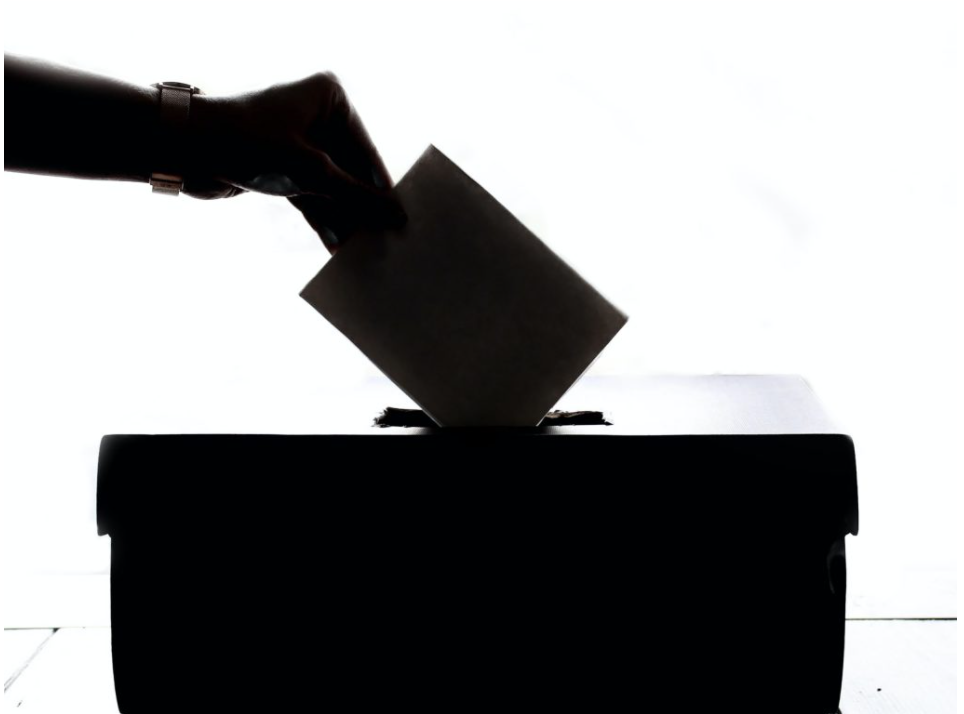
URL-http://irinaspage.com/wp-content/uploads/2019/12/GSK\_headerimage\_1280x720.png

|  |  |
| --- | --- |
|  |  |

**TABLE OF CONTENTS:**

PROJECT 1:

ELECTION DATA



https://keough.nd.edu/sobering-lessons-for-american-democracy-and-the-road-ahead/

SUMMARY

We are hired by one of the leading news channels CNBE who wants to analyse recent elections. This survey was conducted on 1525 voters with 9 variables. We 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.

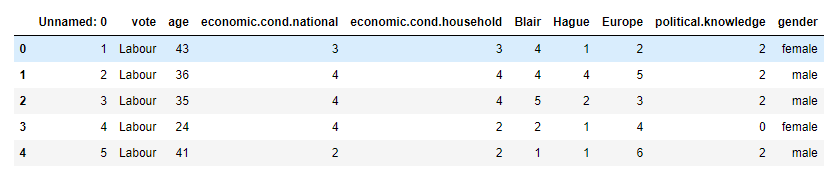
|  |  |
| --- | --- |
| **DATA DICTIONARY** |  |
| Vote | Party choice: Conservative or Labour |
| Age | in years |
| economic. cond. national | Assessment of current national economic conditions, 1 to 5. |
| economic. cond. household | Assessment of current household economic conditions, 1 to 5. |
| Blair | Assessment of the Labour leader, 1 to 5. |
| Hague | Assessment of the Conservative leader, 1 to 5. |
| Europe | An 11-point scale that measures respondents' attitudes toward European integration. High scores represent ‘Eurosceptic’ sentiment. |
| political. knowledge | Knowledge of parties' positions on European integration, 0 to 3. |
| gender | The vote casted by a male or female |

INTRODUCTION:

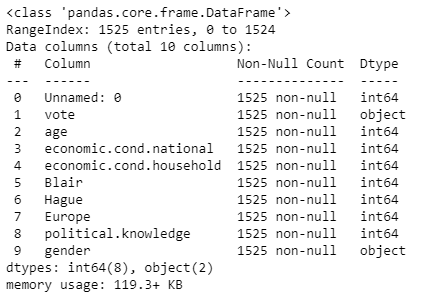
1.1) Read the dataset. Describe the data briefly. Interpret the inferences for each. Initial steps like head () .info (), Data Types, etc . Null value check, Summary stats, Skewness must be discussed.

We have the above data dictionary through which we will start on predicting overall win and seats covered by a particular party. The data set has the 9 columns and 1525 rows:



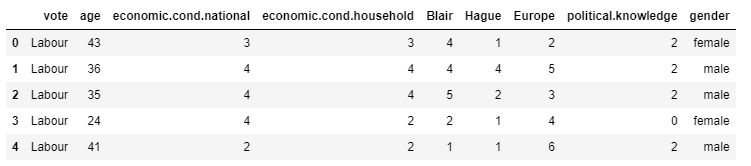


The data has the following data information as to which column is an integer type or object type:



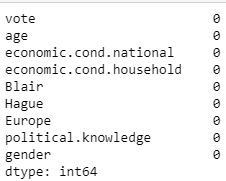
* The data types can be seen as 2 object or categorical columns and rest all as integer or numeric columns.
* As we can see from the data information above, we have only Vote and Gender as a categorical data type rest all being under numeric data type. Since the memory usage is paid in real world data, we can drop on the unnecessary column which is not adding any information to our data which is the index column:

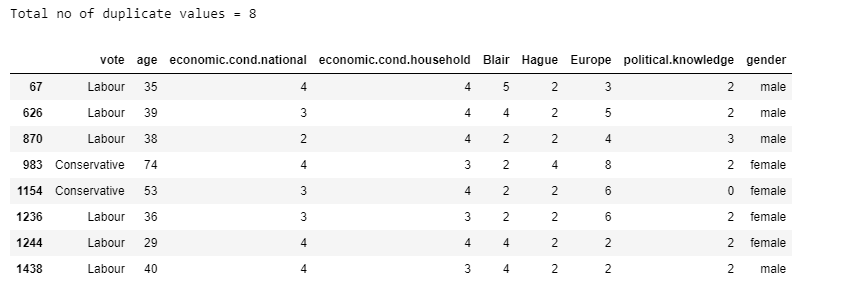
New dataset looks post dropping:



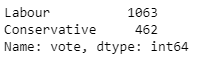
New memory usage of our file: 

The file of data has no null or empty rows hence we can proceed with the data to check duplicity:

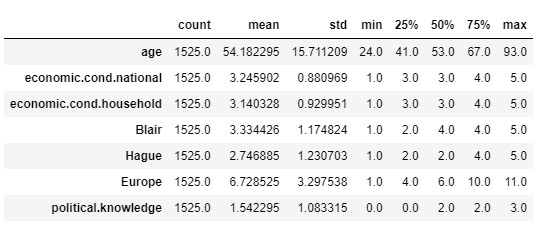




INFERENCE: Since the data has 8 duplicates which we won’t remove as we see all are different to each other as per each row when compared.



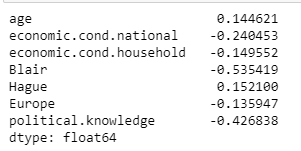
INFERENCE: We will check the votes casted by Labour Class and the ones by Conservative in order to know whose input in the data is more. Here it’s the labour class.



INFERENCE:

* As per the data description we can see the counts in all the columns is 1525.
* The mean being the highest for mean, standard deviation, minimum values and quartile values.
* The lowest is for having political knowledge in all the data description headings.

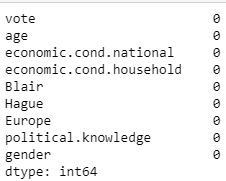
The skewness of the data being as below:

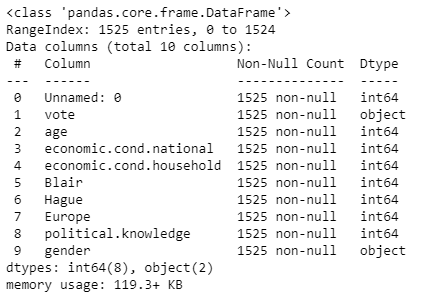


* We can infer that the data is fairly symmetrical.

1.2) Perform EDA (Check the null values, Data types, shape, Univariate, bivariate analysis). Also check for outliers (4 pts). Interpret the inferences for each (3 pts) Distribution plots(histogram) or similar plots for the continuous columns. Box plots, Correlation plots. Appropriate plots for categorical variables. Inferences on each plot. Outliers proportion should be discussed, and inferences from above used plots should be there. There is no restriction on how the learner wishes to implement this but the code should be able to represent the correct output and inferences should be logical and correct.

As checked, we have no null values in the data set.

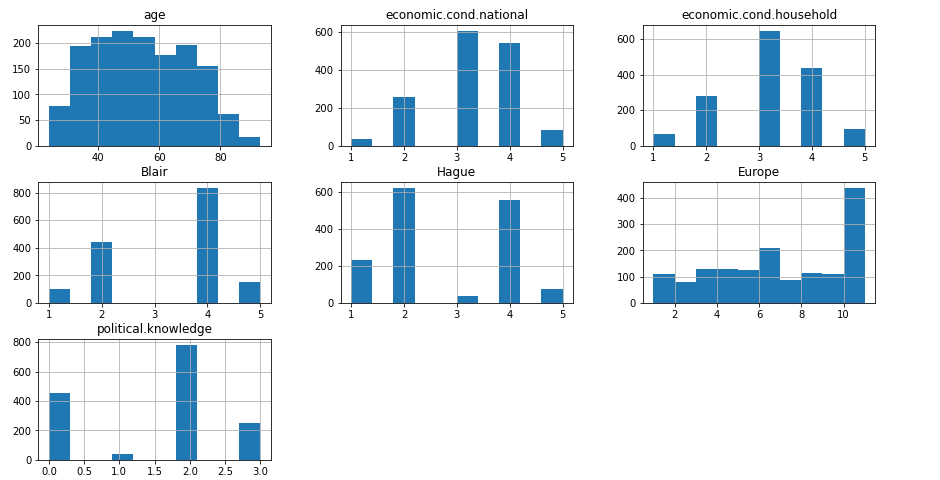




Also as seen alongside the data types is 2 object and 8 integer type values.

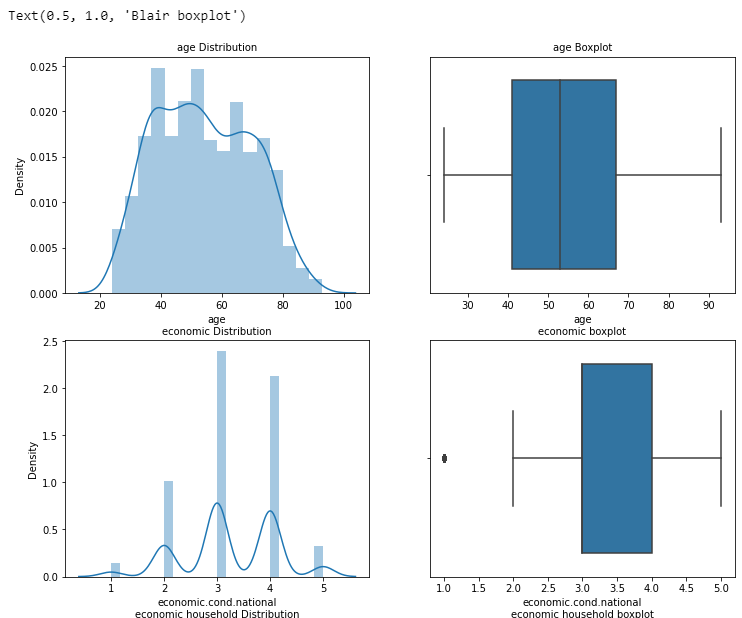
Going ahead with rows, columns in the data (1525, 9), we will perform UNIVARIATE, BIVARIATE and MULTIVARIATE ANALYSIS.

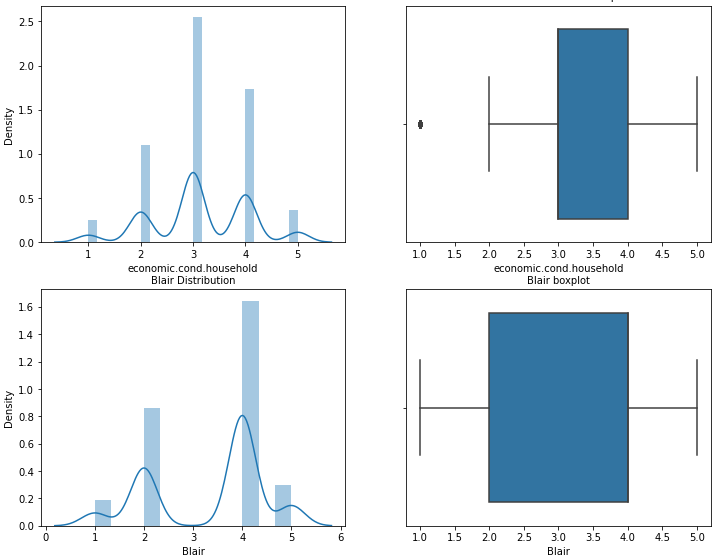
Let’s see the spread of data:



INFERENCE:

* We can see that maximum data is in between age of 40-75 years.
* For economic. Cond.national: Assessment of current national economic conditions, 1 to 5.
* UNIVARIATE ANALYSIS with Checking Outliers as well:





INFERENCE:

* As we can see