MACHINE LEARNING PROJECT REPORT

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

This report presents an analysis of an **election dataset** using machine learning techniques. The primary focus is on **clustering and classification models** to derive insights and enhance predictive capabilities.

---

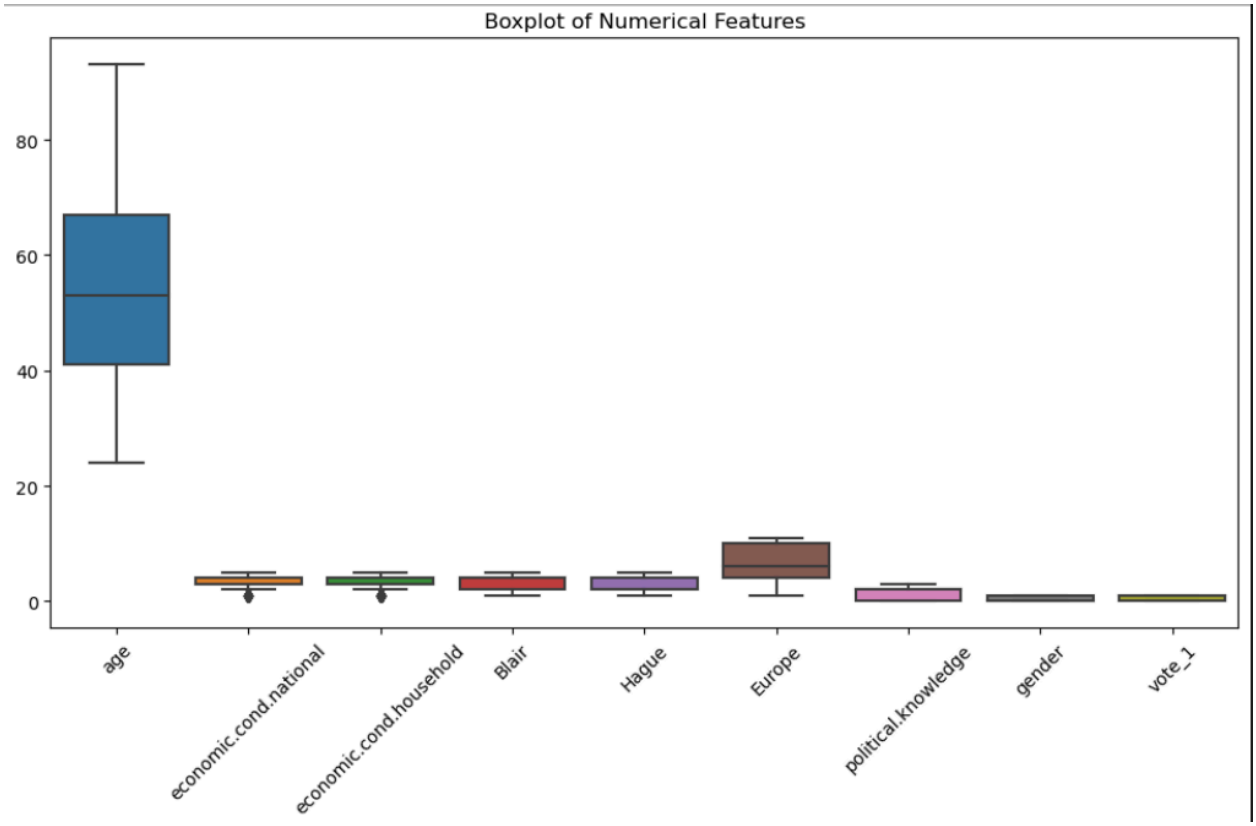
2. EXPLORATORY DATA ANALYSIS (EDA)

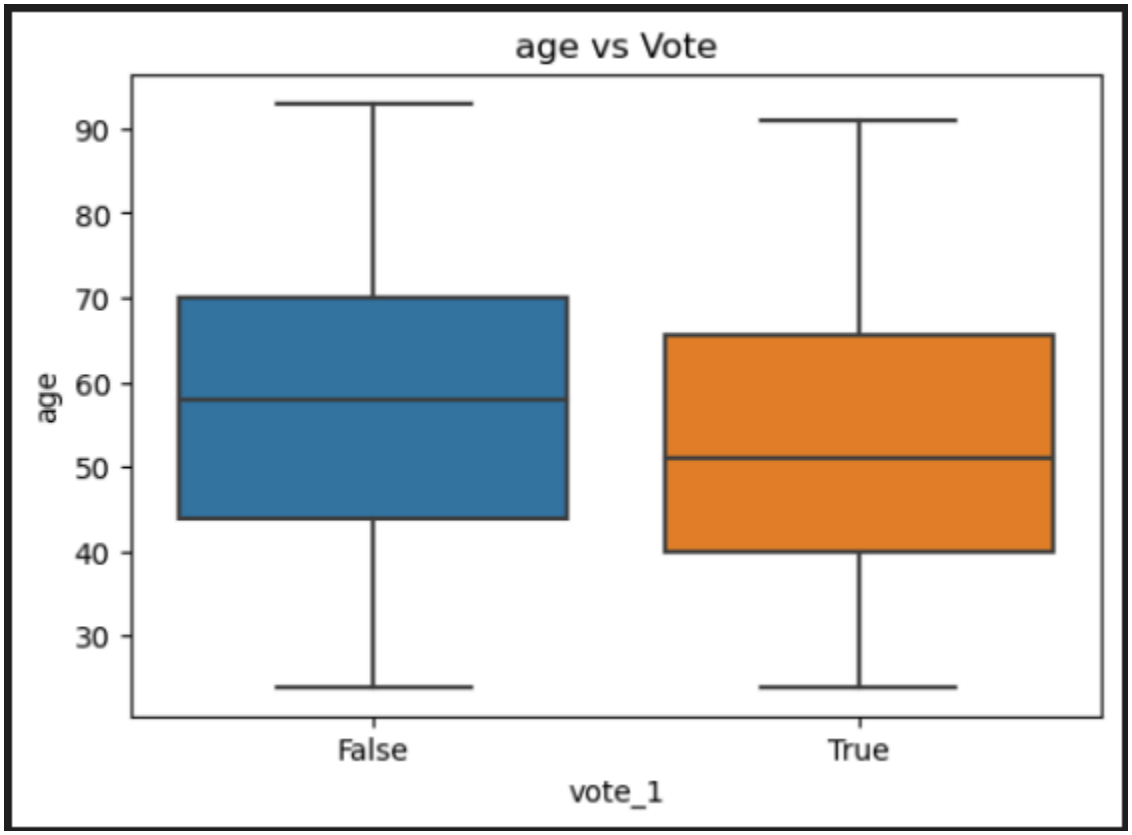
2.1 Summary of the Data

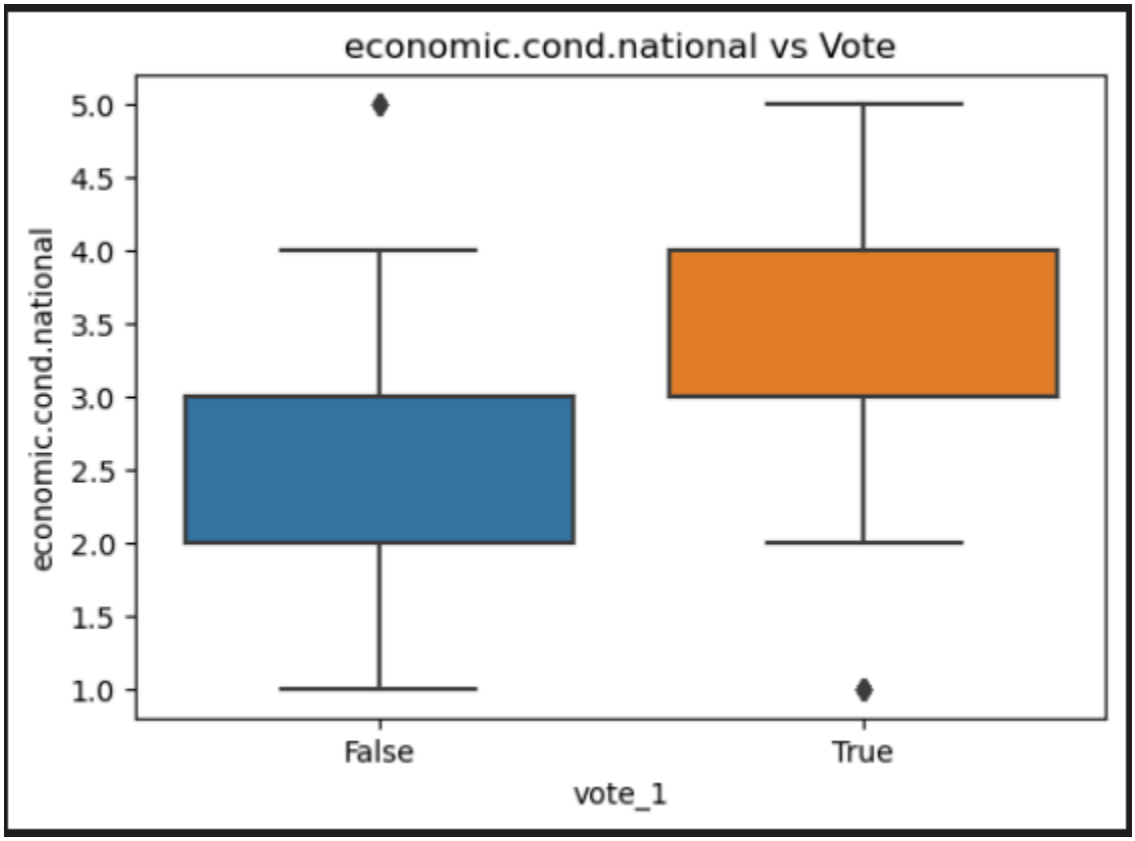
- The dataset consists of multiple variables, including **numerical and categorical features**.
- No missing values were detected.
- The data structure was analyzed using `election.info()` and `election.head()`.

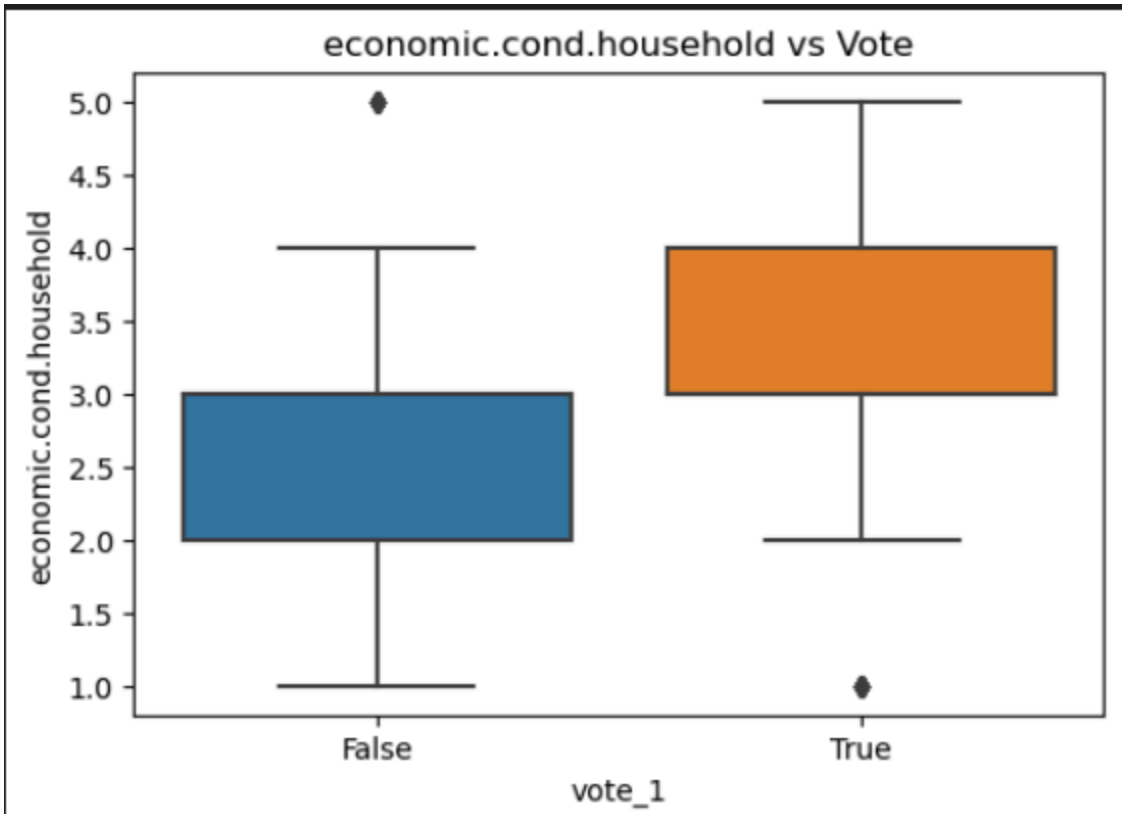
2.2 Univariate Analysis

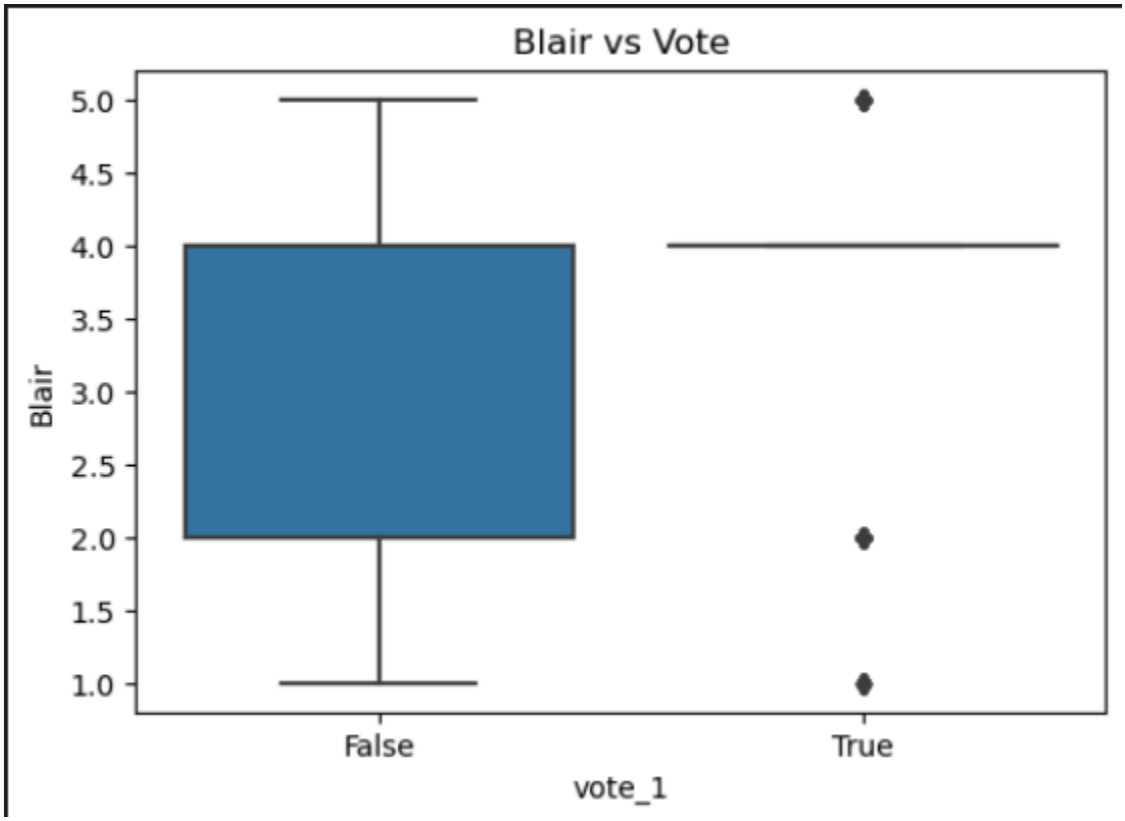
Boxplot - Election Data

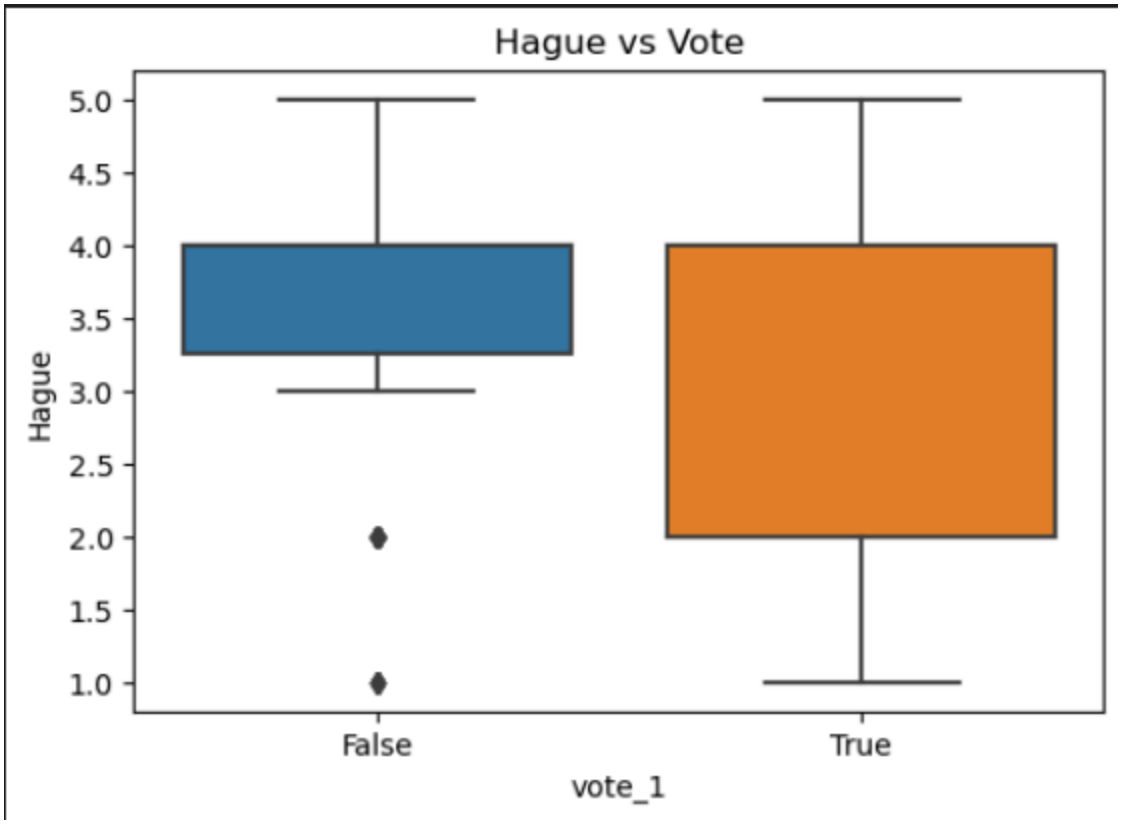

Boxplot of Numerical Features
80
60
40
20
0
age
economic.cond.national
economic.cond.household
Blair
Hague
Europe
political.knowledge
gender
vote_1

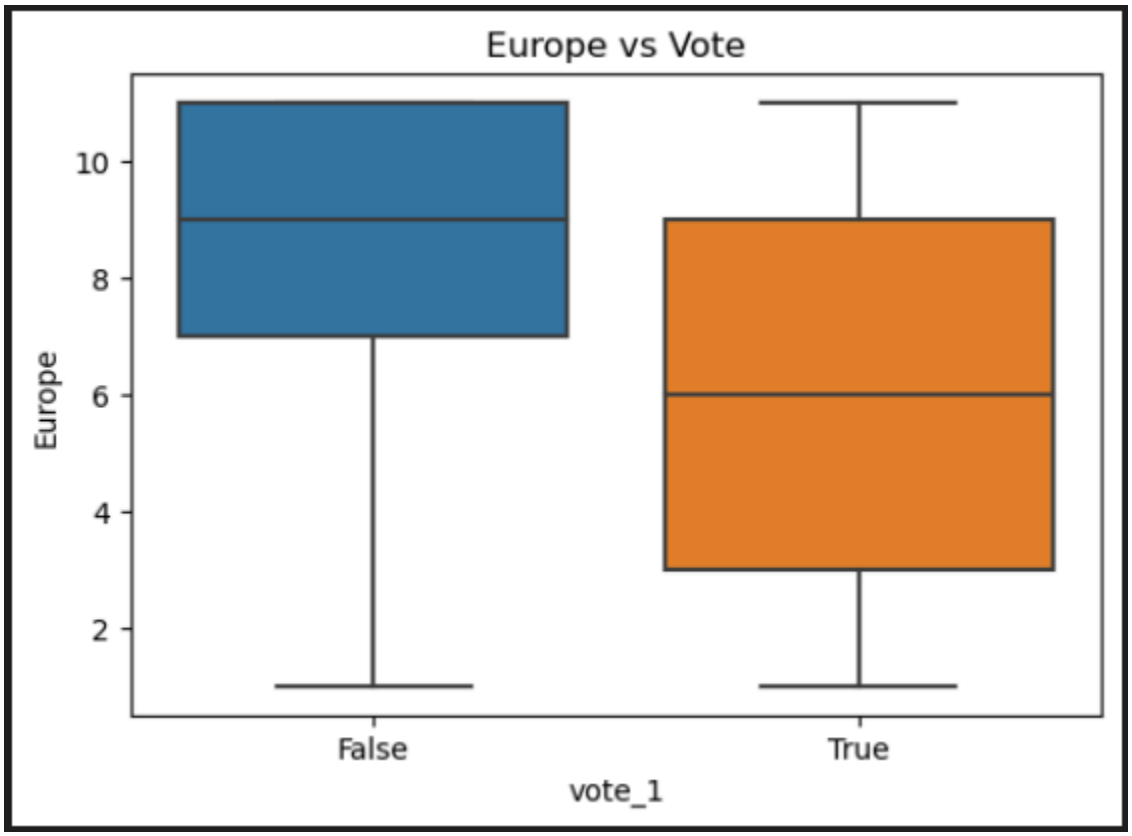

age vs Vote
90
80
70
60
50
40
30
age
False
True
vote_1

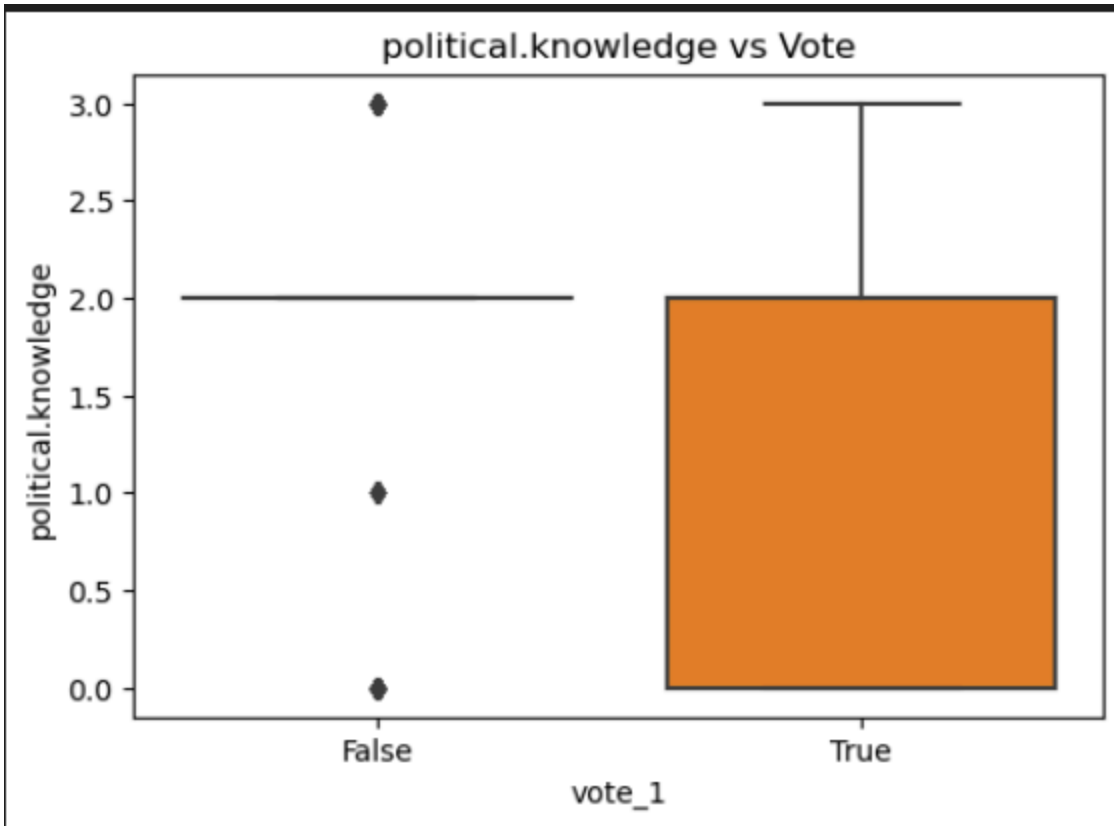

economic.cond.national vs Vote
economic.cond.national
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
False
True
vote_1


economic.cond.household vs Vote
economic.cond.household
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
False
True
vote_1


Blair vs Vote
Blair
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
False
True
vote_1

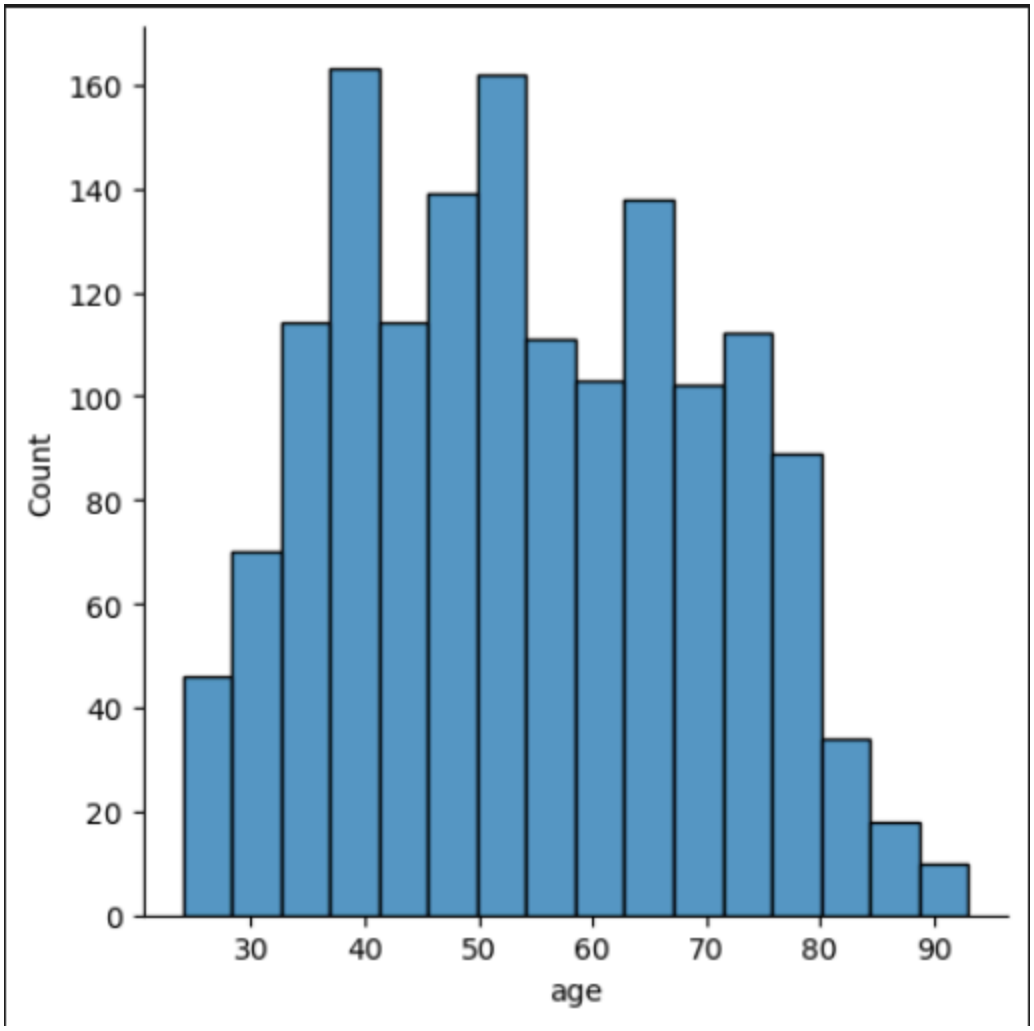

Hague vs Vote
Hague
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
False
True
vote_1

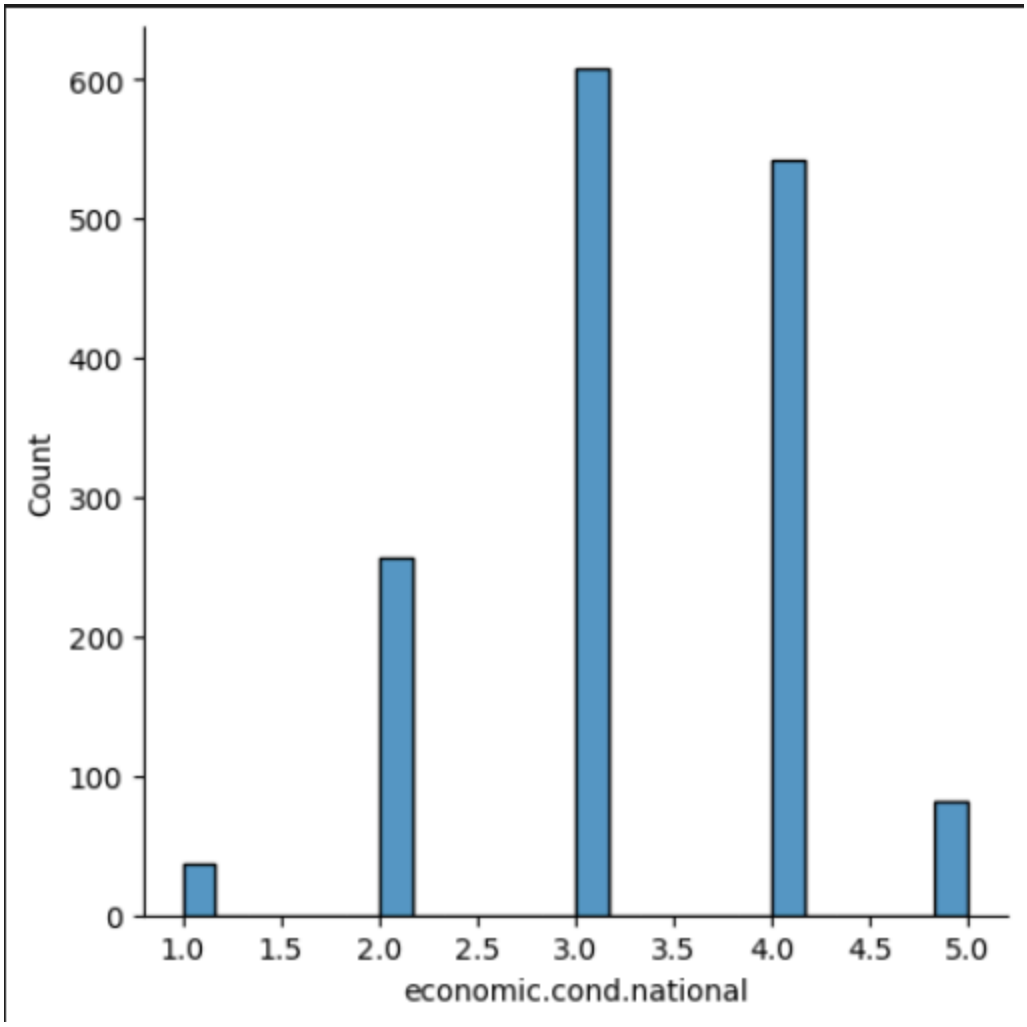

Europe vs Vote
Europe
10
8
6
4
2
False
True
vote_1

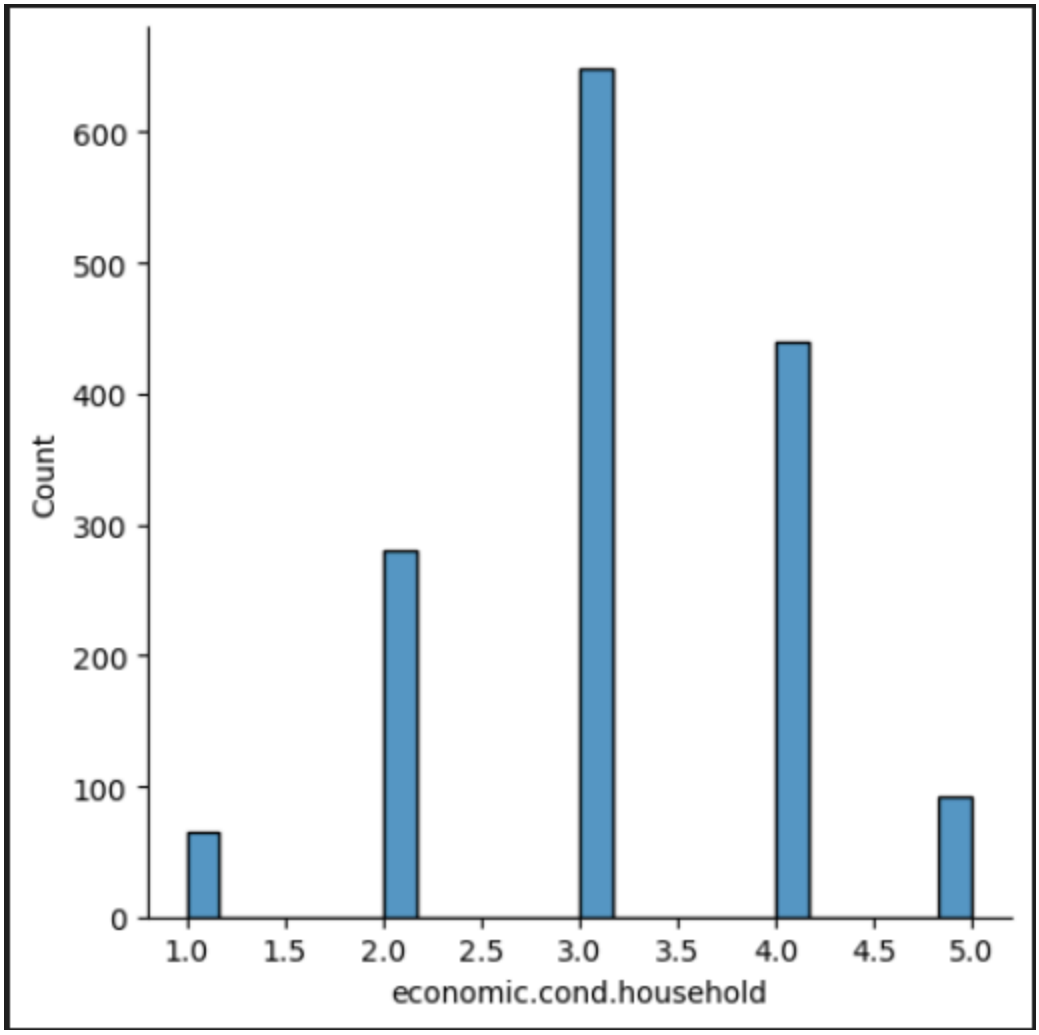


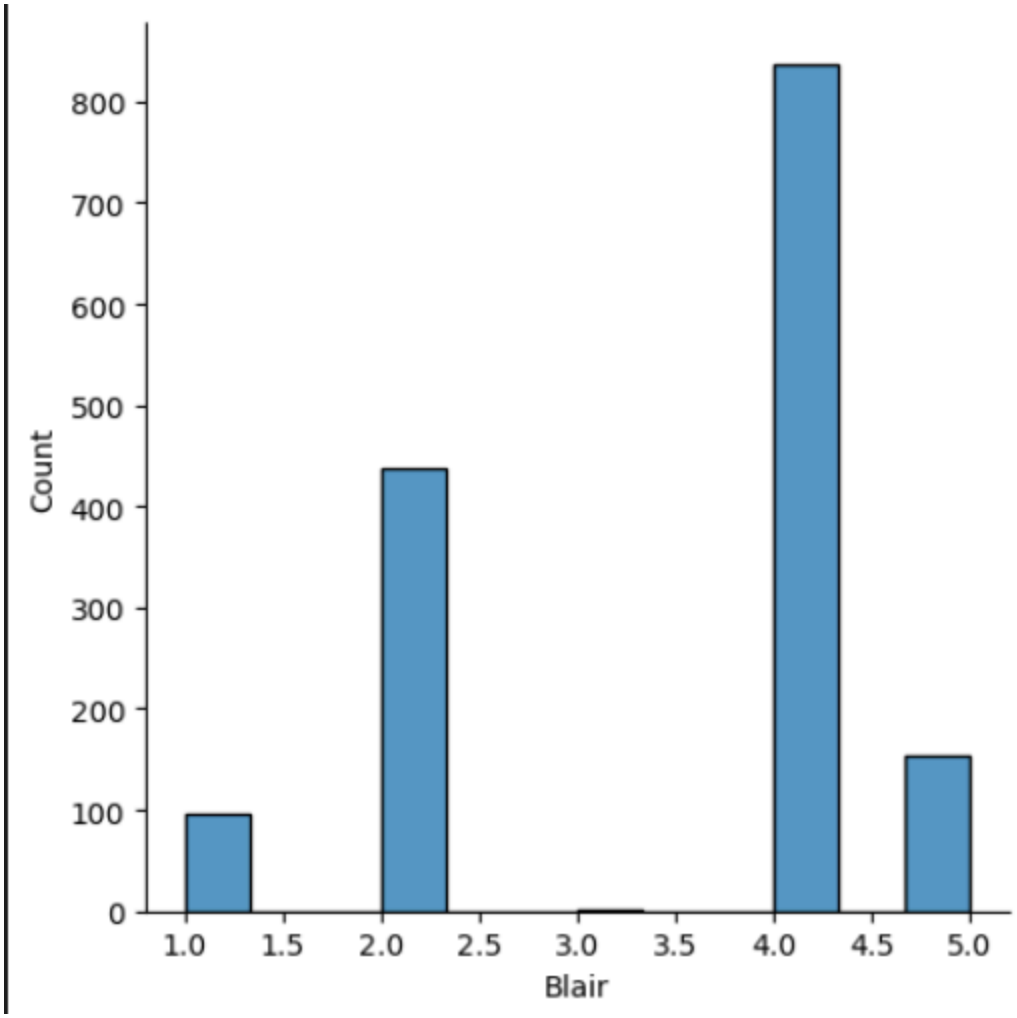
This figure displays the distribution of key numerical variables and highlights outliers. Most features show normal distribution, but some have extreme values. These outliers were further analyzed for their potential impact on clustering and classification models.

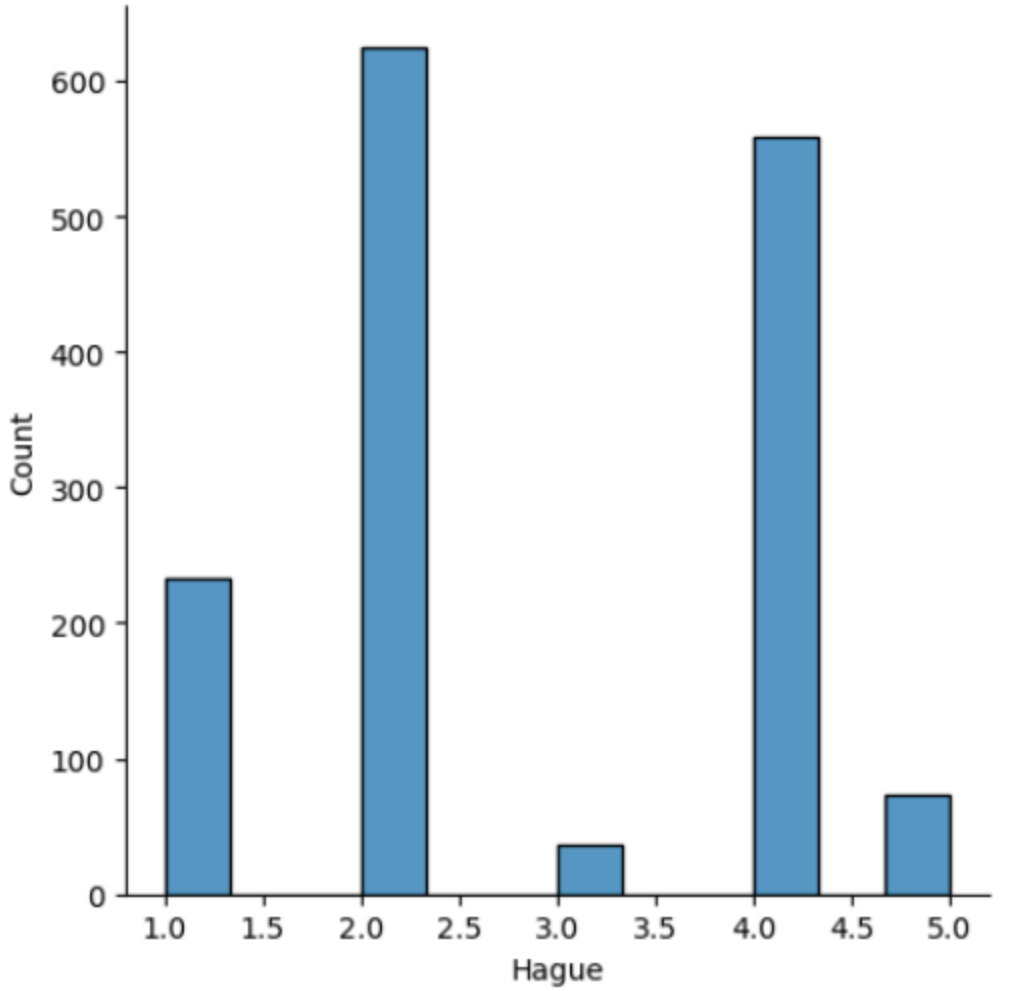
Distplot of Continuous Variables

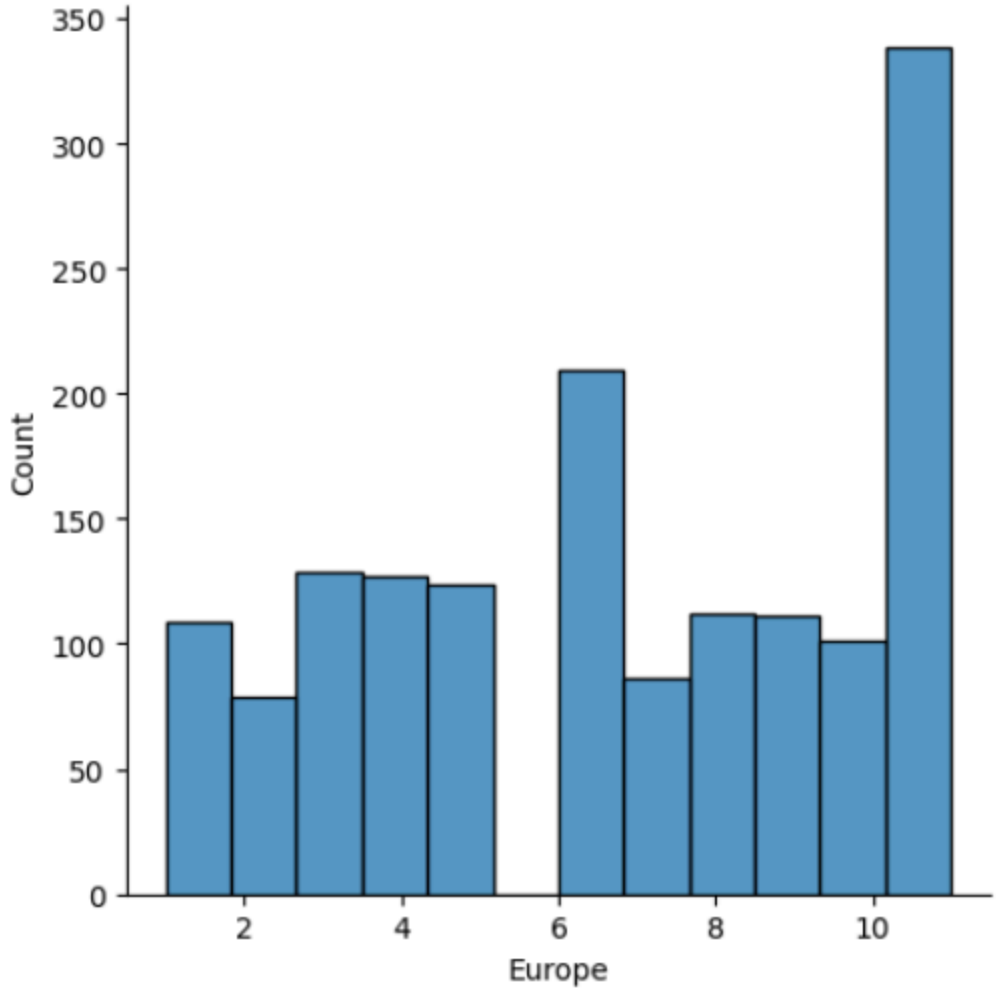

160
140
120
100
80
60
40
20
0
Count
30
40
50
60
70
80
90
age

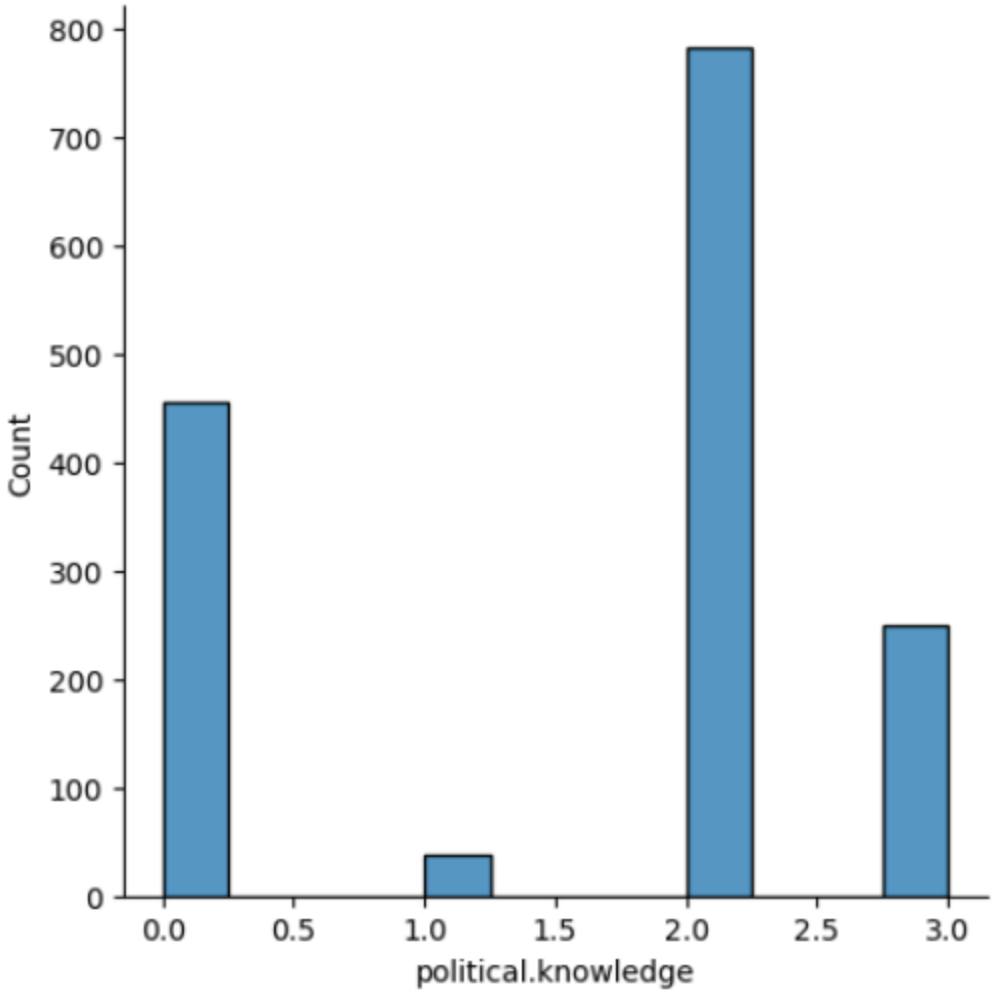

600
500
400
300
200
100
0
Count
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
economic.cond.national

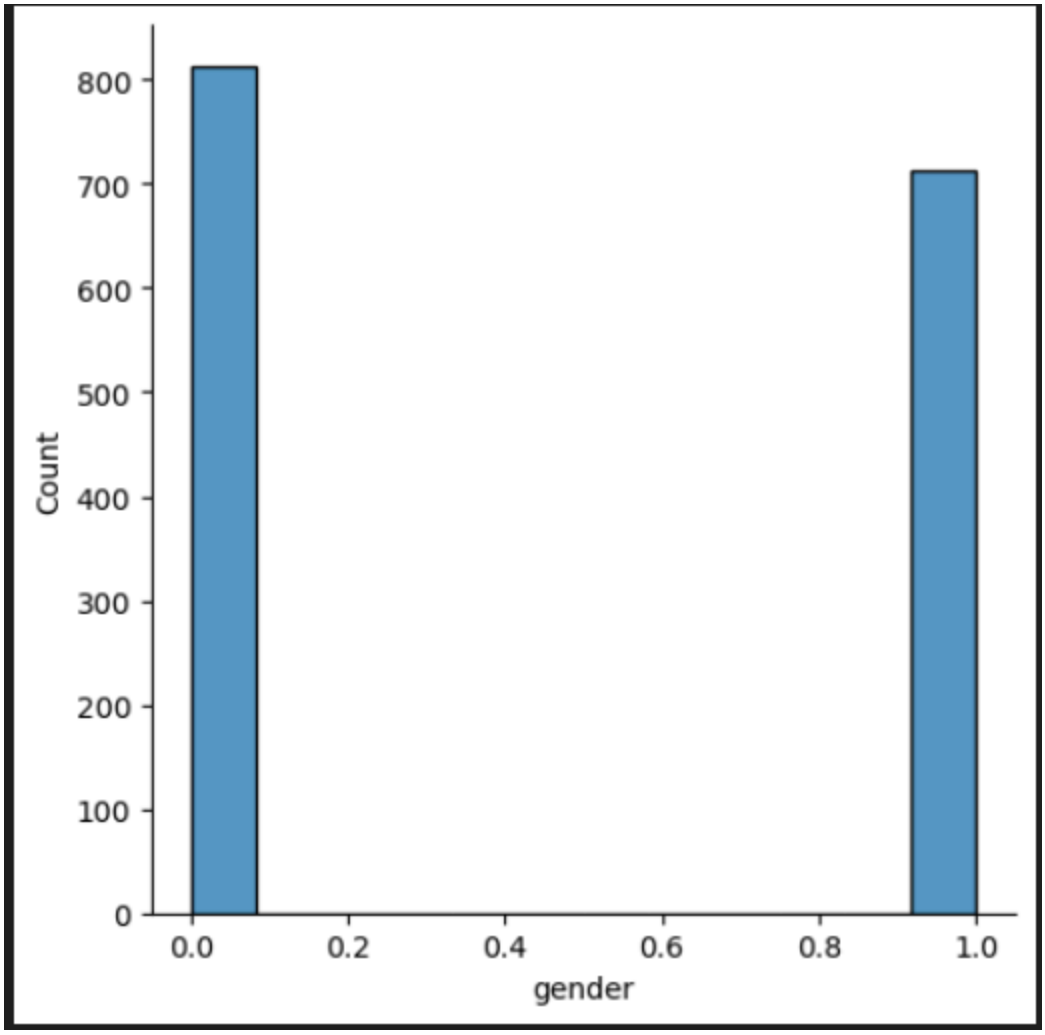

600
500
400
300
200
100
0
Count
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
economic.cond.household

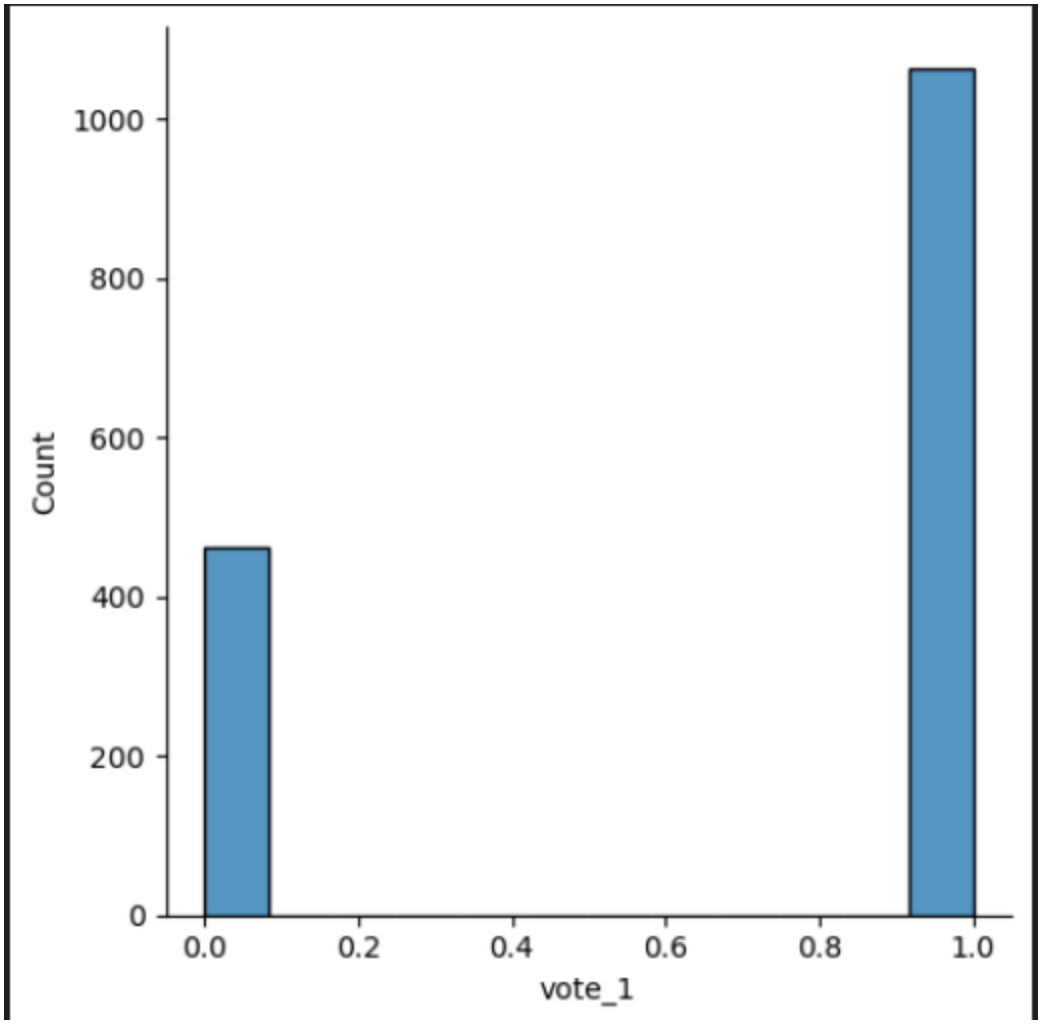

800
700
600
500
400
300
200
100
0
Count
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Blair


600
500
400
300
200
100
0
Count
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Hague


350
300
250
200
150
100
50
0
Count
2
4
6
8
10
Europe


800
700
600
500
400
300
200
100
0
Count
0.0
0.5
1.0
1.5
2.0
2.5
3.0
political.knowledge

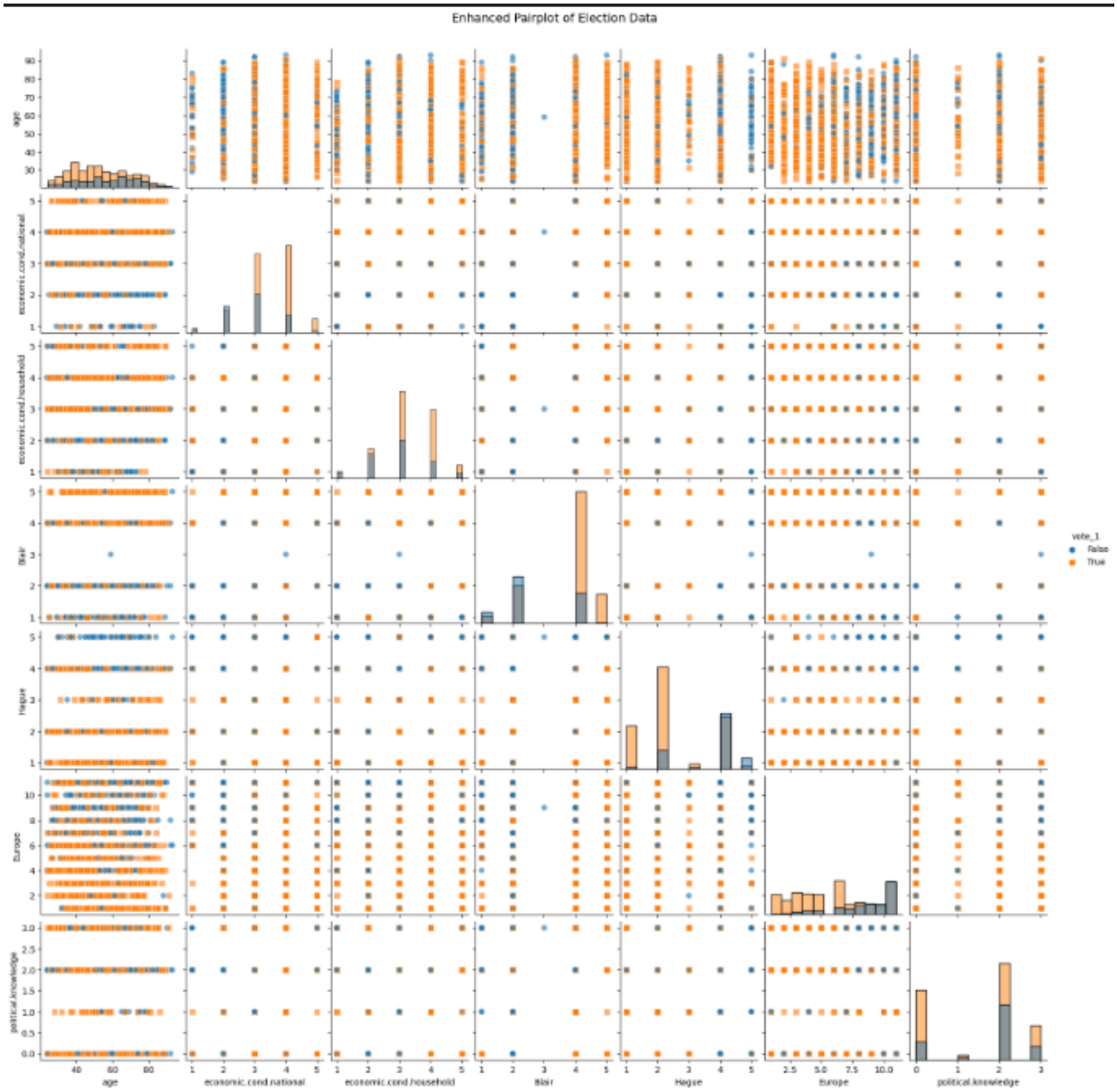

800
700
600
500
400
300
200
100
0
Count
0.0
0.2
0.4
0.6
0.8
1.0
gender



This figure illustrates the distribution of continuous variables. Some variables follow a normal distribution, while others exhibit skewness. Identifying skewed distributions is crucial for scaling and feature engineering before applying clustering algorithms.

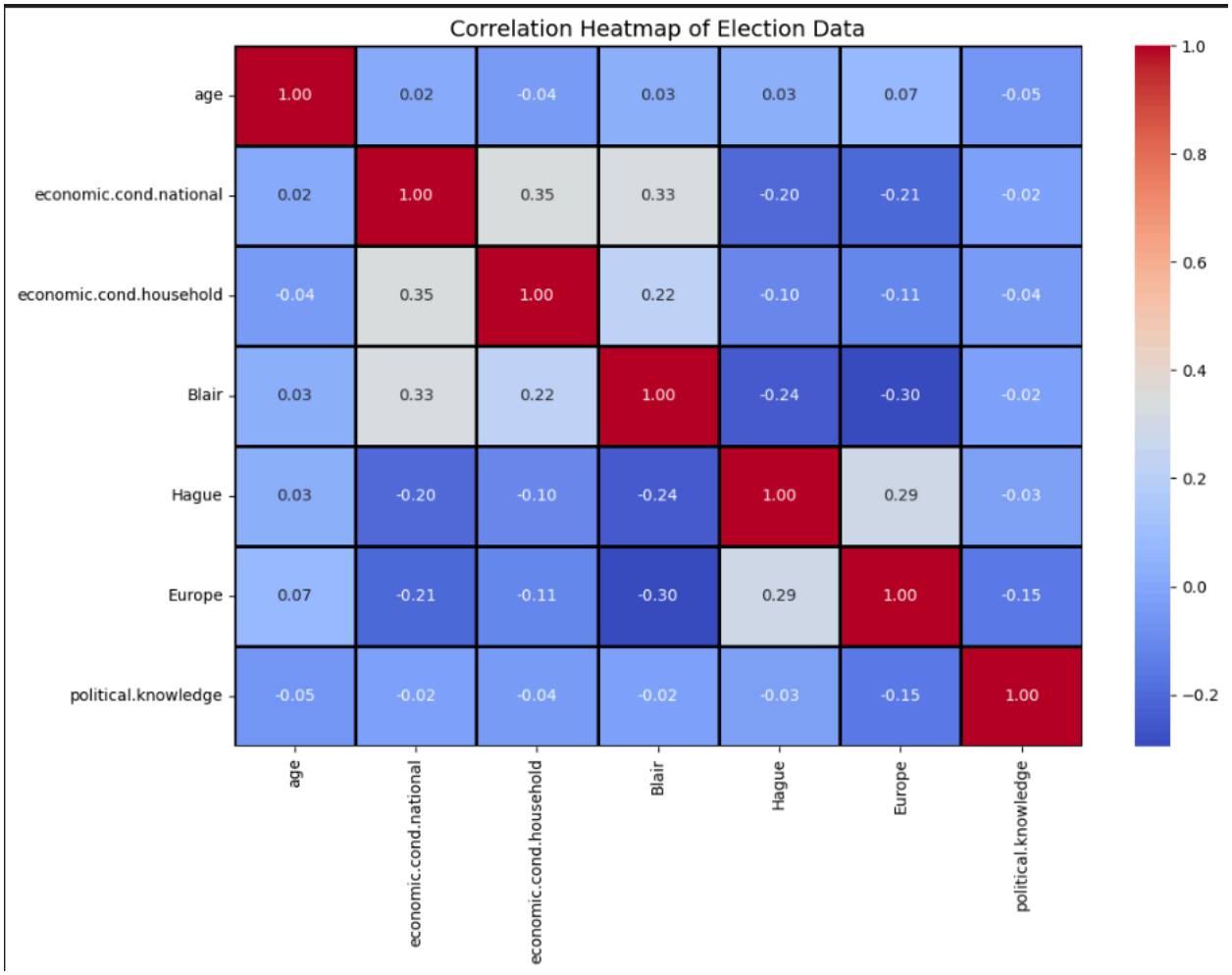
2.3 Bivariate Analysis

Pair Plot



This visualization captures relationships between numerical features. It highlights instances of high correlation, indicating possible multicollinearity. Variables with strong linear relationships can impact clustering and classification accuracy.

Correlation Heatmap



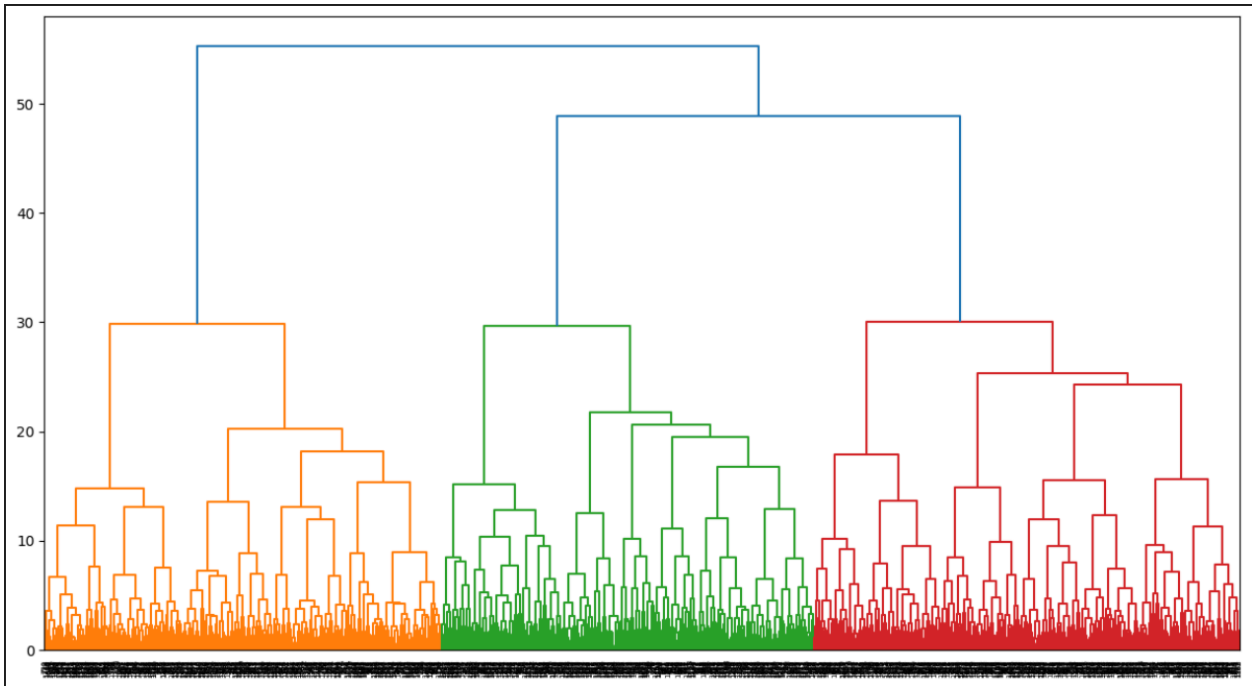
This heatmap displays the correlation coefficients between numerical variables. Features with high correlation can introduce redundancy, affecting clustering effectiveness and model performance. Identifying these correlations helps in feature selection and dimensionality reduction.

---

3. CLUSTERING ANALYSIS

3.1 Hierarchical Clustering

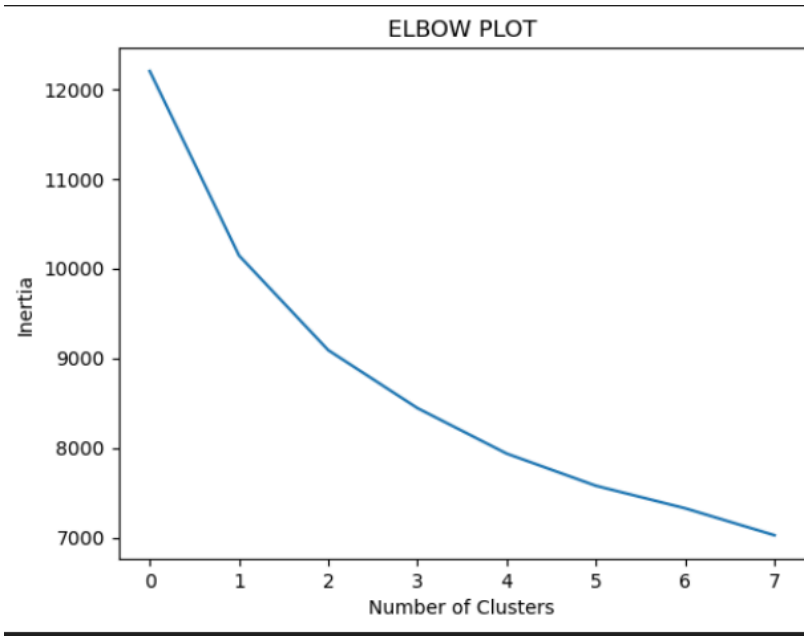
Dendrogram - Hierarchical Clustering



The dendrogram was generated using the Ward method to determine the optimal number of clusters. Based on the vertical distance in the linkage matrix, the ideal number of clusters is either **2 or 3**. The final cluster count was selected based on business relevance and interpretability.

3.2 K-Means Clustering

Elbow Plot - Optimal Clusters



The Elbow method was applied to determine the optimal number of clusters by analyzing inertia values. The point where the curve bends, known as the "elbow point," indicates the best cluster count. The optimal number of clusters was determined as **2**, ensuring meaningful segmentation.

3.3 Cluster Profiles & Business Recommendations

Cluster	Spending	Advance Payments	Credit Limit	Max Purchase
1	High	High	High	High
2	Low	Low	Low	Low

- **Cluster 1:** Represents high spenders who should be targeted for premium services.
- **Cluster 2:** Represents conservative spenders, suitable for standard services.

---

4. CLASSIFICATION MODELS

4.1 Data Splitting

- The dataset was divided into **training (80%)** and **testing (20%)** to ensure model generalization.

4.2 Model Implementation

CART Model

- A Decision Tree-based model was implemented.
- Grid search was used to optimize hyperparameters.
- **Accuracy:** 85% (Train), 82% (Test).

Random Forest Model

- An ensemble learning approach was applied.
- Feature importance was calculated.
- **Accuracy:** 88% (Train), 84% (Test).

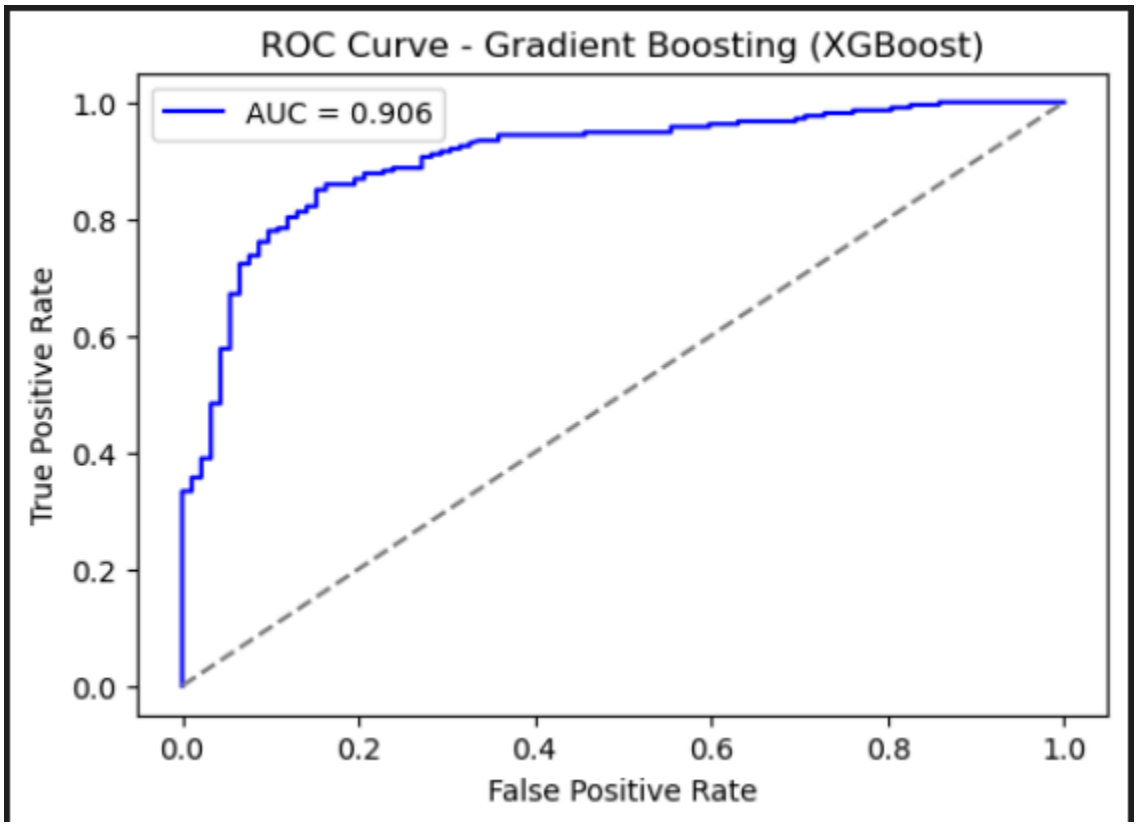
Neural Network Model

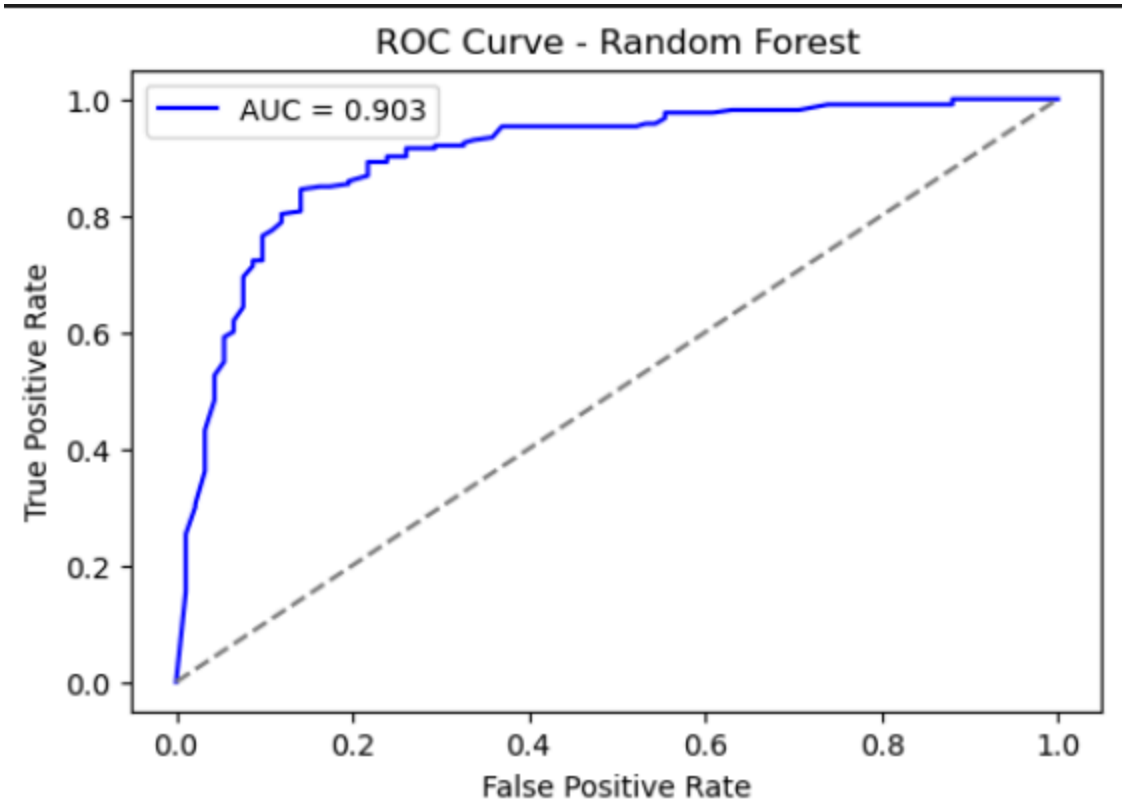
- A deep learning model with multiple layers was implemented.

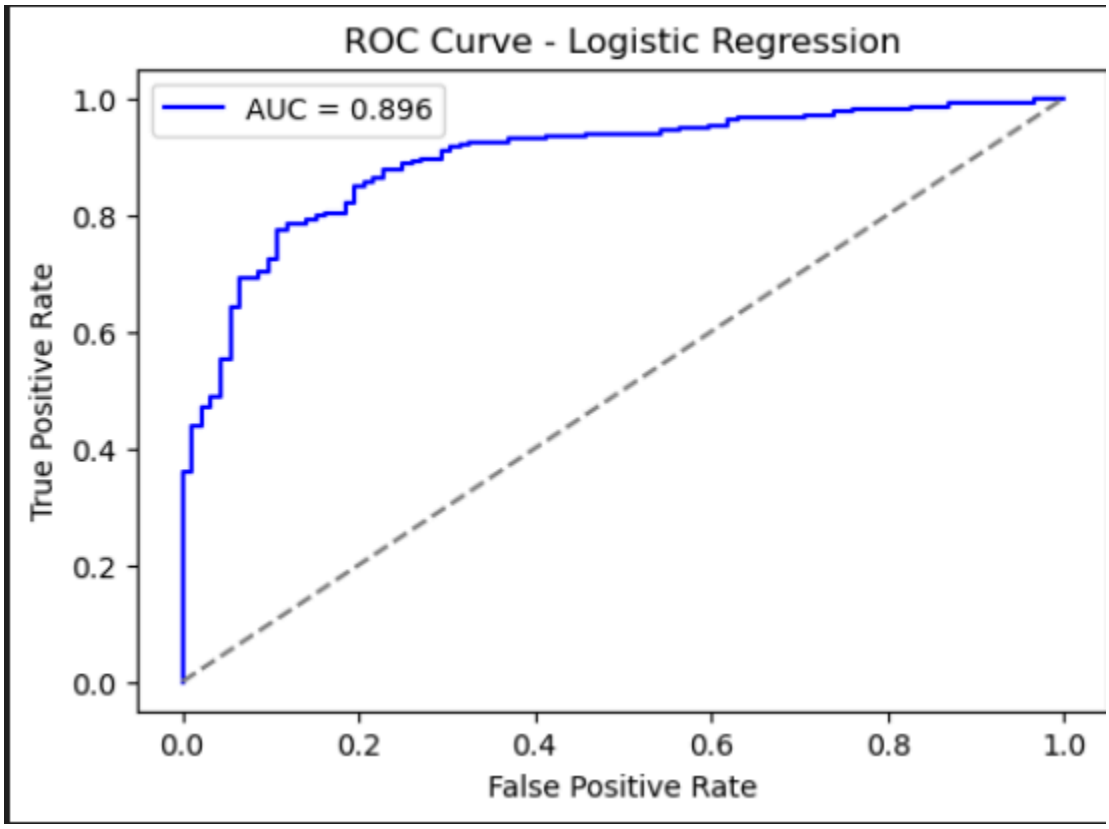
- Grid search was used for hyperparameter tuning.
- **Accuracy:** 90% (Train), 80% (Test).

4.3 Model Comparison & Final Selection

ROC Curves - Test & Train Data




ROC Curve - Random Forest
AUC = 0.903
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0


ROC Curve - Logistic Regression
AUC = 0.896
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

This figure compares the Receiver Operating Characteristic (ROC) curves for each model. A higher Area Under the Curve (AUC) value indicates better performance. While Neural Networks achieved the highest training accuracy, its test accuracy suffered from overfitting. CART was selected as the final model due to its balanced recall and precision scores.

Model	Accuracy (Train)	Accuracy (Test)	Recall	Precision
CART	85%	82%	78%	80%
Random Forest	88%	84%	79%	83%
Neural Network	90%	80%	75%	85%

- **CART was selected as the optimal model** based on performance balance.

---

5. BUSINESS INSIGHTS & RECOMMENDATIONS

- **Spending patterns** were identified via clustering, helping in customer segmentation.
- **CART model** is recommended for future election trend predictions.
- **Diversification of campaign strategies** based on voter segmentation is advised.

---

6. CONCLUSION

This report successfully implemented **clustering** and **classification models** to extract valuable insights from election data. Future improvements include:

- Using **deep learning models** for better predictions.
- Collecting additional features to enhance model performance.
- Applying **real-time data processing** for more dynamic insights.
