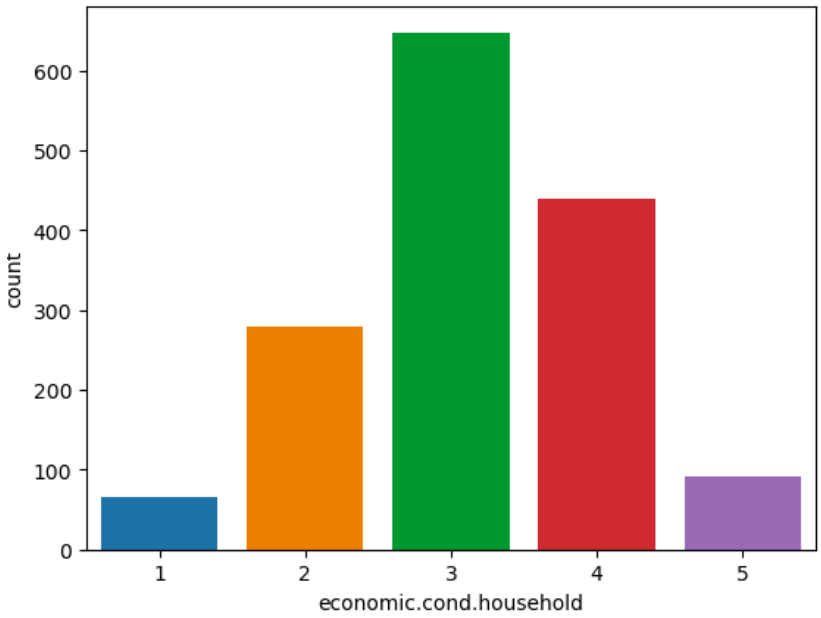


Figure 13

- Assessments of household economic conditions also range from 1 to 5.
- The mean assessment is approximately 3.14.
- The most common assessment is 3, followed by 4, indicating a relatively balanced distribution.

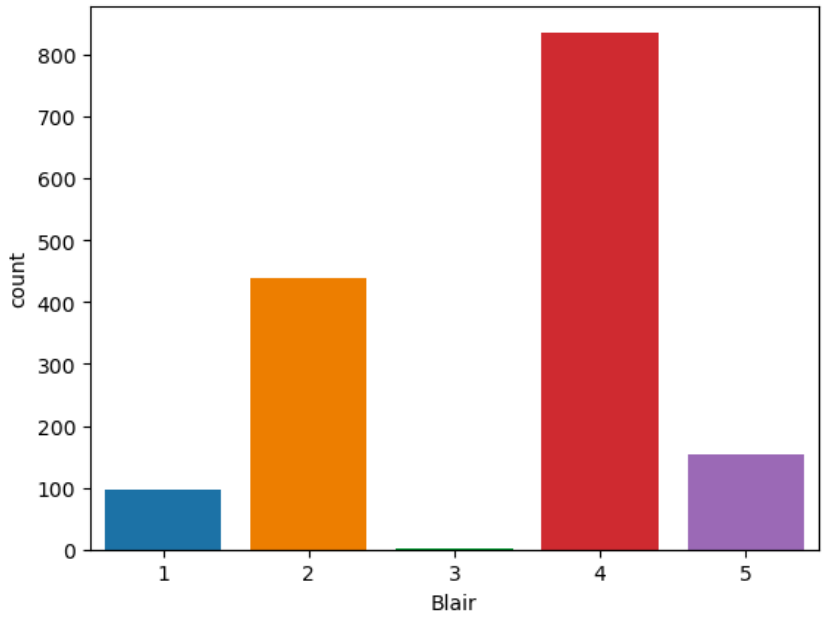


Figure 14

- Assessments range from 1 (poor) to 5 (excellent).
- The mean assessment is approximately 3.33.
- Assessment 4 is the most common, occurring 836 times, although there is an unusual count of 1 for assessment 3.

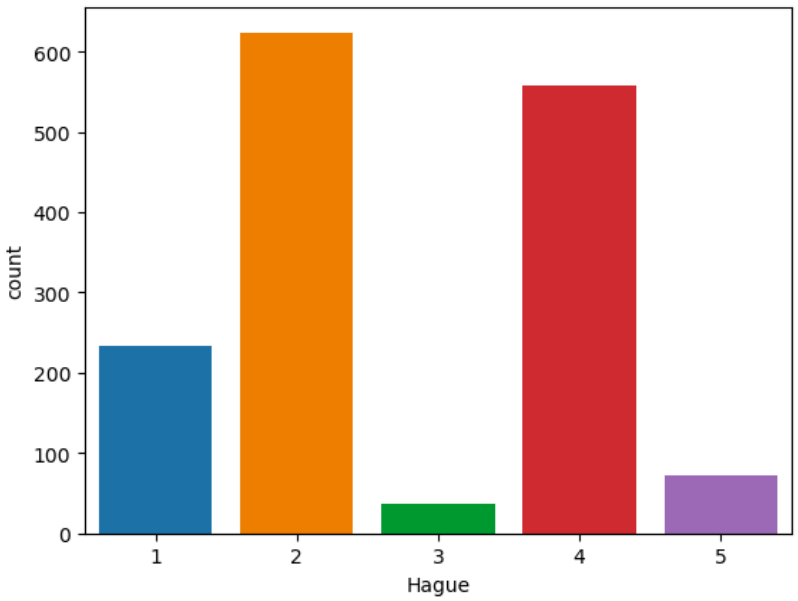


Figure 15

- Like the Blair variable, assessments of the Conservative leader range from 1 to 5.
- The mean assessment is approximately 2.75.
- Assessment 2 is the most common, with 624 occurrences.

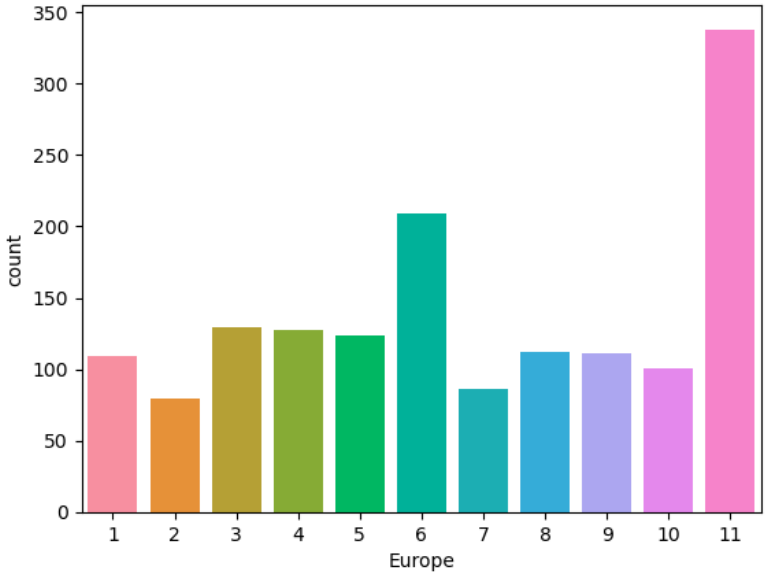


Figure 16

- The variable is based on an 11-point scale measuring attitudes toward European integration.
- The mean score is approximately 6.73, and the standard deviation is about 3.30.
- Attitude 11 is the most frequently observed, with varying frequencies across the scale.

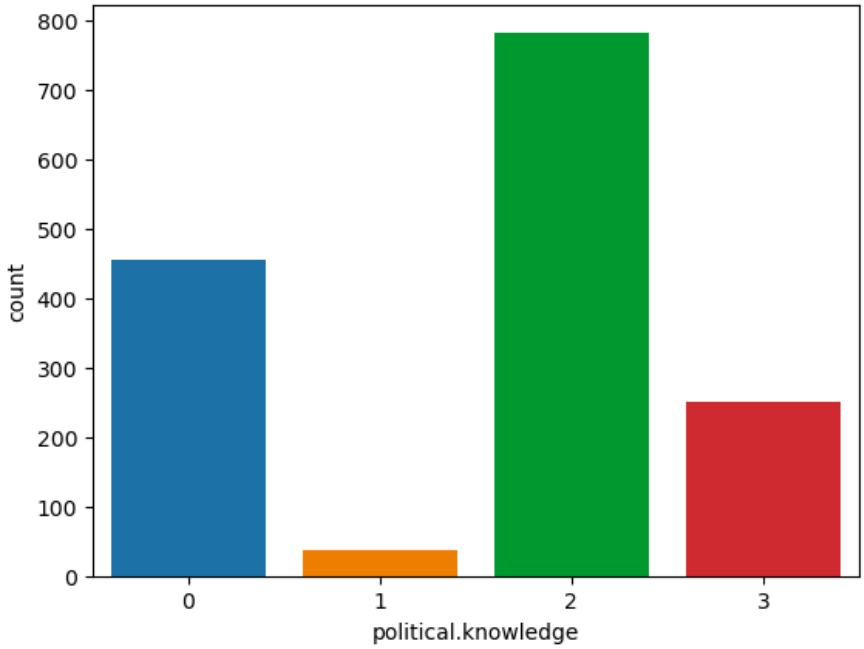


Figure 17

- Knowledge levels range from 0 to 3, reflecting different levels of understanding.
- The mean knowledge level is approximately 1.54.
- Level 2 is the most common, with 782 occurrences.

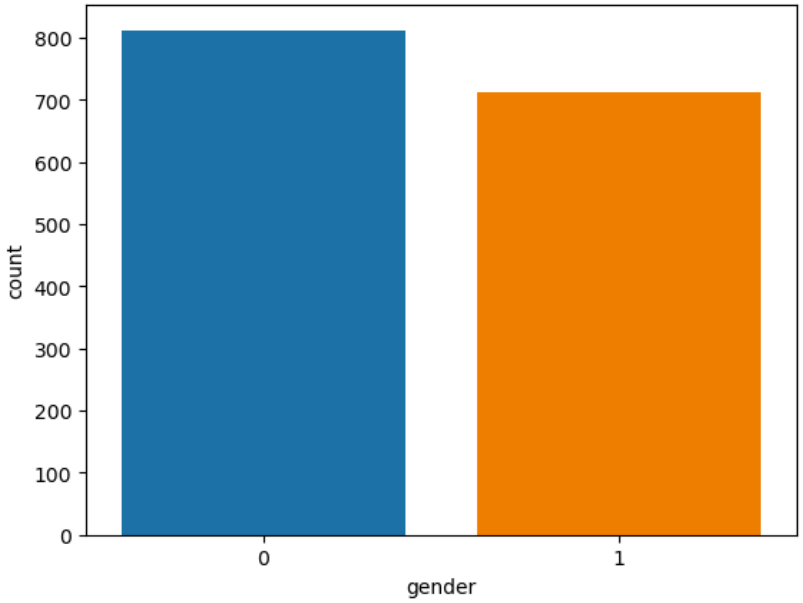
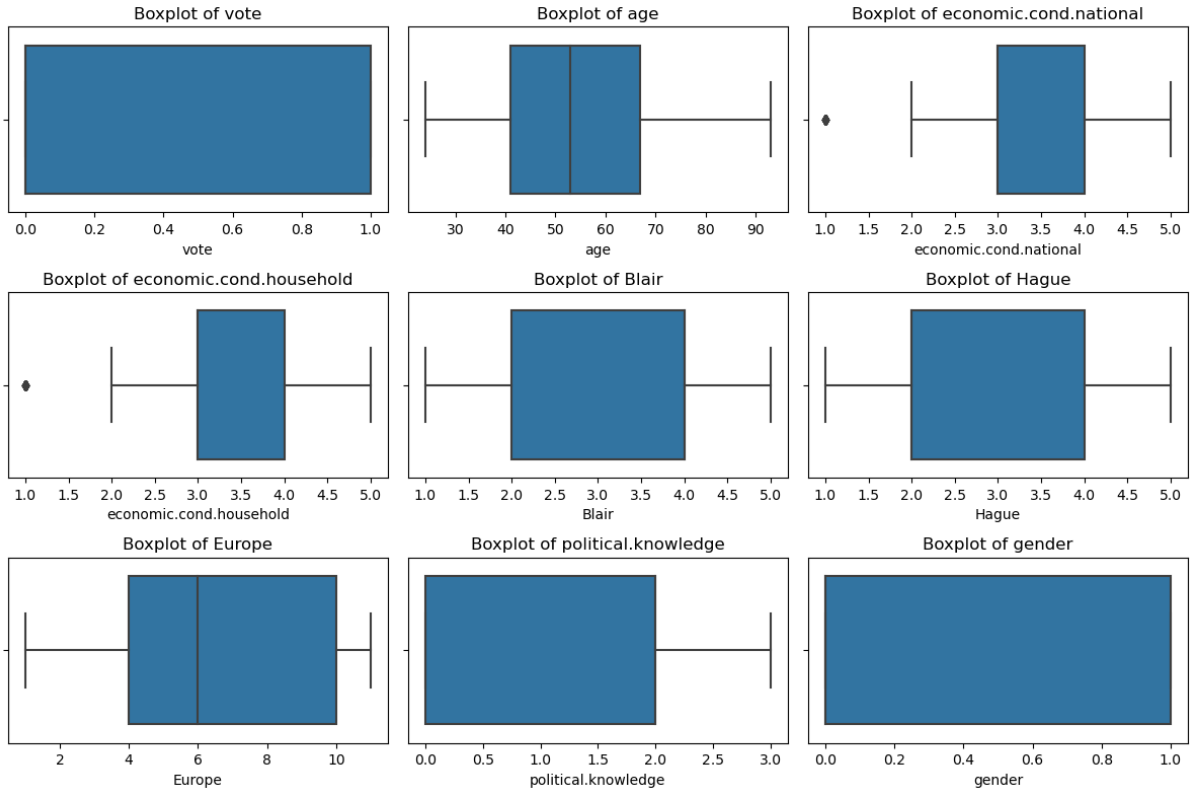


Figure 18

- There are two unique values: 0 (likely female) and 1 (likely male).
- "Female" (0) is the most frequently occurring gender, with 812 observations, while "male" (1) has 713 observations.

Outliers:



It's observed that **outliers** are present in both the "Economic Condition - National" and "Economic Condition - Household" variables. Specifically, these outliers are values of 1, which correspond to poor economic conditions at both the national and household levels.

Given that these outliers represent valid and meaningful data points (i.e., poor economic conditions), **there is no need to treat or remove them**. Retaining these values is important for maintaining the integrity of the dataset and accurately reflecting the distribution of economic conditions, including cases where conditions are poor.

Bivariate Analysis:

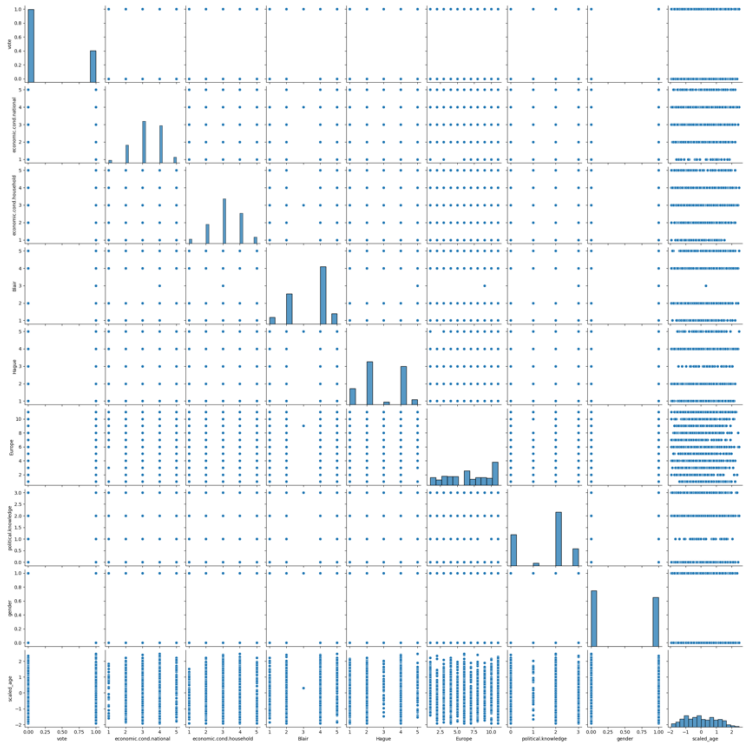


Figure 19

Independence of Variables: The lack of visible patterns, trends, or relationships in the pairplot indicates that the variables may not be strongly dependent on each other.
And the below heatmap also suggests that these variables are largely independent. This information is valuable for understanding the dataset's characteristics and planning further analyses or modeling approaches.

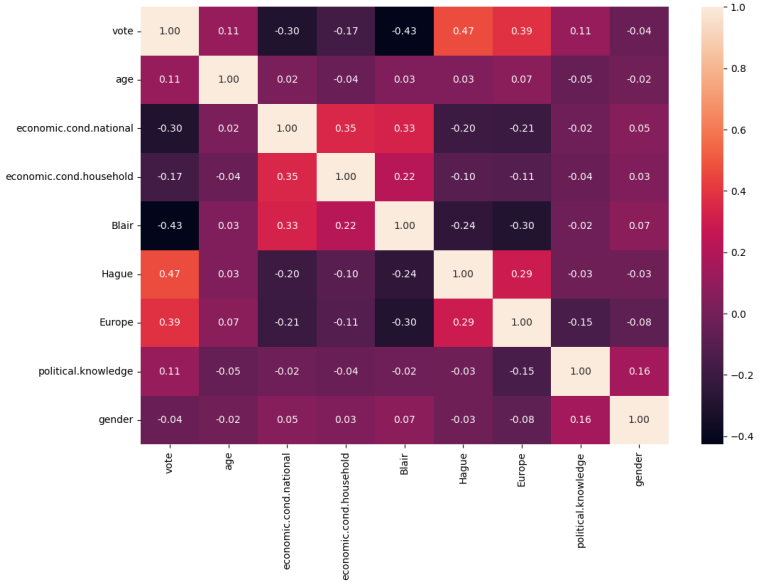


Figure 20

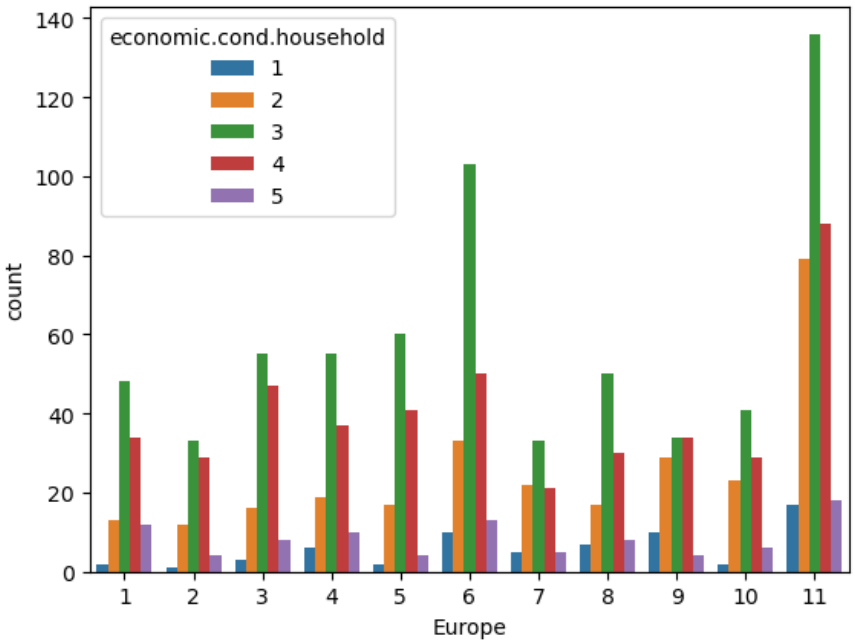
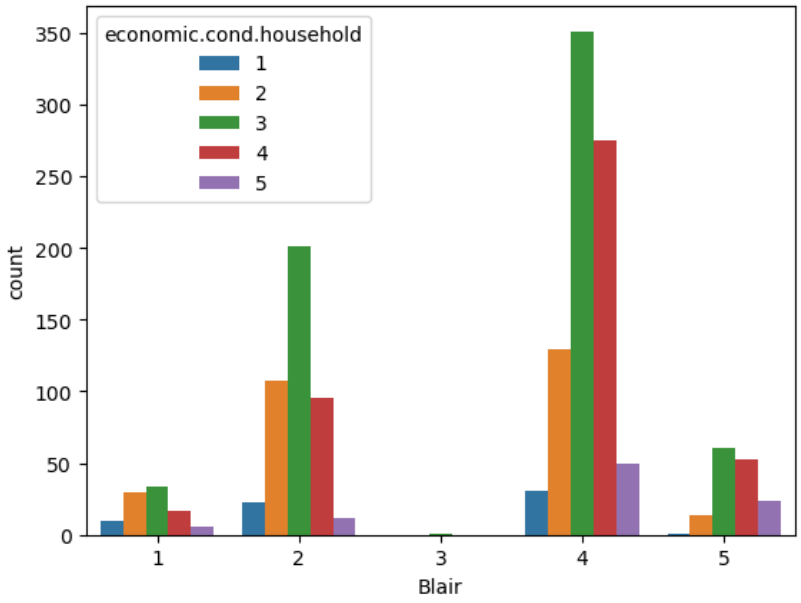
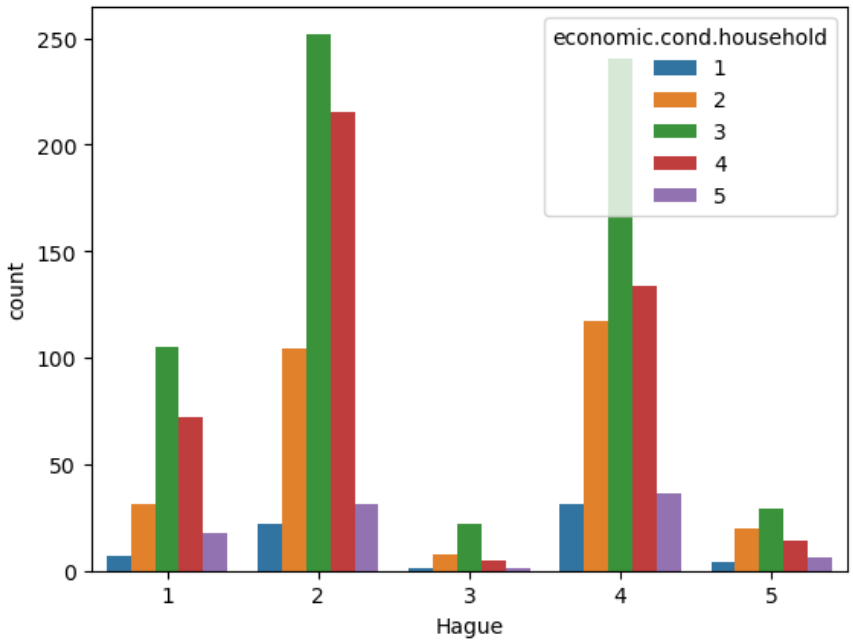


Figure 21

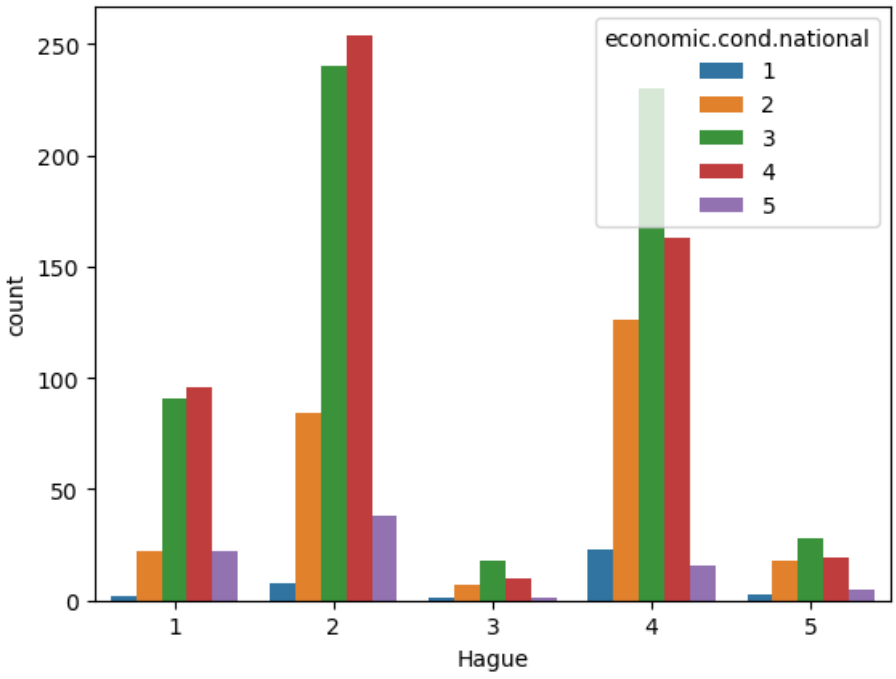
The 'Europe' variable assesses individuals' attitudes toward European integration, employing an 11-point scale. It's noteworthy that the most frequently occurring attitude, marked as '11,' reflects strong sentiments in favor of European integration. Additionally, it's interesting to observe that individuals who exhibit the highest level of support for European integration, as indicated by an attitude score of '11,' often hold assessments of their household economic conditions with a predominant score of '3.'"



The highest rating received is 4 . Notably, the Labour leader garners significant support coming from those who rate their economic conditions as either 3 or 4.

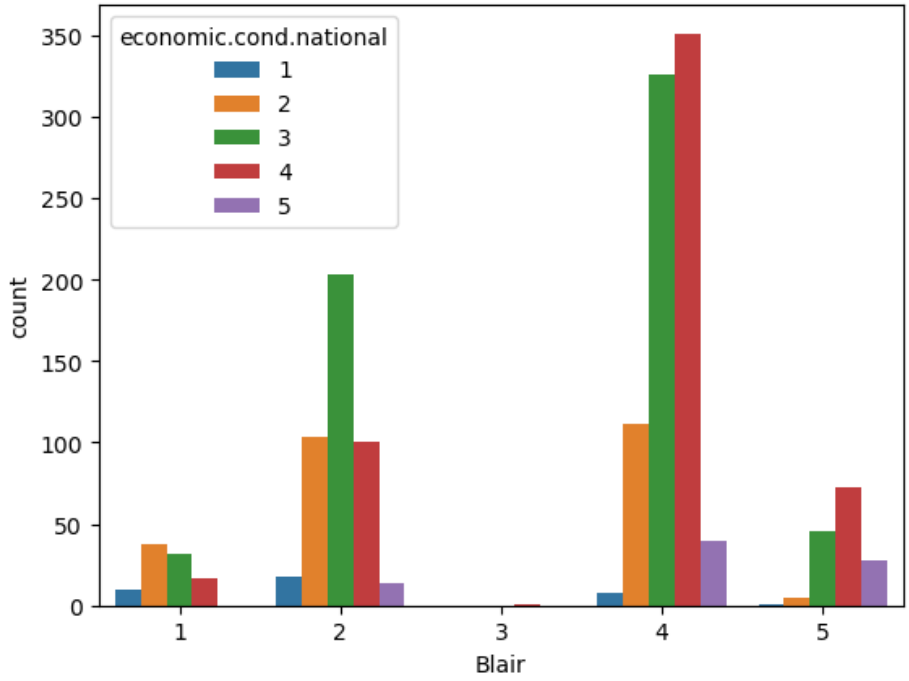


The distribution of assessments for the Conservative leader (Hague) is bimodal, indicating that voters have predominantly rated their support as either 2 or 4. Conversely, the assessments for the Conservative leader are less frequent among voters who have rated their household economic conditions as either 1 or 5, suggesting that the least support is coming from individuals with the most extreme economic condition assessments.

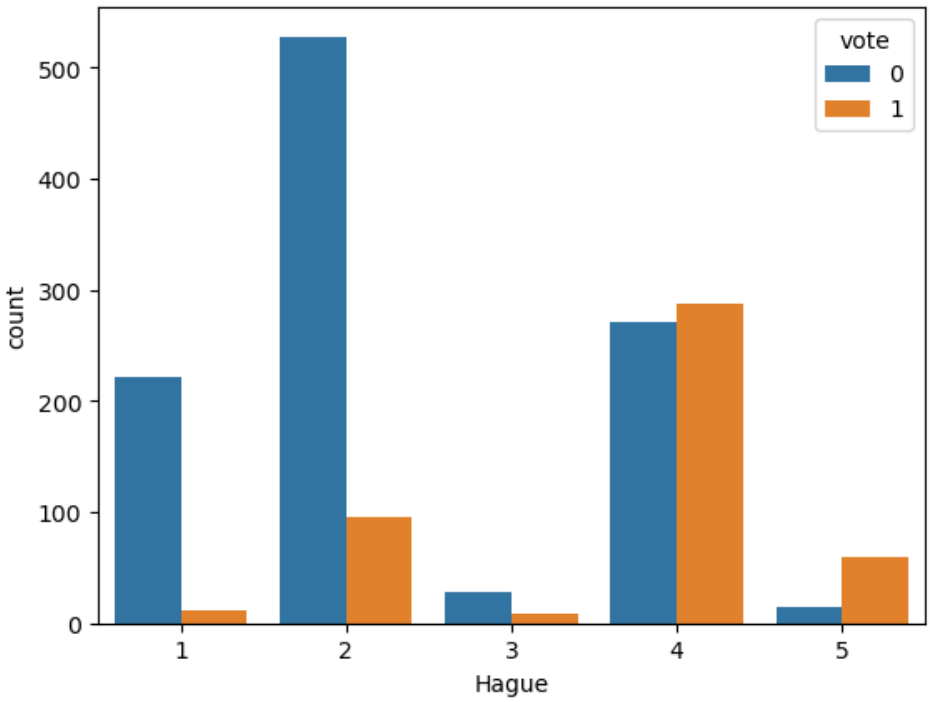


When assessing the Conservative leader (Hague), it is evident that the majority of assessments tend to be lower, falling within the range of 1 to 3. Notably, a

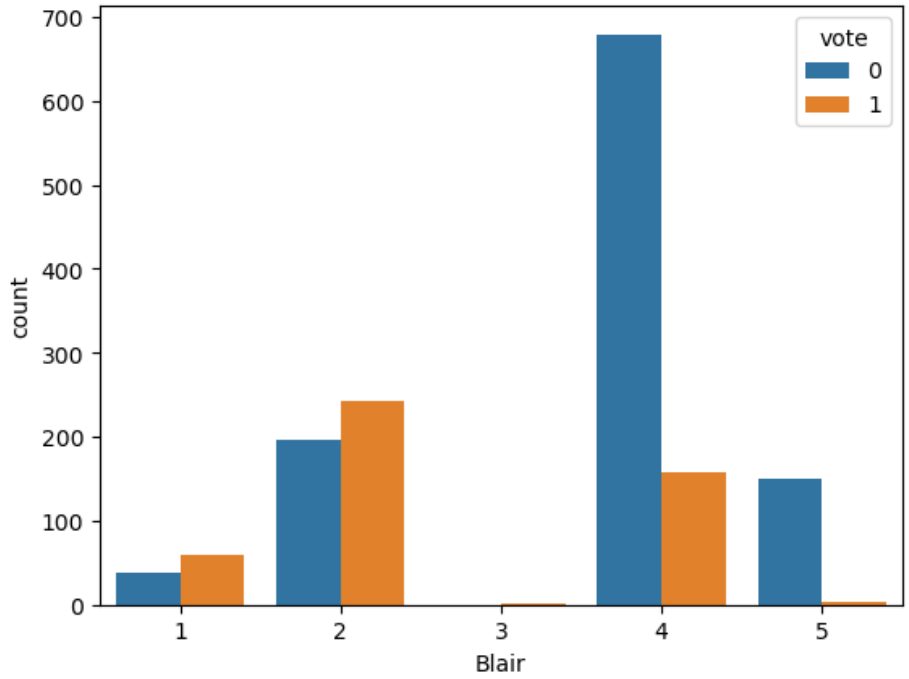
significant portion of individuals who have assessed the Conservative leader with a score of 2 belong to the category of national economic conditions rated as 4.



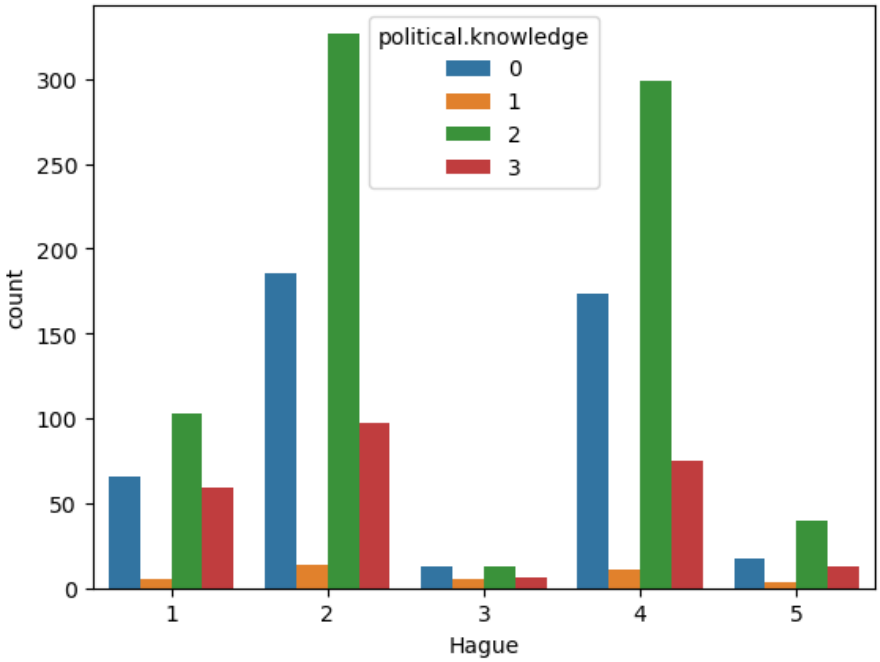
Blair has received the highest rating of 4, signifying strong support from voters. This strong support is particularly evident among individuals who belong to the category of national economic conditions assessed as either 3 or 4. Conversely, the least support for Blair comes from individuals who fall into the category of national economic conditions rated as 1.



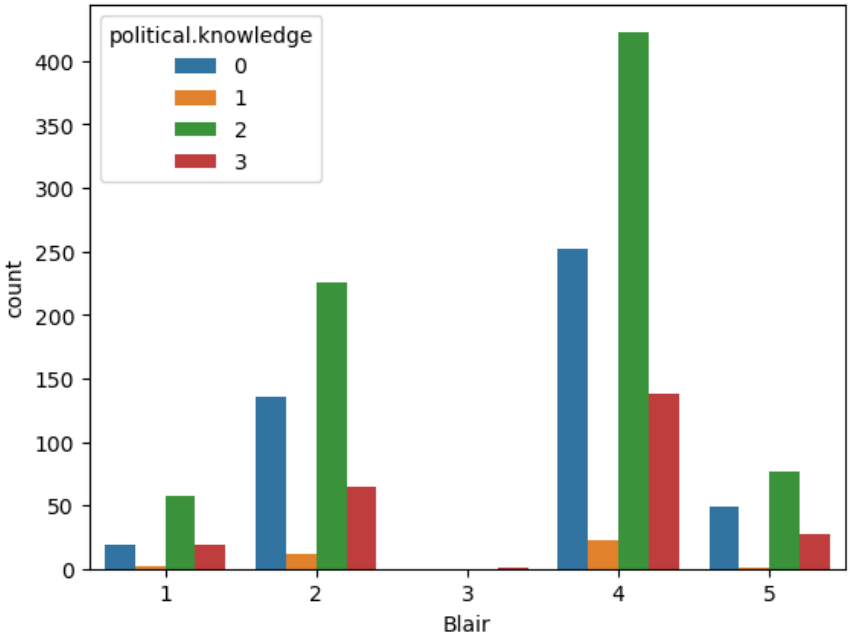
Individuals who rated the Conservative leader (Hague) with assessments of 1 and 2 have shown a preference for the Labour party in their voting choices.



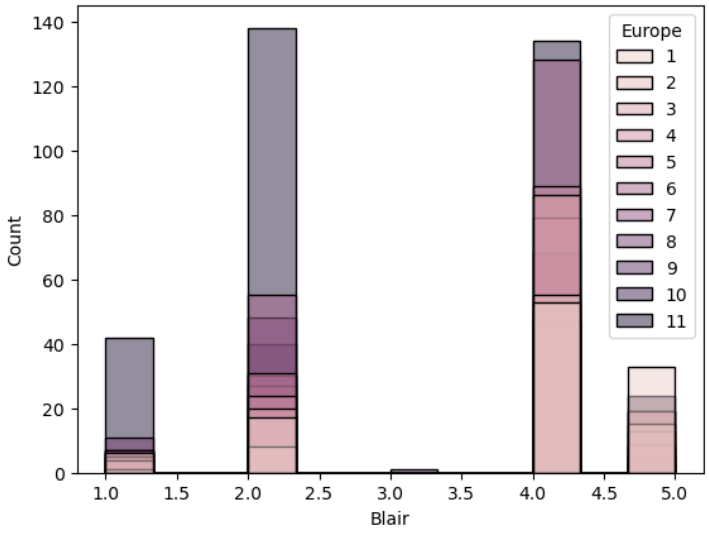
Regardless of their assessments for the Labour leader, those who voted for the Labour party consistently supported the party. Notably, individuals who voted for the Labour party often rated Blair with high assessments of 4 or 5.



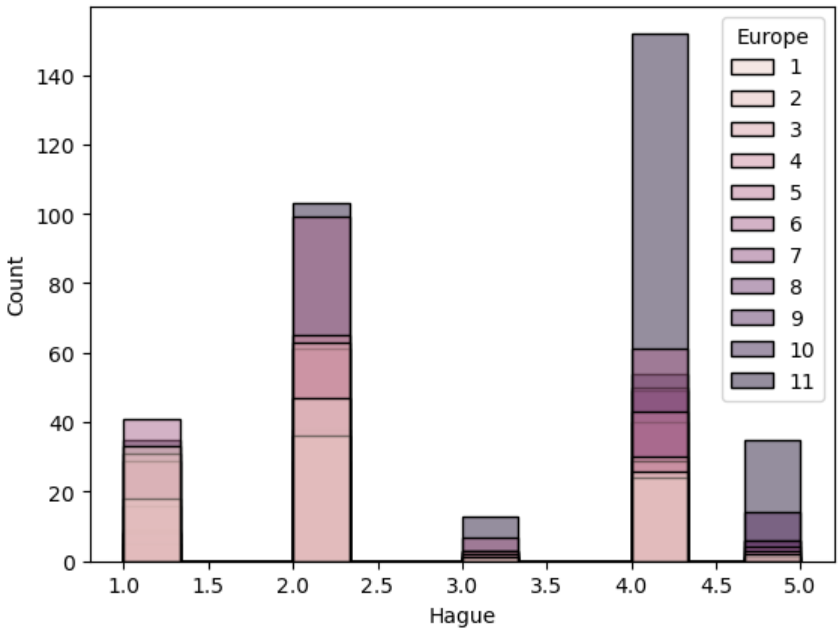
Individuals with a high level of political knowledge have often rated the Conservative leader (Hague) with assessments of 2 or 4. Conversely, those with little to no political knowledge tend to provide similar ratings for Hague.



Individuals with good political knowledge and those with no political knowledge consistently rated the Labour party leader (Blair) with an assessment of 4.



Individuals who hold the highest attitudes toward European integration tend to rate Blair with assessments of 2 and 1.



Individuals who hold the highest attitudes toward European integration tend to rate Hague with assessments of 4 and 5.

1.3 Encoding the data for modeling typically involves converting categorical variables (string values) into numerical representations, such as one-hot encoding or label encoding, so that machine learning algorithms can work with them. In this case, as we have two categorical variables: "vote" (Party choice: Conservative or Labour) and "gender" (female or male).

Here's how we can encode these variables:

1. **Vote Variable (Party Choice):** You can use label encoding to convert "Conservative" to 1 and "Labour" to 0. This allows the model to work with these values.
2. **Gender Variable:** You can use label encoding as well, where "female" is encoded as 0, and "male" is encoded as 1.

The data is already encoded. **(Refer figure 6,7)**

As for **scaling**, it's typically necessary when you have numerical features with different ranges or units. Scaling ensures that all features have a similar scale, which can be important for many machine learning algorithms, particularly those that rely on distance metrics or gradient-based optimization.

Since all the variables are on the same scale except for "age," we can focus on scaling the "age" variable and proceed with your modeling steps. (Refer Figure 22)

	vote	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender	scaled_age
1	0	3	3	4	1	2	2	0	-0.711973
2	0	4	4	4	4	5	2	1	-1.157661
3	0	4	4	5	2	3	2	1	-1.221331
4	0	4	2	2	1	4	0	0	-1.921698
5	0	2	2	1	1	6	2	1	-0.839313
...	...	...	...	...	...	...	...	...	...
1521	1	5	3	2	4	11	3	1	0.816100
1522	1	2	2	4	4	8	2	1	1.198118
1523	0	3	3	5	4	2	2	1	-1.093992
1524	1	3	3	1	4	11	2	1	0.434081
1525	1	2	3	2	4	11	0	0	1.261787

1525 rows × 9 columns

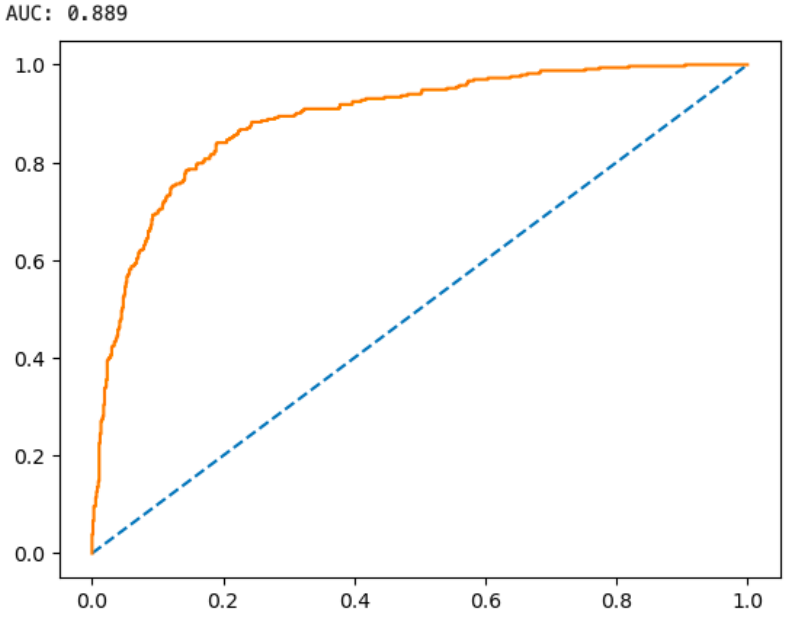
Figure 22

The data is split in the ratio 70:30:

- **Training Data (70%):** This portion of the data (70%) is used to train your machine learning model. During training, the model learns the relationships and patterns in the data.
- **Testing Data (30%):** The remaining 30% of the data is reserved for testing and evaluating the model's performance. It serves as a holdout dataset that the model has never seen during training.

1.4 Logistic Regression :

Training Data Evaluation:
- The logistic regression model achieved a training accuracy score of approximately 84.07%, indicating that it correctly classified about 84.07% of the training data.
- The area under the curve (AUC) for the receiver operating characteristic (ROC) curve is approximately 0.889, suggesting good discrimination between classes.

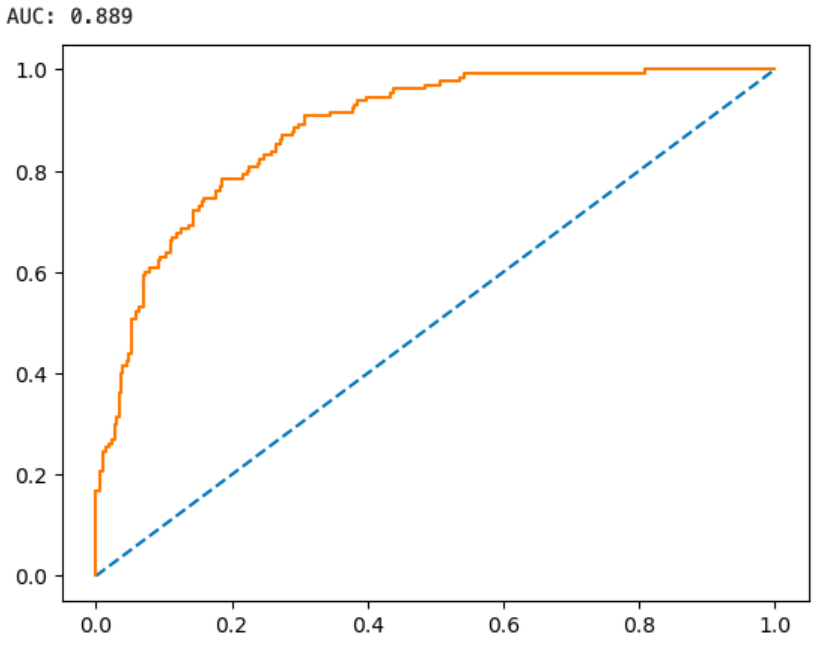


- The confusion matrix for the training data reveals the following:
  - True Positives (TP): 230
  - True Negatives (TN): 667
  - False Positives (FP): 68
  - False Negatives (FN): 102

- The precision, recall, and F1-score for class 0 (Conservative) and class 1 (Labour) are provided. Class 0 has higher precision (0.87), while class 1 has lower precision (0.77). Class 0 also has higher recall (0.91) compared to class 1 (0.69).
- The macro and weighted averages for precision, recall, and F1-score are also provided.

	precision	recall	f1-score	support
0	0.87	0.91	0.89	735
1	0.77	0.69	0.73	332
accuracy			0.84	1067
macro avg	0.82	0.80	0.81	1067
weighted avg	0.84	0.84	0.84	1067

Testing Data Evaluation:
- The logistic regression model achieved a testing accuracy score of approximately 82.31%, indicating that it correctly classified about 82.31% of the testing data.
- The AUC for the ROC curve is consistent with the training data, approximately 0.889.



- The confusion matrix for the testing data reveals the following:
  - True Positives (TP): 85
  - True Negatives (TN): 292

  - False Positives (FP): 36
  - False Negatives (FN): 45
- The precision, recall, and F1-score for class 0 and class 1 are provided. Similar to the training data, class 0 (Conservative) has higher precision (0.87) and recall (0.89) compared to class 1 (Labour), which has lower precision (0.70) and recall (0.65).
- The macro and weighted averages for precision, recall, and F1-score are also provided.

	precision	recall	f1-score	support
0	0.87	0.89	0.88	328
1	0.70	0.65	0.68	130
accuracy			0.82	458
macro avg	0.78	0.77	0.78	458
weighted avg	0.82	0.82	0.82	458

Conclusion:
- The logistic regression model demonstrates reasonably good performance in classifying voters' choices between the Conservative and Labour parties.
- The model has achieved consistent AUC values between the training and testing datasets, indicating robustness.
- Class 0 (Conservative) tends to have higher precision and recall compared to class 1 (Labour), suggesting that the model is more accurate in predicting Conservative votes.
- The model provides reliable predictions and could be used for making inferences about voter choices in the given context.

LDA:

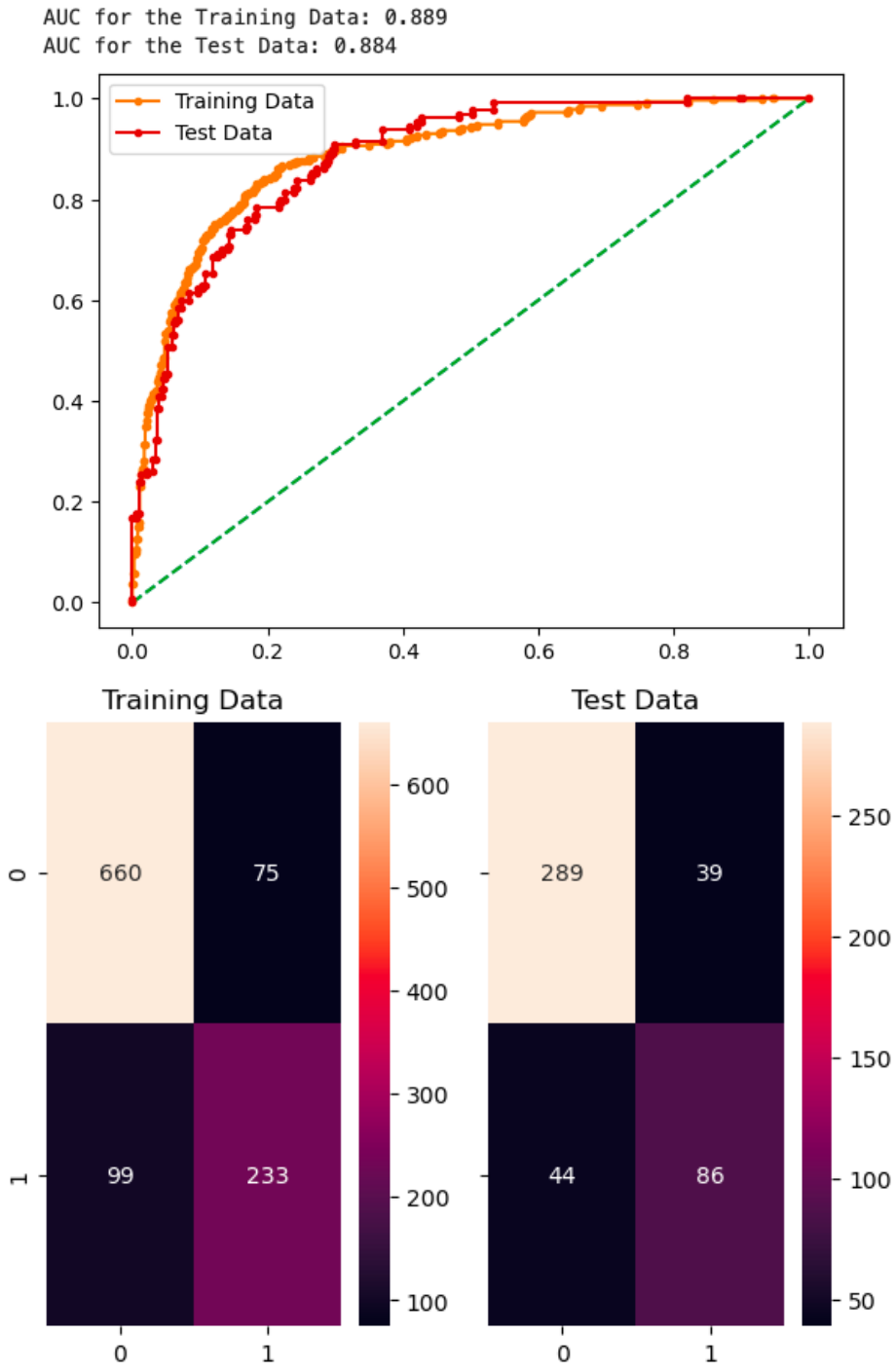
Classification Report of the training data:

	precision	recall	f1-score	support
0	0.87	0.90	0.88	735
1	0.76	0.70	0.73	332
accuracy			0.84	1067
macro avg	0.81	0.80	0.81	1067
weighted avg	0.83	0.84	0.84	1067

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.87	0.88	0.87	328
1	0.69	0.66	0.67	130
accuracy			0.82	458
macro avg	0.78	0.77	0.77	458
weighted avg	0.82	0.82	0.82	458

Training Data Evaluation:
- LDA achieved an accuracy of approximately 84% on the training data, indicating that it correctly classified about 84% of the data.
- For class 0 (Conservative), LDA achieved a precision of 0.87 and a recall of 0.90, resulting in an F1-score of 0.88.
- For class 1 (Labour), LDA achieved a precision of 0.76 and a recall of 0.70, resulting in an F1-score of 0.73.
- The macro and weighted averages for precision, recall, and F1-score are also provided. These metrics show good performance on the training data.
- The AUC for the ROC curve on the training data is approximately 0.889, indicating strong discrimination between classes.



Testing Data Evaluation:

- LDA achieved an accuracy of approximately 82% on the test data, indicating that it correctly classified about 82% of the data.
- For class 0 (Conservative), LDA achieved a precision of 0.87 and a recall of 0.88, resulting in an F1-score of 0.87.
- For class 1 (Labour), LDA achieved a precision of 0.69 and a recall of 0.66, resulting in an F1-score of 0.67.

- The macro and weighted averages for precision, recall, and F1-score are also provided. These metrics indicate satisfactory performance on the test data.
- The AUC for the ROC curve on the test data is approximately 0.884, which is consistent with the training data.

Conclusion:
- Linear Discriminant Analysis (LDA) demonstrates good classification performance in predicting voter choices between the Conservative and Labour parties.
- The model is particularly strong in correctly classifying voters who choose the Conservative party (class 0) based on high precision and recall values.
- While the model performs well for the Labour party (class 1), there is room for improvement, as reflected in slightly lower precision and recall values.
- The AUC values for both training and testing data suggest that the model's discrimination ability is robust.
- LDA is a viable method for this classification problem, and its performance is competitive with the logistic regression model. Further fine-tuning or exploring other classification algorithms may lead to even better results.

```
array([[-0.34920682, -0.1503469 , -0.70441972,  0.96718782,  0.25856319,
         0.57234288, -0.24908203,  0.40241615]])
```

Refer the above image, each coefficient represents the change in the log-odds of the dependent variable (voter choice) associated with a one-unit change in the corresponding independent variable while holding all other variables constant. Here's an interpretation of the coefficients for each feature:

1. economic.cond.national: A one-unit increase in the assessment of national economic conditions (going from poor to excellent) is associated with a decrease in the log-odds of choosing the Conservative party (class 1) by approximately 0.349 units, holding all other variables constant.

2. economic.cond.household: A one-unit increase in the assessment of household economic conditions (going from poor to excellent) is associated with a decrease in the log-odds of choosing the Conservative party (class 1) by approximately 0.150 units, holding all other variables constant.

3. Blair: An increase in the assessment of the Labour leader (going from poor to excellent) is associated with a significant decrease in the log-odds of choosing the Conservative party (class 1) by approximately 0.704 units, holding all other variables constant.

4. Hague: An increase in the assessment of the Conservative leader (going from poor to excellent) is associated with a significant increase in the log-odds of choosing the Conservative party (class 1) by approximately 0.967 units, holding all other variables constant.

5. Europe: An increase in attitudes toward European integration is associated with an increase in the log-odds of choosing the Conservative party (class 1) by approximately 0.259 units, holding all other variables constant.

6. political.knowledge: An increase in political knowledge is associated with an increase in the log-odds of choosing the Conservative party (class 1) by approximately 0.572 units, holding all other variables constant.

7. gender: Being male (gender = 1) is associated with a decrease in the log-odds of choosing the Conservative party (class 1) by approximately 0.249 units compared to being female (gender = 0), holding all other variables constant.

8. scaled_age: A one-unit increase in scaled age is associated with an increase in the log-odds of choosing the Conservative party (class 1) by approximately 0.402 units, holding all other variables constant.

These coefficients help you **understand the direction and strength of the relationship** between each feature and the probability of

choosing the Conservative party. Positive coefficients indicate an increase in the likelihood of choosing the Conservative party, while negative coefficients indicate a decrease in that likelihood, all else being equal.

1.5 KNN Model:

For the K-Nearest Neighbors (KNN) model, the following accuracy results have been obtained:

Train Accuracy: 0.84 (84%)
Test Accuracy:0.81 (81%)

These accuracy scores indicate how well the KNN model performs in classifying voter choices between the Conservative and Labour parties. An accuracy of 84% on the training data and 81% on the test data suggests that the model is making correct predictions for a substantial portion of the dataset.

It's worth noting that there is a slight drop in accuracy from the training set to the test set, which is expected and indicates that the model is generalizing reasonably well.

To make a comprehensive assessment of the model's performance, it's essential to consider other evaluation metrics such as precision, recall, F1-score, and AUC, as well as to compare it with other classification models. The choice of the most suitable model may depend on specific project requirements and objectives.

```
KNN Train Accuracy: 0.84
KNN Test Accuracy: 0.81
Naïve Bayes Train Accuracy: 0.83
Naïve Bayes Test Accuracy: 0.83
```

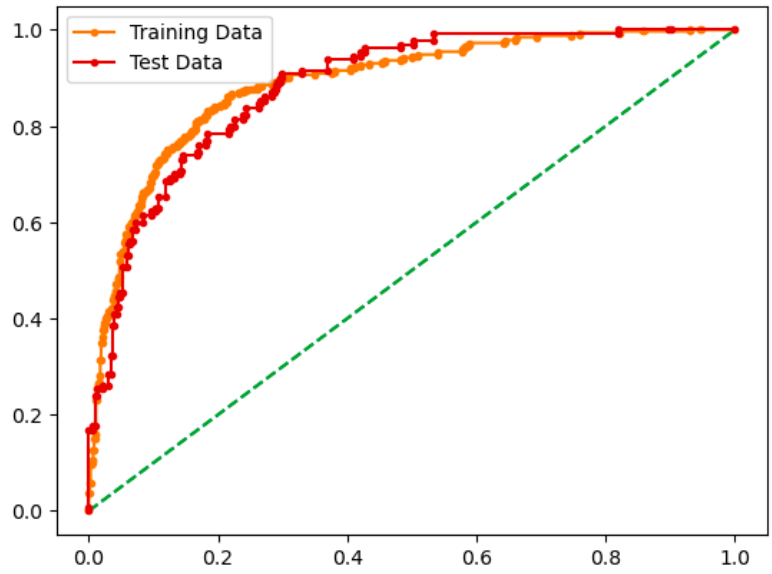
Naïve Bayes Model:

The Naïve Bayes model (GaussianNB) has been evaluated on both the training and test datasets. Here's a summary of the evaluation results:

Test Data Evaluation:
- Confusion Matrix:
  - True Positives (TP): 94
  - True Negatives (TN): 284
  - False Positives (FP): 36
  - False Negatives (FN): 44
- Precision for class 0 (Conservative): 0.87
- Recall for class 0 (Conservative): 0.89
- F1-score for class 0 (Conservative): 0.88
- Precision for class 1 (Labour): 0.72
- Recall for class 1 (Labour): 0.68
- F1-score for class 1 (Labour): 0.70
- Overall Test Accuracy: 0.83 (83%)
- The AUC for the ROC curve on the test data is approximately 0.884, indicating good discrimination between classes.

```
GaussianNB()
[[284  36]
 [ 44  94]]
              precision    recall  f1-score   support

           0       0.87      0.89      0.88       320
           1       0.72      0.68      0.70       138

    accuracy                           0.83       458
   macro avg       0.79      0.78      0.79       458
weighted avg       0.82      0.83      0.82       458


[[649  92]
 [ 86 240]]
              precision    recall  f1-score   support

           0       0.88      0.88      0.88       741
           1       0.72      0.74      0.73       326

    accuracy                           0.83      1067
   macro avg       0.80      0.81      0.80      1067
weighted avg       0.83      0.83      0.83      1067
```

AUC for the Training Data: 0.889
AUC for the Test Data: 0.884



Training Data Evaluation:
- Confusion Matrix:
  - True Positives (TP): 240
  - True Negatives (TN): 649
  - False Positives (FP): 92
  - False Negatives (FN): 86
- Precision for class 0 (Conservative): 0.88
- Recall for class 0 (Conservative): 0.88
- F1-score for class 0 (Conservative): 0.88
- Precision for class 1 (Labour): 0.72
- Recall for class 1 (Labour): 0.74
- F1-score for class 1 (Labour): 0.73
- Overall Train Accuracy: 0.83 (83%)
- The AUC for the ROC curve on the training data is approximately 0.889, indicating strong discrimination between classes.

Conclusion:
- The Naïve Bayes model, specifically GaussianNB, demonstrates reasonably good performance in classifying voter choices between the Conservative and Labour parties.
- The model is particularly strong in correctly classifying voters who choose the Conservative party (class 0) based on high precision, recall, and F1-score values.

- While the model performs well for the Labour party (class 1), there is a slightly lower precision value, indicating room for improvement.
- The AUC values for both training and testing data suggest that the model's discrimination ability is robust.
- Overall, Naïve Bayes is a viable method for this classification problem, and its performance is competitive with the logistic regression and K-Nearest Neighbors (KNN) models. Further fine-tuning or exploring other classification algorithms may lead to even better results.

1.6 Bagging

Bagging, which is performed using a decision tree classifier, has shown significant improvement in classification accuracy when compared to a standalone model. Here's a summary of the results:

Standalone Model:
- Array of Accuracy Scores: [0.73442623, 0.78360656, 0.75737705, 0.76721311, 0.73770492]
- Mean Accuracy: 0.75606557

Bagging Classifier (Using Decision Tree):
- Train Accuracy: 0.81630
- Test Accuracy: 0.812227

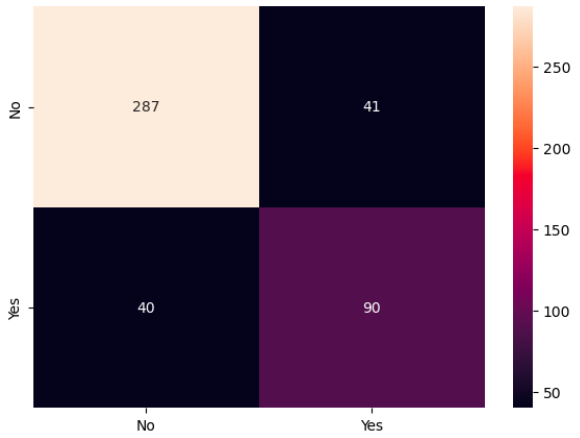
After Cross-Validation:
- Array of Accuracy Scores: [0.80983607, 0.81311475, 0.8295082, 0.82622951, 0.82295082]
- Mean Accuracy: 0.820327

Conclusion:
Bagging, which combines multiple models to improve predictive accuracy, has demonstrated its effectiveness in this scenario. The Bagging Classifier using a decision tree as the base model has significantly increased the accuracy of the model when compared to a standalone decision tree model.

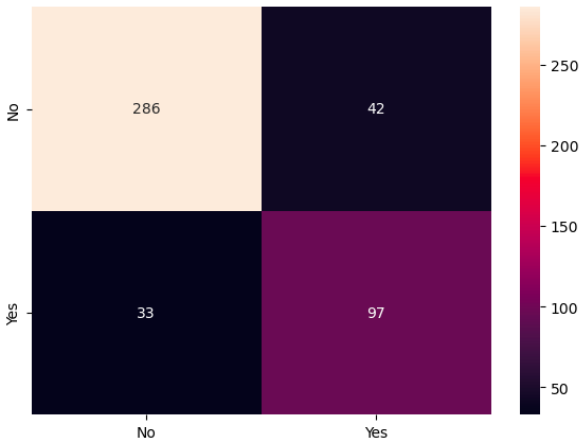
The mean accuracy of the bagging classifier is approximately 0.820327, which is higher than the mean accuracy of the standalone model (0.75606557). This indicates that bagging has improved the model's ability to make accurate predictions.

Boosting:

AdaBoostClassifier and GradientBoostingClassifier have been applied to the dataset, and their test and post-GridSearchCV results are as follows:



AdaBoostClassifier: Confusion matrix



GradientBoostingClassifier: Confusion matrix

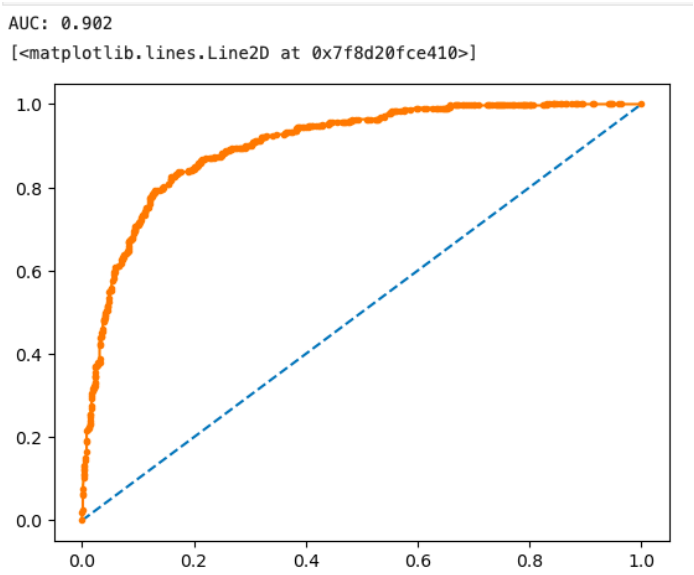
```
0.8350515463917526
[[674  61]
 [115 217]]
              precision    recall  f1-score   support

           0       0.85      0.92      0.88       735
           1       0.78      0.65      0.71       332

    accuracy                           0.84      1067
   macro avg       0.82      0.79      0.80      1067
weighted avg       0.83      0.84      0.83      1067
```

AdaBoostClassifier: Train classification report

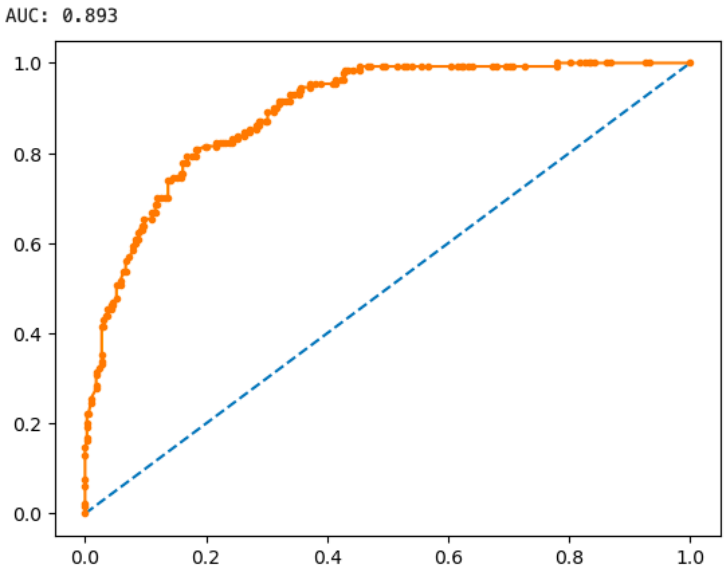
```
0.8318777292576419
[[296  32]
 [ 45  85]]
              precision    recall  f1-score   support

           0       0.87      0.90      0.88       328
           1       0.73      0.65      0.69       130

    accuracy                           0.83       458
   macro avg       0.80      0.78      0.79       458
weighted avg       0.83      0.83      0.83       458
```

AdaBoostClassifier: Test classification report

Test:



Test:



```
The best parameters are {'max_features': 2, 'n_estimators': 150} with a score of 0.83
```

- Both AdaBoostClassifier and GradientBoostingClassifier are ensemble methods that have demonstrated good performance in classifying voter choices between the Conservative and Labour parties.
- The test accuracy for AdaBoostClassifier is approximately 0.823144, while for GradientBoostingClassifier, it is approximately 0.8362445. These accuracy scores indicate that both models make correct predictions for a significant portion of the test dataset.
- After performing GridSearchCV on GradientBoostingClassifier, the train accuracy improved to approximately 0.8350515463917526. The precision, recall, and F1-score for both classes are competitive.
- The test accuracy for GradientBoostingClassifier after GridSearchCV is approximately 0.8318777292576419, and it maintains a balance between precision and recall for both classes.

Based on these metrics, GradientBoostingClassifier tends to have slightly higher accuracy and balanced precision and recall scores for both classes. However, the difference in accuracy between AdaBoostClassifier and GradientBoostingClassifier is not substantial. In terms of accuracy and overall performance, GradientBoostingClassifier appears to be the better choice.

1.8 Insights:

In simple terms, let's understand how each feature impacts the likelihood of someone choosing the Conservative party (class 1) as opposed to the Labour party (class 0):

1. Economic Conditions (National and Household):
   - If voters have a more positive perception of both national and household economic conditions, they are less likely to choose the Conservative party.

2. Assessment of the Labour Leader (Blair):
   - A better assessment of the Labour leader (Blair) leads to a decreased likelihood of choosing the Conservative party.

3. Assessment of the Conservative Leader (Hague):
   - A better assessment of the Conservative leader (Hague) increases the likelihood of choosing the Conservative party.

4. Attitudes Toward European Integration (Europe):
   - More positive attitudes toward European integration are associated with a higher likelihood of choosing the Conservative party.

5. Political Knowledge:
   - Having more political knowledge increases the likelihood of choosing the Conservative party.

6. Gender:
   - Being male decreases the likelihood of choosing the Conservative party compared to being female.

7. Age:
   - As age increases, the likelihood of choosing the Conservative party also increases.

In summary, the assessment of party leaders, attitudes toward European integration, political knowledge, and age play a role in determining whether someone supports the Conservative party. Conversely, a more positive view of economic conditions and being female are associated with a higher likelihood of choosing the Labour party. These insights can be valuable for political campaigns and strategies.

Problem: 2

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

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

Text Analytics:

Text analysis is the process of using computer systems to read and understand human-written text for business insights. Text analysis software can independently classify, sort, and extract information from text to identify patterns, relationships, sentiments, and other actionable knowledge.

NLTK:

NLTK, or Natural Language Toolkit, is a Python package that you can use for NLP. A lot of the data that you could be analyzing is unstructured data and contains human-readable text. Before you can analyze that data programmatically, you first need to preprocess it.

2.1: Number of characters, words and sentences for the mentioned documents.

Data used: [Speeches]

	Name	Speech
0	Roosevelt	On each national day of inauguration since 178...
1	Kennedy	Vice President Johnson, Mr. Speaker, Mr. Chief...
2	Nixon	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...

Dataset overview

	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

Represents word count in each speeches

	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

Represents number of character count in each speeches

	Speech	sentence_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

The number of sentences in each speeches.

2.2 Remove all the stopwords from all three speeches:

	Speech	avg_word
0	On each national day of inauguration since 178...	4.783825
1	Vice President Johnson, Mr. Speaker, Mr. Chief...	4.626100
2	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	4.713397

Shows average length of the word in the whole speech

	Speech	stopwords
0	On each national day of inauguration since 178...	632
1	Vice President Johnson, Mr. Speaker, Mr. Chief...	618
2	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	899

The table provides information about the number of stopwords in three different speeches. The first speech contains 632 stopwords, while the second speech contains 618 stopwords, and the third speech contains 899 stopwords.

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

Contains 0 hashtags

	Speech	numerics
0	On each national day of inauguration since 178...	14
1	Vice President Johnson, Mr. Speaker, Mr. Chief...	7
2	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	10

	Speech	UpperCase
0	On each national day of inauguration since 178...	1
1	Vice President Johnson, Mr. Speaker, Mr. Chief...	5
2	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	13

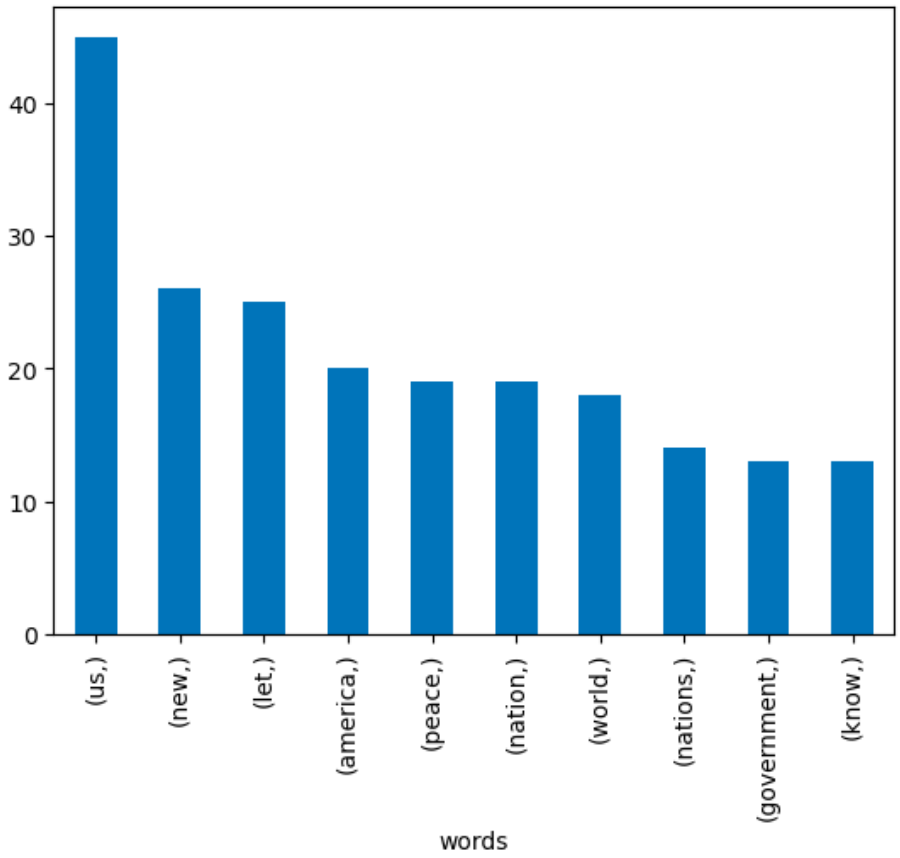
2.3: Which word occurs the most number of times in his inaugural address for each president? Mention the top three words. (after removing the stopwords)

words	
us	45
new	26
let	25
america	20
peace	19
nation	19
world	18
nations	14
government	13
know	13

The table provides a list of words and the number of times they occur in the speeches. Some of the words that occur frequently in the speeches are:

1. "us" - 45 times
2. "new" - 26 times
3. "let" - 25 times
4. "America" - 20 times

These words appear repeatedly in the speeches and may indicate their significance in the context of the addresses.



The top 3 words :

1. Roosevelt:

```
The top three words in Roosevelt's Speech(after removing the stopwords) are :
 [('nation', 17), ('know', 10), ('peopl', 9), ('spirit', 9), ('life', 9), ('democraci', 9)]
```

2.Kennedy:

```
The top three words in Kennedy's Speech(after removing the stopwords) are :
 [('let', 16), ('us', 12), ('power', 9)]
```

3. Nixon:

```
The top three words in Nixon's Speech(after removing the stopwords) are :
 [('us', 26), ('let', 22), ('america', 21)]
```

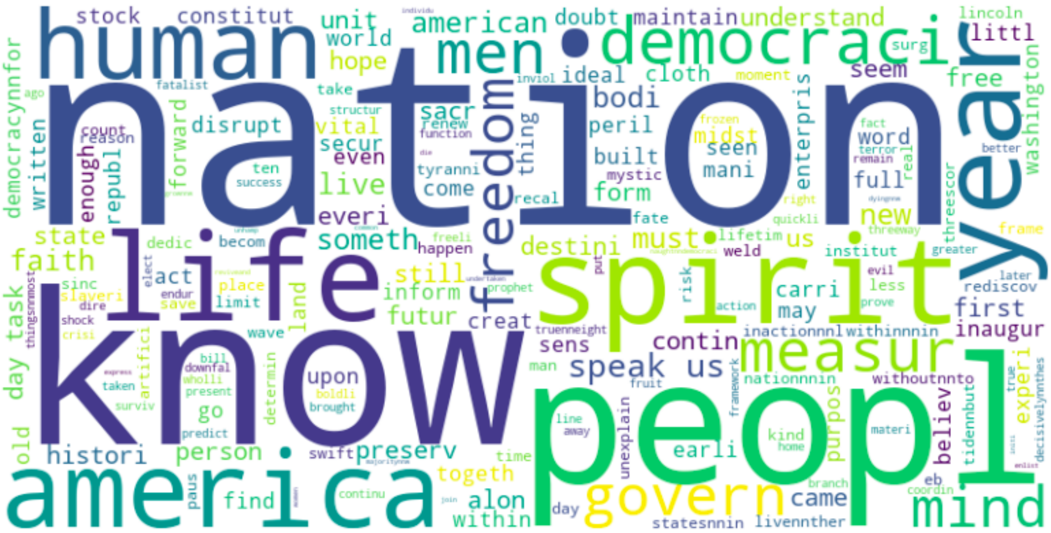
2.4 The word clouds for the inaugural speeches of Presidents Roosevelt, Kennedy, and Nixon provide some insights into the key themes and focus areas of their speeches:

1. Roosevelt:
   - Major words: human, nation, people, life, know, America
   - Themes: The words "human," "nation," "people," and "life" suggest a strong emphasis on addressing the needs and well-being of the American people. "America" is a common patriotic term.
   - Conclusion: Roosevelt's speech appears to focus on issues related to the welfare and unity of the American people, emphasizing a sense of national identity and purpose.

2.Kennedy:
   - Major words: side, power, new, pledge, nation, world
   - Themes: The words "power," "new," "pledge," "nation," and "world" indicate a speech that emphasizes the role and responsibilities of the United States on the global stage. "Pledge" suggests a commitment to specific goals.
   - Conclusion: Kennedy's speech appears to focus on the nation's role and responsibilities in the world, particularly in the context of the Cold War and global politics.

3. Nixon:
   - Major words: response, peace, new, America, nation
   - Themes: The words "response," "peace," "new," "America," and "nation" suggest a speech focused on addressing the challenges of the era, particularly the Vietnam War, and striving for peace and renewal.
   - Conclusion: Nixon's speech appears to center on the idea of responding to the nation's challenges, pursuing peace, and renewing the spirit of America.

Word Cloud - Franklin D. Roosevelt 1941



Word Cloud - John F. Kennedy 1961

Word Cloud - Richard Nixon 1973