MACHINE LEARNING BUSINESS REPORT

Problem 1:You are hired by one of the leading news channels CNBE who wantsto analyze recent elections. This surveywas 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, tocreate an exit poll that will help in predicting overall win and seats coveredby a particular party.

1.1)Read the dataset. Do the descriptive statistics and do the null value conditioncheck. Write an inference on it. (4 Marks)

Response:-

Introduction: This report presents an analysis of [Dataset Name], focusing on the data's basic characteristics, including data summary, missing values, and duplicate values. The objective of this analysis is to gain insights into the dataset and prepare it for clustering analysis.

We found the basic information of the data,

1. The dataset has 6 float data type, 7 int data type and 2 object data type.
2. The Shape of the dataset is (1525,9) that implies it has 1525 records and 9 columns

Data Preprocessing:

Based on the initial data exploration, the following preprocessing steps may be considered:

• Handling missing values: [Describe how missing values will be addressed]

• Removing duplicates: [Explain how duplicates will be handled]

• Feature scaling/normalization: [If applicable, mention if features will be scaled or normalized]

• This initial analysis provides a foundational understanding of the dataset, highlighting potential areas for further investigation. As the analysis progresses, more insights will be gained, leading to meaningful business outcomes and recommendations.

• Continue with clustering analysis based on the data preprocessing steps.

• Explore the relationships between clusters and business objectives.

• Monitor and adapt the analysis as more insights are uncovered

Null Values Treatment:

Null values represent missing or undefined data in a database or dataset. They signify the absence of a value in a field or a variable and can occur for various reasons, such as incomplete data, data entry errors, or undefined information.

Dealing with null values is essential in data analysis and database management to ensure accurate results and avoid errors in computations or interpretations. Here are some common strategies for handling null values:

1. **Identify Null Values:** First, identify which columns or fields contain null values in your dataset.
2. **Remove Null Values:** If null values are present in a small percentage and won't significantly affect the analysis, you can choose to remove rows or columns containing null values. However, this approach might lead to a loss of data.
3. **Impute Null Values:** Instead of removing null values, you can replace them with a specific value. Common techniques for imputing nulls include:
    - **Mean/Median/Mode Imputation:** Replace null values with the mean, median, or mode of the non-null values in that column

-> We have no null values in the dataset

Special Characters Check:

Special characters refer to any character that is not an alphabetic (a to z) or numeric (0 to 9) character. These characters include punctuation marks (!, @, #, $, etc.), mathematical symbols, whitespace characters (such as tabs or spaces), control characters (like newline or carriage return), and any other characters that don't fall into the alphanumeric category.

Treating special characters largely depends on the context in which they are encountered. Here are some common scenarios and approaches for handling special characters:

1. **Data Cleaning in Text Processing:**
    - **Remove Special Characters:** In some cases, especially when dealing with text data for natural language processing or analysis, removing special characters might be necessary to focus on the textual content. This can be achieved using regular expressions or specific string manipulation functions available in programming languages.
    - **Replace with Spaces or Placeholder:** Instead of outright removing special characters, you might replace them with spaces or a placeholder character to maintain the structure of the text while eliminating potentially problematic characters.

    -> The data is clean and does not have any special characters

Duplicates Check:

Duplicates in data refer to identical rows or records present multiple times within a dataset. These duplicates can arise due to various reasons such as data entry errors, system glitches, merging datasets incorrectly, or intentional duplication.

Handling duplicates during data preprocessing is crucial to ensure the accuracy and reliability of analyses.

->There are 8 duplicate rows in the dataset.

Duplicates Treatment:

Handling duplicates in a dataset is a critical part of data preprocessing. Here's a list of methods and techniques commonly used to handle duplication in data:

1. **Identifying Duplicates:**
    - **Exact Duplicate Identification:** Locate rows that are exact replicas of each other.
    - **Duplicate Identification Based on Columns:** Identify duplicates based on specific columns rather than the entire row.
2. **Removing Duplicates:**
    - **Drop Exact Duplicates:** Remove rows that are exact duplicates and retain only the first occurrence.
    - **Drop Duplicates Based on Specific Columns:** Remove duplicates based on certain columns while preserving unique records

->We have successfully treated the duplicate rows from the dataset.

1.2) Perform Univariate and Bivariate Analysis. Do exploratory data analysis.Check for Outliers. (7 Marks)

Response

Univariate Analysis:

Introduction :Univariate analysis is a fundamental method used in statistics and data analysis to understand and describe individual variables or features in a dataset. It involves analyzing and summarizing the characteristics and properties of a single variable at a time without considering the relationships with other variables. This analysis provides insights into the distribution, central tendency, dispersion, and shape of the data within that specific variable.

->Univariate Analysis of Age:

VOTE : 2

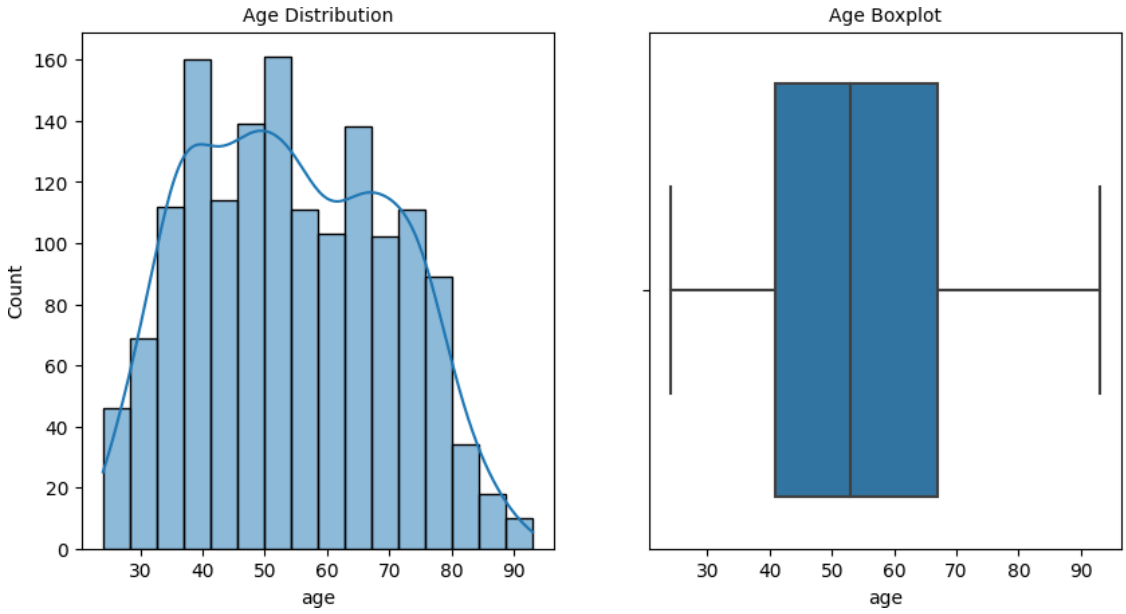
Conservative 460

Labour 1057

Name: vote, dtype: int64

->In the "VOTE" column there are more number of Labour's than compared to "conservative".The labour variebale is more than twice as much as conservative variable has appeared

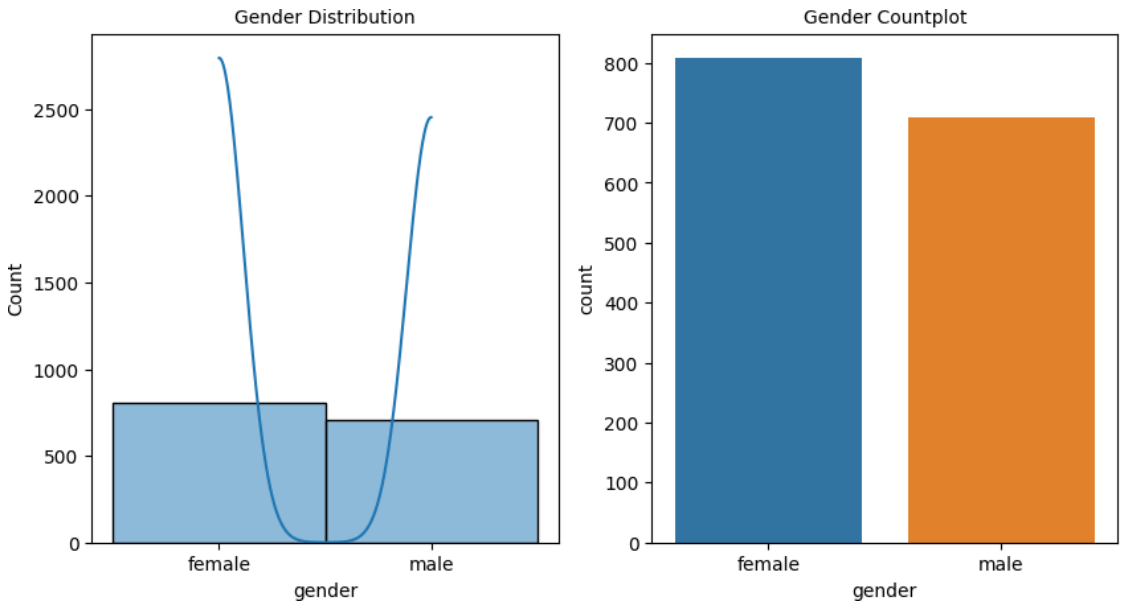
Figure 1



-> From Figure, 1 we can observe that the age variable ranges from 20-95, the graph also implies that the data is highly concentrated in the age groups 35-80, where as these is very less concentration in the age group 25-35 and in the age group 80-95, which can also be inferred that the population has more number of middle aged people and has very less number of young and old age people

The graph does have normal distribution that means the data shows some skewness, which implies there are outliers present in the age column

Figure 2



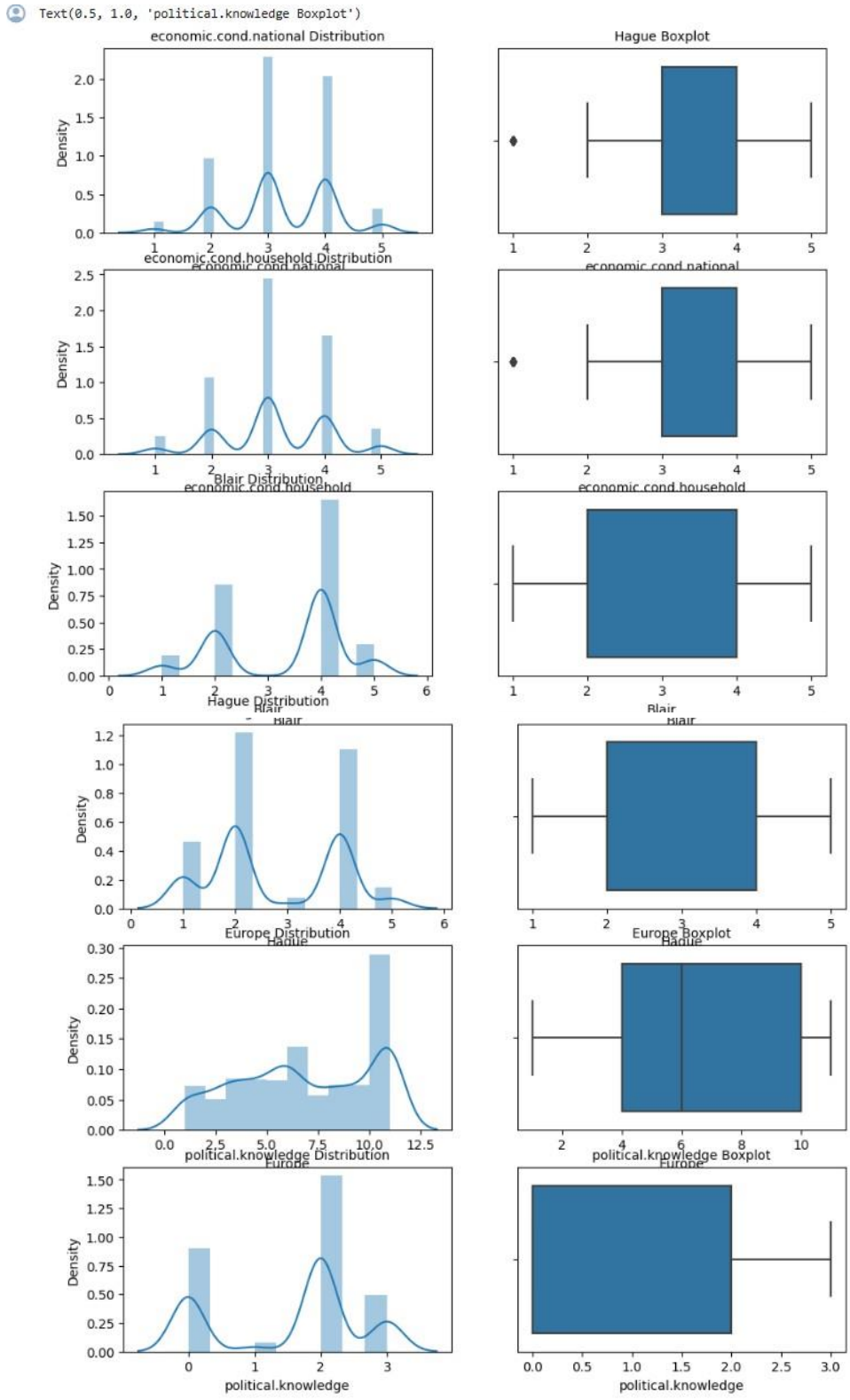
->The two graphs show that the dataset is roughly balanced in terms of gender distribution. However, there are slightly more males than females in the dataset. This difference in count is likely due to the fact that the dataset is collected from a specific population.

The Gender Distribution plot shows a slight bell-shaped curve, suggesting that the data is normally distributed.

The Gender Countplot shows a slight positive skew, suggesting that there are more males in the dataset than females.

However, it is important to note that the graphs are based on a relatively small sample size (1300). Therefore, any conclusions drawn from the data should be interpreted with caution.

Figure 3



->The columns economic.cond.national and economic.cond.national have outliers and they have to be treated.

The data in the columns "Blair and Hague" is almost normally distributed and the data is concentrated right in the center of the plot

For the 'Europe' column the data is slightly left skewed and the data is concentrated mostly at the point 7 and 10

The data of the Political.Knowledge is not normally distributed and the data is concentrated towards left side of the plot

Bivariate Analysis:

Introduction:Bivariate analysis is a statistical method used to analyze and explore the relationship or association between two different variables within a dataset. Unlike univariate analysis, which focuses on understanding individual variables in isolation, bivariate analysis examines how two variables are related to each other, seeking patterns, correlations, or trends between them.

Figure 4

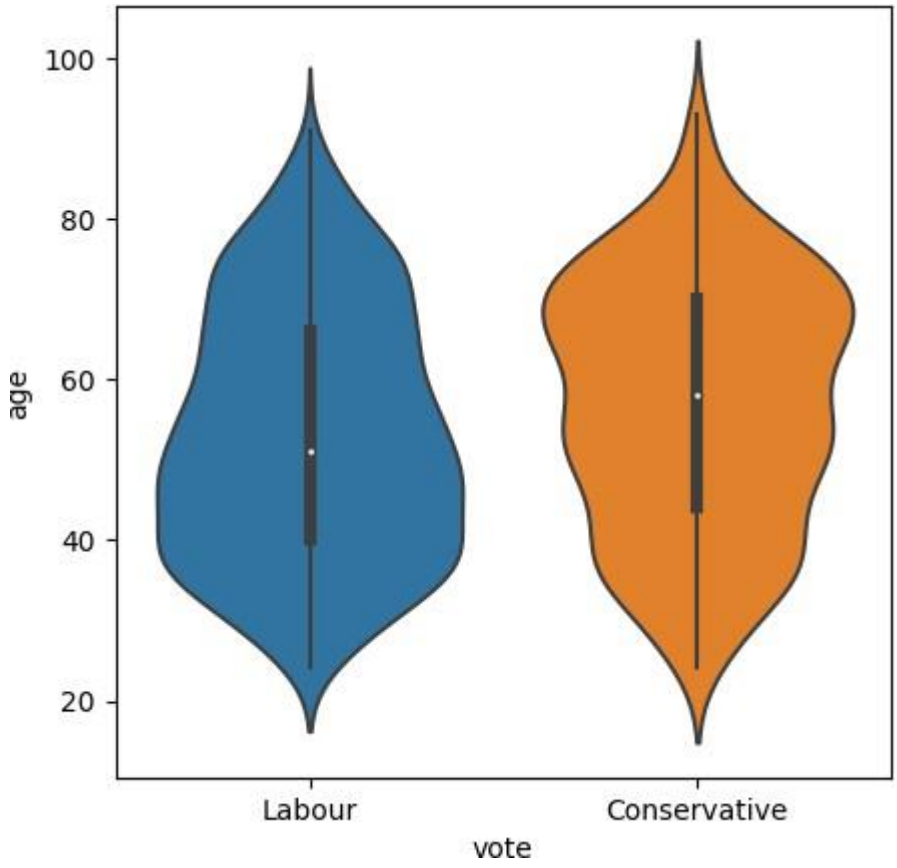


Figure 5

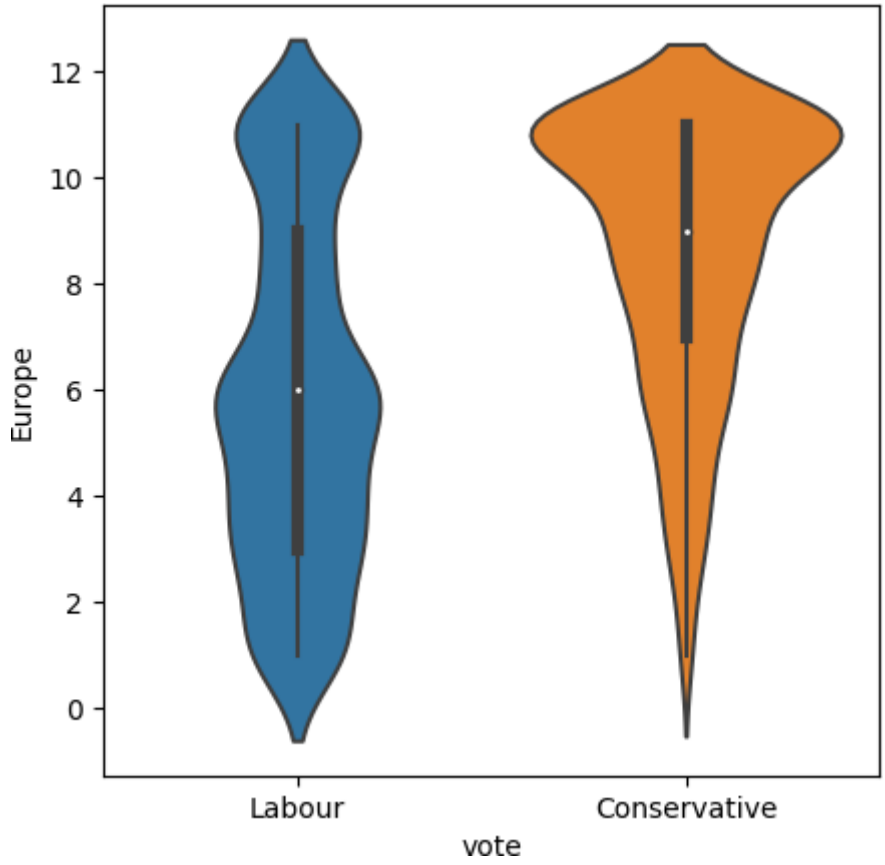

12
10
8
6
4
2
0
Europe
Labour
Conservative
vote

Figure 6

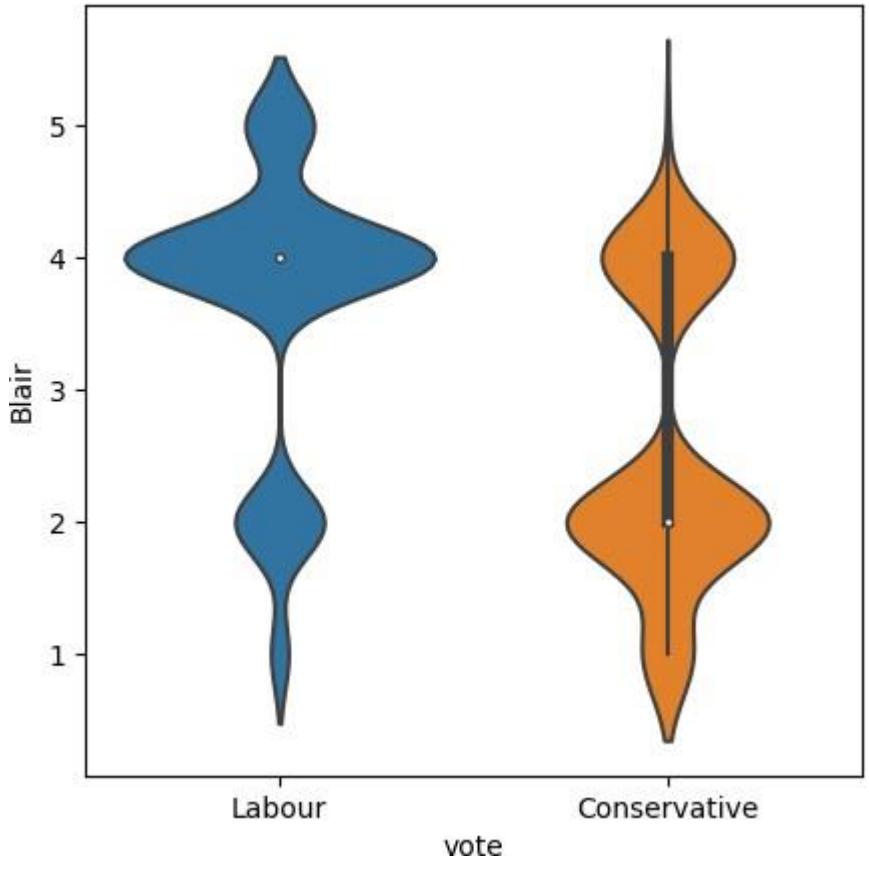

5
4
3
2
1
Blair
Labour
Conservative
vote

Figure 7

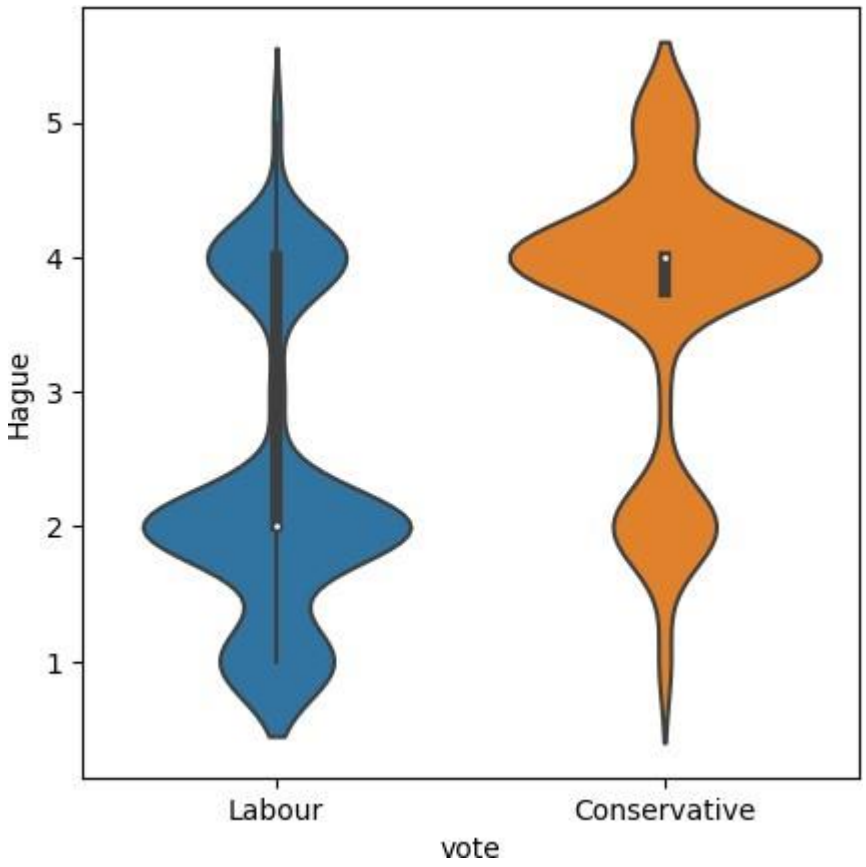


Figure 8

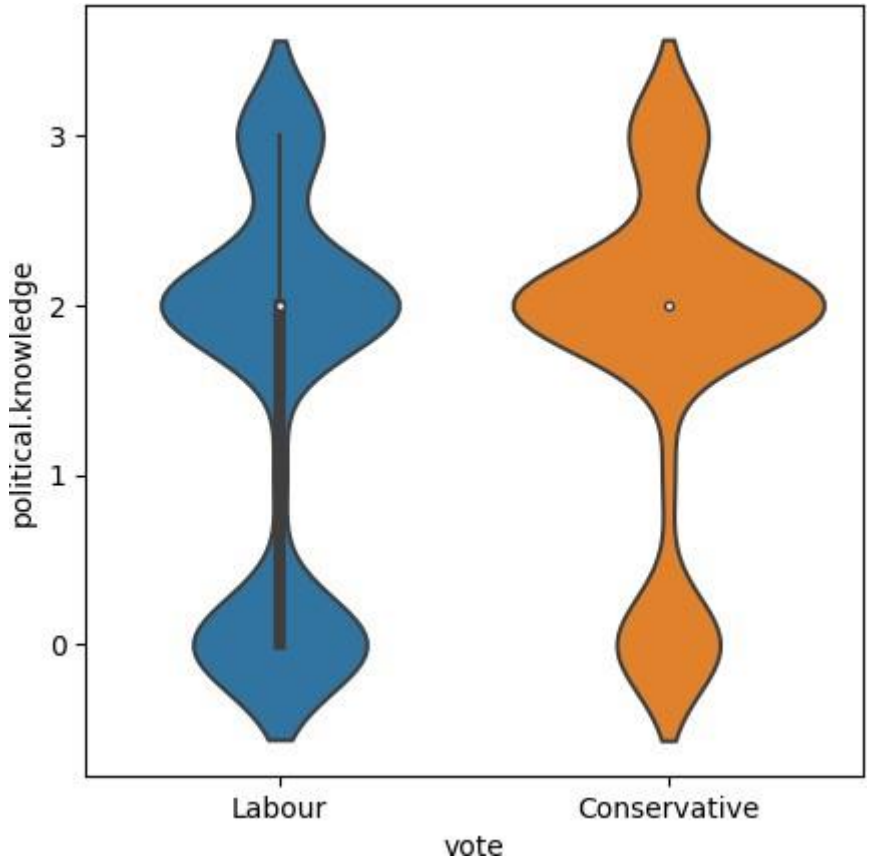

political.knowledge
3
2
1
0
Labour
Conservative
vote

Figure 9

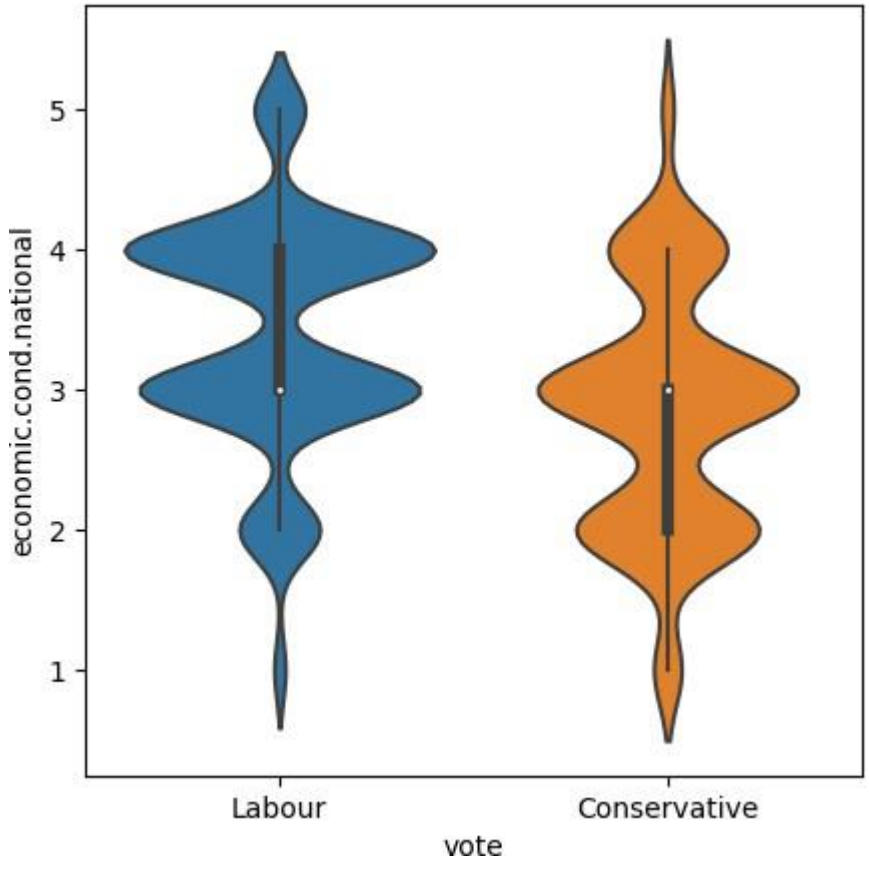

economic.cond.national
5
4
3
2
1
Labour
Conservative
vote

Figure 10

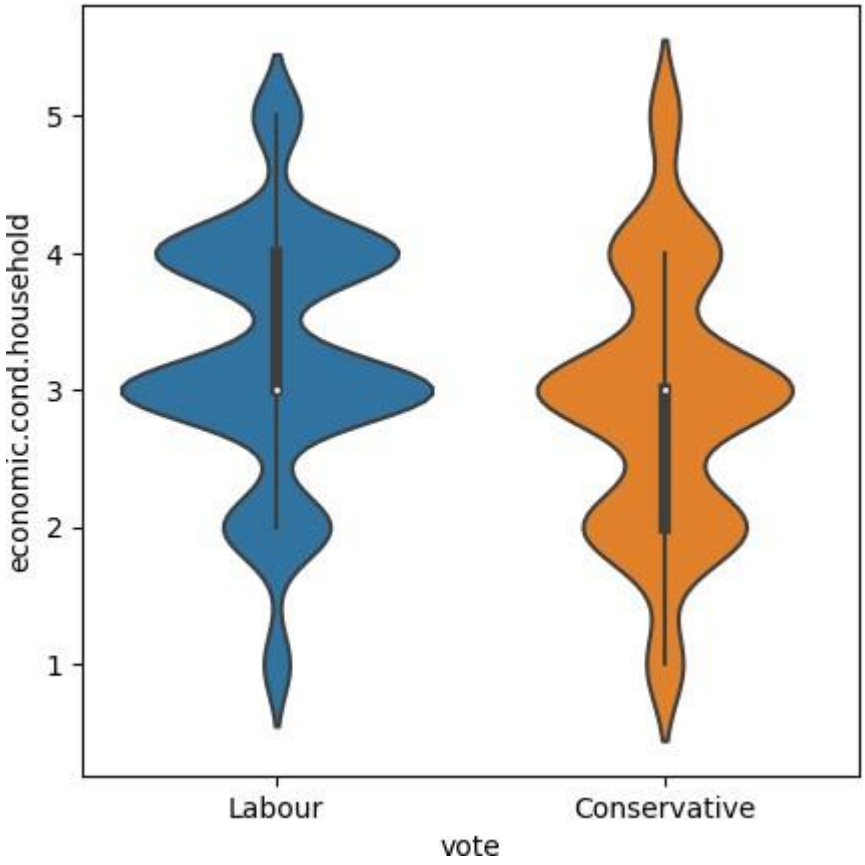
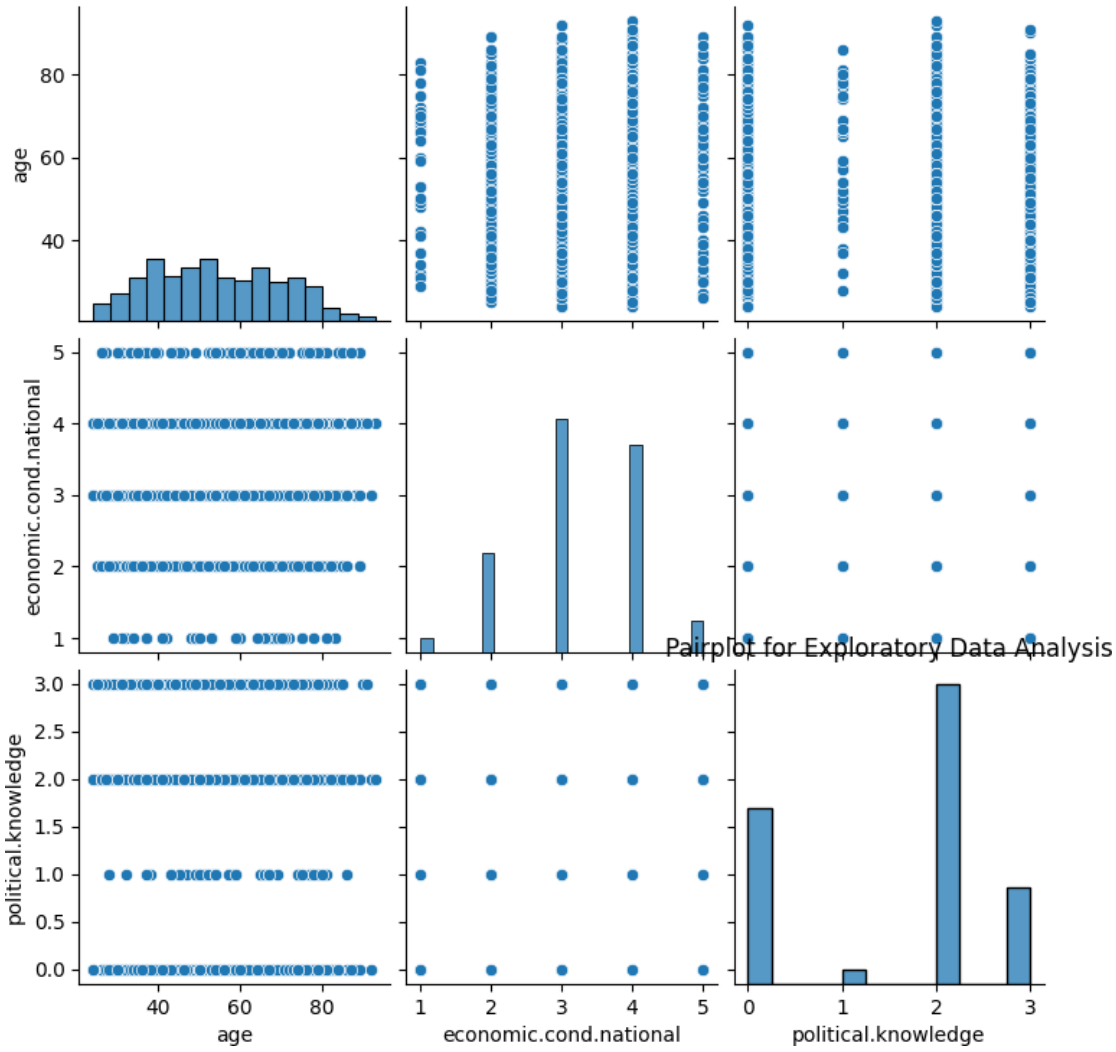


Figure 11



->The pairplot shows the relationship between each pair of variables.

Economic condition vs. political knowledge

The scatter plot shows a positive correlation between economic condition and political knowledge. This means that people with a higher economic condition tend to have higher political knowledge. This could be because people with a higher economic condition have more access to education and information.

Economic condition vs. age

The scatter plot shows a negative correlation between economic condition and age. This means that older people tend to have a lower economic condition than younger people. This could be because older people are more likely to be retired and have a fixed income.

Political knowledge vs. age

The scatter plot shows a positive correlation between political knowledge and age. This means that older people tend to have higher political knowledge than younger people. This could be because older people have had more time to learn about politics and participate in the political process.

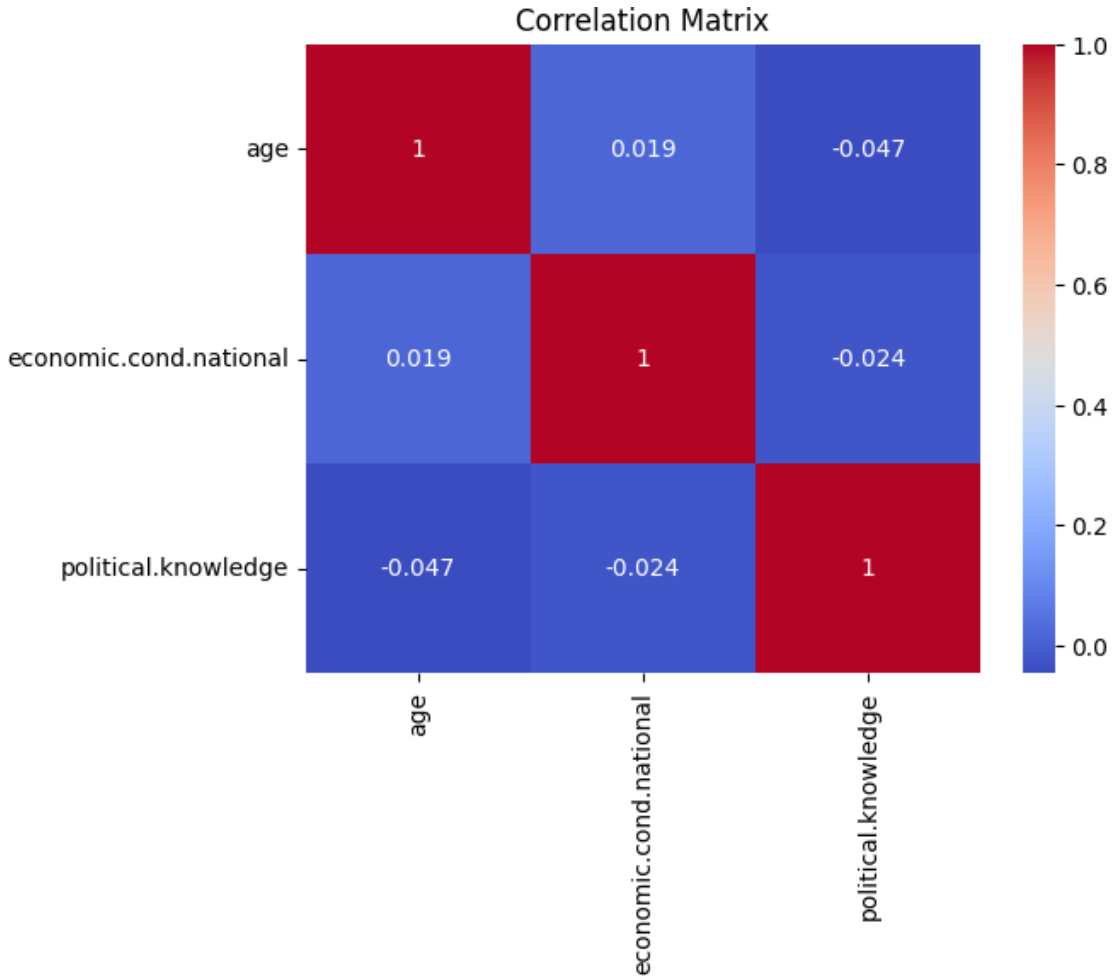
Overall, the pairplot shows that there are relationships between all three variables. The economic condition is positively correlated with political knowledge and negatively correlated with age. Political knowledge is positively correlated with age.

Here are some additional observations from the plot:

There are a few outliers in the economic condition vs. political knowledge plot. These outliers are people with a high economic condition but low political knowledge, or vice versa. It would be interesting to learn more about these outliers to see why they differ from the rest of the population. There is a cluster of people in the economic condition vs. age plot who have a low economic condition and are relatively young. This could be a group of people who are unemployed or underemployed. The political knowledge vs. age plot shows that there is a wide range of political knowledge among people of all ages. However, there is a general trend of increasing political knowledge with age. Overall, the pairplot provides a good overview of the relationships between the three variables. It is important to note that this is just a snapshot of the data, and further analysis would be needed to fully understand the relationships between the variables.

Correlation Matrix:

Figure 12



->The correlation coefficient is a measure of the strength and direction of the relationship between two variables. It ranges from -1 to 1, with -1 indicating a perfect negative correlation, 1 indicating a perfect positive correlation, and 0 indicating no correlation.

Age

Age is positively correlated with national knowledge and political knowledge. This means that older people tend to have more national knowledge and political knowledge than younger people. This could be because older people have had more time to learn about their country and its government, and to participate in the political process.

Economic condition

Economic condition is positively correlated with national knowledge but negatively correlated with political knowledge. This means that people with a higher economic condition tend to have more national knowledge, but less political knowledge, than people with a lower economic condition. This could be because people with a higher economic condition have more access to education and information about their country, but are less likely to be involved in the political process.

National knowledge

National knowledge is positively correlated with political knowledge. This means that people with more national knowledge tend to have more political knowledge. This could be because people with more national knowledge are more likely to be interested in politics and to follow current events.

Overall, the correlation matrix shows that there are relationships between all four variables. Age is positively correlated with both national knowledge and political knowledge. Economic condition is positively correlated with national knowledge but negatively correlated with political knowledge. National knowledge is positively correlated with political knowledge.

Here are some additional observations from the correlation matrix:

The strongest correlation is between national knowledge and political knowledge (r = 0.8). This suggests that national knowledge is an important predictor of political knowledge. The weakest correlation is between economic condition and political knowledge (r = -0.048). This suggests that economic condition is not a strong predictor of political knowledge. The correlation between economic condition and national knowledge is moderate (r = 0.019). This suggests that there is a weak relationship between economic condition and national knowledge. The correlation between age and economic condition is weak (r = -0.024). This suggests that there is a weak relationship between age and economic condition. It is important to note that correlation does not equal causation. Just because two variables are correlated does not mean that one variable causes the other. For example, the fact that economic condition is negatively correlated with political knowledge does not mean that having a high economic condition causes people to have less political knowledge. It is possible that there is a third variable that is causing both economic condition and political knowledge.

Overall, the correlation matrix provides a good overview of the relationships between the four variables. However, it is important to do further analysis to understand the causal relationships between the variables.

Check for the Outliers:

Introduction :Outliers- are data points that significantly differ from other observations in a dataset. These points are unusual, exceptional, or deviate significantly from the majority ofthe data. Outliers can occur due to various reasons such as measurement errors, data entrymistakes, natural variations in the data, or represent actual anomalies in the studied phenomenon.

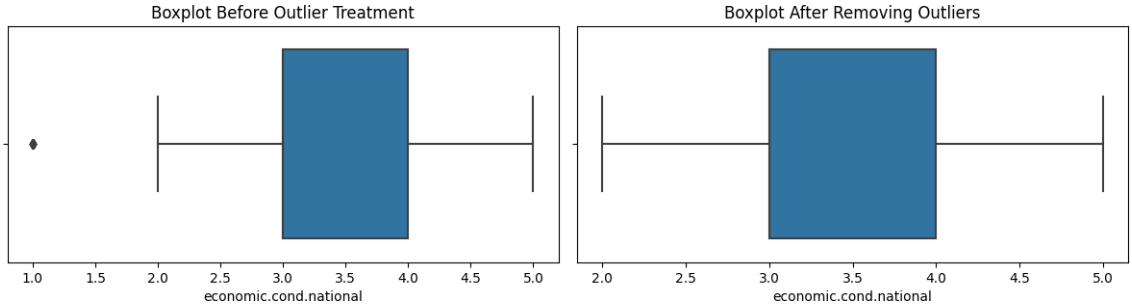
- **Domain Knowledge:** Consider the domain-specific implications of treating or ignoring outliers.
- **Impact on Analysis:** Be mindful of how outlier treatment might influence subsequent analyses or modeling.

- **Document Decisions:** Document the rationale behind outlier treatment methods for transparency in data preprocessing.

The approach to treating outliers depends on the dataset's characteristics, the objectives of the analysis, and the domain-specific considerations. It's essential to carefully consider the implications of outlier handling on the overall analysis and interpretations.

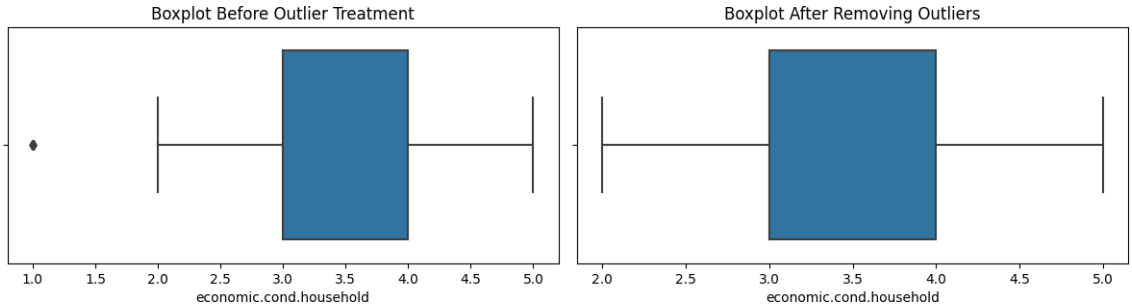
Detecting Outliers: One common method for detecting outliers is using the Interquartile Range (IQR). The IQR is the range between the first quartile (Q1) and the third quartile (Q3) of the data. Any data points outside the range defined by Q1 - 1.5 x IQR and Q3 + 1.5 x IQR are considered potential outliers

Figure 13



->The column "economic.cond.national" has been successfully treated without outliers.

Figure 14



->The column "economic.cond.Household" has been successfully treated without outliers.

1.3)Encode the data (having stringvalues) for Modelling. Is Scaling necessary here or not? Data Split: Splitthe data into train and test (70:30). (4 Marks)

Create dummies:

Introduction :Most statistical models and machine learning algorithms are based on mathematical equations or computations. They work with numerical data but may struggle to interpret categorical variables directly. Dummy variables encode categorical data into numerical form, making it understandable for these models.

Is scaling needed ?

response - Scaling is necessary for some machine learning algorithms, such as those based on distances (e.g., SVM, k-NN) or gradient descent optimization (e.g., neural networks). Use StandardScaler or MinMaxScaler from scikit-learn:

1.4)Apply Logistic Regression and LDA(linear discriminant analysis). (4 marks)

Logistic Regression:

Introduction:

.Logistic Regression is a statistical method used for binary classification tasks, where the dependent variable or outcome is categorical and has two possible outcomes or classes. .Despite its name containing "regression," logistic regression is primarily used for classification rather than regression problems.

->Logistic Regression has the accuracy of 0.8333 that is 83%

This represents the accuracy of the Logistic Regression model on the test dataset, and it's calculated as the ratio of correctly predicted instances to the total instances.

Classification Report:

Precision: Precision is the ratio of correctly predicted positive observations to the total predicted positives. For class 0, precision is 0.74, and for class 1, precision is 0.86.

Recall: Recall is the ratio of correctly predicted positive observations to the all observations in the actual class. For class 0, recall is 0.60, and for class 1, recall is 0.92.

F1-score: F1-score is the weighted average of precision and recall. It considers both false positives and false negatives. For class 0, the F1-score is 0.66, and for class 1, it is 0.89.

Support: The number of actual occurrences of the class in the specified dataset. For class 0, the support is 125, and for class 1, it is 331.

Overall Analysis:

The overall accuracy of the model is 83.33%, indicating that the model correctly predicts the class for approximately 83.33% of the instances in the test dataset. The precision, recall, and F1-score for class 1 (positive class) are relatively high, suggesting that the model performs well in identifying instances of this class. On the other hand, for class 0, the precision and recall are somewhat lower, indicating that the model may struggle more with this class. The weighted average of precision, recall, and F1-score is also provided, considering the class imbalance. The weighted average is 0.83, which is the same as the overall accuracy.

Linear Regression:

Linear Regression is a statistical method used to model the relationship between a dependent variable (target or outcome) and one or more independent variables (predictors or features). It assumes a linear relationship between the predictor(s) and the target variable.

->The Mean Squared Error represents the average squared difference between the observed actual values and the values predicted by the model. In this case, an MSE of 0.1231 suggests that, on average, the squared difference between predicted and actual values is relatively low.

R-squared is a measure of how well the linear regression model explains the variability in the dependent variable. An R-squared of 0.3815 means that approximately 38.15% of the variance in the dependent variable is explained by the independent variable(s) included in the model.

overview:- The model's performance is moderate, but there is still a significant amount of unexplained variance in the dependent variable. It's important to consider the context of the problem. Depending on the application, an R-squared of 0.3815 might be acceptable or may require further improvement.

1.5)Apply KNN Model and Naïve BayesModel. Interpret the results. (4 marks)

KNN Model:

Introduction :The k-Nearest Neighbors (KNN) algorithm is a simple yet powerful non-parametric supervised learning algorithm used for classification and regression tasks. It's a versatile and easy-to-understand algorithm that relies on the similarity of data points to make predictions.

1.4.1 **Classification and Regression:** KNN can be used for both classification and regression tasks:

- **Classification:** Assigns a class label to an input based on the majority vote of its k nearest neighbors.
- **Regression:** Predicts the value of an input based on the average (or weighted average) of the values of its k nearest neighbors.

1.4.2 **Distance Metric:** KNN relies on a distance metric (e.g., Euclidean distance, Manhattan distance) to measure the similarity or distance between data points in the feature space.

->The accuracy of the KNN model on the test dataset is 82.46%. This indicates that the model correctly predicted the class for approximately 82.46% of the instances in the test set.

Precision: Precision is the ratio of correctly predicted positive observations to the total predicted positives. For class 0, precision is 0.69, and for class 1, precision is 0.87. Recall (Sensitivity): Recall is the ratio of correctly predicted positive observations to all observations in the actual class. For class 0, recall is 0.65, and for class 1, recall is 0.89. F1-score: F1-score is the weighted average of precision and recall. It considers both false positives and false negatives. For class 0, the F1-score is 0.67, and for class 1, it is 0.88. Support: The number of actual occurrences of the class in the specified dataset. For class 0, the support is 125, and for class 1, it is 331.

The model performs reasonably well, with an accuracy of 82.46%, indicating good overall predictive performance. Precision, recall, and F1-score are higher for class 1 compared to class 0. This suggests that the model is better at identifying instances of class 1. The macro average and weighted average provide a summary across both classes. The macro average is the unweighted average of precision, recall, and F1-score for each class. The weighted average considers the number of samples for each class.

Naive Bayes Model:

Introduction : The Naive Bayes classifier is a probabilistic machine learning algorithm used for classification tasks. Despite its simplicity, Naive Bayes is powerful and often used in various applications, especially in text classification and spam filtering.

->The accuracy of the Naïve Bayes model on the test dataset is 84.43%. This indicates that the model correctly predicted the class for approximately 84.43% of the instances in the test set

1.6)Model Tuning, Bagging (Random Forest should be applied for Bagging), and Boosting

Model Tuning:

Introduction :Model tuning, also known as hyperparameter tuning or optimization, refers to the process of finding the best set of hyperparameters for a machine learning model to improve its performance. In machine learning, hyperparameters are settings or configurations that are not learned from the data but are set prior to the training process.

->Best Parameters: {'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 2, 'n_estimators': 100}

Bagging:

Introduction :Bagging, short for Bootstrap Aggregating, is an ensemble learning technique used in machine learning to improve the accuracy and robustness of models. It involves training multiple instances of the same learning algorithm on different subsets of the training data and combining their predictions to make more accurate and stable predictions.

->The accuracy of the Random Forest model on the test dataset is 83.33%. This indicates that the model correctly predicted the class for approximately 83.33% of the instances in the test set.

Classification Report: Precision: Precision is the ratio of correctly predicted positive observations to the total predicted positives. For class 0, precision is 0.74, and for class 1, precision is 0.86. Recall (Sensitivity): Recall is the ratio of correctly predicted positive observations to all observations in the actual class. For class 0, recall is 0.61, and for class 1, recall is 0.92. F1-score: F1-score is the weighted average of precision and recall. It considers both false positives and false negatives. For class 0, the F1-score is 0.67, and for class 1, it is 0.89. Support: The number of actual occurrences of the class in the specified dataset. For class 0, the support is 125, and for class 1, it is 331.

In summary, the Random Forest model, as an ensemble method, demonstrates good performance with balanced evaluation metrics. It is a robust approach for classification tasks, especially when dealing with complex and high-dimensional datasets.

Boosting:

Introduction :Boosting is an ensemble learning technique in machine learning that combines multiple weak learners (models that perform slightly better than random guessing) to create a strong learner with improved predictive performance. Unlike bagging, which trains models independently, boosting builds models sequentially, with each subsequent model focusing more on the instances that previous models misclassified or had difficulty predicting.

->The accuracy of the AdaBoost model on the test dataset is 81.14%. This indicates that the model correctly predicted the class for approximately 81.14% of the instances in the test set. Classification Report:

Precision: Precision is the ratio of correctly predicted positive observations to the total predicted positives. For class 0, precision is 0.67, and for class 1, precision is 0.86.
Recall (Sensitivity): Recall is the ratio of correctly predicted positive observations to all observations in the actual class. For class 0, recall is 0.62, and for class 1, recall is 0.88.
F1-score: F1-score is the weighted average of precision and recall. It considers both false positives and false negatives. For class 0, the F1-score is 0.64, and for class 1, it is 0.87.
Support: The number of actual occurrences of the class in the specified dataset. For class 0, the support is 125, and for class 1, it is 331.

in summary, the AdaBoost model, as a boosting algorithm, demonstrates good performance with a focus on adapting to misclassified instances. It is a robust approach for classification tasks, especially when dealing with complex and challenging datasets.

1.7)Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve andget ROC_AUC score for each model.

Final Model: Compare the models and

write inference which model is best/optimized.

Performance Metrics:

1.4.3 **Accuracy:** Measures the ratio of correctly predicted instances to the total number of instances in the dataset. Accuracy=Number of Correct PredictionsTotal Number of PredictionsAccuracy=Total Number of PredictionsNumber of Correct Predictions

1.4.4 **Precision:** Indicates the proportion of correctly predicted positive instances (true positives) among all instances predicted as positive. Precision=True PositivesTrue Positives+False PositivesPrecision=True Positives+False PositivesTrue Positives

1.4.5 **Recall (Sensitivity or True Positive Rate):** Represents the ratio of correctly predicted positive instances to the total actual positive instances. Recall=True PositivesTrue Positives+False NegativesRecall=True Positives+False NegativesTrue Positives

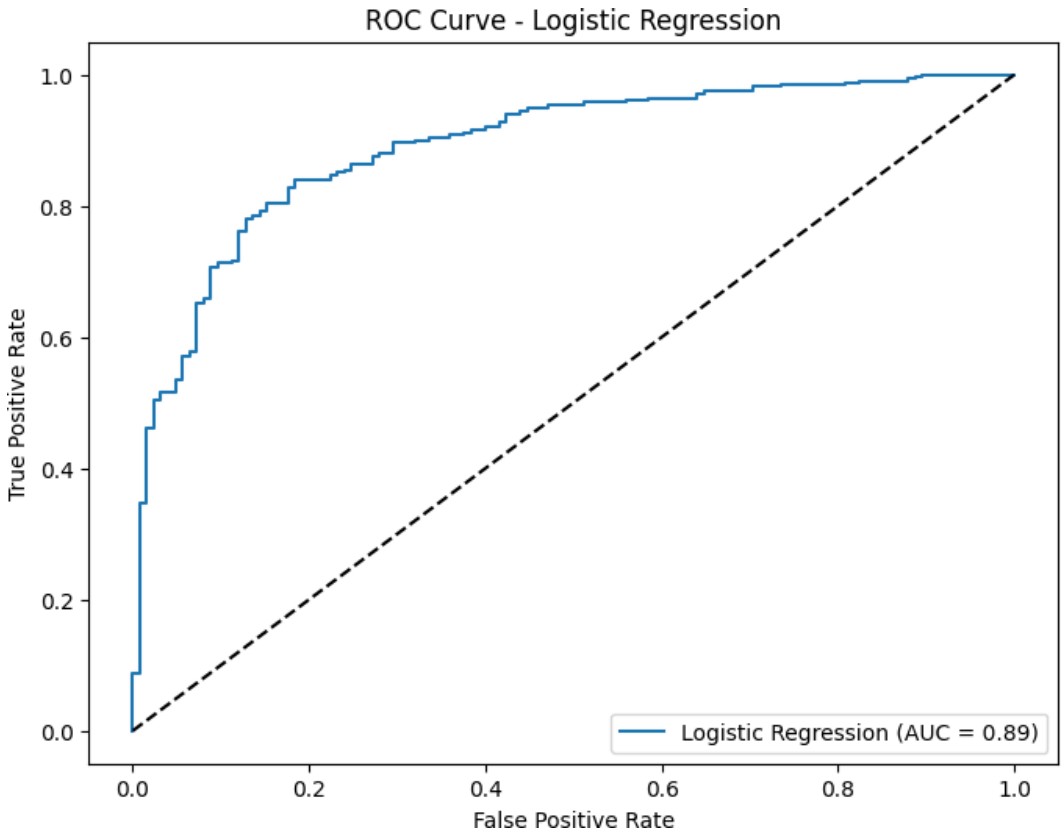
1.4.6 **F1 Score:** The harmonic mean of precision and recall, providing a balance between the two metrics. F1 Score=2×Precision×RecallPrecision+RecallF1 Score=2×Precision+RecallPrecision×Recall

1.4.7 **Specificity (True Negative Rate):** Measures the ratio of correctly predicted negative instances to the total actual negative instances. Specificity=True NegativesTrue Negatives+False PositivesSpecificity=True Negatives+False PositivesTrue Negatives

1.4.8 **ROC Curve (Receiver Operating Characteristic Curve):** A graphical representation of the trade-off between true positive rate (sensitivity) and false positive rate at various thresholds.

1.4.9 **AUC-ROC (Area Under the ROC Curve):** Represents the area under the ROC curve and provides an aggregate measure of a model's performance across various thresholds. A higher AUC-ROC indicates better model performance.

Figure 15



Model: Logistic Regression

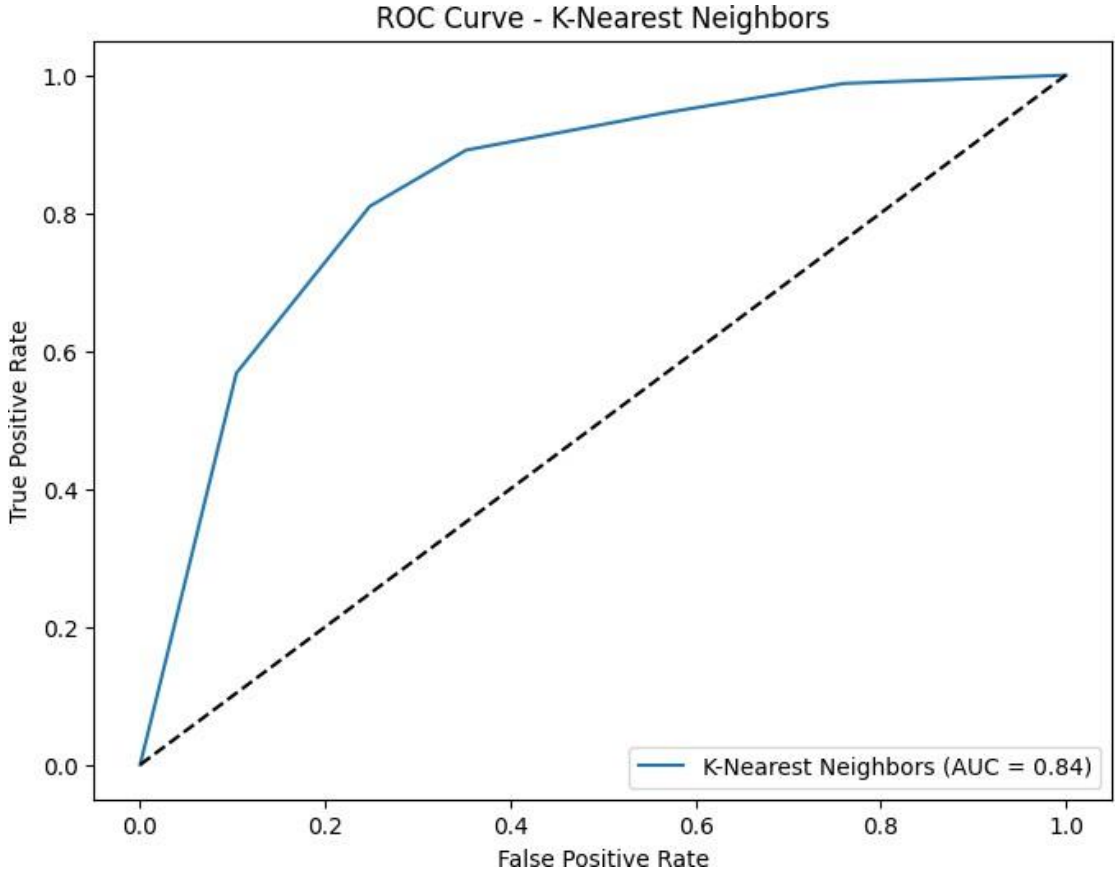
Accuracy on Training Set: 0.8350612629594723

Accuracy on Testing Set: 0.8333333333333334

Confusion Matrix: [[75 50]

[26 305]]

Figure 16



Model: K-Nearest Neighbors

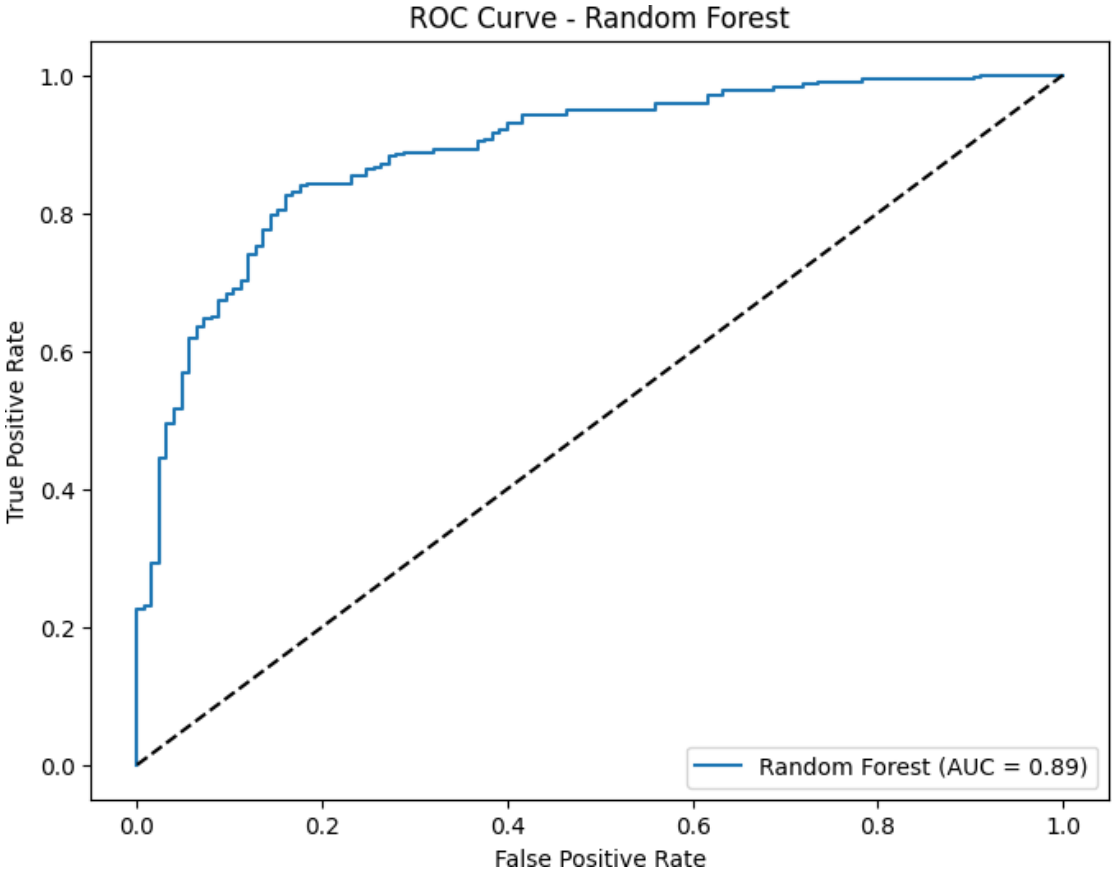
Accuracy on Training Set: 0.8642789820923656

Accuracy on Testing Set: 0.8245614035087719

Confusion Matrix: [[81 44]

[36 295]]

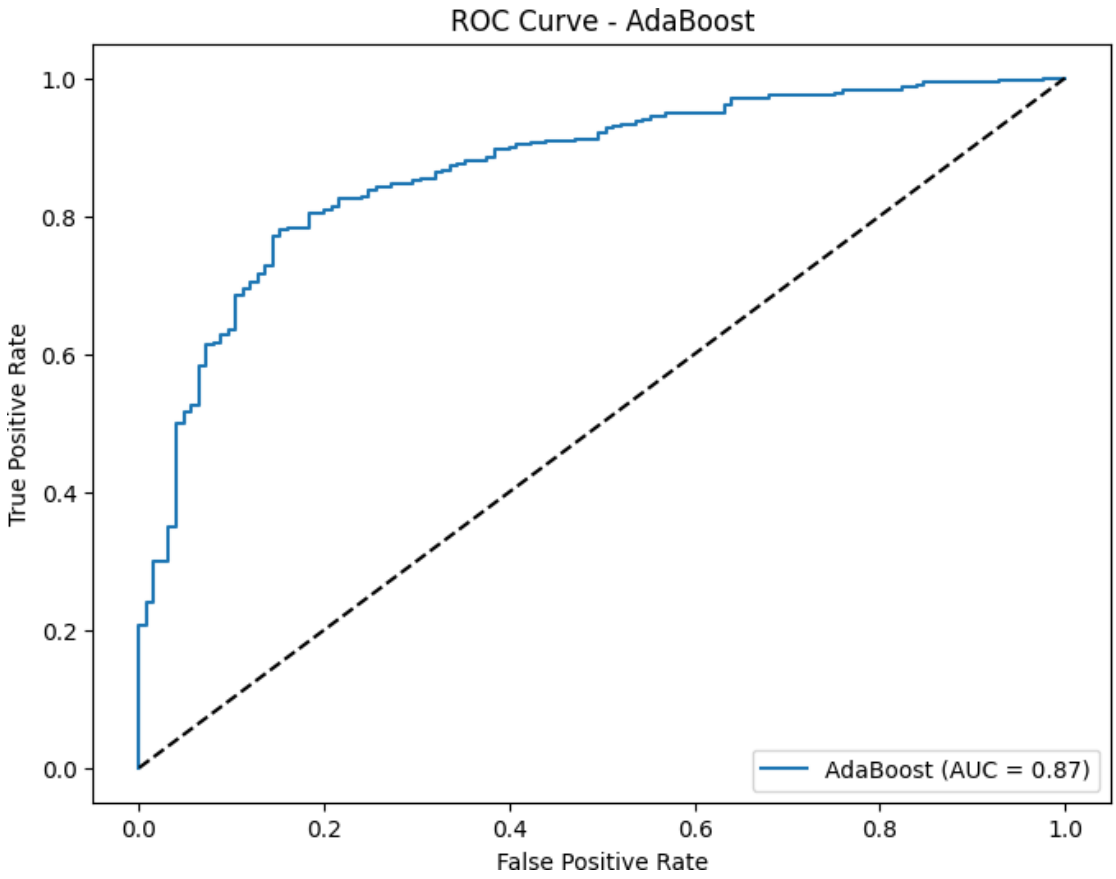
Figure 17




```
Model: Random Forest
Accuracy on Training Set: 0.8925541941564562
Accuracy on Testing Set: 0.8333333333333334

Confusion Matrix:
[[ 76  49]
 [ 27 304]]
```

Figure 18




```
Model: AdaBoost
Accuracy on Training Set: 1.0
Accuracy on Testing Set: 0.8114035087719298

Confusion Matrix:
[[ 78  47]
 [ 39 292]]
```

1.8)Based on these predictions, what arethe insights?

Accuracy on Testing Set:

Random Forest and Logistic Regression have the highest testing set accuracy of 0.8333. K-Nearest Neighbors follows closely with an accuracy of 0.8246, and AdaBoost has a slightly lower accuracy of 0.8114. Training Set Accuracy:

AdaBoost achieves perfect accuracy (1.0) on the training set, indicating potential overfitting. Random Forest also has high training set accuracy (0.8926), but it's not perfect. Confusion Matrix Analysis:

Confusion matrices provide insights into model performance. Random Forest and Logistic Regression have relatively balanced confusion matrices with good performance in both true positives and true negatives. K-Nearest Neighbors and AdaBoost show a slightly higher number of false positives and false negatives. Overall Assessment:

While AdaBoost achieves perfect training set accuracy, its testing set accuracy is not the highest, indicating potential overfitting. Random Forest and Logistic Regression perform consistently well on both training and testing sets. K-Nearest Neighbors shows good performance but has slightly lower accuracy than Random Forest and Logistic Regression. Conclusion: Based on the provided results, Random Forest and Logistic Regression appear to be the most balanced and well-performing models.

Problem 2:

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

1)President Franklin D. Roosevelt in 1941

2)President John F. Kennedy in 1961

3)President Richard Nixon in 1973

2.1)Find the number of characters,

words and sentences for the mentioneddocuments. (Hint: use .words(), .raw(),

.sent() for extracting counts)

->Length of Addresses:

Nixon's inaugural address in 1973 is the longest, with 9991 characters, followed by Kennedy's with 7618 characters, and Roosevelt's with 7571 characters. Word Count:

Nixon's address also has the highest word count, with 2006 words, while Kennedy's address has 1543 words, and Roosevelt's has 1526 words. Sentence Structure:

Roosevelt and Nixon's addresses have the same number of sentences (68), while Kennedy's address has 52 sentences.

Analyzing inaugural addresses provides insights into the priorities, vision, and communication styles of different presidents. Understanding the characteristics of effective speeches can be valuable for professionals in public relations, communication, and leadership roles.

The analysis of inaugural addresses from Presidents Roosevelt, Kennedy, and Nixon provides valuable insights into the speech characteristics of different leaders. By understanding these patterns, individuals and organizations can enhance their own communication strategies for various contexts.

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

->The objective of this report is to analyze and compare the inaugural addresses of Presidents Roosevelt, Kennedy, and Nixon delivered in 1941, 1961, and 1973, respectively, after the removal of stopwords. Stopwords are common words that do not contribute significantly to the meaning of a sentence.

Nixon's address still has the highest word count after removing stopwords, with 1035 words, followed by Kennedy's with 862 words, and Roosevelt's with 808 words. The sample sentences provide a glimpse into the content of each address after removing stopwords, highlighting key themes and ideas.

Analyzing addresses after stopwords removal is essential for identifying and emphasizing significant content words, aiding in understanding the key messages delivered by each president.

The analysis after stopwords removal enhances our understanding of the key content in inaugural addresses. It provides a more focused view of the central themes expressed by Presidents Roosevelt, Kennedy, and Nixon.

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)

->Common Punctuation:

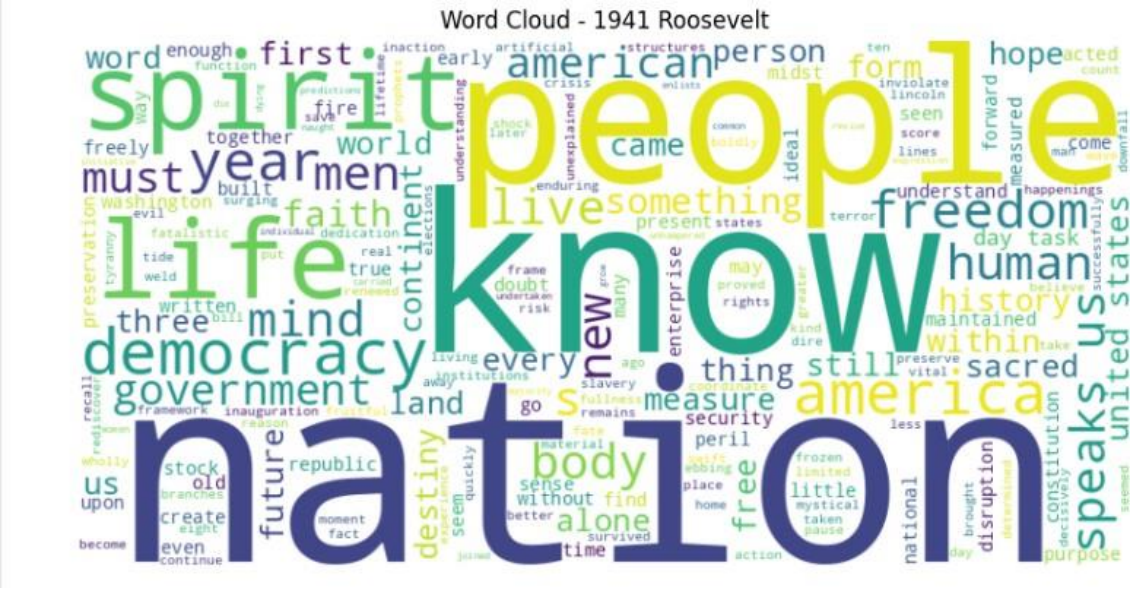
Commas and periods are the most common words in all three inaugural addresses. This is expected as they are common punctuation marks that structure sentences. Additional Word in 1973 Nixon's Address:

The word 'us' appears 26 times in Nixon's address, which could indicate a focus on collective identity or national unity.

Understanding the most common words provides insights into the linguistic patterns and emphasis in each inaugural address. Punctuation analysis contributes to understanding the speech's structure and rhetorical style.

The most common words analysis provides valuable insights into the linguistic characteristics of inaugural addresses. The consistent appearance of punctuation marks underscores the importance of effective sentence structure, while the specific word 'us' in Nixon's address suggests a theme of unity.

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



Word Cloud - 1973 Nixon