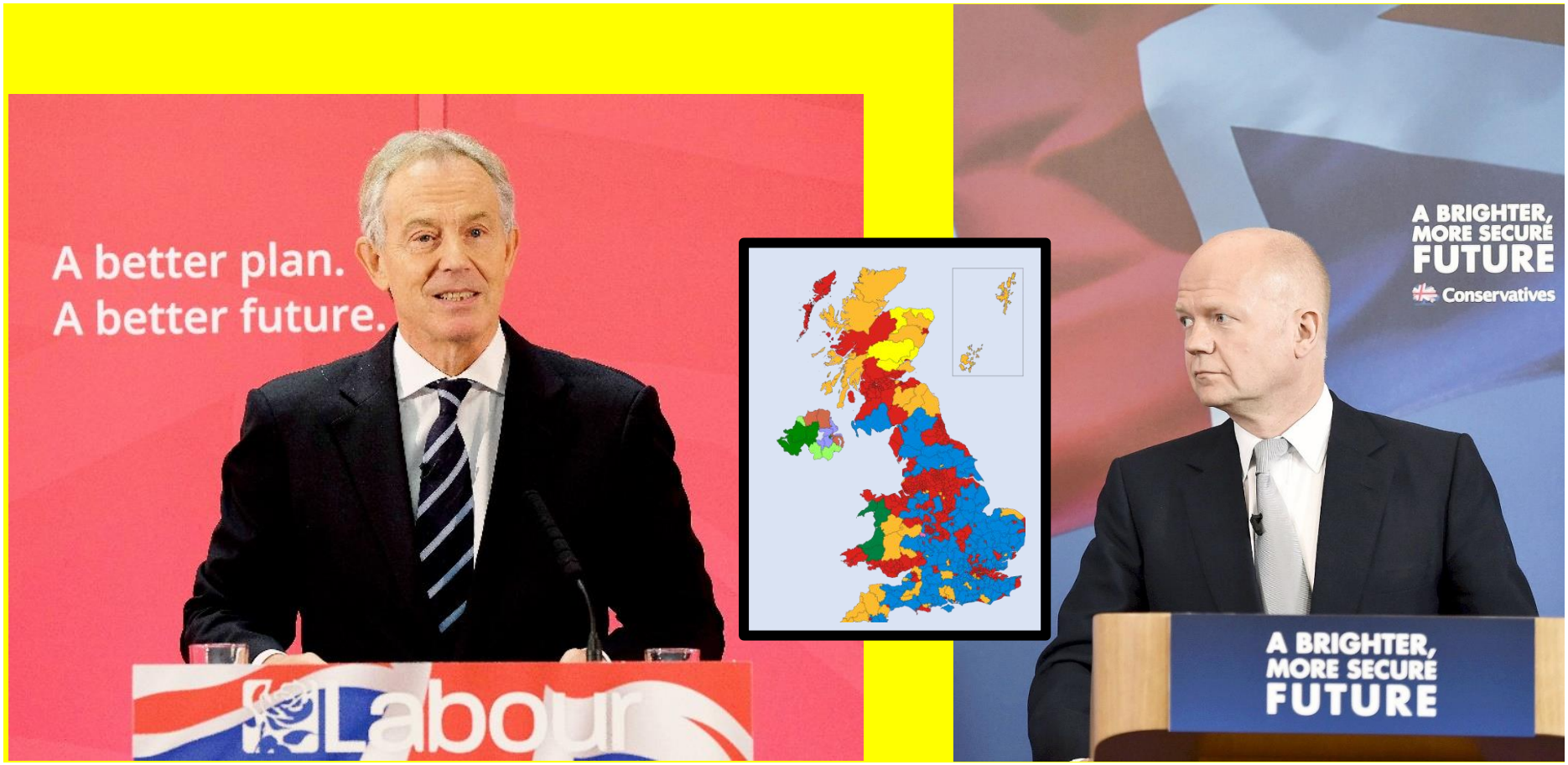
Business Report
Machine Learning
Modelling and Text Mining

UK Elections and US Inaugural Speeches

Filed by Aditya Rishi, PGP-DSBA Sept 5, 2021



Pic 1. **Tony Blair**, Labour. Photo by Ian Forsyth/Getty Images

Pic 2. **William Hague**, Conservative. Photo by Toby Melville/Getty Images

Problem 1:

Executive summary: **You are hired by one of the leading news channels, CNBE, which wants to analyse recent elections. This survey was conducted on 1,525 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.**

1.1 READ THE DATASET. DO THE DESCRIPTIVE STATISTICS AND DO THE NULL VALUE CONDITION CHECK. WRITE AN INFERENCE ON IT.

READING THE DATASET

	Unnamed: 0	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
5	6	Labour	47	3	4	4	4	4	2	male
6	7	Labour	57	2	2	4	4	11	2	male
7	8	Labour	77	3	4	4	1	1	0	male
8	9	Labour	39	3	3	4	4	11	0	female
9	10	Labour	70	3	2	5	1	11	2	male

Tab 1. UK Elections original dataset

UNDERSTANDING THE DATASET

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

Tab 2. Data dictionary

INTRODUCTION

The dataset tells us that there are **two parties, Labour and Conservative,** contesting to win this election. The other variables in the dataset contain information about the **gender and age** of the respondents, **the party they would vote for**, and how each respondent assesses the **current national economic condition** and the **current household economic condition** on a scale of 1 to 5. We have two more variables that talk about the **voters' assessment** of Labour leader **(Tony) Blair** and Conservative leader **(William) Hague**, again on a scale of 1 to 5. This sample survey or opinion poll is, thus, from the 2001 British general election, which Labour had won. Another variable called Europe measures the respondents' **attitude toward European integration** on an 11-point scale. Higher scores represent a Eurosceptic sentiment. Lastly, we come to the variable that talks about the **public's knowledge of the parties' position** on European integration, on a scale of 0 to 3. On the whole, there are 1,525 respondents and 9 variables. **Vote** (Labour or Conservative) is our **target variable**.

DESCRIPTIVE STATISTICS

```
The shape of the dataset: (1525, 10)
```

• The given data-set is of 1,525 rows of observations from an exit poll, across 9 dimensions of type integer and object.

The 10th variable, titled 'Unnamed: 0' is not useful, being just a serial number, and so we'll drop it from the dataset.

DESCRIPTIVE STATISTICS

```
RangeIndex: 1525 entries, 0 to 1524
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
dtypes: int64(7), object(2)
memory usage: 107.4+ KB
```

Tab 3. Data info

5-number summary

	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

Tab 4. 5-number summary

• Of the variables, 7 are of datatype integer64, while 2 are of datatype object. It's a small dataset that consumes just a little more than 107.4 kilobytes of computer memory. • Only the 'age' variable is numerical and continuous, while the rest of the integer-type variables are ordinal in nature, representing the voters' assessment of different factors that influence their vote. • The categorical variables 'vote' and 'gender' represent which party the respondents voted for and whether they are a man or a woman	•The average age of the voters is 54 years and the median is 53 years, which indicates a normal distribution of age. • The youngest voter is 24, which is the minimum voting age in the UK, while the eldest is 93. • An average voter's assessment of the national and household economic condition is around 3. • While all the ordinal factors showed a minimal deviation of about 1%, the factor 'Europe' shows a variance above 3%. • 50% of the voters are between the age of 41 and 67.	• The variable 'Europe' assesses the level of scepticism of a voter on influence of European union, from 1 to 11, where 11 being highly 'Eurosceptic' • While a significant group of voters (338) were highly Eurosceptic (11) an average voter was found to be ranked little above 6 • The voters seem to be less aware of political parties' position on the European integration of the UK, as the factor of 'political knowledge' shows a mean of little above 1, and maximum frequency 2, out of 3.

DESCRIPTIVE STATISTICS

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

Tab 5. Top vote and participating gender group

Null value check

```
vote                    0
age                     0
economic.cond.national  0
economic.cond.household 0
Blair                   0
Hague                   0
Age_Group               0
Europe                  0
political.knowledge     0
gender                  0
dtype: int64
```

Tab 8. Null value check

• No variables were found to be having 'null' values

```
RangeIndex: 1525 entries, 0 to 1524
Data columns (total 10 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   object
 3   economic.cond.household 1525 non-null  object
 4   Blair                  1525 non-null   object
 5   Hague                  1525 non-null   object
 6   Age_Group              1525 non-null   object
 7   Europe                 1525 non-null   object
 8   political.knowledge    1525 non-null   object
 9   gender                 1525 non-null   object
dtypes: int64(1), object(9)
memory usage: 119.3+ KB
```

Tab 6. Data info, after conversion of seven ordinal variables to categorical and addition of a variable called age group.

• The ordinal variables of integer type were converted to object type for ease of analysis and visualisation.

• **A new dimension**, 'Age_Group' was created from the continuous variable 'age'.

• Age grouping was done as follows:-

Order	Age group
1	Up to 35 years
2	36 to 50 years
3	51 to 65 years
4	66 to 80 years
5	Above 60

Tab 7. Age grouping

DESCRIPTIVE STATISTICS

	count	unique	top	freq
vote	1525	2	Labour	1063
economic.cond.national	1525	5	3	607
economic.cond.household	1525	5	3	648
Blair	1525	5	4	836
Hague	1525	5	2	624
Europe	1525	11	11	338
political.knowledge	1525	4	2	782
gender	1525	2	female	812
Age_Group	1525	5	2	479

Tab 9. Summary of data converted to categorical

• An average voter's assessment of the national and household economic condition is around 3, with 607 and 648 respondents, respectively, choosing 3.

• The voters' assessment of the leaders of the Labour and Conservative parties, Tony Blair and William Hague, respectively, draws a contrasting picture. While the majority (836) ranked Blair at 4, Hague was ranked at 2 by 624 voters.

• Based on this exit poll, the Labour party is getting a clear mandate with 1,063 votes out of 1,525 respondents

```
Number of duplicate rows = 8
```

Duplicate rows

• There were only 8 duplicate rows, which were retained

	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Age_Group	Europe	political.knowledge	gender
67	Labour	35	4	4	5	2	1	3	2	male
626	Labour	39	3	4	4	2	2	5	2	male
870	Labour	38	2	4	2	2	2	4	3	male
983	Conservative	74	4	3	2	4	4	8	2	female
1154	Conservative	53	3	4	2	2	3	6	0	female
1236	Labour	36	3	3	2	2	2	6	2	female
1244	Labour	29	4	4	4	2	1	2	2	female
1438	Labour	40	4	3	4	2	2	2	2	male

Tab 10. The eight duplicate rows in the dataset

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


ECONOMIC.COND.NATIONAL :  5
1     37
5     82
2    257
4    542
3    607
Name: economic.cond.national, dtype: int64


ECONOMIC.COND.HOUSEHOLD :  5
1     65
5     92
2    280
4    440
3    648
Name: economic.cond.household, dtype: int64
```

DESCRIPTIVE STATISTICS

```
BLAIR :  5
3      1
1     97
5    153
2    438
4    836
Name: Blair, dtype: int64


HAGUE :  5
3     37
5     73
1    233
4    558
2    624
Name: Hague, dtype: int64


AGE_GROUP :  5
5     62
1    201
4    367
3    416
2    479
Name: Age_Group, dtype: int64
```

```
EUROPE :  11
2      79
7      86
10    101
1     109
9     111
8     112
5     124
4     127
3     129
6     209
11    338
Name: Europe, dtype: int64


POLITICAL.KNOWLEDGE :  4
1     38
3    250
0    455
2    782
Name: political.knowledge, dtype: int64


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

Tabs 11, 12, 13. Value counts of variables

- Labour win 1,063 to 462
- 607 voters rank national economic condition at 3
- 648 voters rank household economic condition at 3
- 836 voters rank Blair at 4
- 624 voters rank Hague at 2
- 479 voters are in age group 2 (36 to 50)
- 338 voters rank 11 on Euroscepticism scale
- 812 women, 713 men voted

```
RangeIndex: 1525 entries, 0 to 1524
Data columns (total 10 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   object
 3   economic_cond_household 1525 non-null  object
 4   Blair                  1525 non-null   object
 5   Hague                  1525 non-null   object
 6   Age_Group              1525 non-null   object
 7   Europe                 1525 non-null   object
 8   political_knowledge    1525 non-null   object
 9   gender                 1525 non-null   object
dtypes: int64(1), object(9)
memory usage: 119.3+ KB
```

Tab 14. Renaming

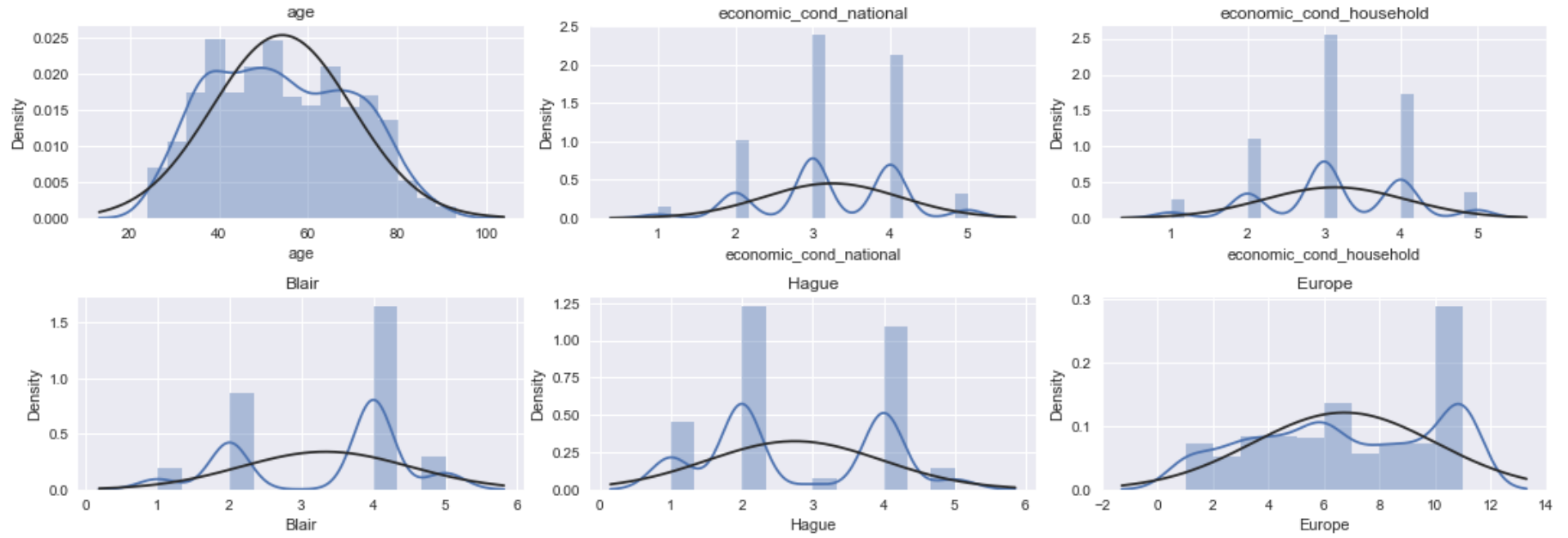
Summary of what we know

Originally, we have two string variables and the rest, except for age, are on a scale basis, so the age variable was cut into bins of age brackets and further labelled 1, 2, 3, etc. to be on a similar scale as the rest. Variables' descriptive numbers:-

- There are more female (53%) respondents than male (47%) respondents.
- Respondents lean more towards the Labour party.
- On an average, people think the economic and household conditions are stable at 3 on a scale of 1 to 5. **Some variables were renamed for easier calculation**.
- Blair has an average rating of 3.3 while Hague's rating is lower at 2.
- In general, respondents seem to be against the European integration, as majority (75%) of the respondents have given really high scores.

1.2 PERFORM UNIVARIATE AND BIVARIATE ANALYSIS. DO EXPLORATORY DATA ANALYSIS. CHECK FOR OUTLIERS.

UNIVARIATE ANALYSIS



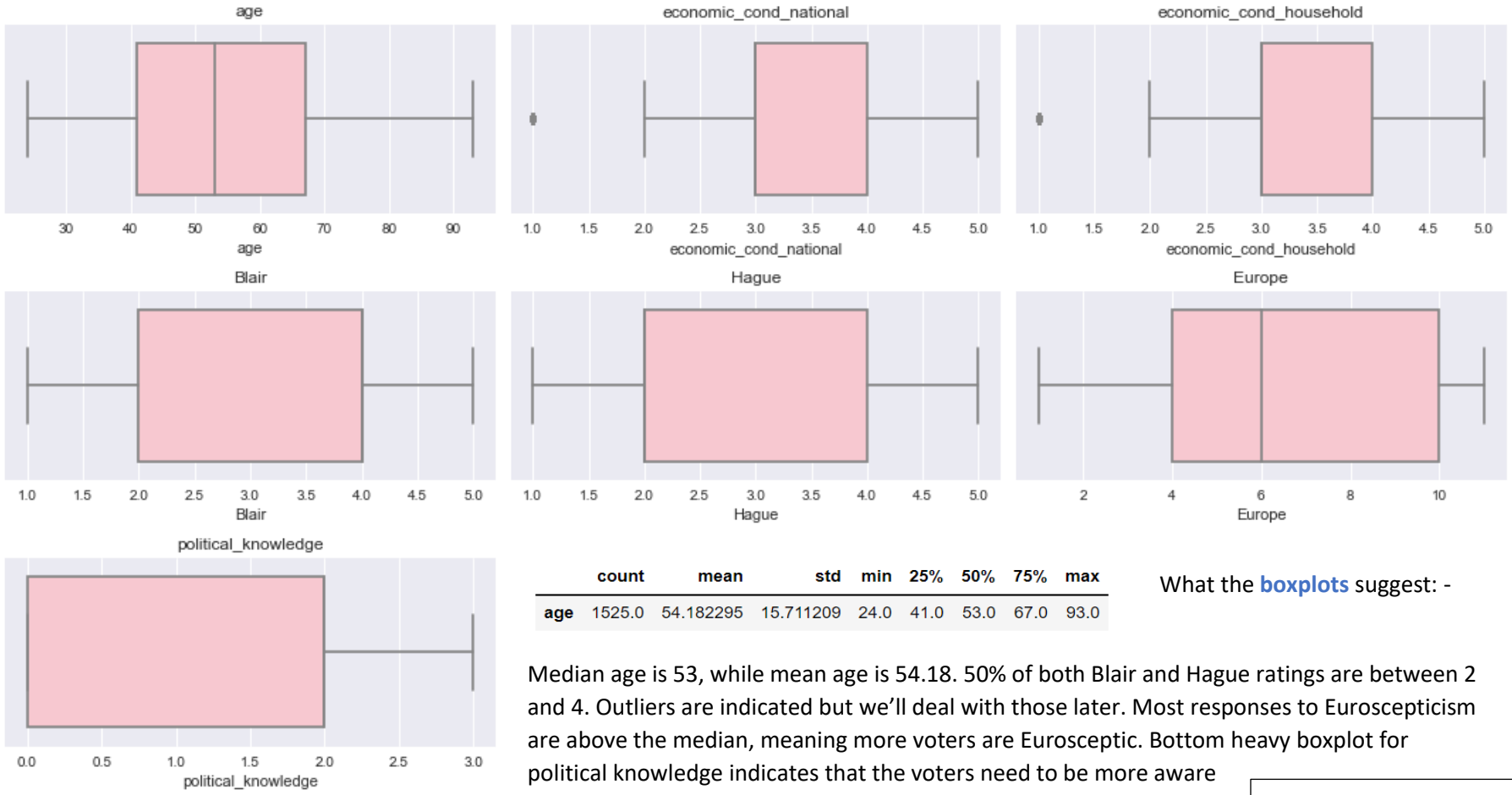
The **distribution plots** with Kernel Density Estimate line suggest the following: -

The multiple peaks in each plot indicate specific clusters. Age is multimodal & distributed normally. Hague, Europe, and Political knowledge are left skewed. Economic condition has almost normal distribution. In a left skewed distribution, the mean is less than the median. In a right skewed distribution, the mean is greater than the median. In a symmetrical or normal distribution, the mean, median, and mode are all equal. Symmetrical distributions are indicated by the black curve for reference, while the actual distribution curves are in blue. A distribution is left skewed if it has a "tail" on the left side of the distribution. A distribution is right skewed if it has a "tail" on the right side of the distribution.

Fig 1. Distribution plots

UNIVARIATE ANALYSIS

Frequency of variables (from the histograms on the distribution plots): We start by looking at the horizontal or x-axis of these histograms to see how the data in each variable is grouped. Then, we look at the vertical or y-axis to see how frequently that data occurs. While the Blair majority ranked him at 4, the Hague majority ranked him at 2. Stable economy (rank 3) has maximum ticks. The sentiment that both national and household economy is stable helps the incumbent government get re-elected, generally. Most voters were rated 2 for their political knowledge of their favourite parties' stand on European integration, and most of them are sceptic about it, suggested by highest frequency for rank 11 (the early seeds of Brexit?). People of around 40, 55, and 65 have voted the most in the opinion poll.



	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

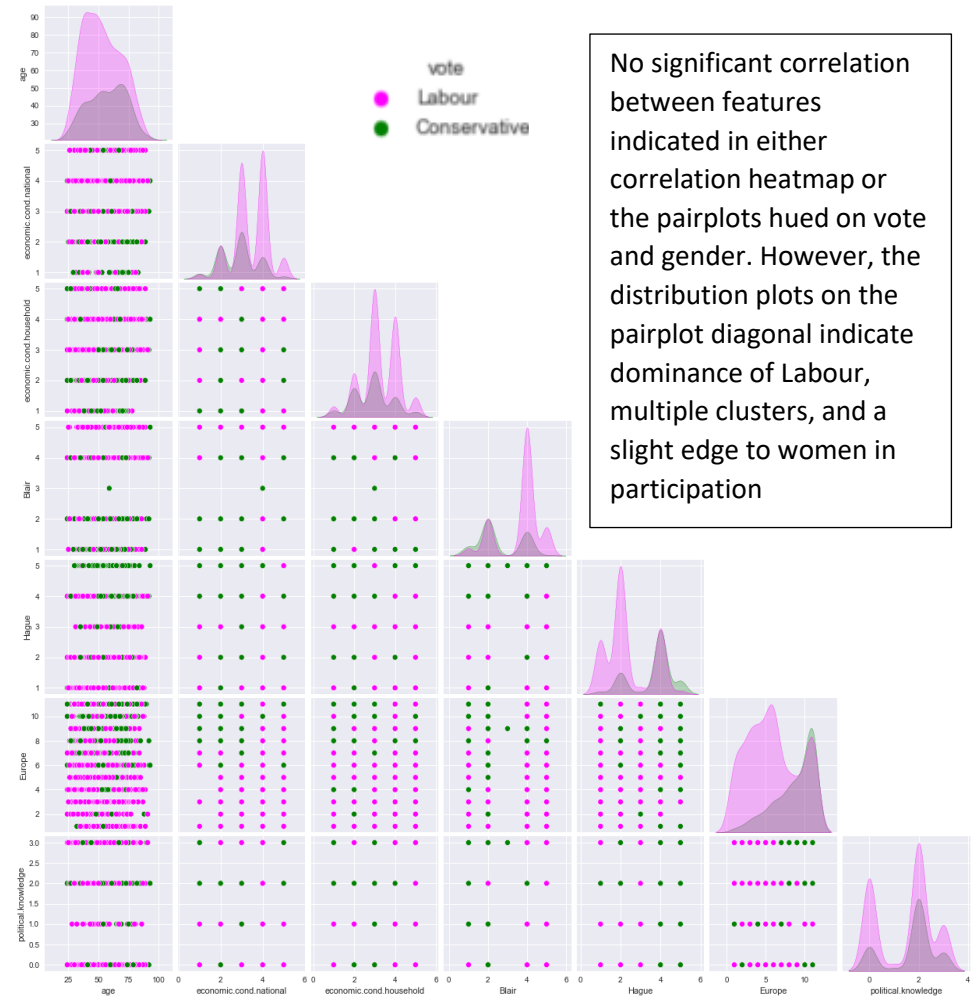
What the **boxplots** suggest: -

Median age is 53, while mean age is 54.18. 50% of both Blair and Hague ratings are between 2 and 4. Outliers are indicated but we'll deal with those later. Most responses to Euroscepticism are above the median, meaning more voters are Eurosceptic. Bottom heavy boxplot for political knowledge indicates that the voters need to be more aware

Fig 2. Boxplots

BIVARIATE ANALYSIS

PAIRPLOTS



No significant correlation between features indicated in either correlation heatmap or the pairplots hued on vote and gender. However, the distribution plots on the pairplot diagonal indicate dominance of Labour, multiple clusters, and a slight edge to women in participation

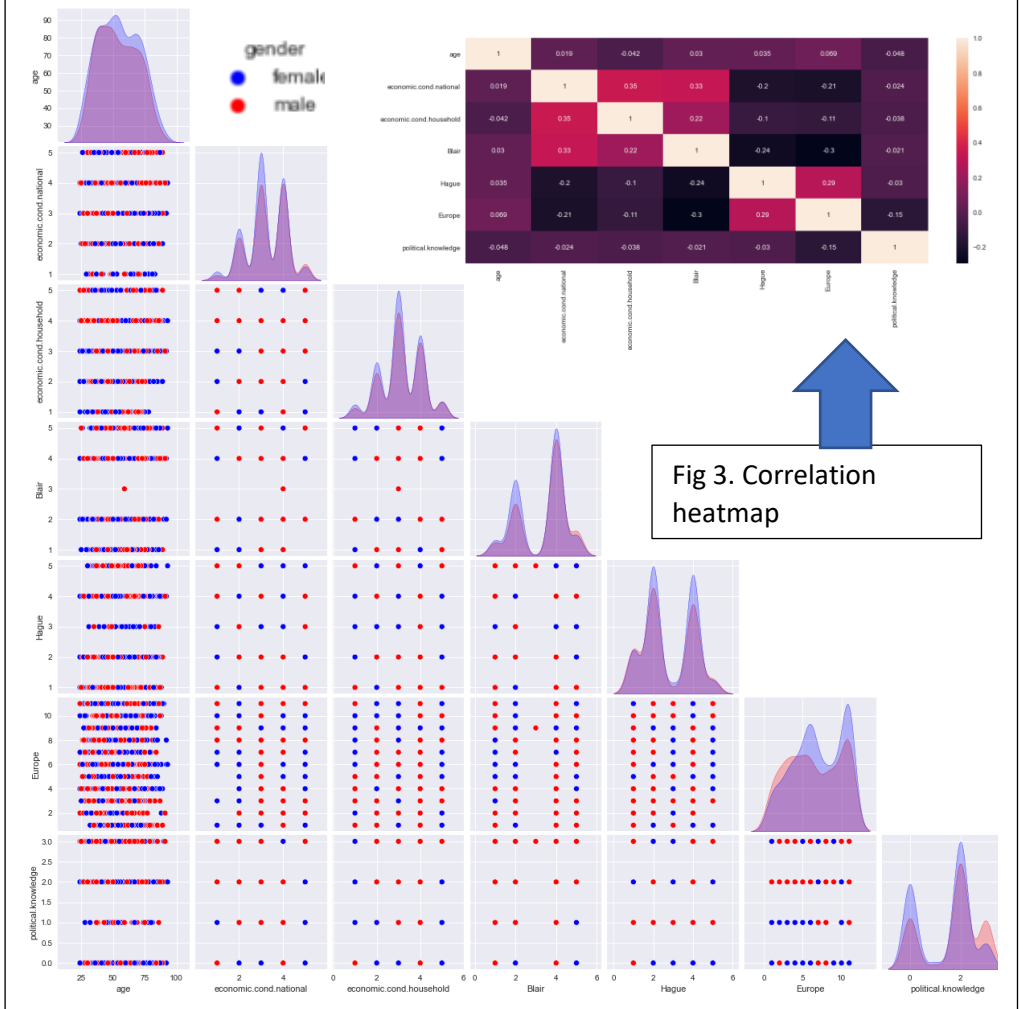


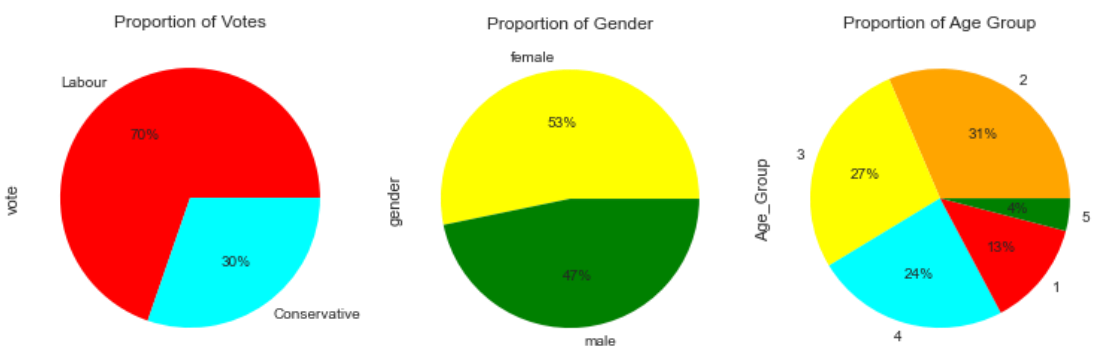
Fig 3. Correlation heatmap

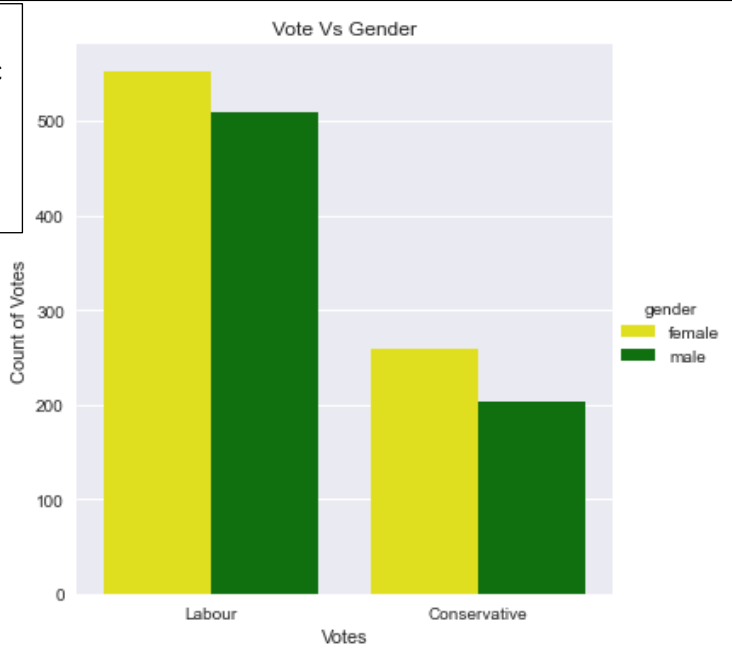
Fig 4. Pairplots hued on vote and gender

EXPLORATORY DATA ANALYSIS (DEMOGRAPHIC FACTORS)

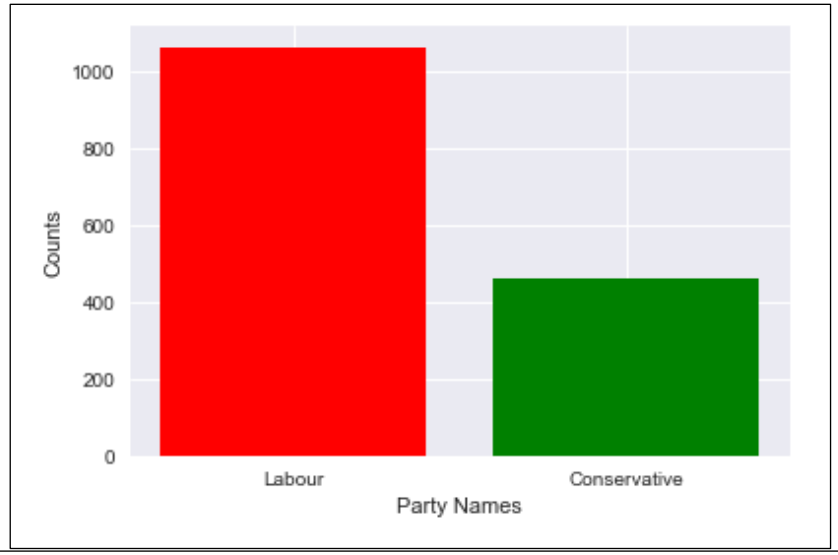
• **A clear mandate to Labour** with 70 % of the votes and a huge 40 % lead over Conservatives
•53 % of the poll respondents are female and 47 % male. Both Labour and Conservatives got **more votes from women** than men. • 57 % of the voters are above 55 years of age
• 31 % of the voters are in the age group of 36 to 50, while those between 24 and 35 are only 13%
• The 51-to-65 age group makes 27 % of the voters. The elderly voters, aged between 66 and 80, make an impressive 24% of the turnout, white those above 80 are just 4%

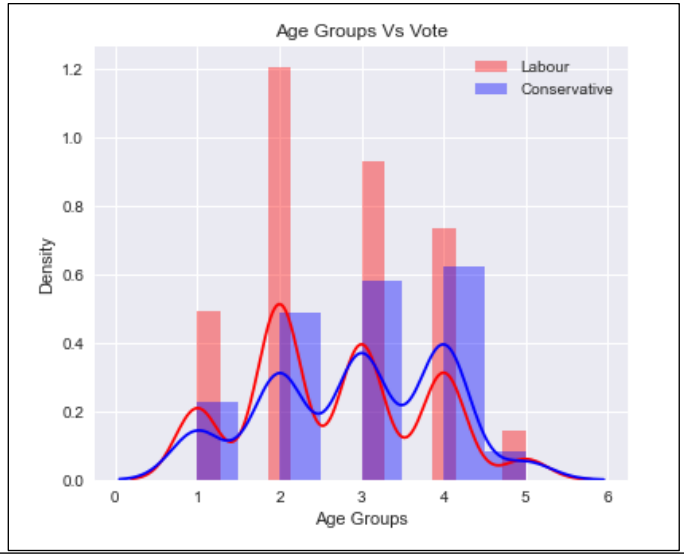
Fig 5. Demographic factors





• Labour got more votes from all age groups
• The **Conservatives are favoured by seniors**. The average age of a Conservative voter is 57
• **Younger voters**, with an average age of 53, **go with Labour**. Those between 36 and 50 makes the pillar stone of Labour's mandate
• 58 % of all voters are aged between 36 and 65 and are **key deciders** of the electoral fortunes





EXPLORATORY DATA ANALYSIS (ECONOMIC FACTORS)

• The features representing voters' assessment of the national and household economic condition gives us an insight into the economic situation of the election time

• 75 % of the Labour voters ranked the national economic condition 3 or above on a scale 1 to 5, 5 being the highest rank

• 68 % of the Conservative voters also ranked the national economic condition 3 or above. Only 35 % of the Tory voters believe the condition is bad

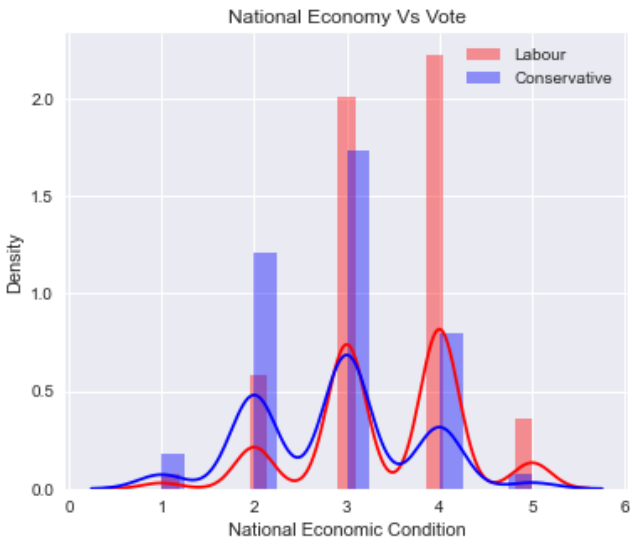
• 81 % of the Labour voters have ranked the household economic condition 3 or above on a scale of 1 to 5

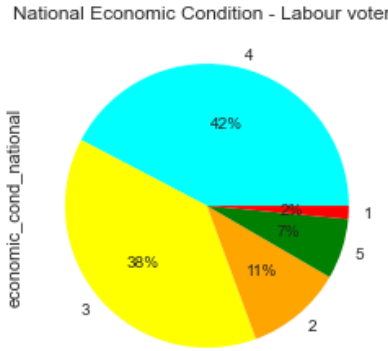
• 67 % of the Conservative voters also ranked the household economic condition 3 or above and only 33 % ranked it below 3

• 81% and 77 % of all the voters have ranked the national and household economic condition, respectively, 3 or above, showing a robust economy

• From the distribution of the economic rankings, it can be inferred that those who ranked both the national and household conditions high (4 or 5) are more likely to vote for Labour

• Those ranked the national and household economic condition to be lower than 3 are more likely to vote for the Conservatives





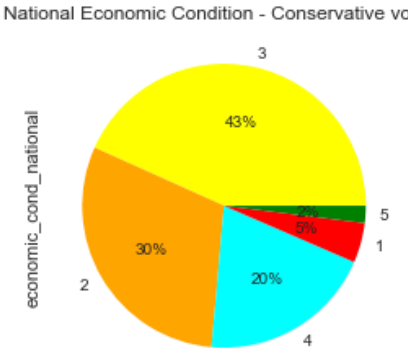
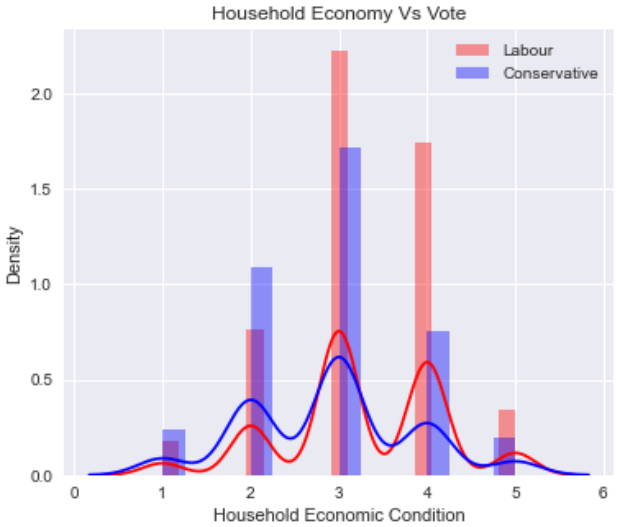
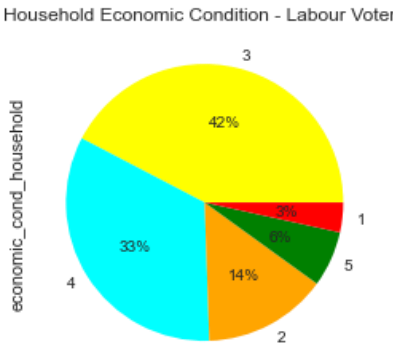


Fig 6. National economy and election





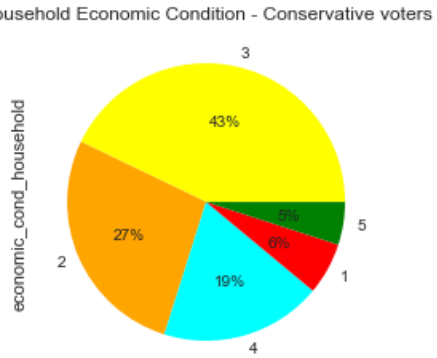


Fig 7. Household economy and election

EDA

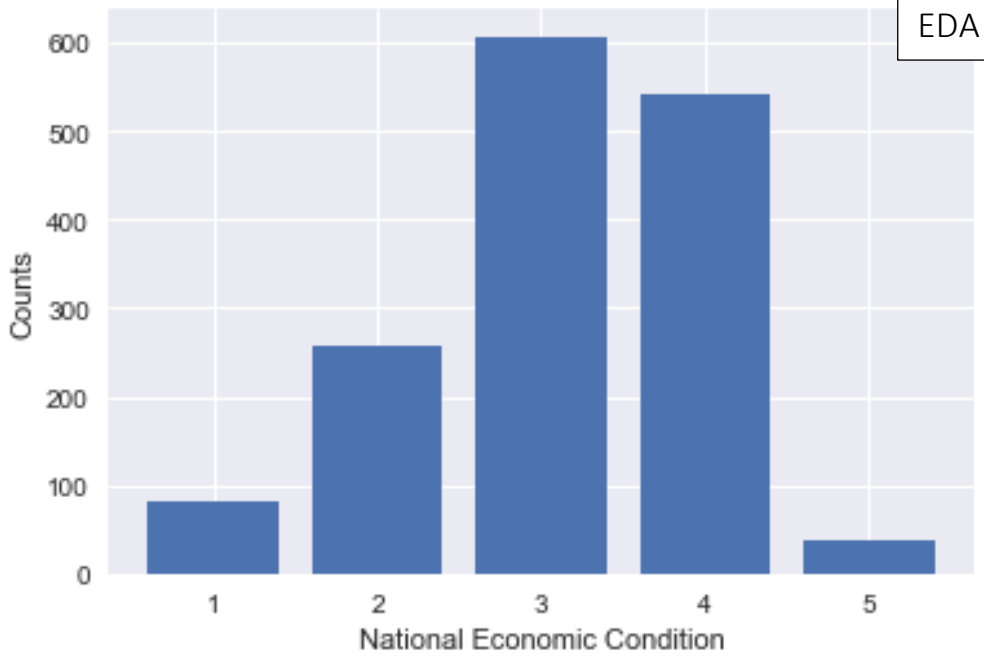


Fig 8. State of national economy
National economic condition for most people belongs to rank 3

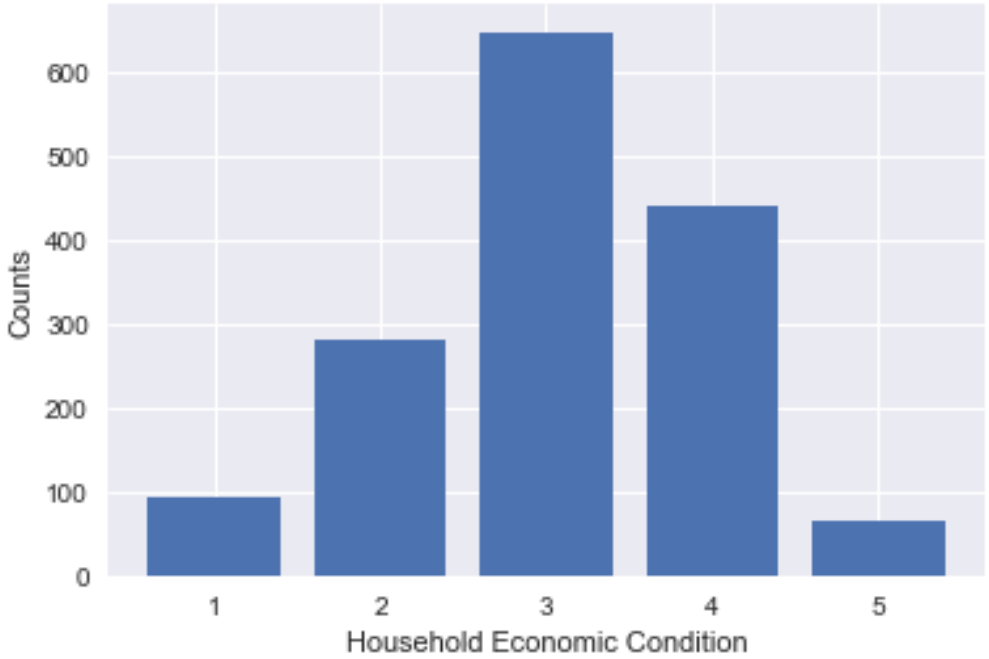
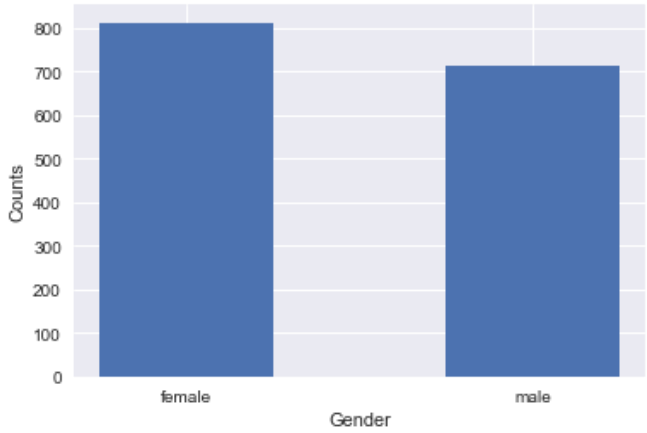
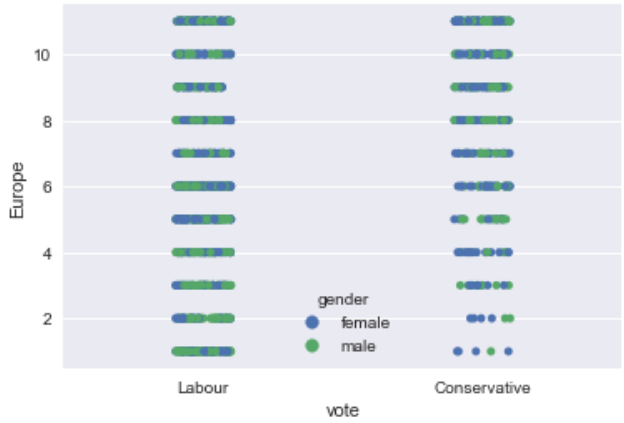


Fig 9. State of household economy
Household economic condition for most people belongs to rank 3





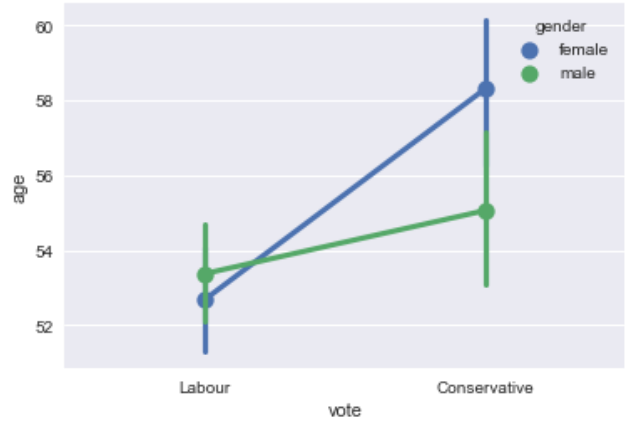


Fig 10. Gender and vote. More women have voted than men. Euroscepticism is decent in both genders, across preferences. The average age of those who voted for Tories is higher for women than men. In contrast, the average age of women and men is close in case of those who voted for Labour.

EDA (TONY BLAIR VERSUS WILLIAM HAGUE)

- The features 'Blair' and 'Hague' represent voters' assessment of the Labour and Conservative party leaders and prime ministerial candidates of the 2001 UK general elections, Tony Blair and William Hague, respectively
- Irrespective of party affiliations, 65 % of voters have rated Blair above 3
- 40 % of the Tory voters and 78 % of Labour voters rated Blair above 3
- 65 % of Conservative voters and 22 % of Labour voters rated Blair's performance as below average
- 55% of all the voters assessed Hague below 3, which shows most of them didn't approve his performance as Conservative leader
- While 71 % of Labour voters rated Hague below 3, only 28% rated him 3 or above
- 75% Tory voters rated Hague above 3, while25 % rated him 3 and below
- Plotting assessment of leaders against age distribution of voters reveals that except among the seniors above 87, Blair is rated above Hague in all age segments
- Blair is rated high among youth and those between 75 and 85, while Hague is rated lowest in 40-to-50 age group

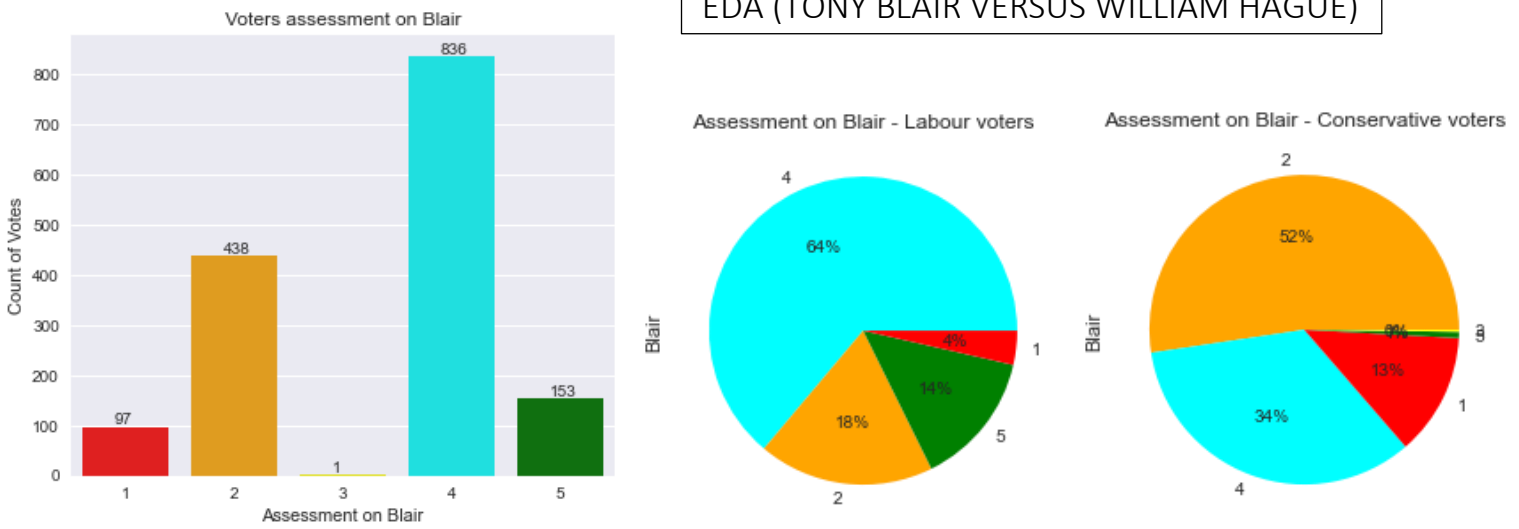


Fig 11. Opinion about Blair

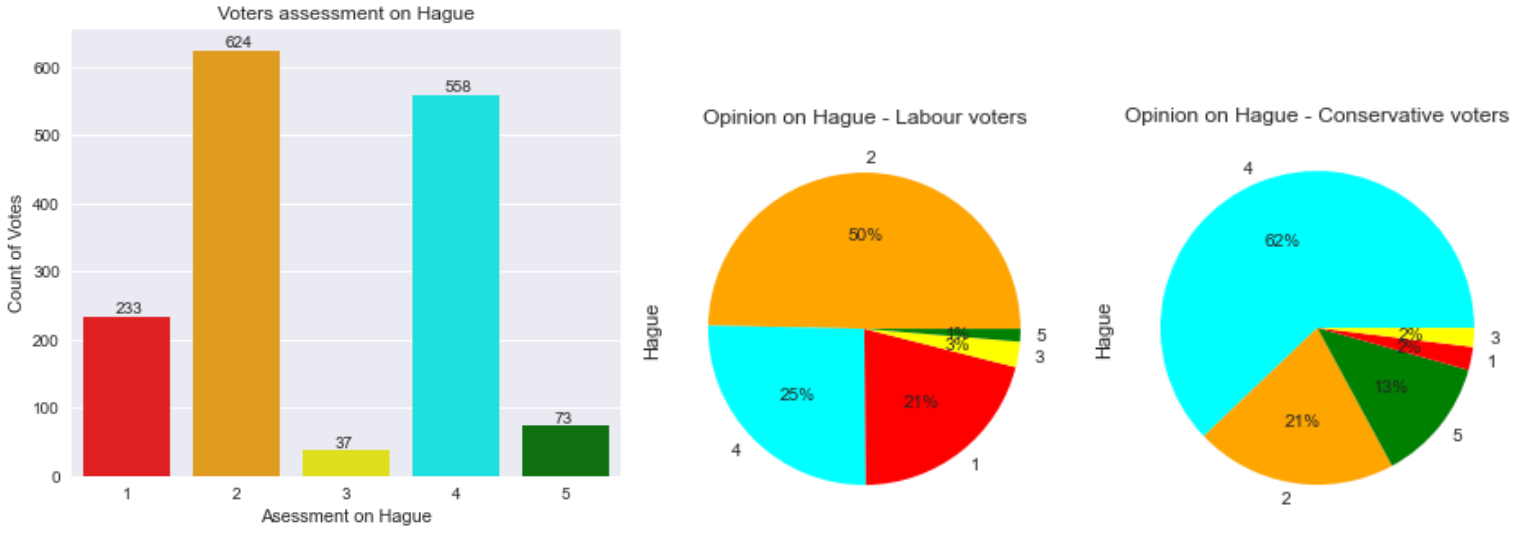


Fig 12. Opinion about Hague

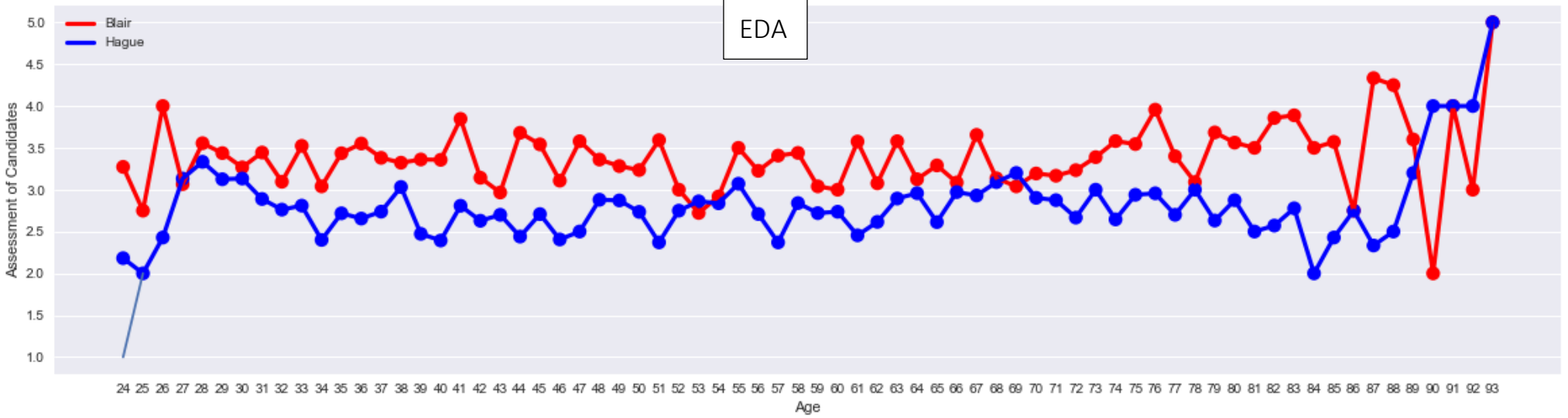


Fig 13. Age of voters and the assessment of candidates (average rating on scale of 1 to 5 for specific age)

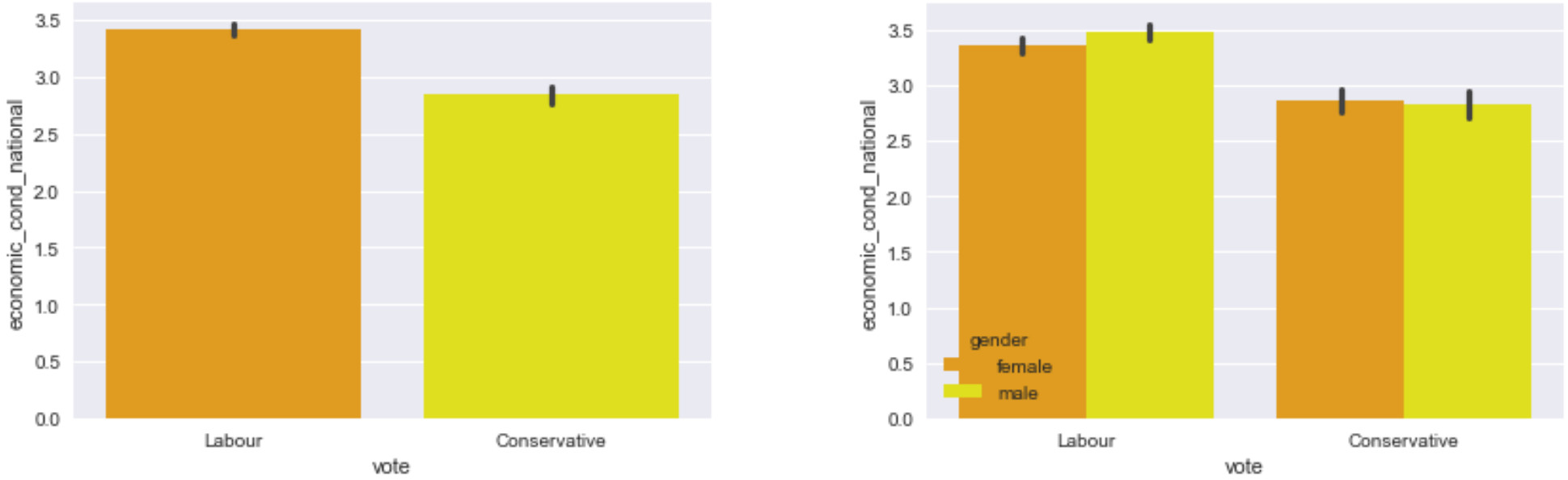
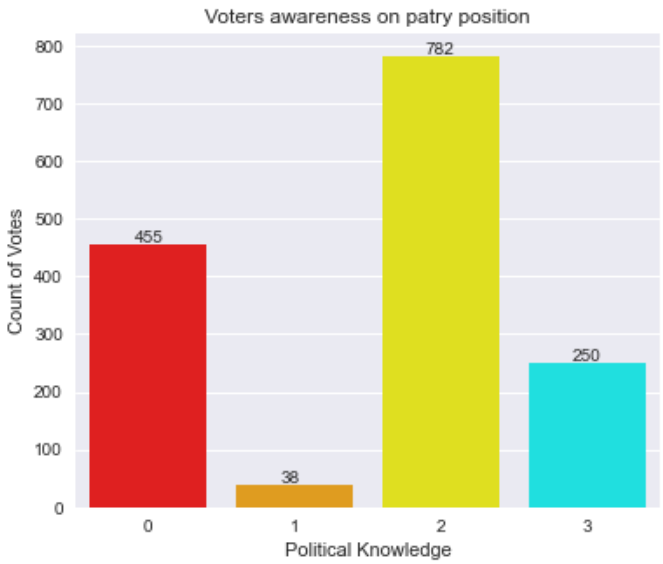


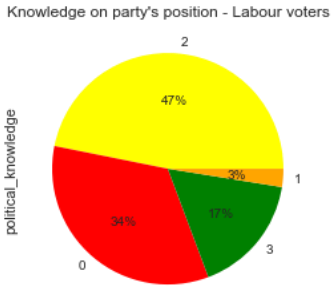
Fig 14. Vote and enconomy. Those who have voted for Labour have a higher assessment of national economic condition. No such clear distinction between genders can be made for this, however.

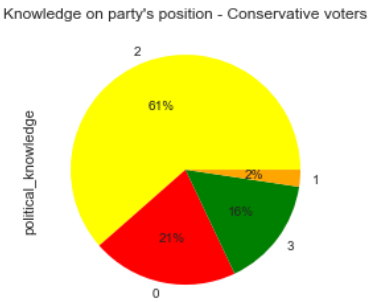
EDA (EUROPEAN INTEGRATION)

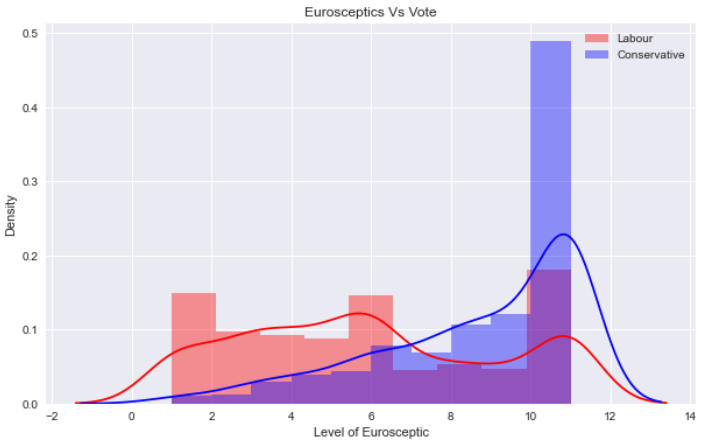
Fig 15. Euroscepticism of voters, genders, age groups

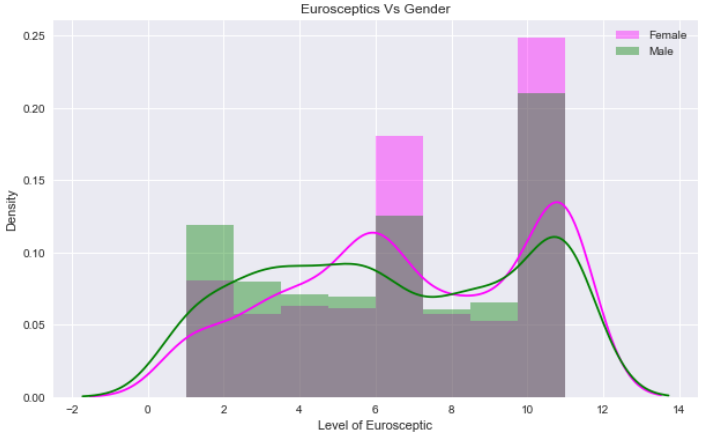
- The increasing influence of the European Union and UK's integration with it is the 2001 poll issue
- The exit poll recorded the voters' knowledge of the political parties' position on European integration. 67% ranked their awareness at 2 or above on a scale of 0 to 3
- Only 32 % of the voters consider self to be unaware of the position of Labour and the Conservatives on European integration
- While 68% of the Labour voters were confident of their awareness, 78% of Tory voters rated themselves high
- The factor 'Europe' gauged the voters' "Euroscepticism" on a scale of 0 to 11, where 11 indicates the person is highly sceptic of the increasing influence of the European Union
- The Conservative voters appears to be highly Eurosceptic than the Labour voters. Average Tory voter is above 8 on this scale, while an average Labour voter is a little below 6
- The women Labour voters are more Eurosceptic than the men, while the Tory voters of both genders are same higher level of sceptics
- Highly Eurosceptic voters are more likely to vote for Tories, while a Labour voter is more likely to be less Eurosceptic
- Young voters appear to be less concerned about the issue than the seniors, among the Conservative voters, especially
- European integration and increasing influence of EU aren't a poll issue for the Labour voters, whereas for the Tory voters, these are a major consideration

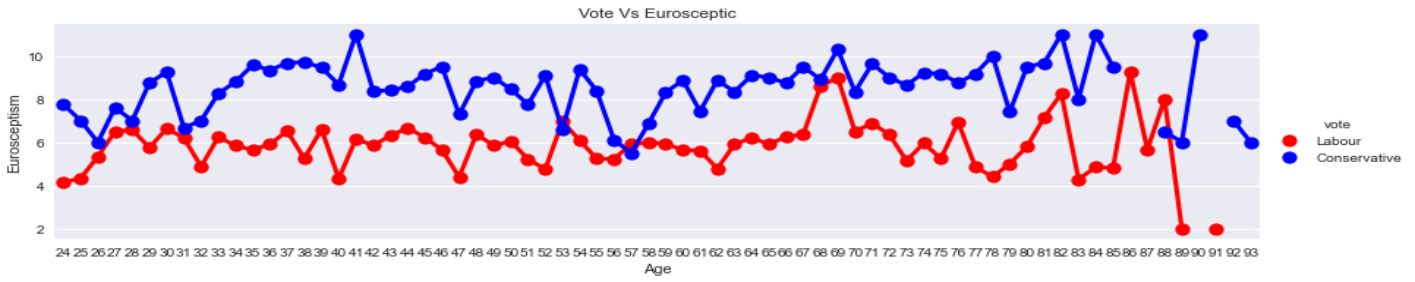












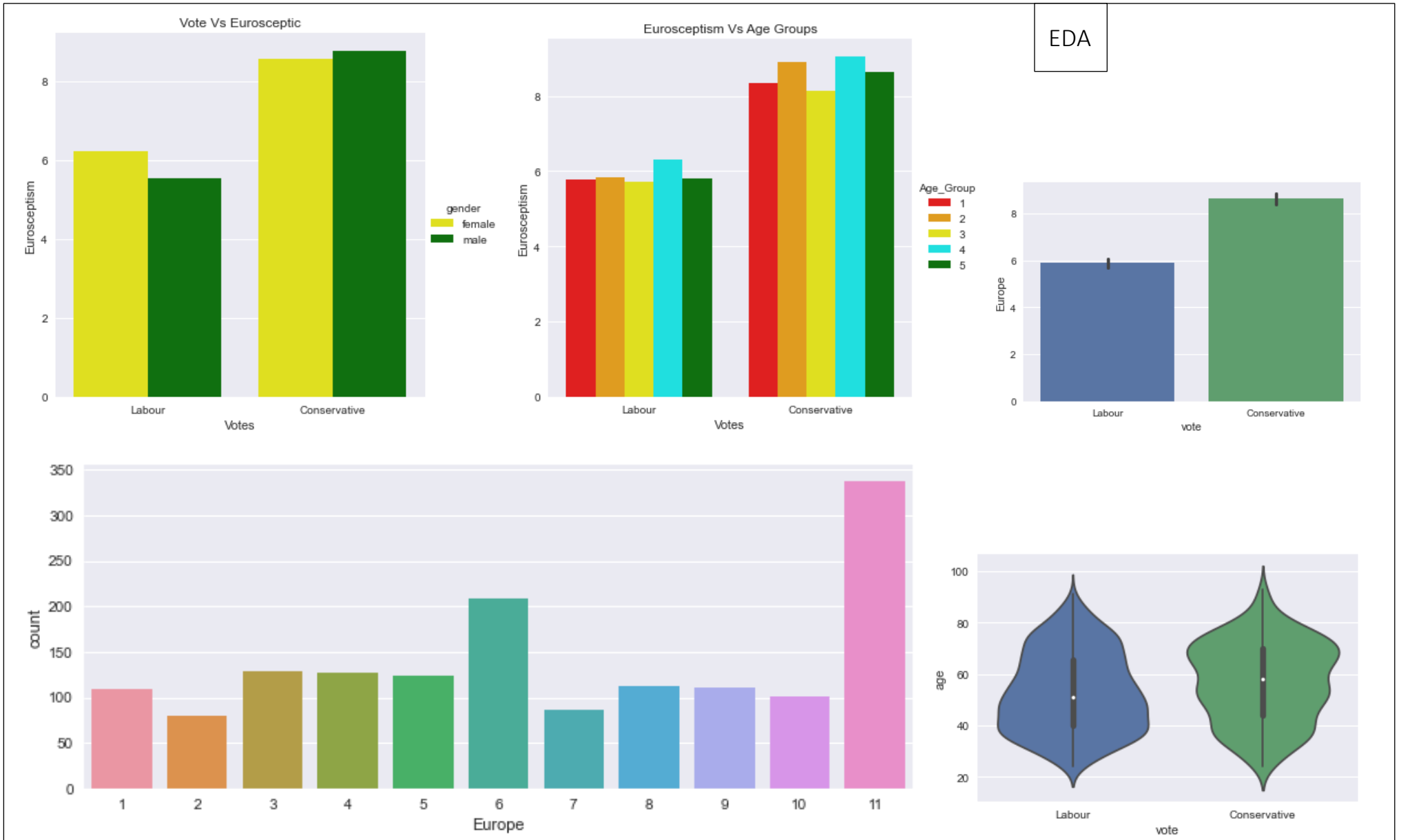


Fig 16. Euroscepticism and age distrubution of the voters. Number of counts corresponding to low ranks is highest, which shows that most voters are in support of European integration. Looking at the width of violin plot, those who have voted for Labour, most are aged 40 to 50, while the Tory voters are aged mostly 60 to 70.

EDA SUMMARY & OUTLIER CHECK

• The age distribution of voters follows a normal distribution with mean of 54 years and median of 53 years. There are no actual outliers

• Those who rated national and household economic condition below 2 are an exception, thus reflected as outliers

From the boxplot, we can see that national and household economic condition columns have some outliers. But we know that these two columns are ordinal, so we will not treat those outliers.

• A voter between 30 and 50 is more likely to vote for Labour, whereas a voter between 65 and 75 is more likely to be a Tory voter

• A voter below 30 is not as decisive as other age groups, is more unlikely to vote

• A voter who rated the economic conditions to be lower is more likely to vote for Tories, where high-rating givers may vote for Labour

• Blair got higher acceptance among both Labour and Conservative voters, whereas Hague got poor ratings among the Labour

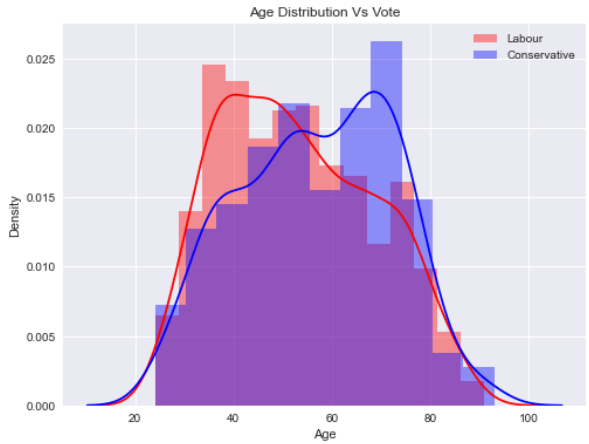


Fig 17. Age distribution versus vote

• Hague got higher acceptance from the Conservatives, understandably so

• A highly Eurosceptic voter is highly probable to vote for Conservatives, whereas a lesser Eurosceptic voter is more likely to vote for Labour

• For a Labour voter, a flourishing economy and acceptance of Blair as Prime Minister are a big motivation to vote for Labour

• Concerns about EU's increasing influence are a major consideration for a Tory voter

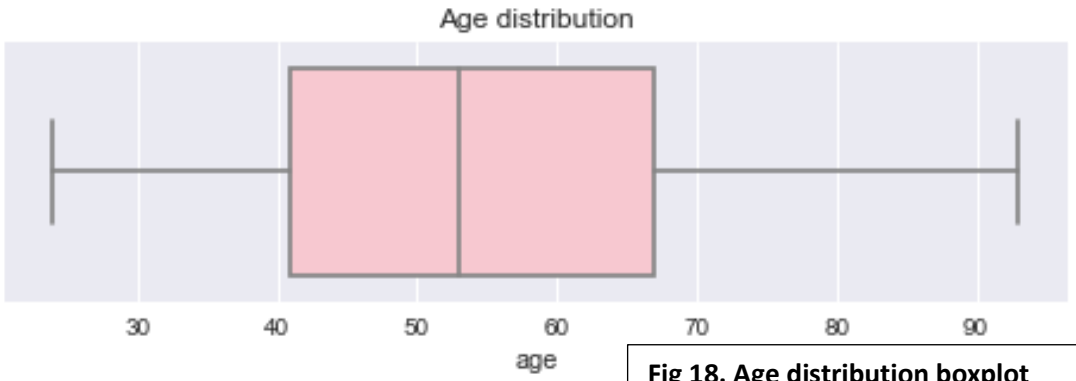


Fig 18. Age distribution boxplot

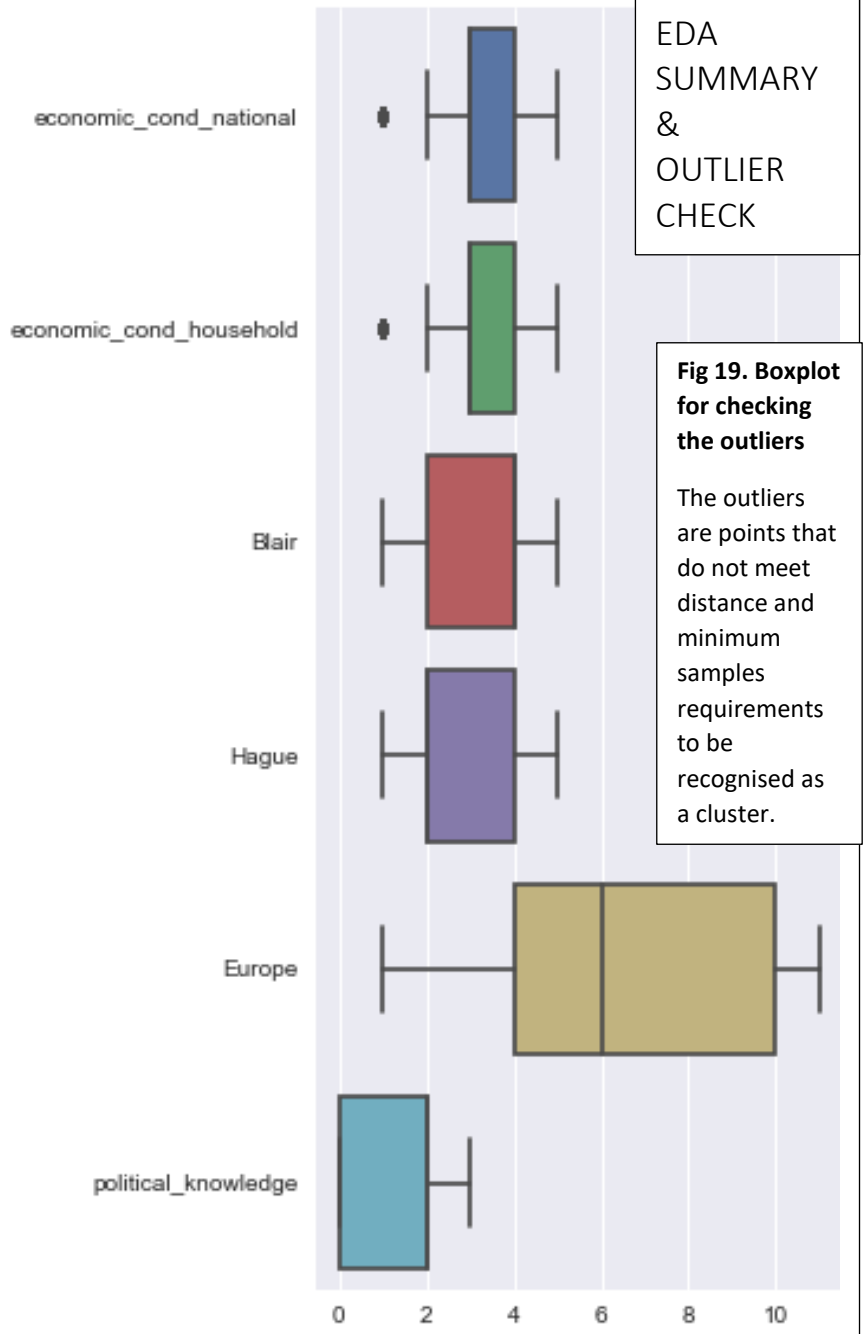


Fig 19. Boxplot for checking the outliers

The outliers are points that do not meet distance and minimum samples requirements to be recognised as a cluster.

1.3 ENCODE THE DATA (HAVING STRING VALUES) FOR MODELLING. IS SCALING NECESSARY HERE OR NOT? DATA SPLIT: SPLIT THE DATA INTO TRAIN AND TEST (70:30).

DATA PRE-PROCESSING

BINNING

• The only numerical variable in the data set 'age' is binned to create the categoric variable 'age_group' to classify the voters into 5 different groups in increasing order of their age

• The variable age is dropped after binning the values into a categoric variable, as keeping both the variables will bring collinearity and redundancy between these variables

Order	Age group
1	Up to 35 years
2	36 to 50 years
3	51 to 65 years
4	66 to 80 years
5	Above 60

```
AGE_GROUP :  5
5     62
1    201
4    367
3    416
2    479
Name: Age_Group, dtype: int64
```

Tab 15. Binning of age

DATA ENCODING

• One hot encoding method is used to encode the categoric variables in the dataset in order to improve the explainability of the models using feature importance and coefficients

• Except 'gender', rest of the categoric variables are encoded with drop_first = False to retain all the categories in the variables including the first category, so that all categories are explainable in the feature set

• The final dataset, including the target variable, has 42 features

	vote	age	economic_cond_national	economic_cond_household	Blair	Hague	Age_Group	Europe	political_knowledge	gender
0	1	43	3	3	4	1	2	2	2	female
1	1	36	4	4	4	4	2	5	2	male
2	1	35	4	4	5	2	1	3	2	male
3	1	24	4	2	2	1	1	4	0	female
4	1	41	2	2	1	1	2	6	2	male

Tab 16, 17. Encoding vote into 1 for Labour, 2 for Conservatives

```
Data columns (total 10 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   vote                    1525 non-null   int8
 1   age                     1525 non-null   int64
 2   economic_cond_national  1525 non-null   object
 3   economic_cond_household 1525 non-null   object
 4   Blair                   1525 non-null   object
 5   Hague                   1525 non-null   object
 6   Age_Group               1525 non-null   object
 7   Europe                  1525 non-null   object
 8   political_knowledge     1525 non-null   object
 9   gender                  1525 non-null   object
dtypes: int64(1), int8(1), object(8)
memory usage: 108.8+ KB
```

SCALING

• Scaling is not required in this business case as there are no numerical variables in different scales present in the dataset after dropping 'age' from the dataset after binning and one-hot encoding

• Binning and one-hot encoding has ensured that all factors are on uniform scale and all models are trained using the same data Data Splitting

• Data has been split into Train and Test sets in 70:30 ratio, with random_state=27

vote	Age_Group_1	Age_Group_2	Age_Group_3	Age_Group_4	Age_Group_5
1	0	1	0	0	0
1	0	1	0	0	0
1	1	0	0	0	0
1	1	0	0	0	0
1	0	1	0	0	0

economic_cond_household_1	economic_cond_household_2	economic_cond_household_3	economic_cond_household_4	...	Europe_8
0	0	1	0	...	0
0	0	0	1	...	0
0	0	0	1	...	0
0	1	0	0	...	0
0	1	0	0	...	0

Europe_9	Europe_10	Europe_11	Hague_1	Hague_2	Hague_3	Hague_4	Hague_5	gender_male
0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	1	0	1
0	0	0	0	1	0	0	0	1
0	0	0	1	0	0	0	0	0
0	0	0	1	0	0	0	0	1

Tab 18. Final encoded dataset for predictive modelling

Data split

```
Shape of original dataset : (1525, 42)
Shape of input - training set (1067, 41)
Shape of output - training set (1067,)
Shape of input - testing set (458, 41)
Shape of output - testing set (458,)
Training dataset after split: 70 %
Test dataset after split: 30 %
```

- After split, the training set has 1067 records, while the test set has 458 records.
- Training set has (1067/1525)*100 = 70% records
- Test set has (458/1525)*100 = 30% records.
- Hence, the split is as desired

× 42 columns

Tab 19. Data columns of the final encoded set —>

ENCODING

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1525 entries, 0 to 1524
Data columns (total 42 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   vote                      1525 non-null   int8
 1   Age_Group_1               1525 non-null   uint8
 2   Age_Group_2               1525 non-null   uint8
 3   Age_Group_3               1525 non-null   uint8
 4   Age_Group_4               1525 non-null   uint8
 5   Age_Group_5               1525 non-null   uint8
 6   economic_cond_household_1 1525 non-null   uint8
 7   economic_cond_household_2 1525 non-null   uint8
 8   economic_cond_household_3 1525 non-null   uint8
 9   economic_cond_household_4 1525 non-null   uint8
 10  economic_cond_household_5 1525 non-null   uint8
 11  Blair_1                   1525 non-null   uint8
 12  Blair_2                   1525 non-null   uint8
 13  Blair_3                   1525 non-null   uint8
 14  Blair_4                   1525 non-null   uint8
 15  Blair_5                   1525 non-null   uint8
 16  political_knowledge_0     1525 non-null   uint8
 17  political_knowledge_1     1525 non-null   uint8
 18  political_knowledge_2     1525 non-null   uint8
 19  political_knowledge_3     1525 non-null   uint8
 20  economic_cond_national_1  1525 non-null   uint8
 21  economic_cond_national_2  1525 non-null   uint8
 22  economic_cond_national_3  1525 non-null   uint8
 23  economic_cond_national_4  1525 non-null   uint8
 24  economic_cond_national_5  1525 non-null   uint8
 25  Europe_1                  1525 non-null   uint8
 26  Europe_2                  1525 non-null   uint8
 27  Europe_3                  1525 non-null   uint8
 28  Europe_4                  1525 non-null   uint8
 29  Europe_5                  1525 non-null   uint8
 30  Europe_6                  1525 non-null   uint8
 31  Europe_7                  1525 non-null   uint8
 32  Europe_8                  1525 non-null   uint8
 33  Europe_9                  1525 non-null   uint8
 34  Europe_10                 1525 non-null   uint8
 35  Europe_11                 1525 non-null   uint8
 36  Hague_1                   1525 non-null   uint8
 37  Hague_2                   1525 non-null   uint8
 38  Hague_3                   1525 non-null   uint8
 39  Hague_4                   1525 non-null   uint8
 40  Hague_5                   1525 non-null   uint8
 41  gender_male               1525 non-null   uint8
dtypes: int8(1), uint8(41)
memory usage: 62.7 KB
```

1.4 APPLY LOGISTIC REGRESSION AND LDA (LINEAR DISCRIMINANT ANALYSIS).

• The final model selected after all the iterations of model tuning is as follows

• The minority class (Conservative) and the majority class (Labour) is in the ratio 1:2.3, which is used to set the class_weight parameter in Logistic Regression

• The function logspace() is used to find the regularisation parameter 'C' which returned a value on log scale

• The GridsearchCV() returned hyperparameters for 'penalty' and 'solver' algorithm, L2 and newton-cg, respectively

LOGISTIC REGRESSION MODEL

```
LogisticRegression(C=0.0379269019073225, class_weight={0: 2, 1: 1},
                   solver='newton-cg')
```

• The proportional values to be applied for class weight parameter is finalised after trial-and-error iterations, where the ideal performance matrices were evaluated for both minority and majority classes. The feature coefficients are plotted to generate insights based on feature ranking

• The resulting model is a right fit one with acceptable score of recall of the minority class

• GridsearchCV() was applied on the LDA model to identify the right solver algorithm and tolerance value

• The default algorithm SVD and the default tolerance value was returned as the ideal hyperparameter

LDA MODEL

```
{'solver': 'svd', 'tol': 0.0001}
```

• Thus, both the first and the final iteration returned the same set of performance matrices

• The model scores are found to be right fit, but the recall of the minority class is lower. The details of model tuning and overall model selection criteria are discussed later

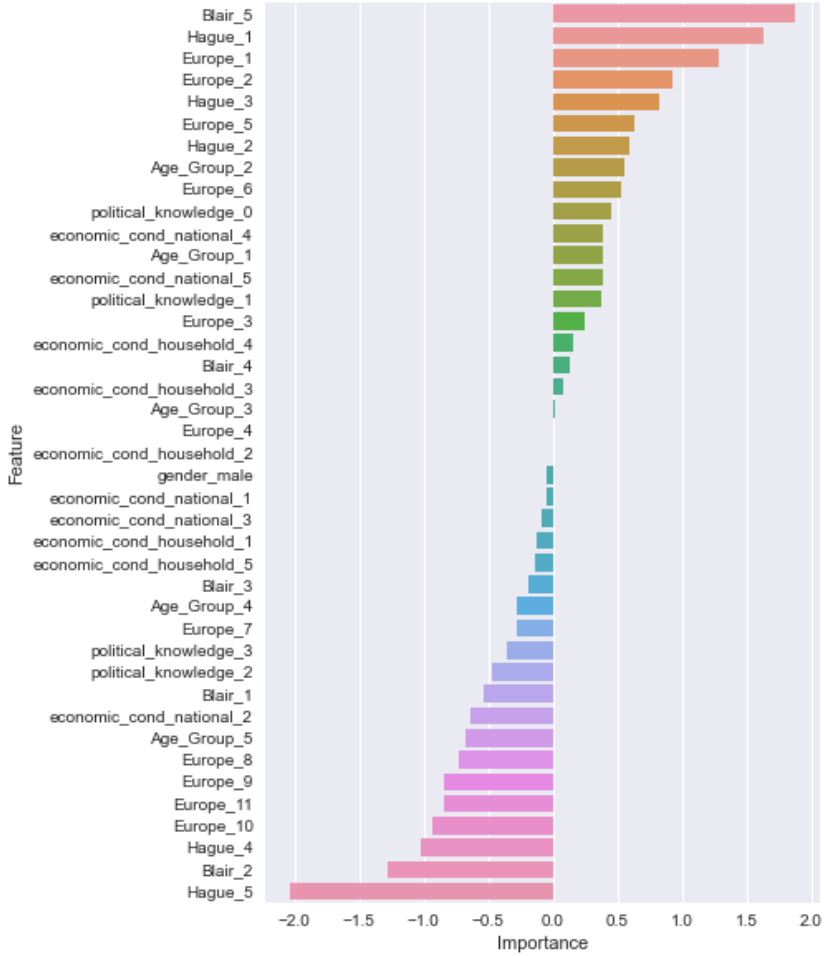


Fig 20. Feature importance from logistic regression model

1.5 APPLY KNN MODEL AND NAÏVE BAYES MODEL. INTERPRET THE RESULTS.

For tuning the hyperparameters of the KNN model GridsearchCV() was applied to find ideal values for n_neighbours, weights and matric parameters

• Though 'distance' was found to be the ideal value for the weight function, the model found to be highly overfitting with the model capturing almost 100% of the variance in train

• Hence 'uniform' weight is used to create a right fit model and default 'minkowski' metric is used for distance measure. Only n_neighbors value is optimised using GridsearchCV()

• The model was found to be right fit, but the recall of minority class was lowest

• For building a Naïve Bayes model, Multinomial Naïve Bayes algorithm is used. The reason being that the dataset contain only nominal factors and it meets the basic assumption of Naïve Bayes that all the factors are non-colinear

• Neither GridsearchCV() nor any further tuning was applied on the model as the function does not have too many hyperparameters to be optimised

• The resulting model was found to be right fit, with acceptable recall rate of both classes

K-NEAREST NEIGHBORS MODEL

```
{'metric': 'minkowski', 'n_neighbors': 19, 'weights': 'uniform'}
```

Neighbors-based classification is a type of instance-based learning or non-generalizing learning: it does not attempt to construct a general internal model, but simply stores instances of the training data. Classification is computed from a simple majority vote of the nearest neighbors of each point: a query point is assigned the data class which has the most representatives within the nearest neighbors of the point.

source: scikit-learn

MNB probabilities

	0	1
35	0.835208	0.164792
335	0.968165	0.031835
16	0.801727	0.198273
260	0.929230	0.070770
447	0.095340	0.904660
310	0.029035	0.970965
395	0.084160	0.915840
326	0.579172	0.420828
227	0.050440	0.949560
171	0.932719	0.067281

MULTINOMIAL NAIVE BAYES (MNB) MODEL

```
MultinomialNB()
```

Choosing Multinomial Naive Bayes as the data is multinomial in nature

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable. Bayes' theorem states the following relationship, given class variable y and dependent feature vector x_1 through x_n

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots x_n \mid y)}{P(x_1, \ldots, x_n)}$$

Advantages:

In spite of their apparently over-simplified assumptions, naive Bayes classifiers have worked quite well in many real-world situations, famously document classification and spam filtering. They require a small amount of training data to estimate the necessary parameters.

Naive Bayes learners and classifiers can be extremely fast compared to more sophisticated methods. The decoupling of the class conditional feature distributions means that each distribution can be independently estimated as a one dimensional distribution. This in turn helps to alleviate problems stemming from the curse of dimensionality.

GaussianNB implements the Gaussian Naive Bayes algorithm for classification

source: scikit-learn

1.6 MODEL TUNING, BAGGING (RANDOM FOREST SHOULD BE APPLIED FOR BAGGING), AND BOOSTING.

• Unlike usual business cases where there is a positive and negative class outcomes, in this case the classifications are neither positive nor negative and both holds equal importance in terms of the predictability of the model

• **Class imbalance:** The majority class (Labour) is 70% and the minority class (Conservative) is 30% of the total target class (ratio of 2.3:1). The machine learning classification algorithms tend towards the majority class in prediction resulting in what is known as accuracy paradox

• **Accuracy paradox** is when the model gives a very high accuracy score but the underlying class distribution is skewed favouring a higher recall and precision of the majority class

• **The approach taken** for model tuning and selection is to first balance the precision and recall of the minority class with that of the majority class in train and test datasets before accuracy and area under curve of the model is assessed

• The **model tuning** measures applied are as follows:

1. SMOTE

• Synthetic minority oversampling technique was applied on the train dataset in second iteration to assess if the performance of the models will improve.

• However, it was found that the model performance remained same or got depreciated in all models (accuracy and recall of minority class in test)

Tab 20. Before and after SMOTE

Model	Before SMOTE		After SMOTE	
Logit	0.82	0.68	0.82	0.71
LDA	0.82	0.71	0.7	0
KNN	0.8	0.63	0.81	0.65
MNB	0.83	0.74	0.75	0.45

Accuracy
Recall

2. Class weights

• The algorithms such as Linear Regression and Random Forest (for Bagging) allow us to set weightage on the classes such that the performance matrices of the minority class can be improved • A dictionary of class weightage was initially set based on the original class ratio (1:2.3) and later tweaked iteratively by validating the performance metrices and not letting the model overfit on train

3. GridsearchCV

• Except for the Multinomial Naïve Bayes and Bagging classifier, GridsearchCV was applied to optimise the hyperparameters of all the models

• Logspace() was used in Logistic regression and SVM models to find the optimal value of the regularisation parameter 'C' in log scale

• The model accuracy and recall of minority class after applying class_weight and GridsearchCV is as below

4. Cross validation

• 10-fold cross validation of all the final selected models were applied to find out the mean accuracy score and standard deviation of the

• Mean accuracy score and standard deviation of the scores will be used in final model selection

Tab 21. Before and after tuning

Model	Before tuning		After tuning	
Logit	0.82	0.68	0.83	0.79
KNN	0.8	0.63	0.82	0.66

Accuracy
Recall

1.6 BAGGING AND BOOSTING

Random Forest Classifier

• The ratio of the target classes (1:2.3) was used and later tuned iteratively to set the class_weight parameter of the Random Forest Classifier. The ideal weightage for classes 0 and 1 was finally set as 4 and 1.5 respectively

• The values for min_samples_leaf and min_samples_split was also tuned iteratively based on the accuracy score and recall of classes

• The Random Forest model was used as base_estimator for the Bagging Classifier and n_estimator was optimised so that model wont overfit

• Without the class_weight values the model was initially found to be highly overfitting and the recall of the minority class was found to be too low

• The final Bagging classifier model was found to be slightly overfit but acceptable as it returned equally high recall rate for both majority and minority classes

RANDOM FOREST CLASSIFIER MODEL

```
BaggingClassifier(base_estimator=RandomForestClassifier(class_weight={0: 4,
                                                                       1: 1.5},
                                                        min_samples_leaf=2,
                                                        min_samples_split=4),
                  n_estimators=50, random_state=1)
```

AdaBoost

• The module sklearn.ensemble includes the popular boosting algorithm AdaBoost, introduced in 1995 by Freund and Schapire. The core principle of AdaBoost is to fit a sequence of weak learners (i.e., models that are only slightly better than random guessing, such as small decision trees) on repeatedly modified versions of the data. The predictions from all are then combined through a weighted majority vote (or sum) for the final prediction.

• The number of weak learners is controlled by the parameter n_estimators. The learning_rate parameter controls the contribution of the weak learners in the final combination. By default, weak learners are decision stumps. Different weak learners can be specified through the base_estimator parameter. The main parameters to tune to obtain good results are n_estimators and the complexity of the base estimators (e.g., its depth max_depth or minimum required number of samples to consider a split min_samples_split).

• AdaBoost Classifier model was built with SAMME.R algorithm, learning rate of 0.1, and 500 as n_estimators. • Its accuracy dropped by one percent and it recall improved by 4 percent in test. Not much improvement was noticed.

ADABOOST MODEL

```
AdaBoostClassifier(learning_rate=0.1, n_estimators=500, random_state=0)
```

XGBoost

• Extreme Gradient Boosting or known as XGBoost is used to build a Boosting model

• GridsearchCV was initially used to optimize the hyperparameters of the model. But it was found to have created an extremely overfitting model with no significant improvement of performance in test

• The hyperparameters are tuned iteratively so that the model is right-fit and the recall of minority class is acceptable along with the majority class

• The final model accuracy score was slightly overfit but acceptable, but the recall in test was lowest of all models

• XGBoost is observed to be more appropriate for very large datasets than smaller ones as in the given business case

EXTREME GRADIENT BOOST MODEL

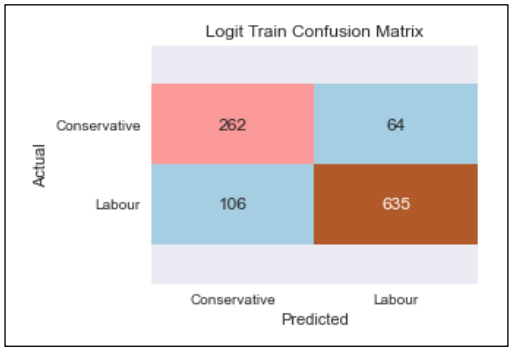
```
XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
              colsample_bynode=1, colsample_bytree=1, gamma=0, gpu_id=-1,
              importance_type='gain', interaction_constraints='',
              learning_rate=0.01, max_delta_step=0, max_depth=5,
              min_child_weight=3, missing=nan, monotone_constraints='()',
              n_estimators=1000, n_jobs=8, num_parallel_tree=1, random_state=0,
              reg_alpha=0, reg_lambda=1, scale_pos_weight=1, subsample=1,
              tree_method='exact', validate_parameters=1, verbosity=None)
```

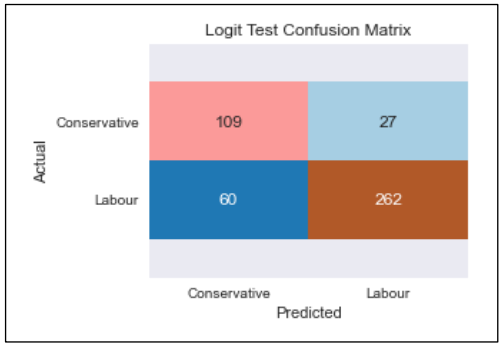
1.7 PERFORMANCE METRICS: CHECK THE PERFORMANCE OF PREDICTIONS ON TRAIN AND TEST SETS USING ACCURACY, CONFUSION MATRIX, PLOT ROC CURVE AND GET ROC_AUC SCORE FOR EACH MODEL.

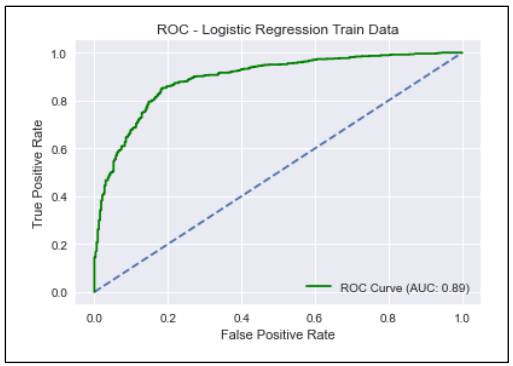
Train **Logistic Regression** Test

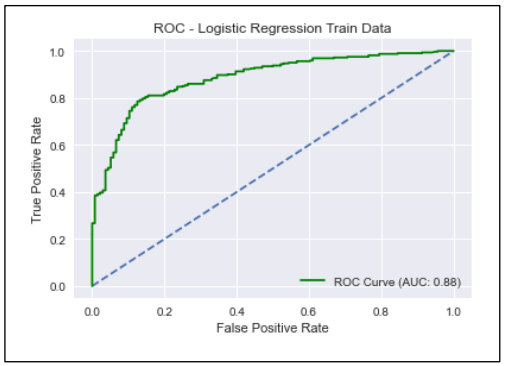
	precision	recall	f1-score	support
0	0.711957	0.803681	0.755043	326.000000
1	0.908441	0.856950	0.881944	741.000000
accuracy	0.840675	0.840675	0.840675	0.840675

	precision	recall	f1-score	support
0	0.644970	0.801471	0.714754	136.000000
1	0.906574	0.813665	0.857610	322.000000
accuracy	0.810044	0.810044	0.810044	0.810044





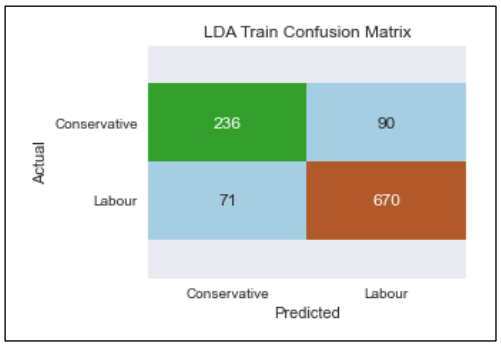


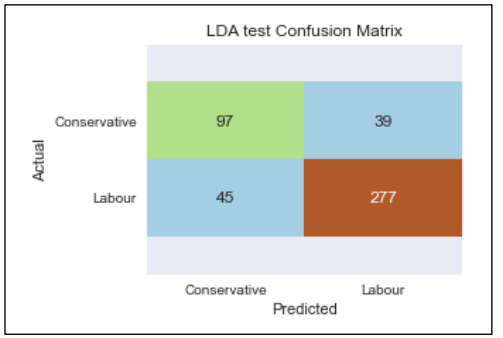


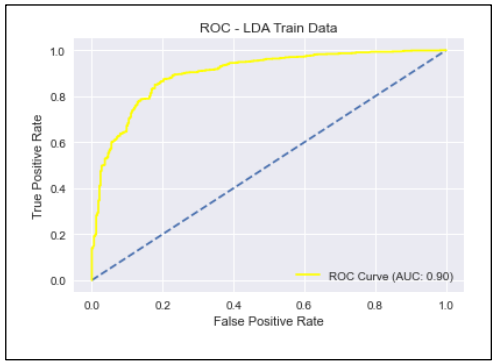
Train **LDA** Test

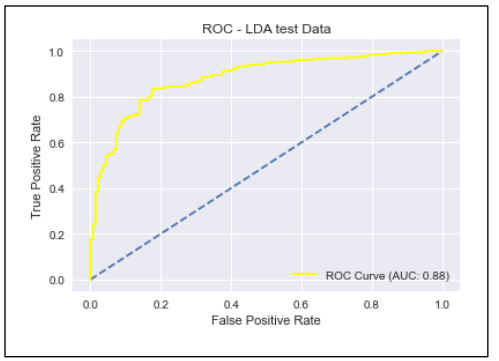
	precision	recall	f1-score	support
0	0.768730	0.723926	0.745656	326.00000
1	0.881579	0.904184	0.892738	741.00000
accuracy	0.849110	0.849110	0.849110	0.84911

	precision	recall	f1-score	support
0	0.683099	0.713235	0.697842	136.000000
1	0.876582	0.860248	0.868339	322.000000
accuracy	0.816594	0.816594	0.816594	0.816594









Confusion Matrix is a performance measure of a 2X2 or other square-matrix shapes, and lists the True Positive, False Positive, True Negative, and False Negative for the model. Various other metrics can be extracted from the same

Fig 21 and 22. Performance metrics for Logistic Regression and LDA

1.7 PERFORMANCE METRICS:

Train	**K-Nearest Neighbours**	Test

	precision	recall	f1-score	support
0	0.798046	0.751534	0.774092	326.000000
1	0.893421	0.916329	0.904730	741.000000
accuracy	0.865979	0.865979	0.865979	0.865979
macro avg	0.845733	0.833932	0.839411	1067.000000
weighted avg	0.864281	0.865979	0.864816	1067.000000

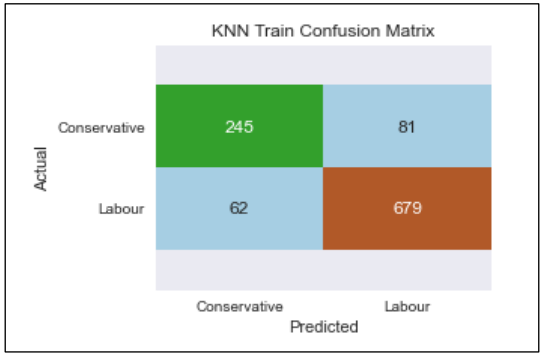
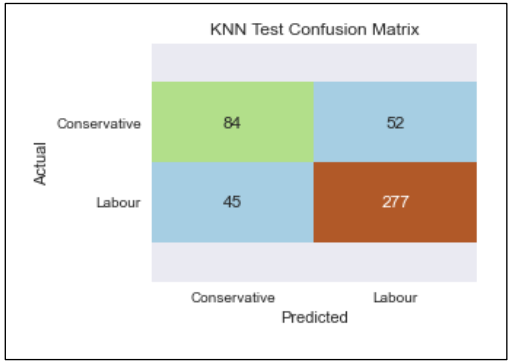
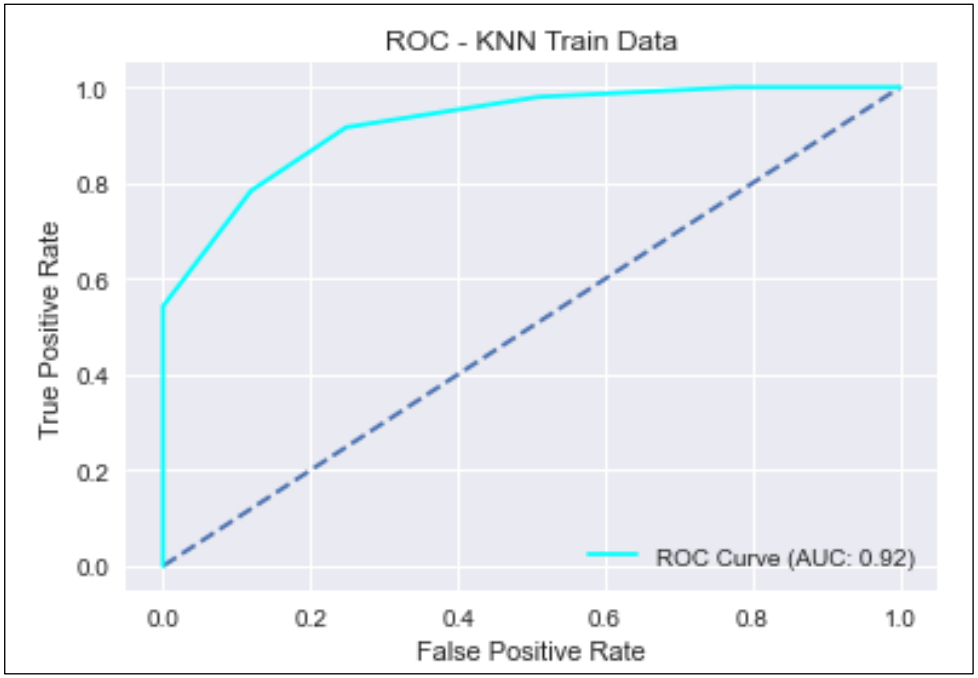
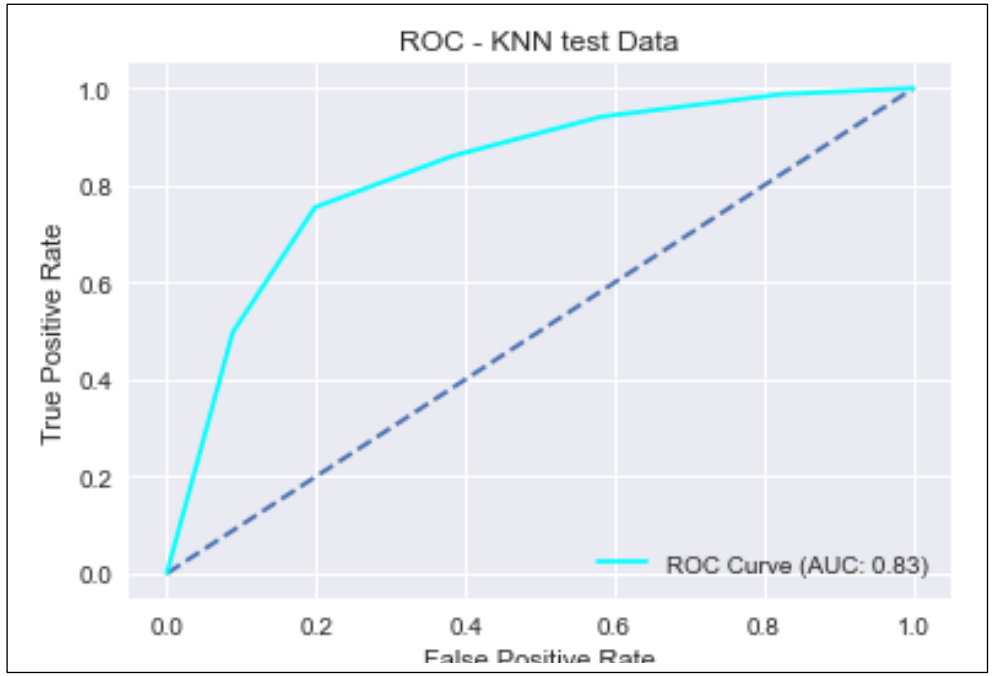


Fig 23 and 24. Performance metrics for KNN

	precision	recall	f1-score	support
0	0.651163	0.617647	0.633962	136.00000
1	0.841945	0.860248	0.850998	322.00000
accuracy	0.788210	0.788210	0.788210	0.78821
macro avg	0.746554	0.738948	0.742480	458.00000
weighted avg	0.785294	0.788210	0.786551	458.00000



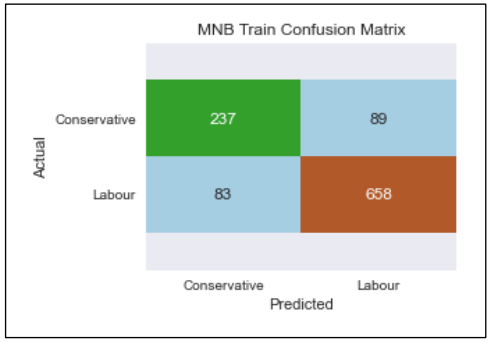




1.7. PERFORMANCE METRICS:

Train | **Multinomial Naïve Bayes** | Test

	precision	recall	f1-score	support
0	0.740625	0.726994	0.733746	326.0000
1	0.880857	0.887989	0.884409	741.0000
accuracy	0.838800	0.838800	0.838800	0.8388
macro avg	0.810741	0.807492	0.809077	1067.0000
weighted avg	0.838012	0.838800	0.838377	1067.0000



	precision	recall	f1-score	support
0	0.706294	0.742647	0.724014	136.000000
1	0.888889	0.869565	0.879121	322.000000
accuracy	0.831878	0.831878	0.831878	0.831878
macro avg	0.797591	0.806106	0.801568	458.000000
weighted avg	0.834668	0.831878	0.833063	458.000000

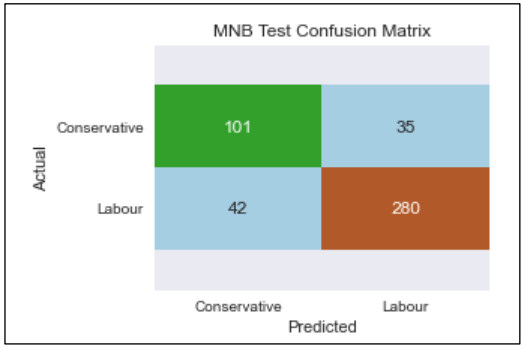
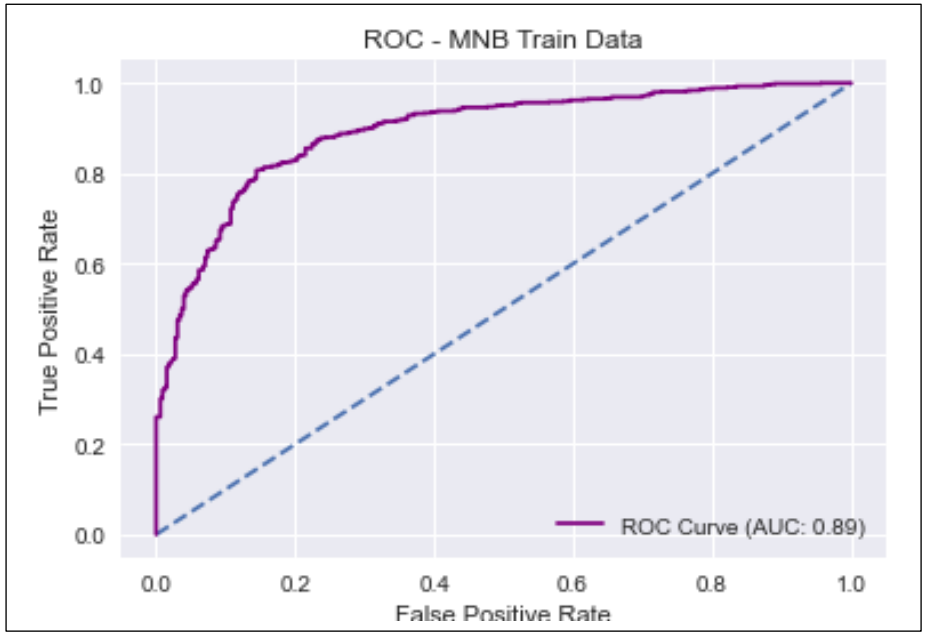
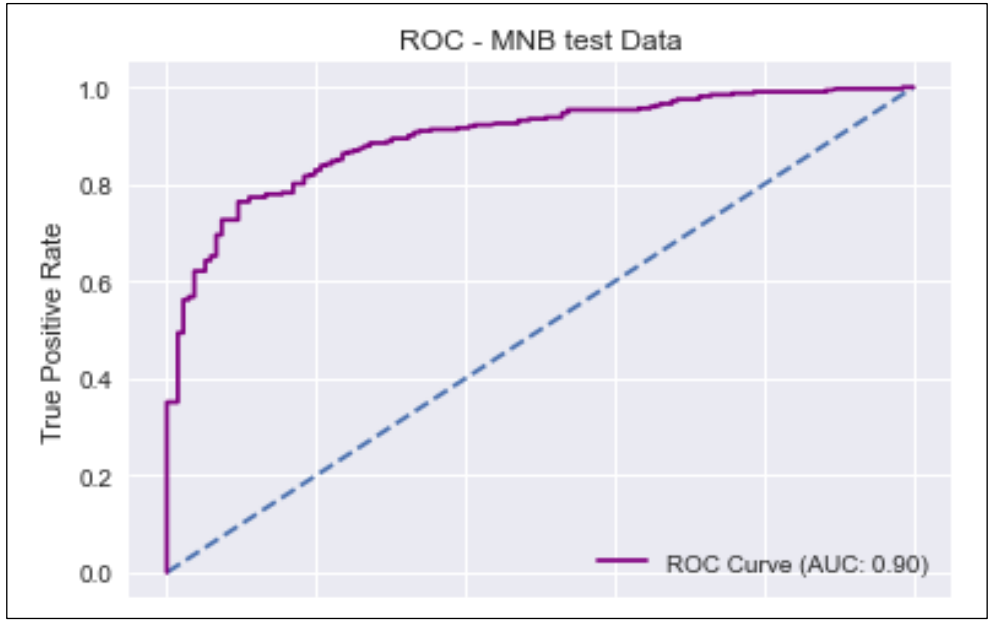


Fig 24 and 26. Performance metrics for Multinomial Naïve Bayes





1.7 PERFORMANCE METRICS:

Train | **Bagging with Random Forest** | Test

	precision	recall	f1-score	support
0	0.801136	0.865031	0.831858	326.000000
1	0.938462	0.905533	0.921703	741.000000
accuracy	0.893158	0.893158	0.893158	0.893158
macro avg	0.869799	0.885282	0.876781	1067.000000
weighted avg	0.896505	0.893158	0.894253	1067.000000

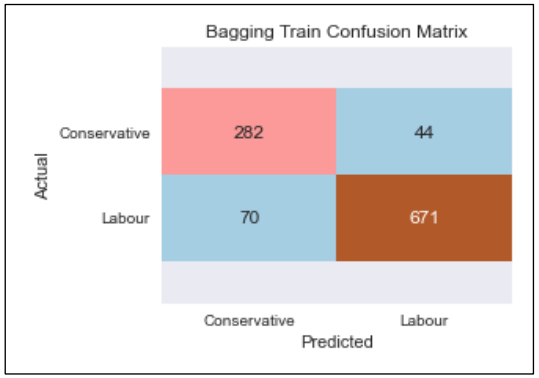
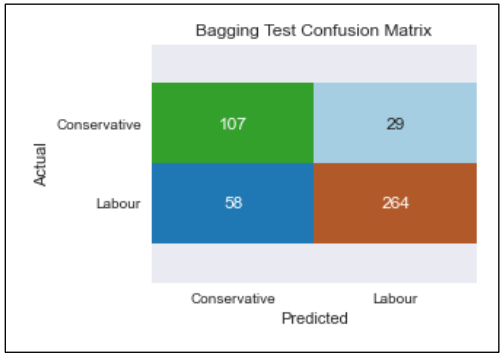
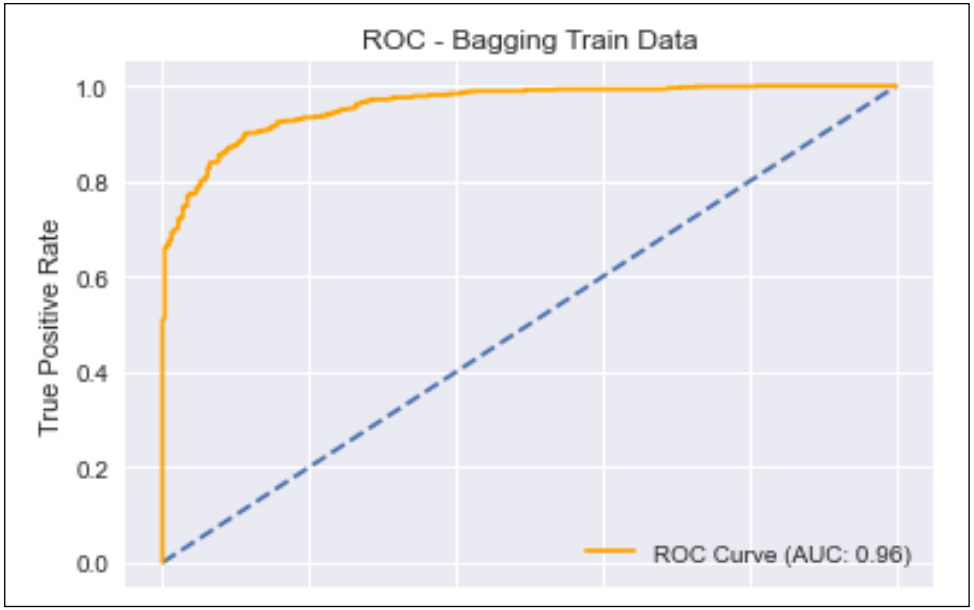
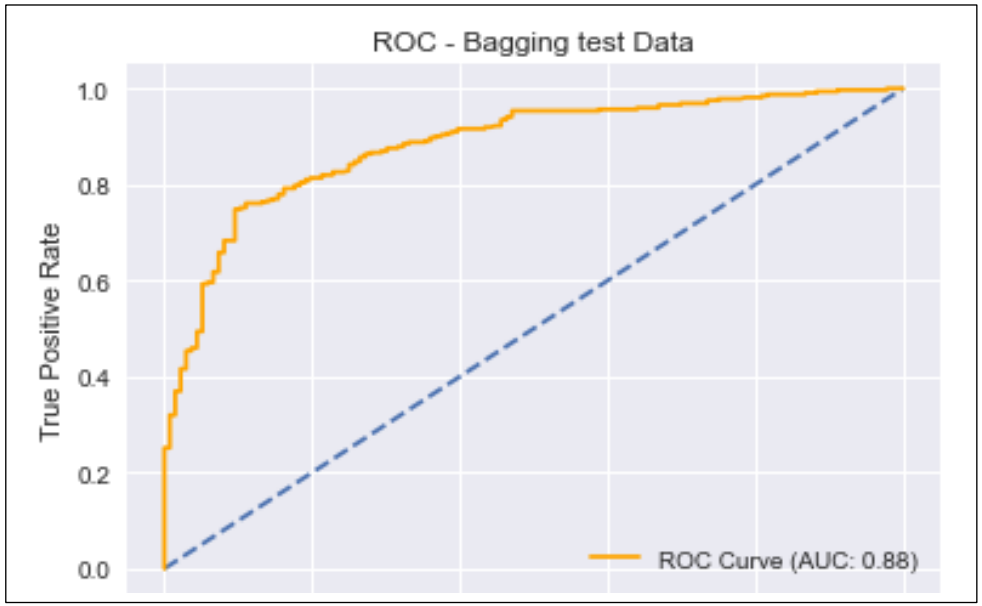


Fig 27 and 28. Performance metrics for bagging with random forest

	precision	recall	f1-score	support
0	0.648485	0.786765	0.710963	136.000000
1	0.901024	0.819876	0.858537	322.000000
accuracy	0.810044	0.810044	0.810044	0.810044
macro avg	0.774754	0.803320	0.784750	458.000000
weighted avg	0.826034	0.810044	0.814716	458.000000







1.7 PERFORMANCE METRICS:

Train **Adaptive Boosting** Test

	precision	recall	f1-score	support
0	0.756014	0.674847	0.713128	326.000000
1	0.863402	0.904184	0.883322	741.000000
accuracy	0.834114	0.834114	0.834114	0.834114
macro avg	0.809708	0.789515	0.798225	1067.000000
weighted avg	0.830592	0.834114	0.831323	1067.000000

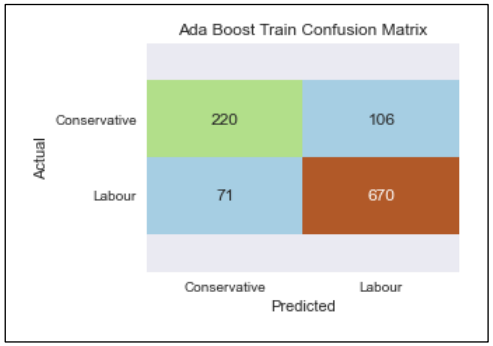
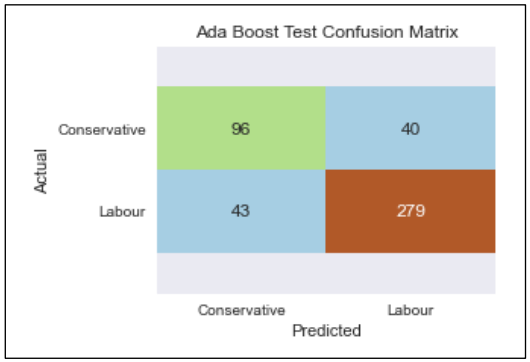
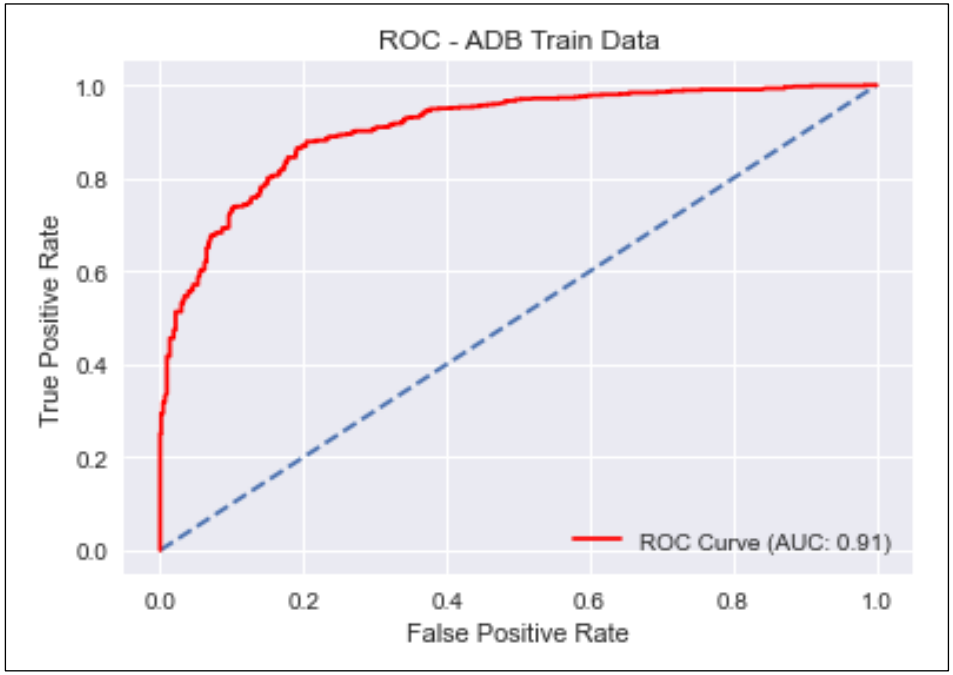
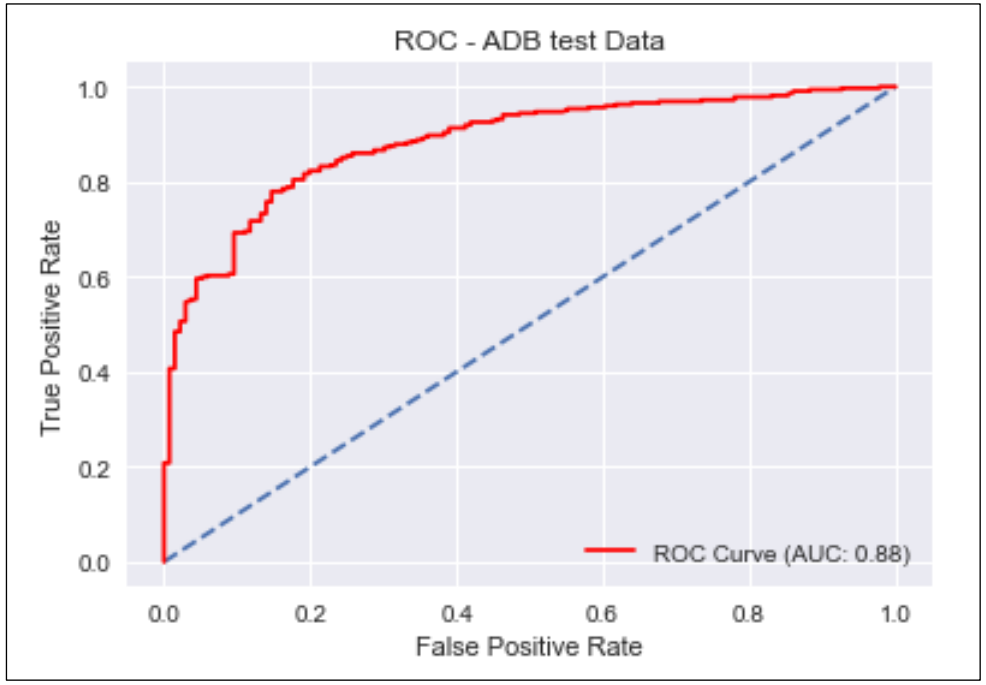


Fig 29 and 30. Performance metrics for Adaptive Boosting

	precision	recall	f1-score	support
0	0.690647	0.705882	0.698182	136.000000
1	0.874608	0.866460	0.870515	322.000000
accuracy	0.818777	0.818777	0.818777	0.818777
macro avg	0.782628	0.786171	0.784348	458.000000
weighted avg	0.819982	0.818777	0.819342	458.000000



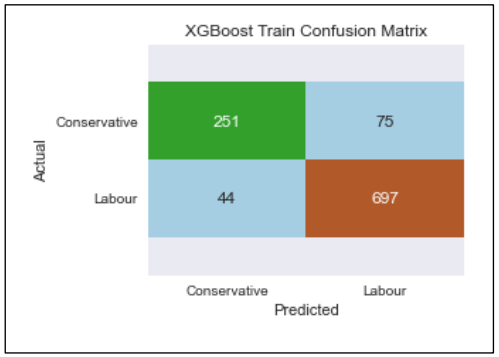




1.7 PERFORMANCE METRICS

Train | **Extreme Gradient Boost** | Test

	precision	recall	f1-score	support
0	0.850847	0.769939	0.808374	326.000000
1	0.902850	0.940621	0.921348	741.000000
accuracy	0.888472	0.888472	0.888472	0.888472
macro avg	0.876849	0.855280	0.864861	1067.000000
weighted avg	0.886962	0.888472	0.886831	1067.000000



	precision	recall	f1-score	support
0	0.704545	0.683824	0.694030	136.000000
1	0.868098	0.878882	0.873457	322.000000
accuracy	0.820961	0.820961	0.820961	0.820961
macro avg	0.786322	0.781353	0.783743	458.000000
weighted avg	0.819532	0.820961	0.820177	458.000000

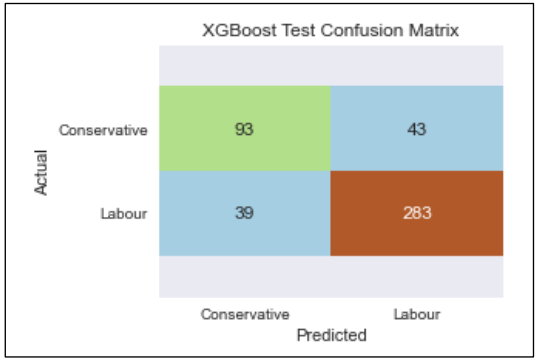
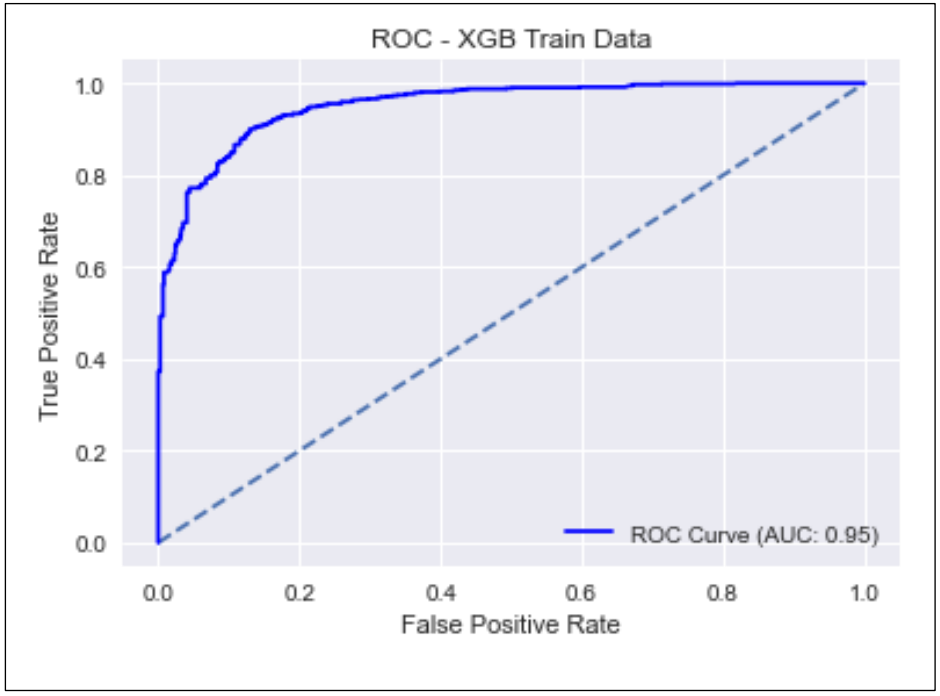
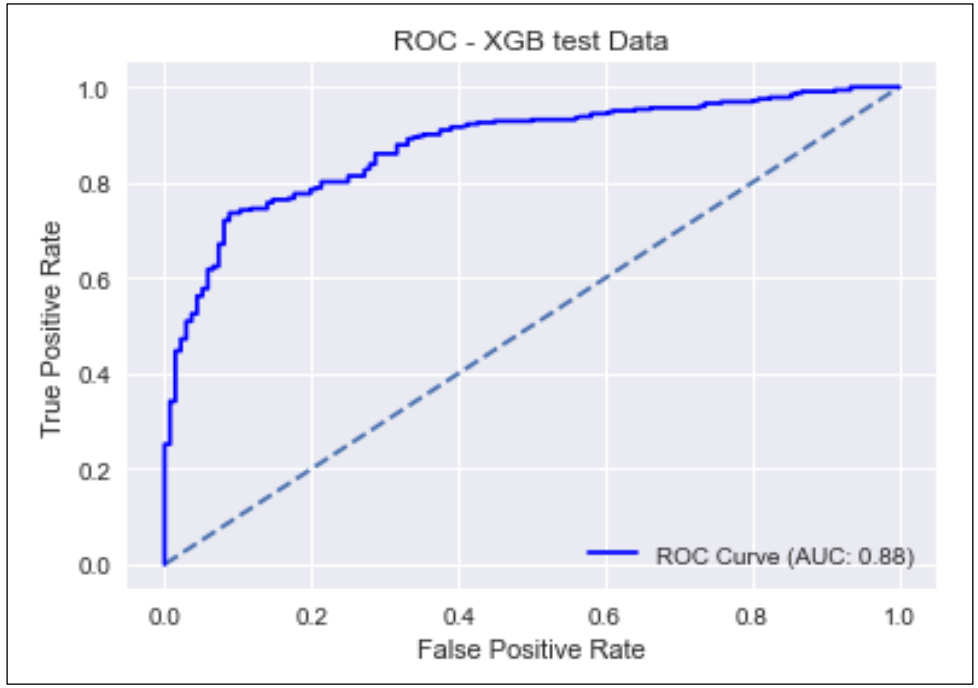


Fig 31 and 32. Performance metrics for Extreme gradient boost





1.7 FINAL MODEL SELECTION

Accuracy

• While Bagging and XGBoost classifiers delivered the highest accuracy of 89% in train, Logistic Regression and Naïve Bayes gave the highest accuracy of 83% in test

• KNN is least with 83 % in train and, in test, Bagging is least at 81%

Recall

• In train, Bagging has the highest recall score for the minority class at 86 % , followed by SVM at 84 %. For the majority class, XGBoost got 94 % recall, followed by KNN AT 91 %

• In test, Logit gave highest recall of minority class at 80% , followed by Bagging at 79 %. For the majority class, KNN gave an 90% recall, followed by XGB (88%) and MNB (87%)

• Logit and Bagging gave most balanced recall of both classes than other models

F 1 Score

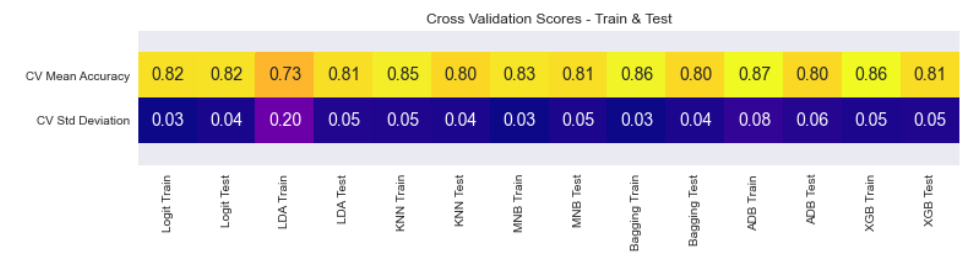
• The harmonic mean of recall and precision, in train is high for Bagging and XGBoost at 83 % and 81 % for minority class and 92 % for majority class • In test, Logit is highest for minority class at 73 % and 87 % by Logit, LDA, KNN and XGBoost for majority class

Cross Validation

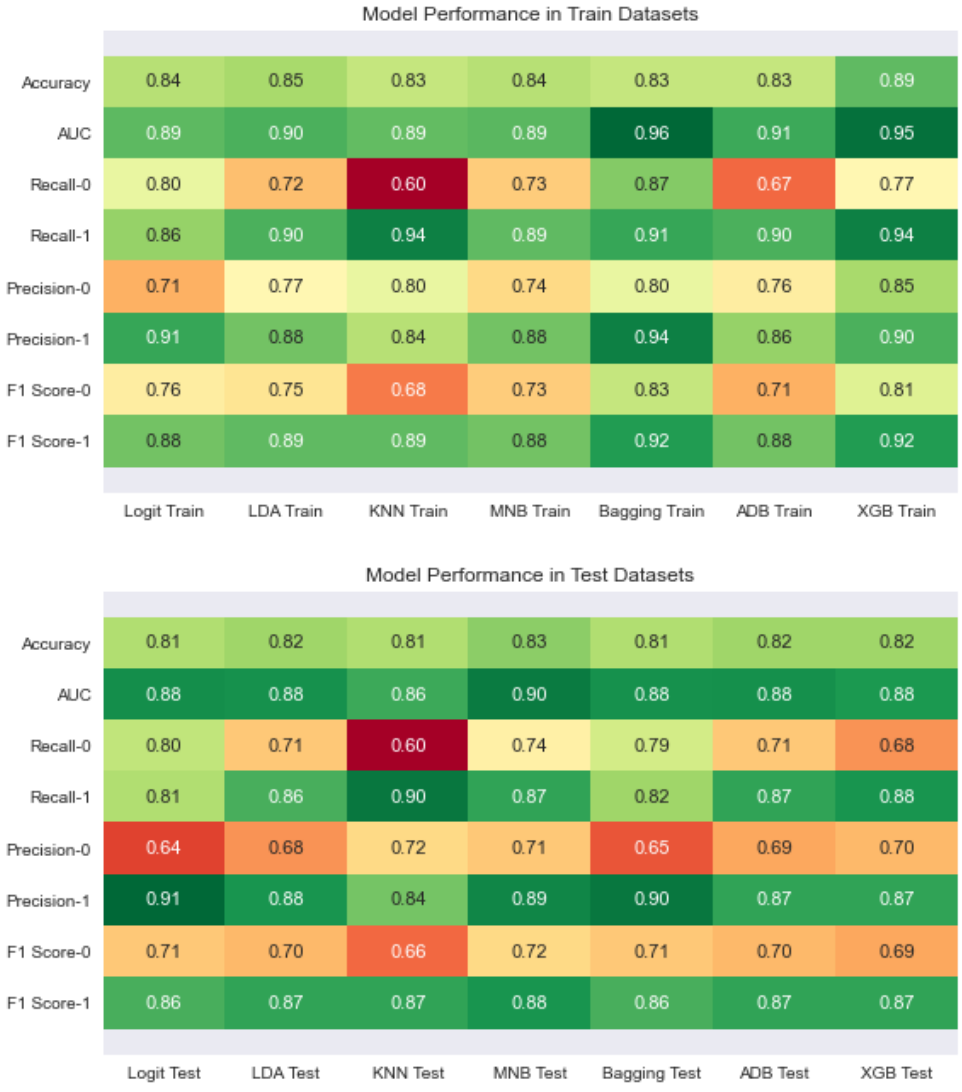
• The 10-fold cross validation shows the highest mean accuracy in train for XGBoost (87%) and lowest for LDA (73%)

• In test, except Bagging (80%) all other models gave a mean accuracy of 81%

• LDA model got highest inconsistency with 20 % std deviation. Logit, MNB, & Bagging got lowest std dev of 3 % • While in test KNN & XGB got 4% deviation and rest at 5 % and 6 % (ADB)



Tab 22. Cross-validation scores for train and test



Tab 23 and 24. Model performance for train and test

1.7 FINAL MODEL SELECTION

AUC score

• Bagging and XGBoost gave the highest AUC score on train, 96 % and 95 % respectively, but on test they gave 88 % each

• On test, Naïve Bayes delivered the highest AUC score at 90% , followed by SVM and Logit at 89 %

• All the models have consistently given an AUC score above 89% in train and above 88 % in test, except KNN. KNN gave the least AUC score in test of 86%

Conclusion

• Bagging and XGBoost models tend towards defining high variance in train, but fail to perform in test, creating overfit models

• Gradient Boosting models like XGBoost would perform extremely well with large datasets than the smaller ones as given in this case

• Bootstrap aggregation (Bagging) require higher computational resources and may pose low performance challenges

• Whereas parametric models like Naïve Bayes and Logit showed better consistency in class predictions and computational performance Final model: Multinomial Naïve Bayes

• LDA showed least consistency in 10-fold cross validation and KNN gave the least recall of minority class in train & test

• Considering the robustness of model on cross validation, recall across both the target classes and consistent accuracy on both train and test datasets, Logistic Regression, SVM and MNB could be chosen as the final model

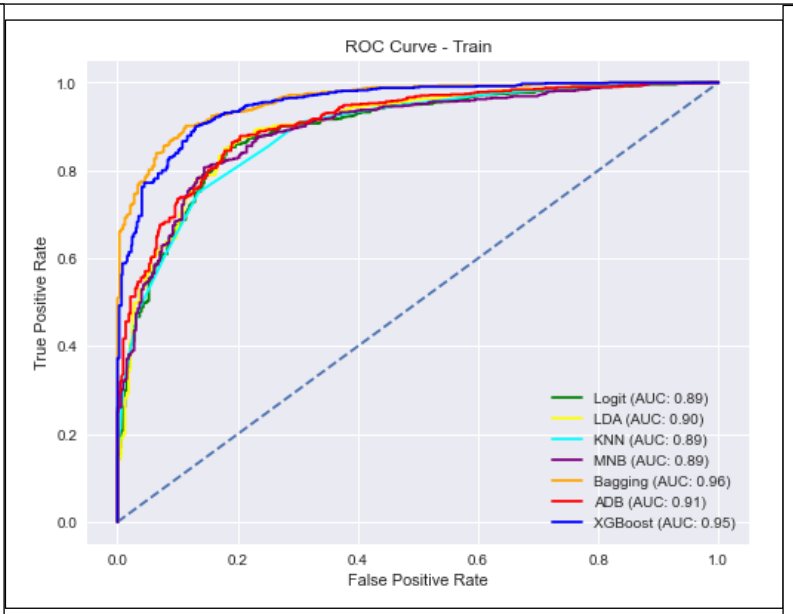
• MNB has delivered highest AUC score in test and right -fit in train and test without any complex model tuning required. The recall rate of the target classes is also on par with Logit

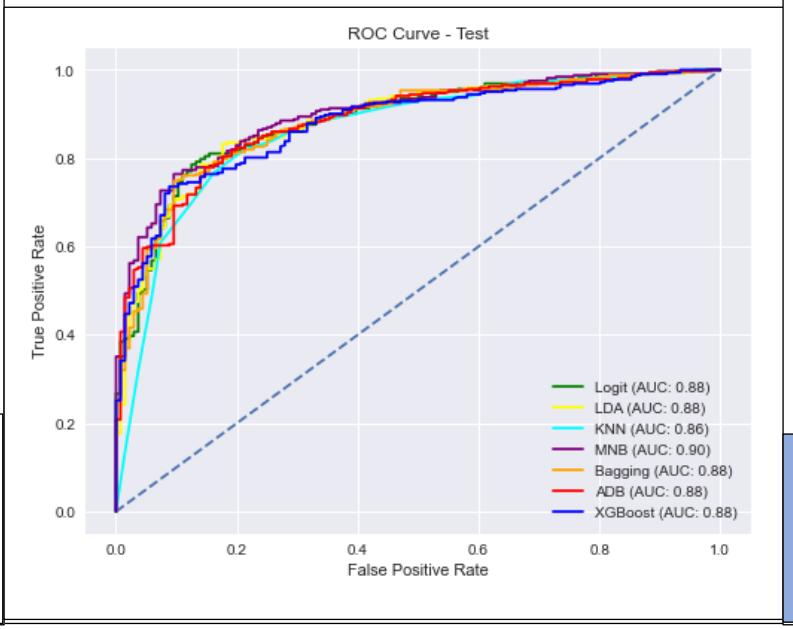
• AdaBoost model doesn't have a very high recall score on minority class, although the recall of majority class is decent

• Logit model has produced very consistent and balanced performance in terms of minority and majority class prediction and overall accuracy in train and test

• Multinomial Naïve Bayes is recommended as the final model considering the low size of the given dataset and the model being fast and highly scalable

Final model: Multinomial Naïve Bayes





The curves have a significant overlap. The multinomial Naïve Bayes curve marked in deep purple is better in certain cases but only marginally, while it is difficult to distinguish significantly from the other curves, as AUC is exactly the same (0.88) for all models except multinomial Naïve Bayes. The AUC score has improved in test for MNB, hence that helped us pick this as our final model.

Figure 33 and 34. All-model ROC curves for Test and Train

1.8 INSIGHTS

• The models applied in the exercise such as Logistic Regression and Gradient Boosting allow us to rank features based on feature coefficient and feature importance values of the respective models

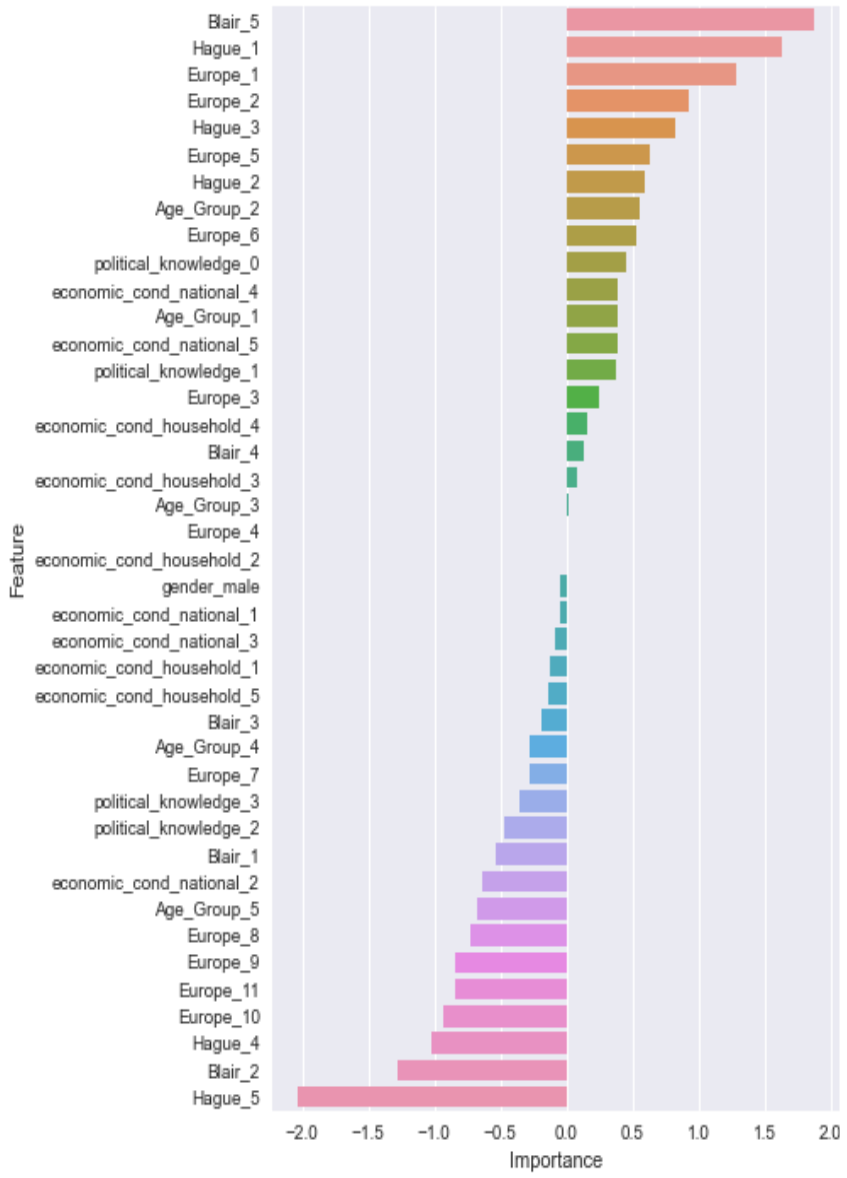
• Here, feature coefficients from the Logistic Regression is used to draw insights on the predictions, by plotting the values as a barplot. Key inferences are as follows.

• The assessment/opinion of voters on the Labour and Conservative leaders, Blair and Hague is the most decisive factor in the poll than any other factors, making the election a mandate on these two political leaders

• A voter who rated Blair as 5 and Hague as 1 is more likely to vote for Labour, whereas the one who rated Hague as 5 and Blair as 2 or 1 is more likely to vote for the Conservatives

• The most important political factor of the poll is observed to be European integration of the UK, but it is second only to voters' confidence and opinion of Blair and Hague

• While the Eurosceptic voters chose to vote for Conservative party, the lesser sceptics or those who are not concerned on the increasing influence of EU chose to vote for Labour Feature Ranking from



• Those who are confident of their awareness of political parties' position on European integration are most likely to be Conservative voters, whereas for the Labour voters, the issue was least of their concerns

• The feature importance from the XGBoost also returned similar patterns of feature ranking (not shown here) where the rating on Mr Blair and Mr Hague stood out to be the most significant factor of the election mandate

• Using the prediction probability from classification, with 83% confidence it can confirmed that Labour would get 70% votes

• The national and household economic condition prevailed during the period is found to be a non-issue during the poll, which is an indictor of the fledging economic situation of the period

• The age of the voter appears to be a decisive factor in predicting the votes, as those from group 2 (36 to 45 years) and lesser (35 and below) appears to have favoured Labour, while those from group 5 and 4 (66 and above) are more likely to vote for the Conservative party

• The gender of the voter doesn't appear to have a significant influence on the outcome of the election

Fig 35. Feature ranking from logistic regression model

1.8 INSIGHTS

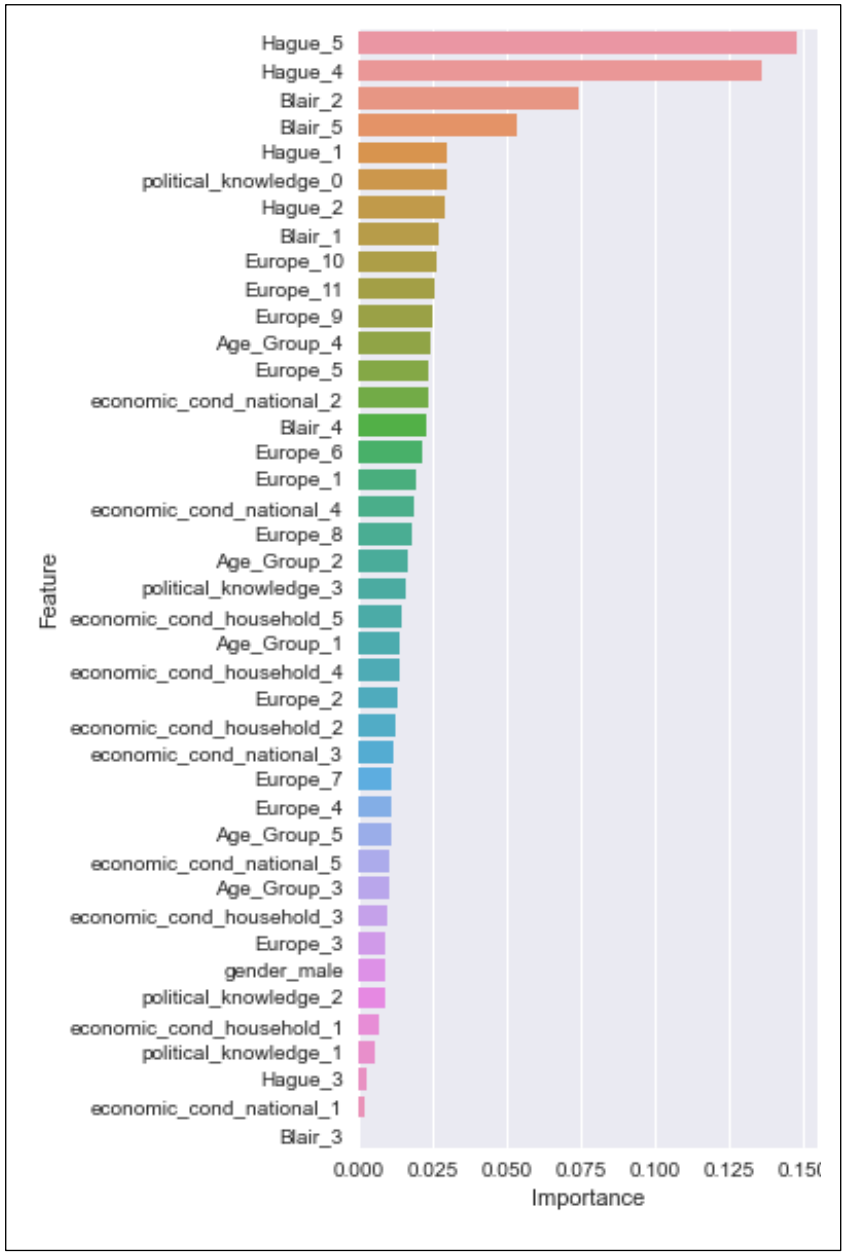


Fig 36. Feature ranking from the XGBoost model

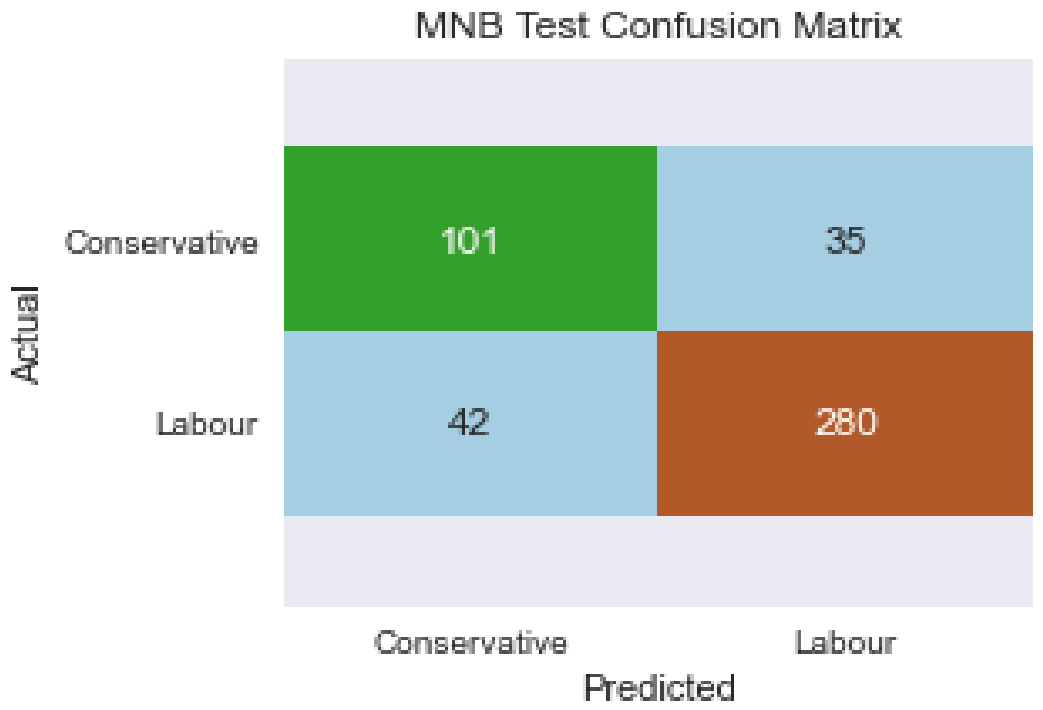


Fig 37. Confusion matrix for test data for the final selected model: MNB

- 101 conservative votes were called truly.
- 280 Labour votes were also called truly.
- 42 Labour votes were given to the Conservative party by error of prediction
- 35 actual Conservative votes were given to the Labour party by error of prediction
- The false positives and negatives don't affect the result at all, since an accurate call would have given the Conservatives 101 votes to 322 of the Labour party. Tories still would have lost

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: -

President Franklin D. Roosevelt in 1941

Photo by Hulton Archive/Getty Images

President John F. Kennedy in 1961

Photo by Alfred Eisenstaedt for Pix Inc. The LIFE Picture Collection via Getty Images

President Richard Nixon in 1973

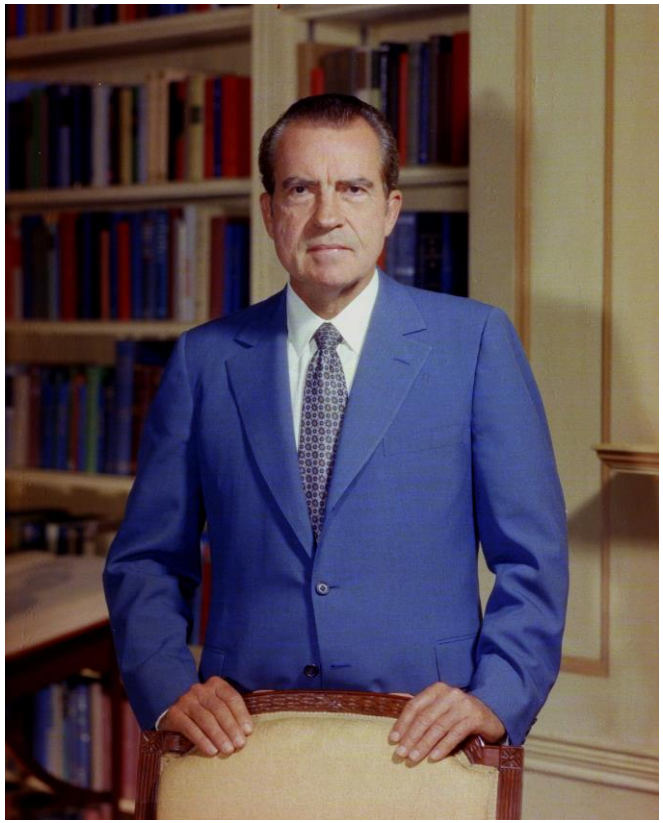


Photo by Bachrach for Getty Images

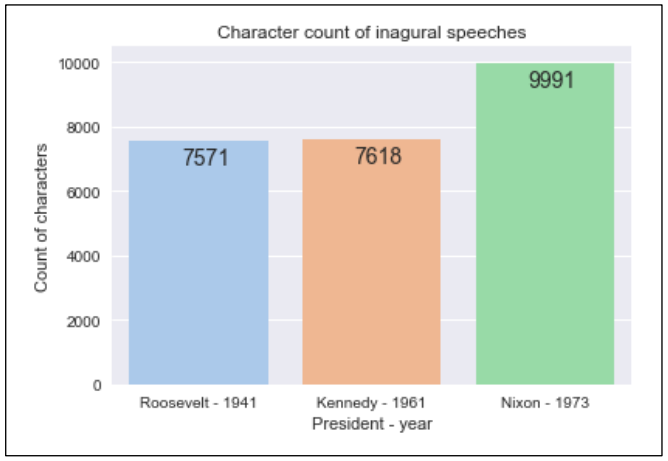
2.1 FIND THE NUMBER OF CHARACTERS, WORDS, AND SENTENCES FOR THE DOCUMENTS.

Inaugural speeches	Roosevelt	Kennedy	Nixon
['1789-Washington.txt', '1793-Washington.txt', '1797-Adams.txt', '1801-Jefferson.txt', '1805-Jefferson.txt', '1809-Madison.txt', '1813-Madison.txt', '1817-Monroe.txt', '1821-Monroe.txt', '1825-Adams.txt', '1829-Jackson.txt', '1833-Jackson.txt', '1837-VanBuren.txt', '1841-Harrison.txt', '1845-Polk.txt', '1849-Taylor.txt', '1853-Pierce.txt', '1857-Buchanan.txt', '1861-Lincoln.txt', '1865-Lincoln.txt', '1869-Grant.txt', '1873-Grant.txt', '1877-Hayes.txt', '1881-Garfield.txt', '1885-Cleveland.txt', '1889-Harrison.txt', '1893-Cleveland.txt', '1897-McKinley.txt', '1901-McKinley.txt', '1905-Roosevelt.txt', '1909-Taft.txt', '1913-Wilson.txt', '1917-Wilson.txt', '1921-Harding.txt', '1925-Coolidge.txt', '1929-Hoover.txt', '1933-Roosevelt.txt', '1937-Roosevelt.txt', '1941-Roosevelt.txt', '1945-Roosevelt.txt', '1949-Truman.txt', '1953-Eisenhower.txt', '1957-Eisenhower.txt', '1961-Kennedy.txt', '1965-Johnson.txt', '1969-Nixon.txt', '1973-Nixon.txt', '1977-Carter.txt', '1981-Reagan.txt', '1985-Reagan.txt', '1989-Bush.txt', '1993-Clinton.txt', '1997-Clinton.txt', '2001-Bush.txt', '2005-Bush.txt', '2009-Obama.txt', '2013-Obama.txt', '2017-Trump.txt']	['On each national day of inauguration since 1789, the people have renewed their sense of dedication to the United States.\n\nIn Washington\'s day the task of the people was to create and weld together a nation.\n\nIn Lincoln\'s day the task of the people was to preserve that Nation from disruption from within.\n\nIn this day the task of the people is to save that Nation and its institutions from disruption from without.\n\nTo us there has come a time, in the midst of swift happenings, to pause for a moment and take stock -- to recall what our place in history has been, and to rediscover what we are and what we may be. If we do not, we risk the real peril of inaction.\n\nLives of nations are determined not by the count of years, but by the lifetime of the human spirit. The life of a man is three-score years and ten: a little more, a little less. The life of a nation is the fullness of the measure of its will to live.\n\nThere are men who doubt this. There are men who beli	['Vice President Johnson, Mr. Speaker, Mr. Chief Justice, President Eisenhower, Vice President Nixon, President Truman, reverend clergy, fellow citizens, we observe today not a victory of party, but a celebration of freedom -- symbolizing an end, as well as a beginning -- signifying renewal, as well as change. For I have sworn I before you and Almighty God the same solemn oath our forebears prescribed nearly a century and three quarters ago.\n\nThe world is very different now. For man holds in his mortal hands the power to abolish all forms of human poverty and all forms of human life. And yet the same revolutionary beliefs for which our forebears fought are still at issue around the globe -- the belief that the rights of man come not from the generosity of the state, but from the hand of God.\n\nWe dare not forget today that we are the heirs of that first revolution. Let the word go forth from this time and place, to friend and foe alike, that the torch has been passed to a new generation of Americans -- born in this century, tempered by war, disciplined by a hard and bitter peace, proud of our ancient	['Mr. Vice President, Mr. Speaker, Mr. Chief Justice, Senator Cook, Mrs. Eisenhower, and my fellow citizens of this great and good country we share together:\n\nWhen we met here four years ago, America was bleak in spirit, depressed by the prospect of seemingly endless war abroad and of destructive conflict at home.\n\nAs we meet here today, we stand on the threshold of a new era of peace in the world.\n\nThe central question before us is: How shall we use that peace? Let us resolve that this era we are about to enter will not be what other postwar periods have so often been: a time of retreat and isolation that leads to stagnation at home and invites new danger abroad.\n\nLet us resolve that this will be what it can become: a time of great responsibilities greatly borne, in which we renew the spirit and the promise of America as we enter our third century as a nation.\n\nThis past year saw far-reaching results from our new policies for peace. By continuing to revitalize our traditional friendships, and by our missions to Peking and to Moscow, we were able to establish th

2.1 FIND THE NUMBER OF CHARACTERS, WORDS, AND SENTENCES

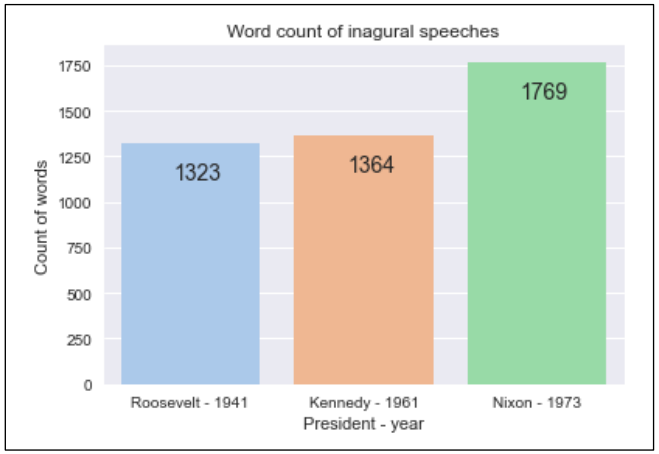
Characters

- President Roosevelt's inaugural speech in 1941 consists of 7571 characters
- In 1961 President Kennedy gave similarly long inaugural speech of 7,618 character long
- President Nixon appears to have given the longest inaugural speech of the three, which is of 9,991 characters long



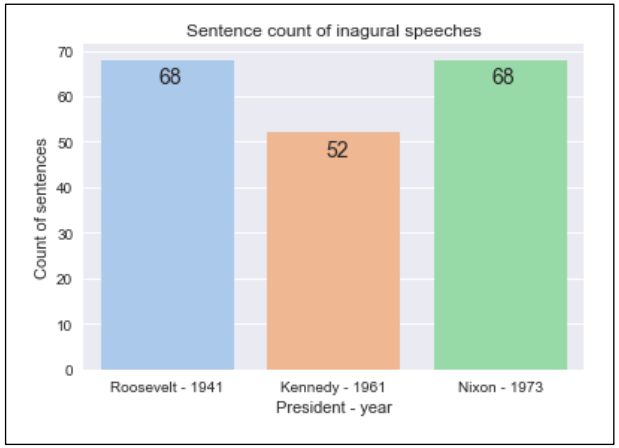
Words

- In terms of number of words used in the speech, President Nixon stands out with 1769 words used in his 1973 inaugural speech
- Followed by President Kennedy's inaugural speech in 1961 with 1364 words
- President Roosevelt's inaugural speech in 1841 consists of 1323 words



Sentences

- It looks like President Nixon favours longer sentences while delivering his speeches, because with larger word base he delivered equal number of sentences as other Presidents
- Both Presidents Roosevelt's and Nixon's inaugural speeches contained 68 sentences
- Followed by President Kennedy with 52 sentences in delivering his inaugural speech



2.2 REMOVE THE STOPWORDS

- Before removing the stopwords all the alphabets in the inaugural speech texts were converted to lower case
- After lower case conversion the punctuations including all special characters were removed from the speeches
- A set of words were added next to the corpus of stop words, as shown here
- Then stopwords are removed from the speeches, sample screen shots of the final texts are as below

	president	text	char_count	word_count	sents_count
1941-Roosevelt	Roosevelt - 1941	On each national day of inauguration since 178...	7571	1323	68
1961-Kennedy	Kennedy - 1961	Vice President Johnson, Mr. Speaker, Mr. Chief...	7618	1364	52
1973-Nixon	Nixon - 1973	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	9991	1769	68

Punctuation marks

```
'!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~'
```

ALL THE STOPWORDS

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'y
ourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself',
'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those',
'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'a
n', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'b
etween', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'of
f', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both',
'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very',
's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'ar
en', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "have
n't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "should
n't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't", '!', '"', '#', '$', '%', '&', "'", '(', ')',
'*', '+', ',', '-', '.', '/', ':', ';', '<', '=', '>', '?', '@', '[', '\\', ']', '^', '_', '`', '{', '|', '}', '~']
```

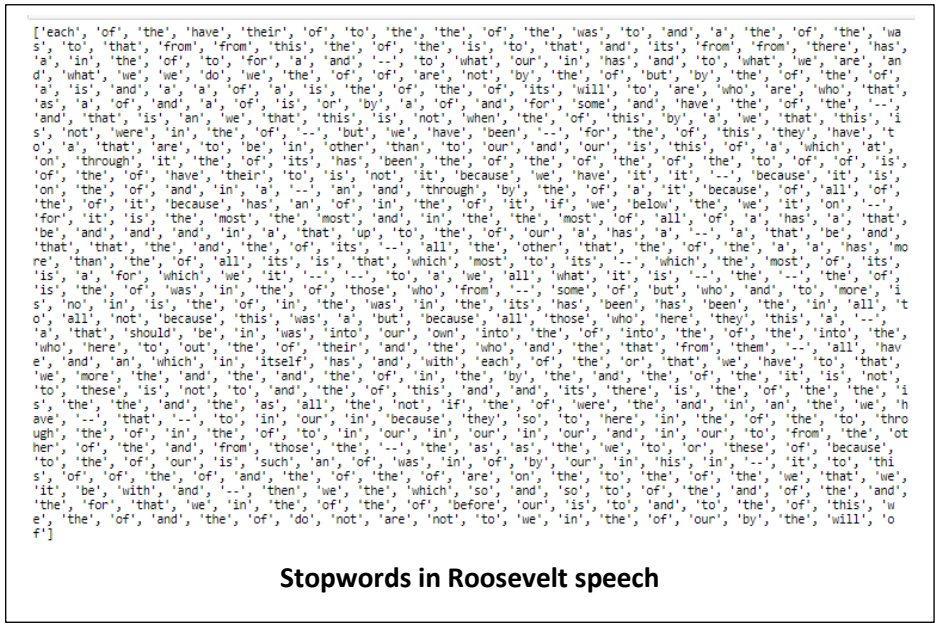
1. Text converted to lower case

```
1941-Roosevelt    on each national day of inauguration since 178...
1961-Kennedy      vice president johnson, mr. speaker, mr. chief...
1973-Nixon        mr. vice president, mr. speaker, mr. chief jus...
Name: text, dtype: object
```

2. Punctuation marks removed

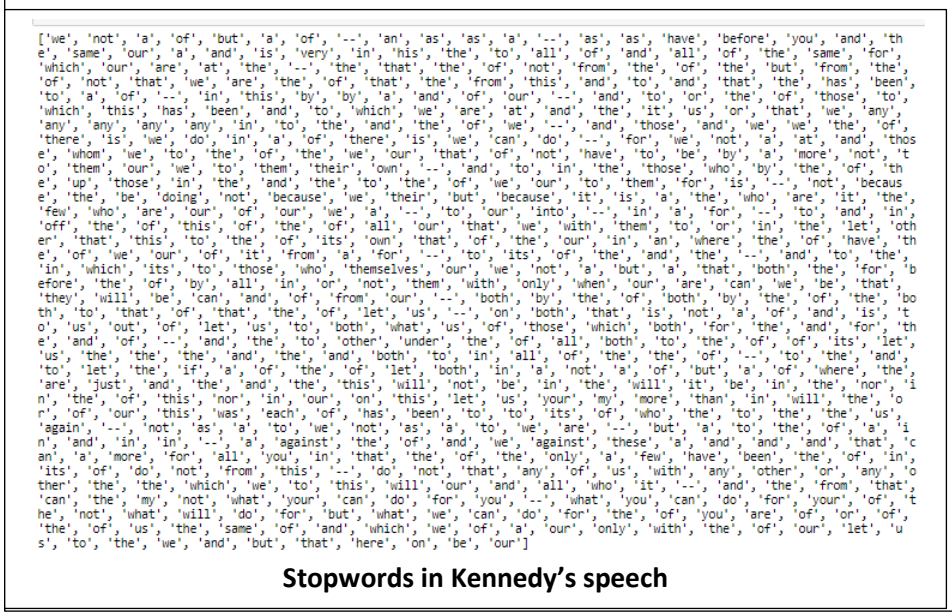
```
1941-Roosevelt    on each national day of inauguration since 178...
1961-Kennedy      vice president johnson mr speaker mr chief jus...
1973-Nixon        mr vice president mr speaker mr chief justice ...
Name: text, dtype: object
```

2.2 REMOVE THE STOPWORDS



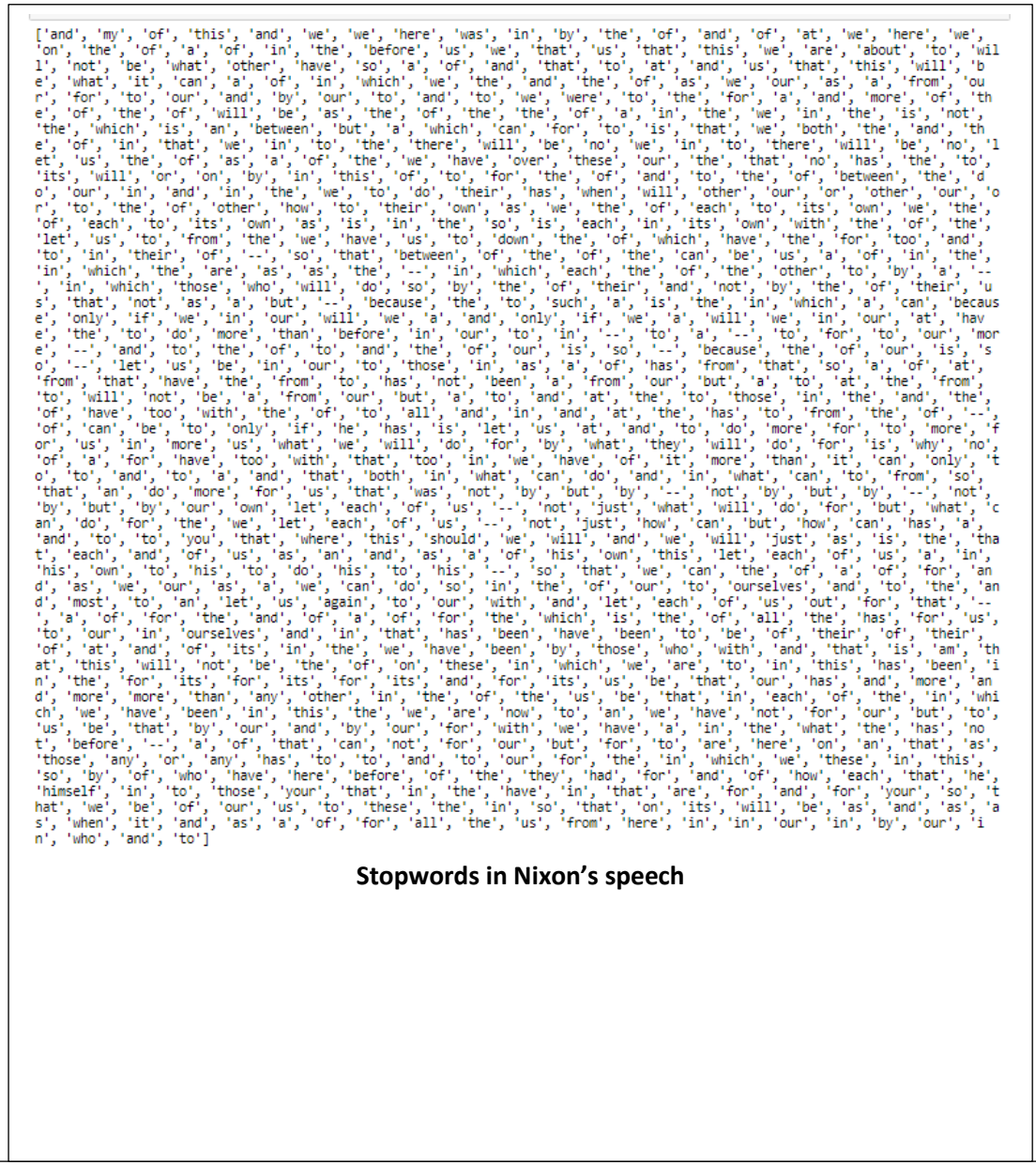
```
['each', 'of', 'the', 'have', 'their', 'of', 'to', 'the', 'the', 'of', 'the', 'was', 'to', 'and', 'a', 'the', 'of', 'the', 'wa
s', 'to', 'that', 'from', 'from', 'this', 'the', 'of', 'the', 'is', 'to', 'that', 'and', 'its', 'from', 'from', 'there', 'has',
'a', 'in', 'the', 'of', 'to', 'for', 'a', 'and', '--', 'to', 'what', 'our', 'in', 'has', 'and', 'to', 'what', 'we', 'are', 'an
d', 'what', 'we', 'we', 'do', 'we', 'the', 'of', 'of', 'are', 'not', 'by', 'the', 'of', 'but', 'by', 'the', 'of', 'the', 'of',
'a', 'is', 'and', 'a', 'a', 'of', 'a', 'is', 'the', 'of', 'the', 'of', 'its', 'will', 'to', 'are', 'who', 'are', 'who', 'that',
'as', 'a', 'of', 'and', 'a', 'of', 'is', 'or', 'by', 'a', 'of', 'and', 'for', 'some', 'and', 'have', 'the', 'of', 'the', '--',
'and', 'that', 'is', 'an', 'we', 'that', 'this', 'is', 'not', 'when', 'the', 'of', 'this', 'by', 'a', 'we', 'that', 'this', 'i
s', 'not', 'were', 'in', 'the', 'of', '--', 'but', 'we', 'have', 'been', '--', 'for', 'the', 'of', 'this', 'they', 'have', 't
o', 'a', 'that', 'are', 'to', 'be', 'in', 'other', 'than', 'to', 'our', 'and', 'our', 'is', 'this', 'of', 'a', 'which', 'at',
'on', 'through', 'it', 'the', 'of', 'its', 'has', 'been', 'the', 'of', 'the', 'of', 'the', 'of', 'the', 'to', 'of', 'of', 'is',
'of', 'the', 'of', 'have', 'their', 'to', 'is', 'not', 'it', 'because', 'we', 'have', 'it', 'it', '--', 'because', 'it', 'is',
'on', 'the', 'of', 'and', 'in', 'a', '--', 'an', 'and', 'through', 'by', 'the', 'of', 'a', 'it', 'because', 'of', 'all', 'of',
'the', 'of', 'it', 'because', 'has', 'an', 'of', 'in', 'the', 'of', 'it', 'if', 'we', 'below', 'the', 'we', 'it', 'on', '--',
'for', 'it', 'is', 'the', 'most', 'the', 'most', 'and', 'in', 'the', 'the', 'most', 'of', 'all', 'of', 'a', 'has', 'a', 'that',
'be', 'and', 'and', 'and', 'in', 'a', 'that', 'up', 'to', 'the', 'of', 'our', 'a', 'has', 'a', '--', 'a', 'that', 'be', 'and',
'that', 'that', 'the', 'and', 'the', 'of', 'its', '--', 'all', 'the', 'other', 'that', 'the', 'of', 'the', 'a', 'a', 'has', 'mo
re', 'than', 'the', 'of', 'all', 'its', 'is', 'that', 'which', 'most', 'to', 'its', '--', 'which', 'the', 'most', 'of', 'its',
'is', 'a', 'for', 'which', 'we', 'it', '--', '--', 'to', 'a', 'we', 'all', 'what', 'it', 'is', '--', 'the', '--', 'the', 'of',
'is', 'the', 'of', 'was', 'in', 'the', 'of', 'those', 'who', 'from', '--', 'some', 'of', 'but', 'who', 'and', 'to', 'more', 'i
s', 'no', 'in', 'is', 'the', 'of', 'in', 'the', 'was', 'in', 'the', 'its', 'has', 'been', 'has', 'been', 'the', 'in', 'all', 't
o', 'all', 'not', 'because', 'this', 'was', 'a', 'but', 'because', 'all', 'those', 'who', 'here', 'they', 'this', 'a', '--',
'a', 'that', 'should', 'be', 'in', 'was', 'into', 'our', 'own', 'into', 'the', 'of', 'into', 'the', 'of', 'the', 'into', 'the',
'who', 'here', 'to', 'out', 'the', 'of', 'their', 'and', 'the', 'who', 'and', 'the', 'that', 'from', 'them', '--', 'all', 'hav
e', 'and', 'an', 'which', 'in', 'itself', 'has', 'and', 'with', 'each', 'of', 'the', 'or', 'that', 'we', 'have', 'to', 'that',
'we', 'more', 'the', 'and', 'the', 'and', 'the', 'of', 'in', 'the', 'by', 'the', 'and', 'the', 'of', 'the', 'it', 'is', 'not',
'to', 'these', 'is', 'not', 'to', 'and', 'the', 'of', 'this', 'and', 'and', 'its', 'there', 'is', 'the', 'of', 'the', 'the', 'i
s', 'the', 'the', 'and', 'the', 'as', 'all', 'the', 'not', 'if', 'the', 'of', 'were', 'the', 'and', 'in', 'an', 'the', 'we', 'h
ave', '--', 'that', '--', 'to', 'in', 'our', 'in', 'because', 'they', 'so', 'to', 'here', 'in', 'the', 'of', 'the', 'to', 'thro
ugh', 'the', 'of', 'in', 'the', 'of', 'to', 'in', 'our', 'in', 'our', 'in', 'our', 'and', 'in', 'our', 'to', 'from', 'the', 'ot
her', 'of', 'the', 'and', 'from', 'those', 'the', '--', 'the', 'as', 'as', 'the', 'we', 'to', 'or', 'these', 'of', 'because',
'to', 'the', 'of', 'our', 'is', 'such', 'an', 'of', 'was', 'in', 'of', 'by', 'our', 'in', 'his', 'in', '--', 'it', 'to', 'thi
s', 'of', 'of', 'the', 'of', 'and', 'the', 'of', 'the', 'of', 'are', 'on', 'the', 'to', 'the', 'of', 'the', 'we', 'that', 'we',
'it', 'be', 'with', 'and', '--', 'then', 'we', 'the', 'which', 'so', 'and', 'so', 'to', 'of', 'the', 'and', 'of', 'the', 'and',
'the', 'for', 'that', 'we', 'in', 'the', 'of', 'the', 'of', 'before', 'our', 'is', 'to', 'and', 'to', 'the', 'of', 'this', 'w
e', 'the', 'of', 'and', 'the', 'of', 'do', 'not', 'are', 'not', 'to', 'we', 'in', 'the', 'of', 'our', 'by', 'the', 'will', 'o
f']
```

Stopwords in Roosevelt speech



```
['we', 'not', 'a', 'of', 'but', 'a', 'of', '--', 'an', 'as', 'as', 'a', '--', 'as', 'as', 'have', 'before', 'you', 'and', 'th
e', 'same', 'our', 'a', 'and', 'is', 'very', 'in', 'his', 'the', 'to', 'all', 'of', 'and', 'all', 'of', 'the', 'same', 'for',
'which', 'our', 'are', 'at', 'the', '--', 'the', 'that', 'the', 'of', 'not', 'from', 'the', 'of', 'the', 'but', 'from', 'the',
'of', 'not', 'that', 'we', 'are', 'the', 'of', 'that', 'the', 'from', 'this', 'and', 'to', 'and', 'that', 'the', 'has', 'been',
'to', 'a', 'of', '--', 'in', 'this', 'by', 'by', 'a', 'and', 'of', 'our', '--', 'and', 'to', 'or', 'the', 'of', 'those', 'to',
'which', 'this', 'has', 'been', 'and', 'to', 'which', 'we', 'are', 'at', 'and', 'the', 'it', 'us', 'or', 'that', 'we', 'any',
'any', 'any', 'any', 'any', 'in', 'to', 'the', 'and', 'the', 'of', 'we', '--', 'and', 'those', 'and', 'we', 'we', 'the', 'of',
'there', 'is', 'we', 'do', 'in', 'a', 'of', 'there', 'is', 'we', 'can', 'do', '--', 'for', 'we', 'not', 'a', 'at', 'and', 'thos
e', 'whom', 'we', 'to', 'the', 'of', 'the', 'we', 'our', 'that', 'of', 'not', 'have', 'to', 'be', 'by', 'a', 'more', 'not', 't
o', 'them', 'our', 'we', 'to', 'them', 'their', 'own', '--', 'and', 'to', 'in', 'the', 'those', 'who', 'by', 'the', 'of', 'th
e', 'up', 'those', 'in', 'the', 'and', 'the', 'to', 'the', 'of', 'we', 'our', 'to', 'them', 'for', 'is', '--', 'not', 'becaus
e', 'the', 'be', 'doing', 'not', 'because', 'we', 'their', 'but', 'because', 'it', 'is', 'a', 'the', 'who', 'are', 'it', 'the',
'few', 'who', 'are', 'our', 'of', 'our', 'we', 'a', '--', 'to', 'our', 'into', '--', 'in', 'a', 'for', '--', 'to', 'and', 'in',
'off', 'the', 'of', 'this', 'of', 'the', 'of', 'all', 'our', 'that', 'we', 'with', 'them', 'to', 'or', 'in', 'the', 'let', 'oth
er', 'that', 'this', 'to', 'the', 'of', 'its', 'own', 'that', 'of', 'the', 'our', 'in', 'an', 'where', 'the', 'of', 'have', 'th
e', 'of', 'we', 'our', 'of', 'it', 'from', 'a', 'for', '--', 'to', 'its', 'of', 'the', 'and', 'the', '--', 'and', 'to', 'the',
'in', 'which', 'its', 'to', 'those', 'who', 'themselves', 'our', 'we', 'not', 'a', 'but', 'a', 'that', 'both', 'the', 'for', 'b
efore', 'the', 'of', 'by', 'all', 'in', 'or', 'not', 'them', 'with', 'only', 'when', 'our', 'are', 'can', 'we', 'be', 'that',
'they', 'will', 'be', 'can', 'and', 'of', 'from', 'our', '--', 'both', 'by', 'the', 'of', 'both', 'by', 'the', 'of', 'the', 'bo
th', 'to', 'that', 'of', 'that', 'the', 'of', 'let', 'us', '--', 'on', 'both', 'that', 'is', 'not', 'a', 'of', 'and', 'is', 't
o', 'us', 'out', 'of', 'let', 'us', 'to', 'both', 'what', 'us', 'of', 'those', 'which', 'both', 'for', 'the', 'and', 'for', 'th
e', 'and', 'of', '--', 'and', 'the', 'to', 'other', 'under', 'the', 'of', 'all', 'both', 'to', 'the', 'of', 'of', 'its', 'let',
'us', 'the', 'the', 'the', 'and', 'the', 'and', 'both', 'to', 'in', 'all', 'of', 'the', 'the', 'of', '--', 'to', 'the', 'and',
'to', 'let', 'the', 'if', 'a', 'of', 'the', 'of', 'let', 'both', 'in', 'a', 'not', 'a', 'of', 'but', 'a', 'of', 'where', 'the',
'are', 'just', 'and', 'the', 'and', 'the', 'this', 'will', 'not', 'be', 'in', 'the', 'will', 'it', 'be', 'in', 'the', 'nor', 'i
n', 'the', 'of', 'this', 'nor', 'in', 'our', 'on', 'this', 'let', 'us', 'your', 'my', 'more', 'than', 'in', 'will', 'the', 'o
r', 'of', 'our', 'this', 'was', 'each', 'of', 'has', 'been', 'to', 'to', 'its', 'of', 'who', 'the', 'to', 'the', 'the', 'us',
'again', '--', 'not', 'as', 'a', 'to', 'we', 'not', 'as', 'a', 'to', 'we', 'are', '--', 'but', 'a', 'to', 'the', 'of', 'a', 'i
n', 'and', 'in', 'in', '--', 'a', 'against', 'the', 'of', 'and', 'we', 'against', 'these', 'a', 'and', 'and', 'and', 'that', 'c
an', 'a', 'more', 'for', 'all', 'you', 'in', 'that', 'the', 'of', 'the', 'only', 'a', 'few', 'have', 'been', 'the', 'of', 'in',
'its', 'of', 'do', 'not', 'from', 'this', '--', 'do', 'not', 'that', 'any', 'of', 'us', 'with', 'any', 'other', 'or', 'any', 'o
ther', 'the', 'the', 'which', 'we', 'to', 'this', 'will', 'our', 'and', 'all', 'who', 'it', '--', 'and', 'the', 'from', 'that',
'can', 'the', 'my', 'not', 'what', 'your', 'can', 'do', 'for', 'you', '--', 'what', 'you', 'can', 'do', 'for', 'your', 'of', 't
he', 'not', 'what', 'will', 'do', 'for', 'but', 'what', 'we', 'can', 'do', 'for', 'the', 'of', 'you', 'are', 'of', 'or', 'of',
'the', 'of', 'us', 'the', 'same', 'of', 'and', 'which', 'we', 'of', 'a', 'our', 'only', 'with', 'the', 'of', 'our', 'let', 'u
s', 'to', 'the', 'we', 'and', 'but', 'that', 'here', 'on', 'be', 'our']
```

Stopwords in Kennedy's speech



```
['and', 'my', 'of', 'this', 'and', 'we', 'we', 'here', 'was', 'in', 'by', 'the', 'of', 'and', 'of', 'at', 'we', 'here', 'we',
'on', 'the', 'of', 'a', 'of', 'in', 'the', 'before', 'us', 'we', 'that', 'us', 'that', 'this', 'we', 'are', 'about', 'to', 'wil
l', 'not', 'be', 'what', 'other', 'have', 'so', 'a', 'of', 'and', 'that', 'to', 'at', 'and', 'us', 'that', 'this', 'will', 'b
e', 'what', 'it', 'can', 'a', 'of', 'in', 'which', 'we', 'the', 'and', 'the', 'of', 'as', 'we', 'our', 'as', 'a', 'from', 'ou
r', 'for', 'to', 'our', 'and', 'by', 'our', 'to', 'and', 'to', 'we', 'were', 'to', 'the', 'for', 'a', 'and', 'more', 'of', 'th
e', 'of', 'the', 'of', 'will', 'be', 'as', 'the', 'of', 'the', 'the', 'of', 'a', 'in', 'the', 'we', 'in', 'the', 'is', 'not',
'the', 'which', 'is', 'an', 'between', 'but', 'a', 'which', 'can', 'for', 'to', 'is', 'that', 'we', 'both', 'the', 'and', 'th
e', 'of', 'in', 'that', 'we', 'in', 'to', 'the', 'there', 'will', 'be', 'no', 'we', 'in', 'to', 'there', 'will', 'be', 'no', 'l
et', 'us', 'the', 'of', 'as', 'a', 'of', 'the', 'we', 'have', 'over', 'these', 'our', 'the', 'that', 'no', 'has', 'the', 'to',
'its', 'will', 'or', 'on', 'by', 'in', 'this', 'of', 'to', 'for', 'the', 'of', 'and', 'to', 'the', 'of', 'between', 'the', 'd
o', 'our', 'in', 'and', 'in', 'the', 'we', 'to', 'do', 'their', 'has', 'when', 'will', 'other', 'our', 'or', 'other', 'our', 'o
r', 'to', 'the', 'of', 'other', 'how', 'to', 'their', 'own', 'as', 'we', 'the', 'of', 'each', 'to', 'its', 'own', 'we', 'the',
'of', 'each', 'to', 'its', 'own', 'as', 'is', 'in', 'the', 'so', 'is', 'each', 'in', 'its', 'own', 'with', 'the', 'of', 'the',
'let', 'us', 'to', 'from', 'the', 'we', 'have', 'us', 'to', 'down', 'the', 'of', 'which', 'have', 'the', 'for', 'too', 'and',
'to', 'in', 'their', 'of', '--', 'so', 'that', 'between', 'of', 'the', 'of', 'the', 'can', 'be', 'us', 'a', 'of', 'in', 'the',
'in', 'which', 'the', 'are', 'as', 'as', 'the', '--', 'in', 'which', 'each', 'the', 'of', 'the', 'other', 'to', 'by', 'a', '--
', 'in', 'which', 'those', 'who', 'will', 'do', 'so', 'by', 'the', 'of', 'their', 'and', 'not', 'by', 'the', 'of', 'their', 'u
s', 'that', 'not', 'as', 'a', 'but', '--', 'because', 'the', 'to', 'such', 'a', 'is', 'the', 'in', 'which', 'a', 'can', 'becaus
e', 'only', 'if', 'we', 'in', 'our', 'will', 'we', 'a', 'and', 'only', 'if', 'we', 'a', 'will', 'we', 'in', 'our', 'at', 'hav
e', 'the', 'to', 'do', 'more', 'than', 'before', 'in', 'our', 'to', 'in', '--', 'to', 'a', '--', 'to', 'for', 'to', 'our', 'mor
e', '--', 'and', 'to', 'the', 'of', 'to', 'and', 'the', 'of', 'our', 'is', 'so', '--', 'because', 'the', 'of', 'our', 'is', 's
o', '--', 'let', 'us', 'be', 'in', 'our', 'to', 'those', 'in', 'as', 'a', 'of', 'has', 'from', 'that', 'so', 'a', 'of', 'at',
'from', 'that', 'have', 'the', 'from', 'to', 'has', 'not', 'been', 'a', 'from', 'our', 'but', 'a', 'to', 'at', 'the', 'from',
'to', 'will', 'not', 'be', 'a', 'from', 'our', 'but', 'a', 'to', 'and', 'at', 'the', 'to', 'those', 'in', 'the', 'and', 'the',
'of', 'have', 'too', 'with', 'the', 'of', 'to', 'all', 'and', 'in', 'and', 'at', 'the', 'has', 'to', 'from', 'the', 'of', '--',
'of', 'can', 'be', 'to', 'only', 'if', 'he', 'has', 'is', 'let', 'us', 'at', 'and', 'to', 'do', 'more', 'for', 'to', 'more', 'f
or', 'us', 'in', 'more', 'us', 'what', 'we', 'will', 'do', 'for', 'by', 'what', 'they', 'will', 'do', 'for', 'is', 'why', 'no',
'of', 'a', 'for', 'have', 'too', 'with', 'that', 'too', 'in', 'we', 'have', 'of', 'it', 'more', 'than', 'it', 'can', 'only', 't
o', 'to', 'and', 'to', 'a', 'and', 'that', 'both', 'in', 'what', 'can', 'do', 'and', 'in', 'what', 'can', 'to', 'from', 'so',
'that', 'an', 'do', 'more', 'for', 'us', 'that', 'was', 'not', 'by', 'but', 'by', '--', 'not', 'by', 'but', 'by', '--', 'not',
'by', 'but', 'by', 'our', 'own', 'let', 'each', 'of', 'us', '--', 'not', 'just', 'what', 'will', 'do', 'for', 'but', 'what', 'c
an', 'do', 'for', 'the', 'we', 'let', 'each', 'of', 'us', '--', 'not', 'just', 'how', 'can', 'but', 'how', 'can', 'has', 'a',
'and', 'to', 'to', 'you', 'that', 'where', 'this', 'should', 'we', 'will', 'and', 'we', 'will', 'just', 'as', 'is', 'the', 'tha
t', 'each', 'and', 'of', 'us', 'as', 'an', 'and', 'as', 'a', 'of', 'his', 'own', 'this', 'let', 'each', 'of', 'us', 'a', 'in',
'his', 'own', 'to', 'his', 'to', 'do', 'his', 'to', 'his', '--', 'so', 'that', 'we', 'can', 'the', 'of', 'a', 'of', 'for', 'an
d', 'as', 'we', 'our', 'as', 'a', 'we', 'can', 'do', 'so', 'in', 'the', 'of', 'our', 'to', 'ourselves', 'and', 'to', 'the', 'an
d', 'most', 'to', 'an', 'let', 'us', 'again', 'to', 'our', 'with', 'and', 'let', 'each', 'of', 'us', 'out', 'for', 'that', '--
', 'a', 'of', 'for', 'the', 'and', 'of', 'a', 'of', 'for', 'the', 'which', 'is', 'the', 'of', 'all', 'the', 'has', 'for', 'us',
'to', 'our', 'in', 'ourselves', 'and', 'in', 'that', 'has', 'been', 'have', 'been', 'to', 'be', 'of', 'their', 'of', 'their',
'of', 'at', 'and', 'of', 'its', 'in', 'the', 'we', 'have', 'been', 'by', 'those', 'who', 'with', 'and', 'that', 'is', 'am', 'th
at', 'this', 'will', 'not', 'be', 'the', 'of', 'on', 'these', 'in', 'which', 'we', 'are', 'to', 'in', 'this', 'has', 'been', 'i
n', 'the', 'for', 'its', 'for', 'its', 'for', 'its', 'and', 'for', 'its', 'us', 'be', 'that', 'our', 'has', 'and', 'more', 'an
d', 'more', 'more', 'than', 'any', 'other', 'in', 'the', 'of', 'the', 'us', 'be', 'that', 'in', 'each', 'of', 'the', 'in', 'whi
ch', 'we', 'have', 'been', 'in', 'this', 'the', 'we', 'are', 'now', 'to', 'an', 'we', 'have', 'not', 'for', 'our', 'but', 'to',
'us', 'be', 'that', 'by', 'our', 'and', 'by', 'our', 'for', 'with', 'we', 'have', 'a', 'in', 'the', 'what', 'the', 'has', 'no
t', 'before', '--', 'a', 'of', 'that', 'can', 'not', 'for', 'our', 'but', 'for', 'to', 'are', 'here', 'on', 'an', 'that', 'as',
'those', 'any', 'or', 'any', 'has', 'to', 'to', 'and', 'to', 'our', 'for', 'the', 'in', 'which', 'we', 'these', 'in', 'this',
'so', 'by', 'of', 'who', 'have', 'here', 'before', 'of', 'the', 'they', 'had', 'for', 'and', 'of', 'how', 'each', 'that', 'he',
'himself', 'in', 'to', 'those', 'your', 'that', 'in', 'the', 'have', 'in', 'that', 'are', 'for', 'and', 'for', 'your', 'so', 't
hat', 'we', 'be', 'of', 'our', 'us', 'to', 'these', 'the', 'in', 'so', 'that', 'on', 'its', 'will', 'be', 'as', 'and', 'as', 'a
s', 'when', 'it', 'and', 'as', 'a', 'of', 'for', 'all', 'the', 'us', 'from', 'here', 'in', 'in', 'our', 'in', 'by', 'our', 'i
n', 'who', 'and', 'to']
```

Stopwords in Nixon's speech

2.2 REMOVE THE STOPWORDS

	Text	word_count	count_stop	stop	upper
0	On each national day of inauguration since 178...	1360	654	[each, of, the, have, their, of, to, the, the,...	3

	Text	word_count	count_stop	stop	upper
0	Vice President Johnson, Mr. Speaker, Mr. Chief...	1390	661	[we, not, a, of, but, a, of, --, an, as, as, a...	5

	Text	word_count	count_stop	stop	upper
0	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	1819	950	[and, my, of, this, and, we, we, here, was, in...	14

THE THREE SPEECHES AFTER THE REMOVAL OF STOPWORDS

Roosevelt-1941

```
['national day inauguration since 178
9 people renewed sense dedication uni
ted states washingtons day task peopl
e create weld together nation lincoln
s day task people preserve nation dis
ruption within day task people save n
ation institutions disruption without
 come time midst swift happenings pau
se moment take stock recall place his
tory rediscover may risk real peril i
naction lives nations determined coun
t years lifetime human spirit life ma
n threescore years ten little little
less life nation fullness measure liv
e men doubt men believe democracy for
m government frame life limited measu
red kind mystical artificial fate une
xplained reason tyranny slavery becom
```

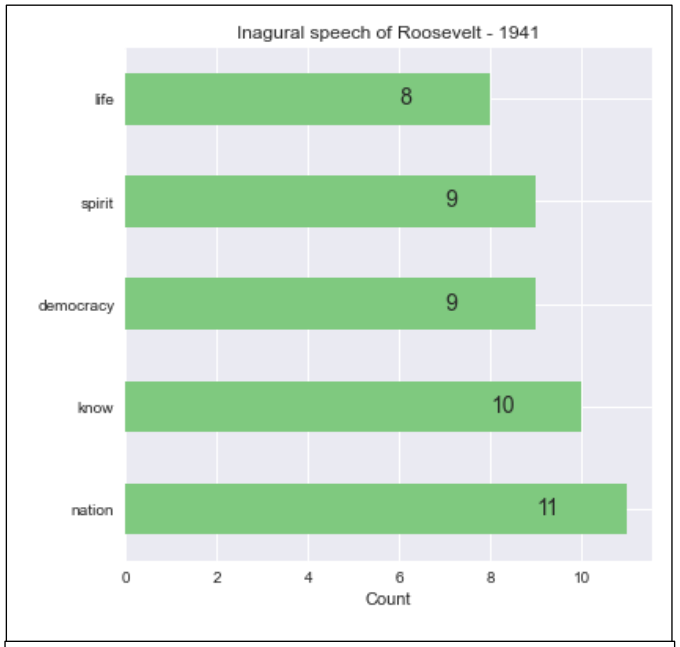
Kennedy-1961

```
['vice president johnson speaker chief j
ustice president eisenhower vice preside
nt nixon president truman reverend clerg
y fellow citizens observe today victory
party celebration freedom symbolizing en
d well beginning signifying renewal well
 change sworn almighty god solemn oath f
orebears l prescribed nearly century thr
ee quarters ago world different man hold
s mortal hands power abolish forms human
 poverty forms human life yet revolution
ary beliefs forebears fought still issue
 around globe belief rights man come gen
erosity state hand god dare forget today
 heirs first revolution word go forth ti
me place friend foe alike torch passed n
ew generation americans born century tem
pered war disciplined hard bitter peace
```

Nixon-1973

```
['vice president speaker chief justi
ce senator cook mrs eisenhower fello
w citizens great good country share
together met four years ago america
bleak spirit depressed prospect seem
ingly endless war abroad destructive
 conflict home meet today stand thre
shold new era peace world central qu
estion use peace resolve era enter p
ostwar periods often time retreat is
olation leads stagnation home invite
s new danger abroad resolve become t
ime great responsibilities greatly b
orne renew spirit promise america en
ter third century nation past year s
aw farreaching results new policies
peace continuing revitalize traditio
nal friendships missions peking mosc
```

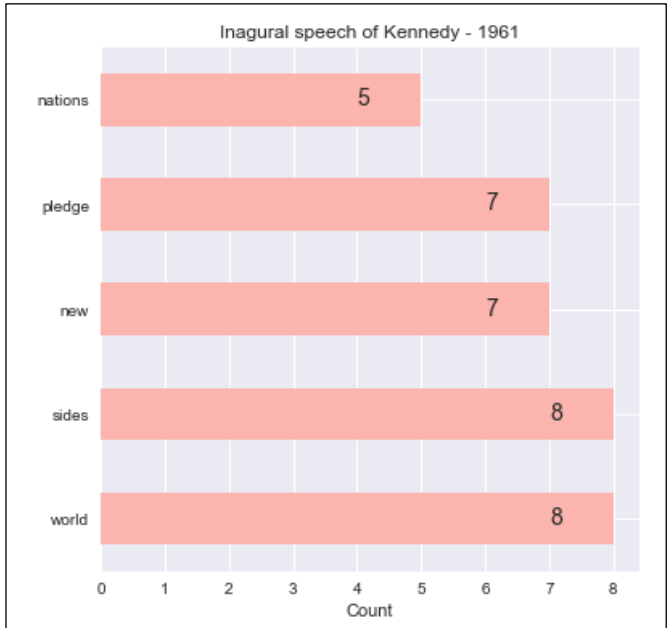
2.3 WHICH WORD OCCURS THE MOST NUMBER OF TIMES IN HIS INAUGURAL ADDRESS FOR EACH PRESIDENT? MENTION THE TOP THREE WORDS. (AFTER REMOVING THE STOPWORDS)



Top 3 words from Roosevelt's speech

```
[('nation', 11), ('know', 10), ('spirit', 9)]
```

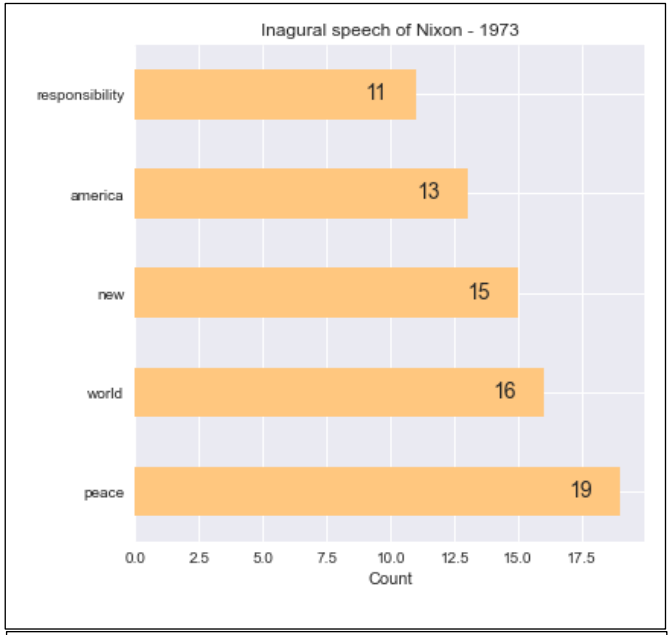
- In President Roosevelt's inaugural speech in 1941, 'nation' is the most often used word – 11 times, followed by 'know' occurring 10 times
- The words spirit and democracy came third, occurring 9 times. It was the Second World War time and he had to keep the nation together



Top 3 words from Kennedy's speech

```
[('world', 8), ('sides', 8), ('new', 7)]
```

- President Kennedy in his 1961 inaugural speech, used 'world' and 'sides' most often, 8 times. It was Cold War and the US was asking the world to take sides
- 'New' and 'pledge' were used 7 times each and 'citizens' used 5 times in the speech

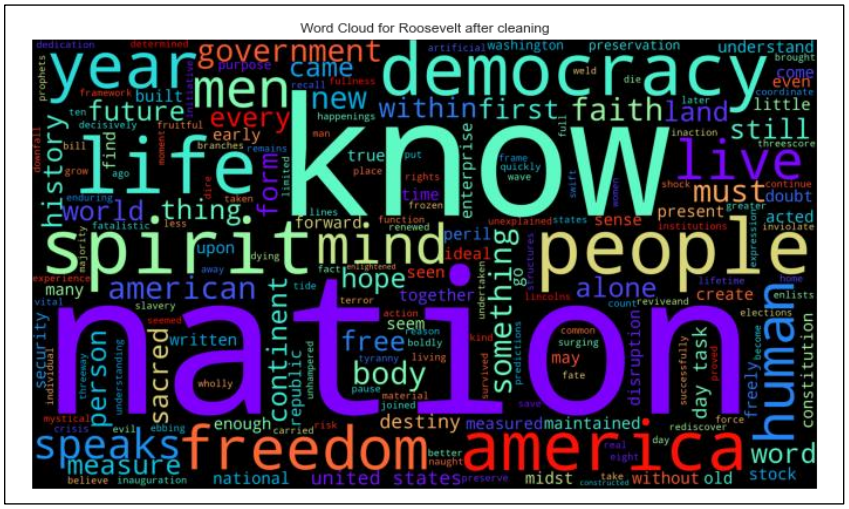


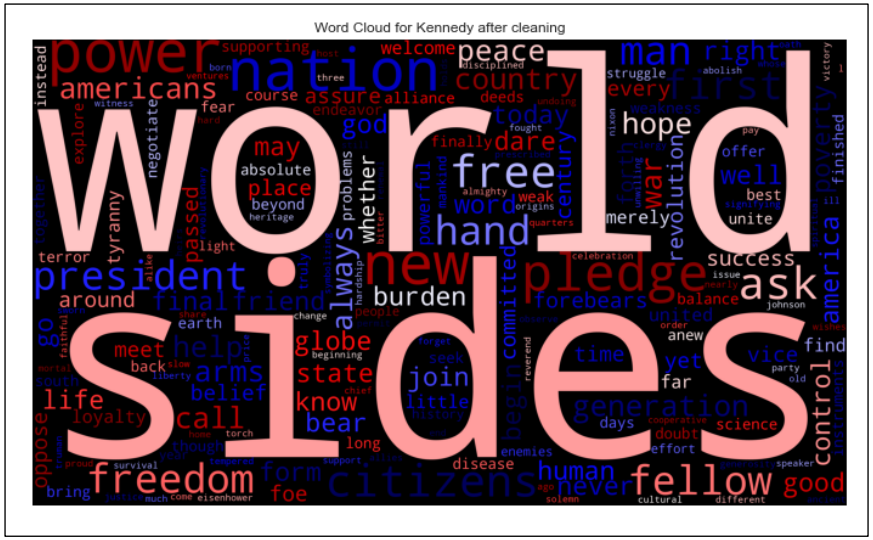
Top 3 words from Nixon's speech

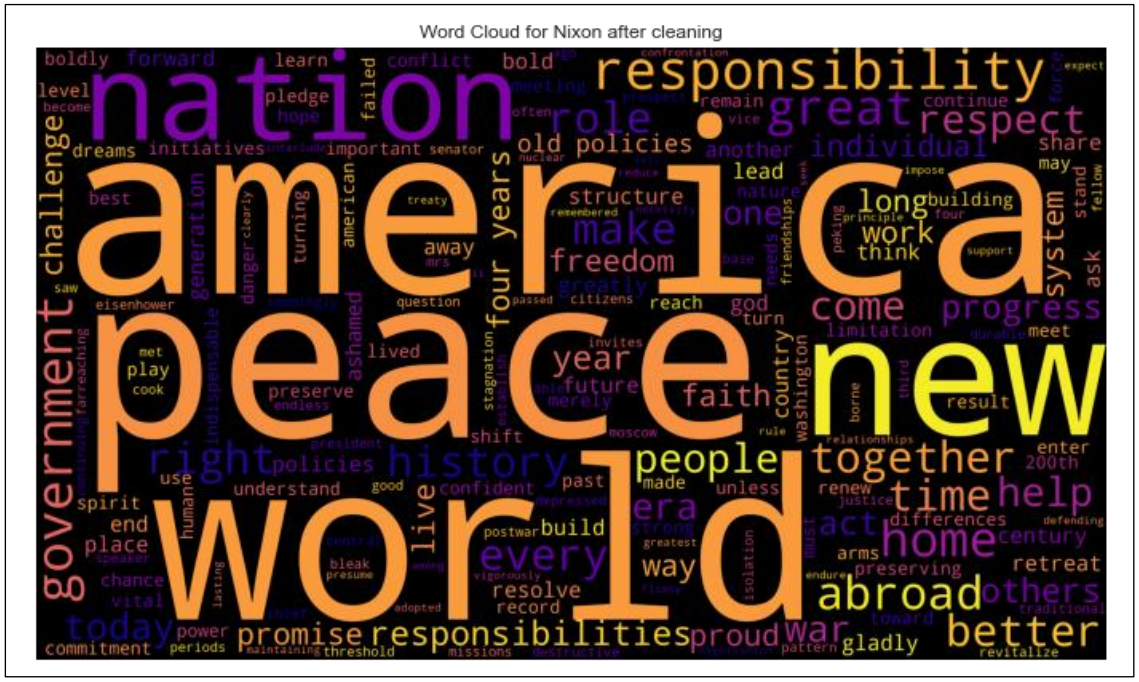
```
[('peace', 19), ('world', 16), ('new', 15)]
```

- In his 1973 inaugural speech, President Nixon used the word 'peace' 19 times, followed by 'world' 16 times and 'new' 15 times. The US was caught in a losing war in Vietnam
- President Nixon used the pronoun 'us' 26 times in this speech

2.4 PLOT THE WORD CLOUD OF EACH OF THE SPEECHES (AFTER REMOVING THE STOPWORDS)





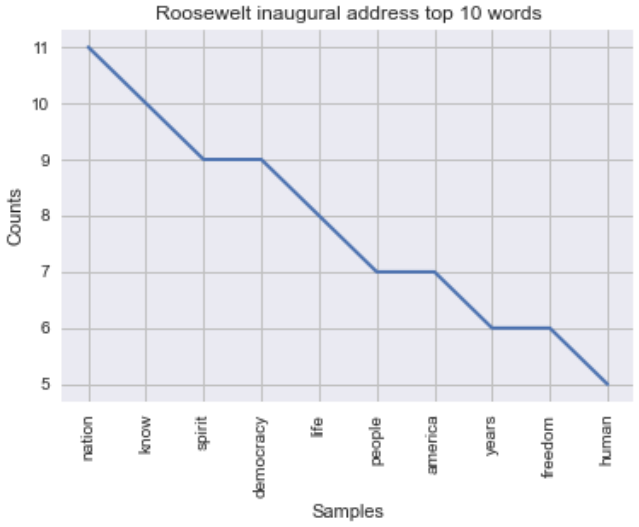


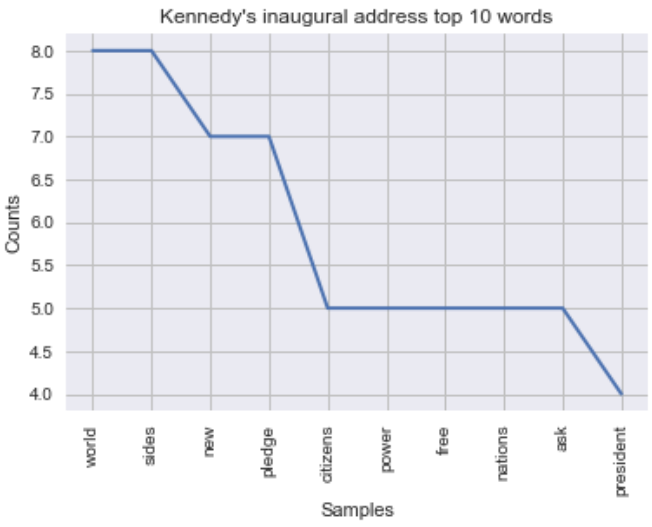
The word-cloud of the given inaugural speeches helps us draw inferences on the prevailing world order and the challenges of times in which these presidents took over the leadership of the USA

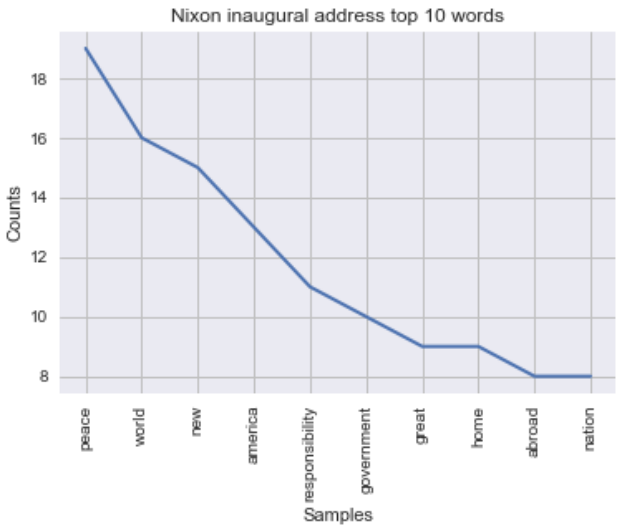
• Some of the most common stated words used by the Presidents Roosevelt, Kennedy and Nixon are ‘freedom’, ‘peace’, ‘world’, ‘nation’, and ‘new’

• When President Roosevelt took oath of office 1941, World War II was looming and America entered the war officially by the end of the year. The words ‘nation’, ‘democracy’, ‘freedom’, ‘spirit’, ‘life’ etc. are reflective of the periods (Roosevelt – 1941, Kennedy – 1961, Nixon – 1973)

• Kennedy took over the presidency at the height of Cold War with the USSR in 1961, where the world order had taken sides as blocs of powers. The words ‘world’, ‘sides’, ‘power, ‘new’, ‘freedom’, ‘peace’, ‘war’, ‘pledge’, ‘hope’ etc. are indicative of the new world order • In 1973 President Nixon won his second term after US called off offensive action in Vietnam. The words ‘peace’, ‘world’, ‘new’, ‘responsibility’, ‘freedom’, ‘history’ etc are reflective of the mandate he won by bringing peace to Vietnam


Roosewelt inaugural address top 10 words
Counts
Samples
nation
know
spirit
democracy
life
people
america
years
freedom
human


Kennedy's inaugural address top 10 words
Counts
Samples
world
sides
new
pledge
citizens
power
free
nations
ask
president


Nixon inaugural address top 10 words
Counts
Samples
peace
world
new
america
responsibility
government
great
home
abroad
nation