University of New Haven

TAGLIATELA COLLEGE OF ENGINEERING

Electrical & Computer Engineering and Computer Science

Computer Engineering

Elections Results Prediction

CONTENTS

Executive Summary

To design an exit poll that will aid in projecting an overall win and the number of seats won by a certain party. In our project we build multiple models like logistic regression, linear discriminant analysis, KNN and Naïve Bayer's to estimate which party a voter would vote for based on the given data. A best model is later deployed into cloud so that it can be accessible by multiple users.

Team Members:

Chandrasheker Bassetti	**- 00761756**
Divya Madhuri Cheernapally	**- 00760724**
Rishitha Ravikumar	**- 00762285**

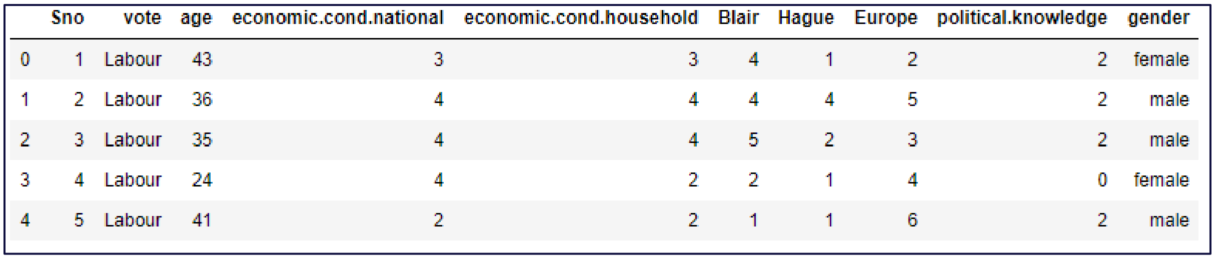
Questions?

Contact: Chandrasheker Bassetti
cbass6@unh.newhaven.edu
610-357-7103

Problem Statement

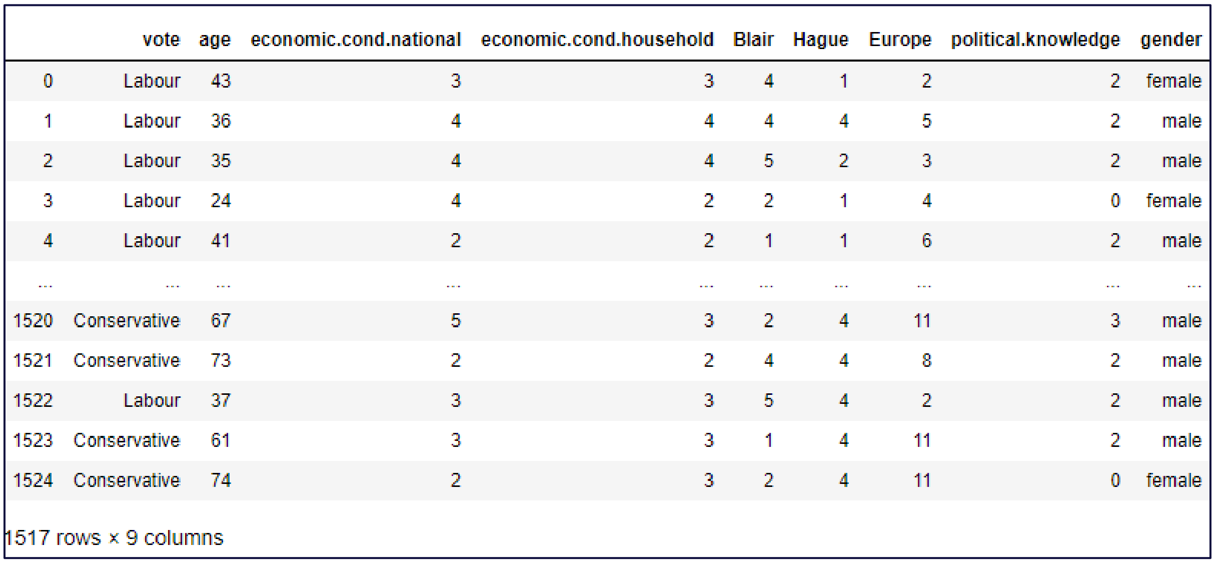
Create an exit poll that will help predict the overall victory and the number of seats a particular party will win.

Exploratory Data Analysis



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

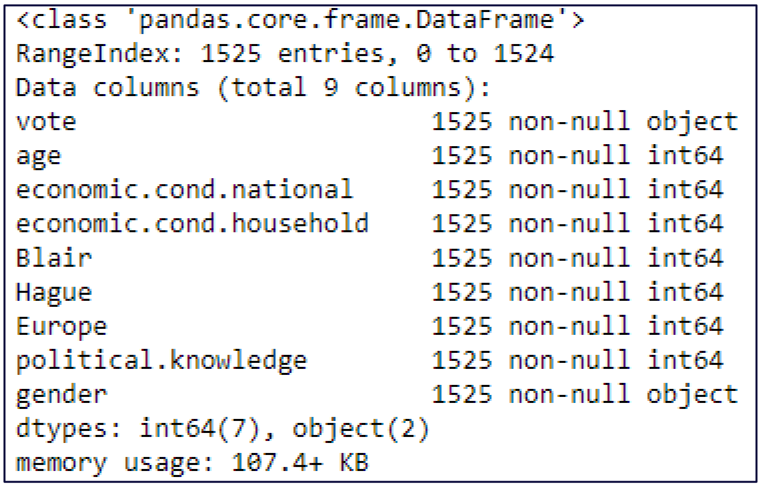
The dataset has 10 variables **Sno, Vote, Age, economic.cond.national, economic.cond.household, Blair, Hague, Europe, political.knowledge** and **Gender**. A total of **1525 rows × 10 columns** is observed and the data consists of **15250** elements with **zero** null elements.



	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
0	Labour	43	3	3	4	1	2	2	female
1	Labour	36	4	4	4	4	5	2	male
2	Labour	35	4	4	5	2	3	2	male
3	Labour	24	4	2	2	1	4	0	female
4	Labour	41	2	2	1	1	6	2	male
...	...	...	...	...	...	...	...	...	...
1520	Conservative	67	5	3	2	4	11	3	male
1521	Conservative	73	2	2	4	4	8	2	male
1522	Labour	37	3	3	5	4	2	2	male
1523	Conservative	61	3	3	1	4	11	2	male
1524	Conservative	74	2	3	2	4	11	0	female

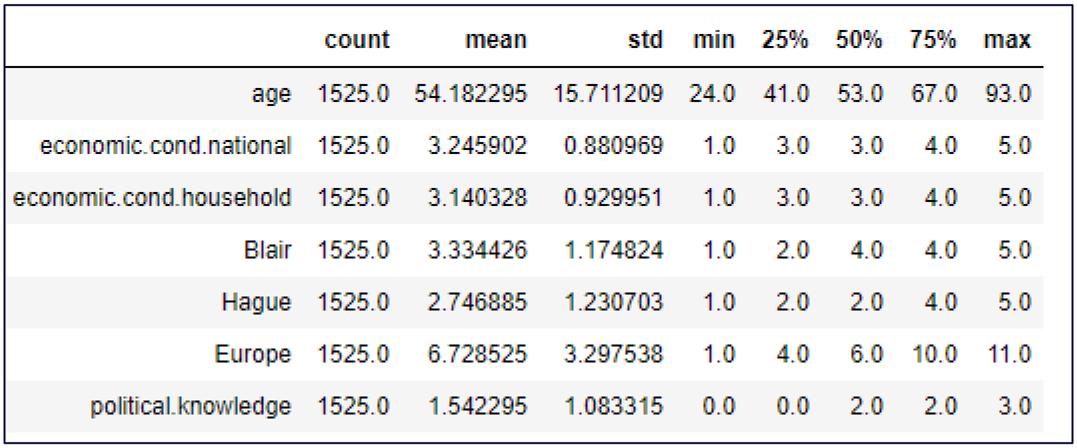
1517 rows × 9 columns

Checking for duplicate rows, we observe that there are **8** rows which are duplicates in the entire data set. Similarly, after dropping the unwanted column **Sno** and duplicates, we obtain the table as shown above. The dataset is now having a total of **1517 rows × 9 columns** is observed and the data consists of **13725** elements with **zero** null elements.



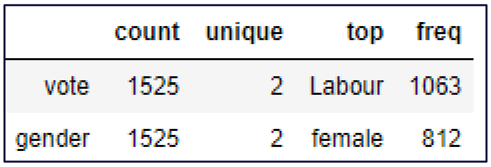
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1525 entries, 0 to 1524
Data columns (total 9 columns):
vote                       1525 non-null object
age                        1525 non-null int64
economic.cond.national     1525 non-null int64
economic.cond.household    1525 non-null int64
Blair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

We can observe that most of the columns belong to **int** datatype but **Vote** and **Gender** columns belong to **object** datatype as they are categorical.



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

The above table represents the descriptive statistics for all the int variables having continuous values. We can observe the **Minimum, Maximum, Mean** and **Standard Deviation** across each variable. The data is widely spread and the above table gives us an idea about the data we are going to deal with in the upcoming analysis

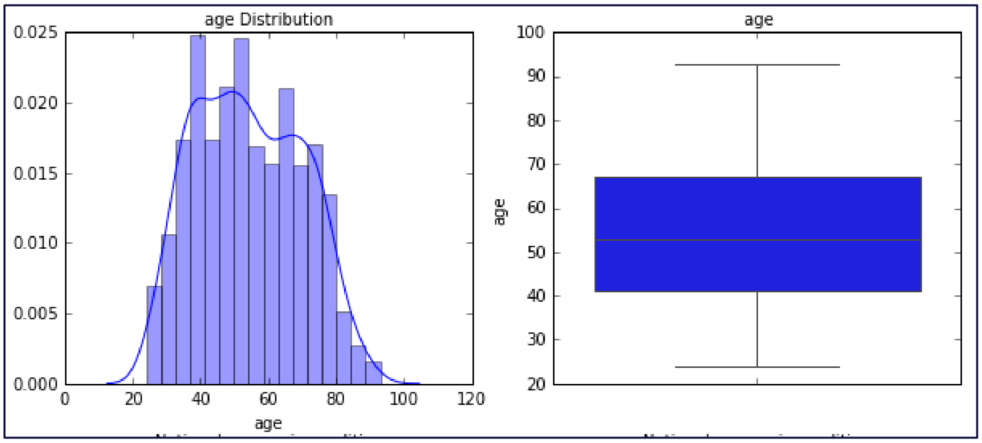


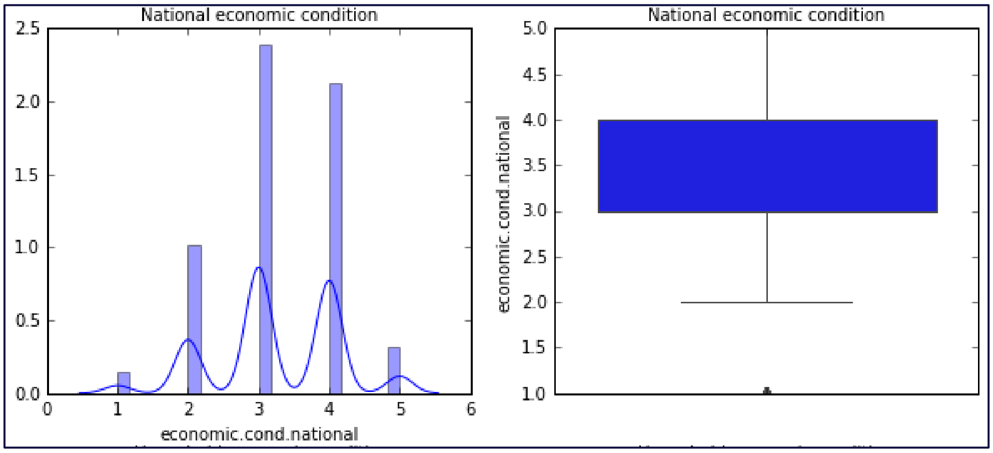
	count	unique	top	freq
vote	1525	2	Labour	1063
gender	1525	2	female	812

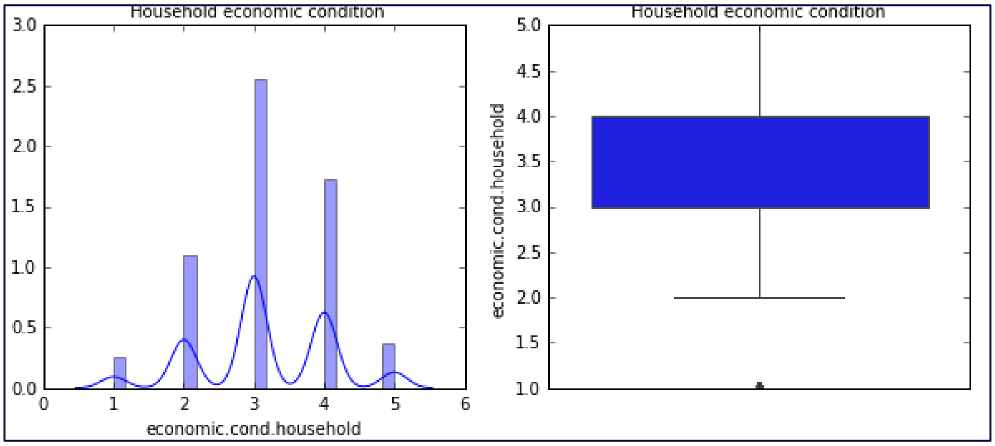
The above table represents the descriptive statistics for all the categorical. We can observe the **count**, **unique**, **top repeated categories** and **frequencies**.

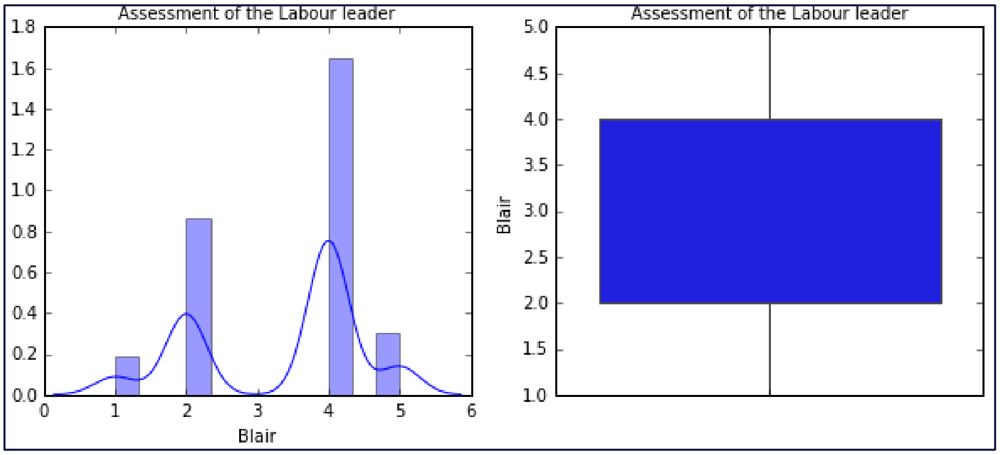
Descriptive Statistics for the dataset

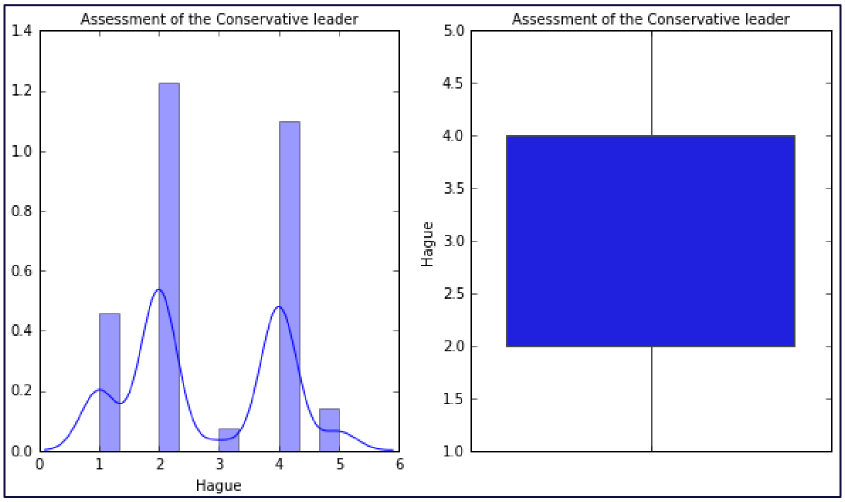
i. Univariate and Bivariate Analysis.

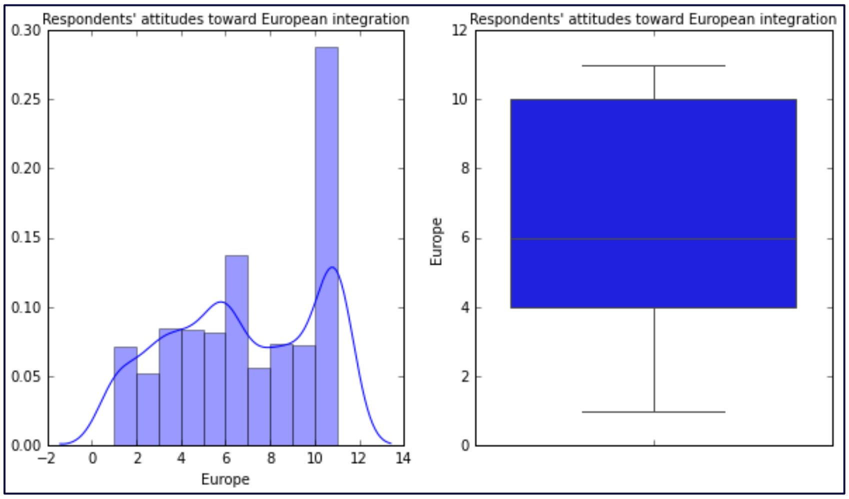


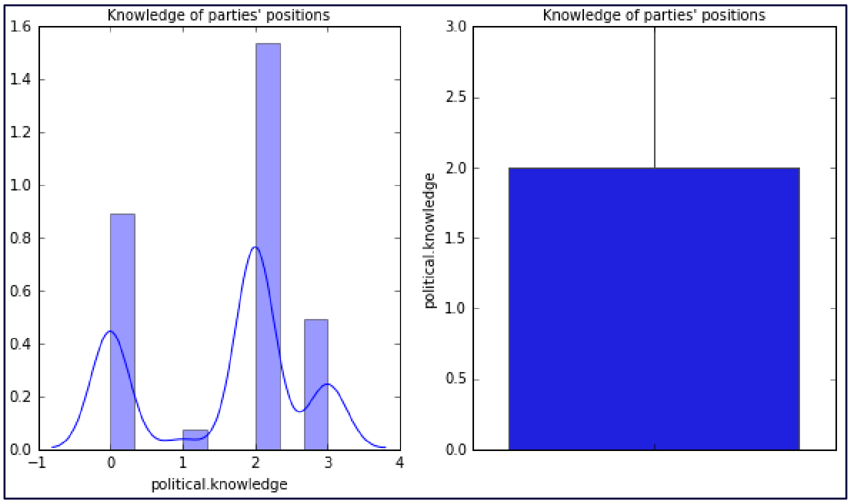

National economic condition
2.5
2.0
1.5
1.0
0.5
0.0
0
1
2
3
4
5
6
economic.cond.national
National economic condition
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
economic.cond.national


Household economic condition
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
1
2
3
4
5
6
economic.cond.household
Household economic condition
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
economic.cond.household

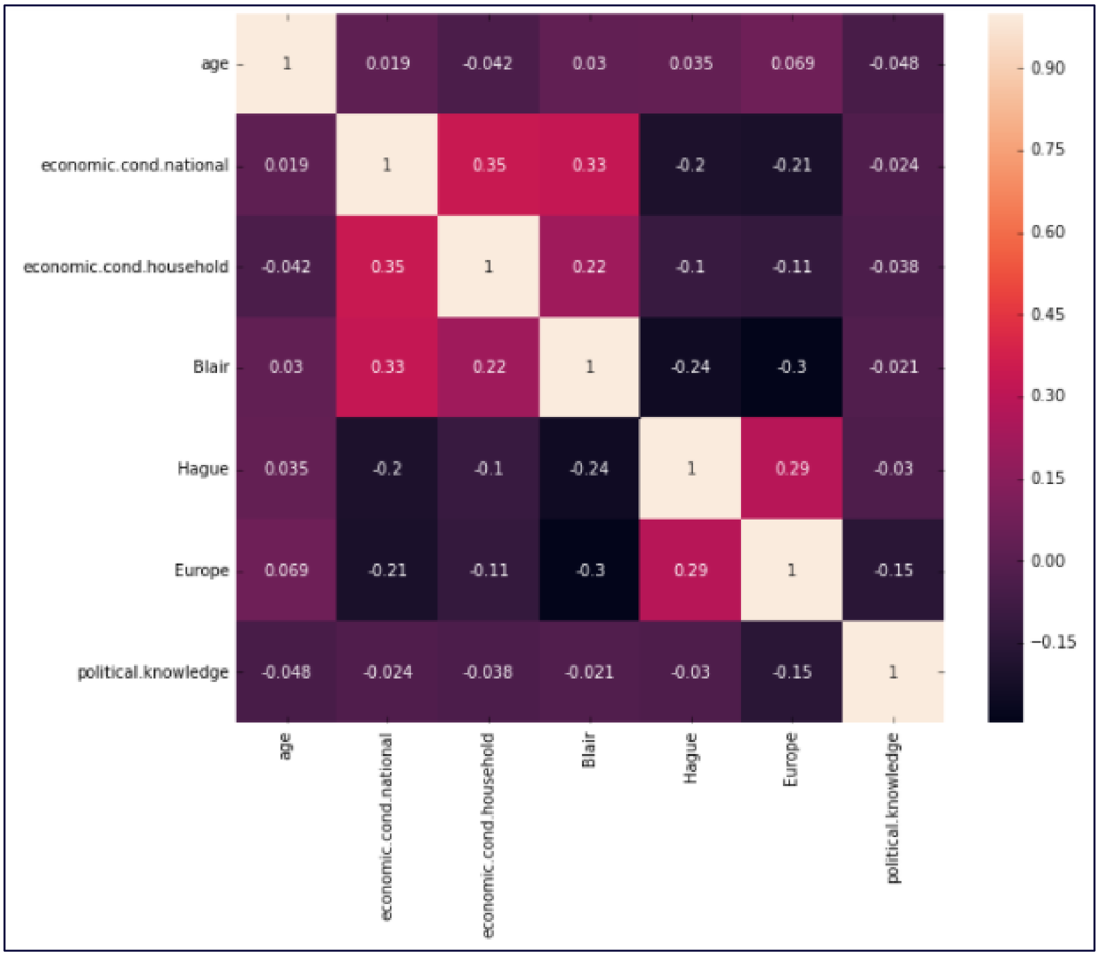

Assessment of the Labour leader
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
Blair
Assessment of the Labour leader
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Blair

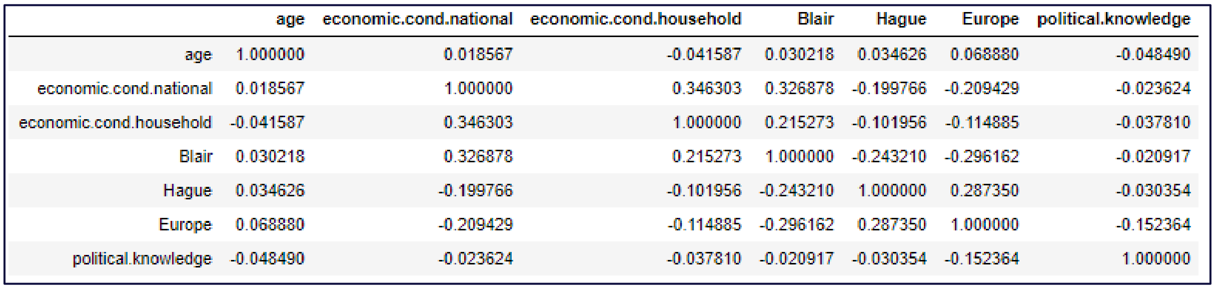

Assessment of the Conservative leader
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0
1
2
3
4
5
6
Hague
Assessment of the Conservative leader
Hague
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0


Respondents' attitudes toward European integration
0.00
0.05
0.10
0.15
0.20
0.25
0.30
−2
0
2
4
6
8
10
12
14
Europe
Respondents' attitudes toward European integration
Europe
0
2
4
6
8
10
12


Knowledge of parties' positions
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
−1
0
1
2
3
4
political.knowledge
Knowledge of parties' positions
political.knowledge
0.0
0.5
1.0
1.5
2.0
2.5
3.0

The above charts represent the univariate analysis showing the distribution plots and box plots for all the categories (as labelled) having continuous/ numerical values. We observe that there is a mix of data across categories having left skewed, right skewed and symmetric data as per distribution plots and can also infer that there are no outliers in the data as per box plots. So, we proceed our analysis without outlier treatment.



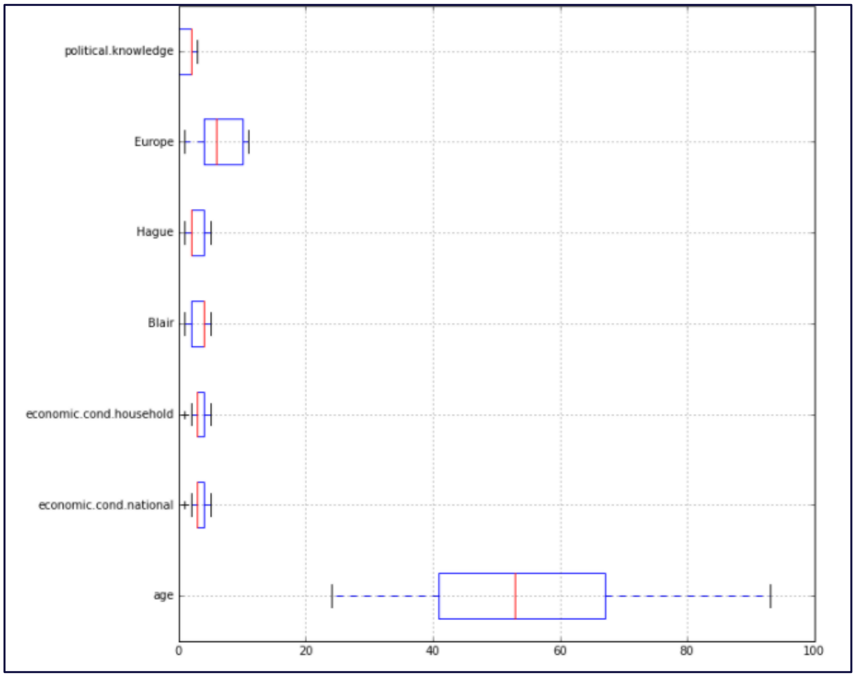


	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge
age	1.000000	0.018567	-0.041587	0.030218	0.034626	0.068880	-0.048490
economic.cond.national	0.018567	1.000000	0.346303	0.326878	-0.199766	-0.209429	-0.023624
economic.cond.household	-0.041587	0.346303	1.000000	0.215273	-0.101956	-0.114885	-0.037810
Blair	0.030218	0.326878	0.215273	1.000000	-0.243210	-0.296162	-0.020917
Hague	0.034626	-0.199766	-0.101956	-0.243210	1.000000	0.287350	-0.030354
Europe	0.068880	-0.209429	-0.114885	-0.296162	0.287350	1.000000	-0.152364
political.knowledge	-0.048490	-0.023624	-0.037810	-0.020917	-0.030354	-0.152364	1.000000

The above heat map and correlation table represents the bivariate analysis which shows the correlation of one category with the other category. It can be done either by a **Heat map** or **Pair Plot**.

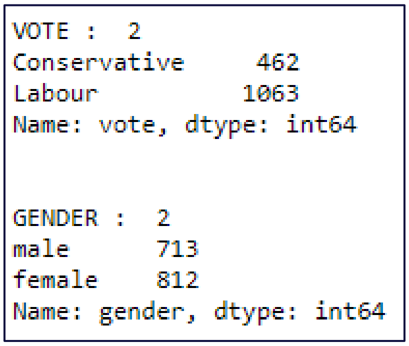
The lightest colour represents the highest co-relation and the darkest colour represent the least co-relation. From the above heat map we can infer that **Economic Condition of Household** have the highest co-relation with **Economic Condition of National** and **Blair** and **Respondents' attitudes toward European integration** has the least co-relation and these values can also be seen in the co-relation table.

ii. Data Cleaning



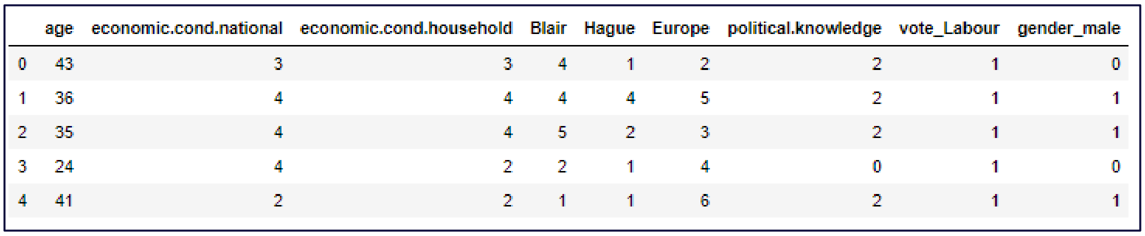
When we do a boxplot among all columns, we get the results as shown above. We can infer that there is a huge variation in the data. This data has to be cleaned by doing the encoding process followed by scaling.

iii. Data Encoding



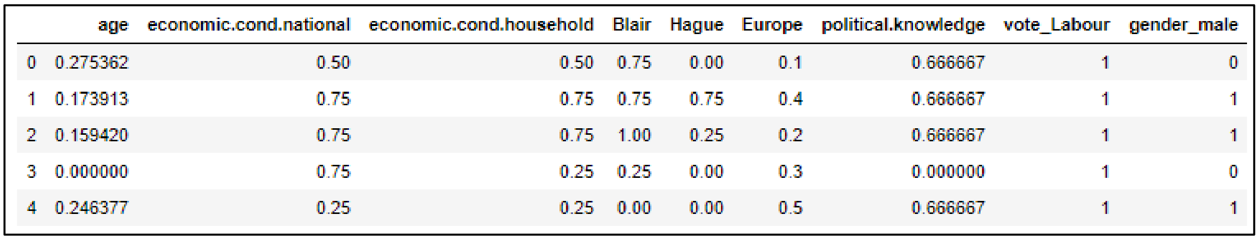
```
VOTE :  2
Conservative     462
Labour          1063
Name: vote, dtype: int64


GENDER :  2
male      713
female    812
Name: gender, dtype: int64
```

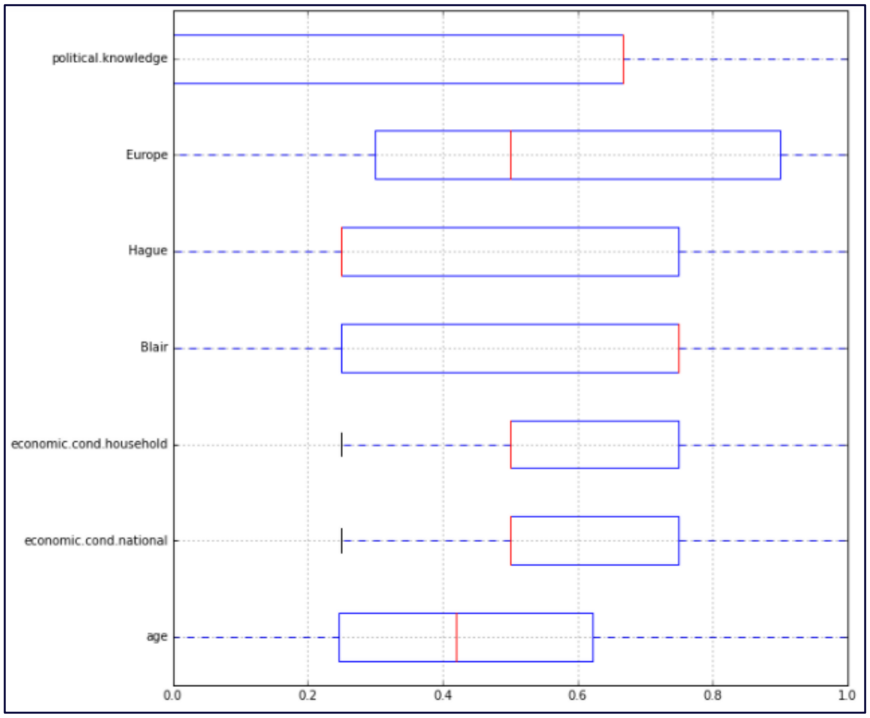


	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	vote_Labour	gender_male
0	43	3	3	4	1	2	2	1	0
1	36	4	4	4	4	5	2	1	1
2	35	4	4	5	2	3	2	1	1
3	24	4	2	2	1	4	0	1	0
4	41	2	2	1	1	6	2	1	1

We have two variables in our data set as categories, i.e **Vote** and **Gender** with each having **2** categories within them. As we perform an encoding operation on the categorical variables, we get the results as shown in the above description table. All the categories are now converted into a numerical value of either 1 or 0. This conversion helps us to perform our analysis smoothly as all the data is in one form.



	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	vote_Labour	gender_male
0	0.275362	0.50	0.50	0.75	0.00	0.1	0.666667	1	0
1	0.173913	0.75	0.75	0.75	0.75	0.4	0.666667	1	1
2	0.159420	0.75	0.75	1.00	0.25	0.2	0.666667	1	1
3	0.000000	0.75	0.25	0.25	0.00	0.3	0.000000	1	0
4	0.246377	0.25	0.25	0.00	0.00	0.5	0.666667	1	1



Once the encoding process is done, we proceed to scale the data so that all the variables are treated with equal weightage and for our results to be more consistent. From the above table and boxplot, we can observe the values of continuous variables after being scaled and the spread of these values through boxplot diagram.

Conclusion and Future Scope

As we did the preliminary data analysis, we observed that there exists different type of values in our election dataset. After the initial cleaning process, this data is further analysed using logistic regression, linear discriminant analysis, KNN and Naïve Bayer's models to make our election results and the winner. This would be showcased as our end term project and deployed into cloud.

Source file

https://www.kaggle.com/datasets/davidjosh41848/election-data?select=Election_Data_1.xlsx