Project Report

on

Machine Learning

Prepared By

Mr. Jaydeep Bhaskar Ashtekar

PGPDSBA: Group 9

This report gives solution for given dataset.

A. Predicting election poll results

1.Exploratory Data Analysis of the dataset for exit poll

2. Performance of different models which predicts an exit poll that will help in predicting overall win and seats covered by a particular party.

B. Text analysis of the Speeches

1.Identifying number of sentences, words, characters etc.

2.Removing stopwords and forming worldcloud

Index

Problem 1:

You are hired by one of the leading news channel CNBE who wants to analyze recent elections. This survey was conducted on 1525 voters with 9 variables. You have to build a model, to predict which party a voter will vote for on the basis of the given information, to create an exit poll that will help in predicting overall win and seats covered by a particular party.

Dataset for Problem: Election_Data.xlsx

Data Ingestion: 12 marks

1. Read the dataset. Do the descriptive statistics and do null value condition check. Write an inference on it. (5 Marks)

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

Data Preparation: 5 marks

1. 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). (5 Marks)

Modelling: 26 marks

1. Apply Logistic Regression and LDA (linear discriminant analysis). (5 marks)

2. Apply KNN Model, Naïve Bayes Model and support vector machine (SVM) model. Interpret the results. (7 marks)

3. Model Tuning, Bagging and Boosting. (7 marks)

4. 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. Final Model: Compare the models and write inference which model is best/optimized. (7 marks)

Inference: 5 marks

1. Based on these predictions, what are the insights? (5 marks)

1.1. Read the dataset. Do the descriptive statistics and do null value condition check. Write an inference on it.

Ans:

1.df = pd.read_excel ('Election_Data.xlsx',sheet_name='Election_Dataset_Two Classes')

2.Data Description:

a.Numeric Data:

	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge
count	1525	1525	1525	1525	1525	1525	1525
mean	54.182295	3.245902	3.140328	3.334426	2.746885	6.728525	1.542295
std	15.711209	0.880969	0.929951	1.174824	1.230703	3.297538	1.083315
min	24	1	1	1	1	1	0
25%	41	3	3	2	2	4	0
50%	53	3	3	4	2	6	2
75%	67	4	4	4	4	10	2
max	93	5	5	5	5	11	3

For age column,mean and 50% has almost same values,so age seems to be normally distributed.Also it will have less/no outliers.

Remaining all are ordinal data type.

economic.cond.national:1 to 5

economic.cond.household:1 to 5

Blair:1 to 5

Hauge:1 to 5

Europe:1 to 11

b.Categorical Data:

	vote	gender

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

3.df.info()

```
<class 'pandas.core.frame.DataFrame'>
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
```

No null values in given data set.

4.df.isna().sum()

0

5.df.shape

```
(1525, 9)
```

```
Data set has 1525 rows and 9 columns
```

6.df.median()

```
age                         53.0
economic.cond.national       3.0
economic.cond.household      3.0
Blair                        4.0
Hague                        2.0
Europe                       6.0
political.knowledge          2.0
dtype: float64
```

7.df.mode()

vote	age	economic.cond. national	economic.cond. household	Blair	Hague	Europe	political.kno wledge	gender
Labour	37	3	3	4	2	11	2	female

Mode is most frequently repeated value in column which is shown in above table.

8.df.var()

Variance:

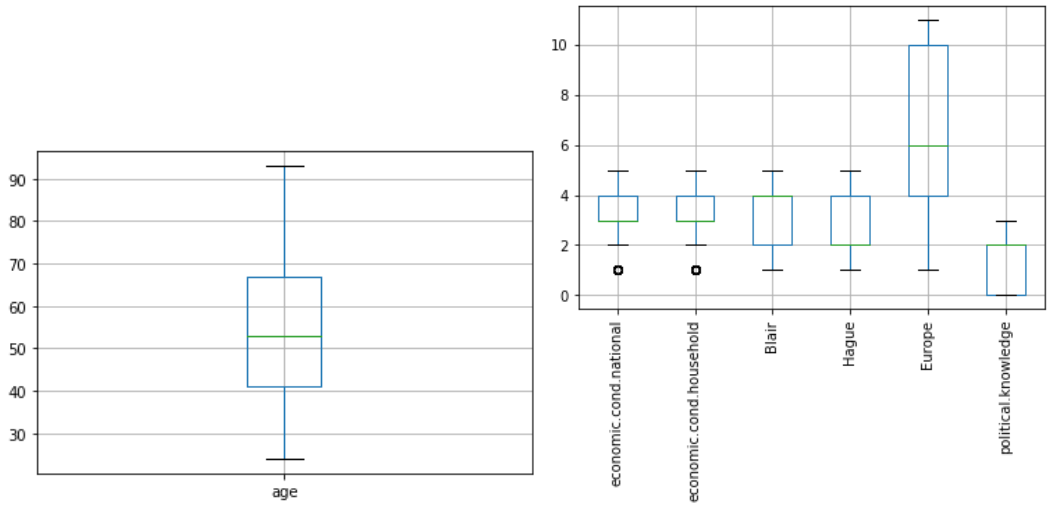
```
age                          246.842075
economic.cond.national         0.776107
economic.cond.household        0.864810
Blair                          1.380212
Hague                          1.514631
Europe                        10.873759
political.knowledge            1.173571
```

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

1.Univariate analysis:

1.boxplot

Few outliers in economi.condition.national and economi.condition.household



2.Countplot:

From following countplot:

1.More than 175 males and 150 females have 'Eurosceptic' sentiment (11 scale)

2.More than 300 males and females have moderate assessment of household economic conditions.

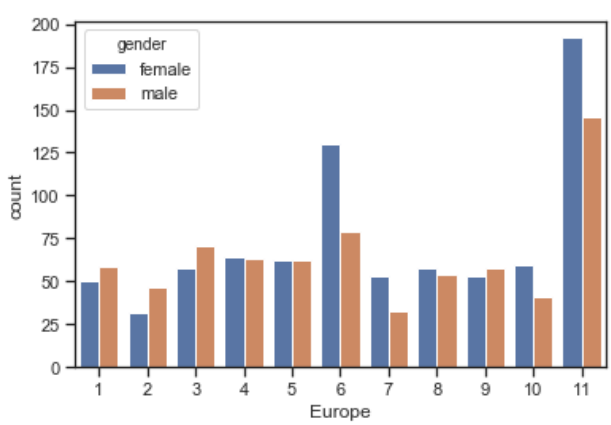
3.Most of males and females have assessment of household economic conditions in of 3 and 4.

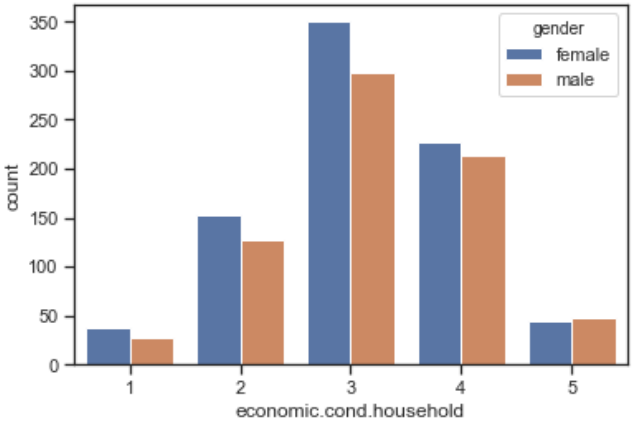
4. Maximum males and females have done Assessment of the Conservative leader for 2 and 4 and very less for 3.

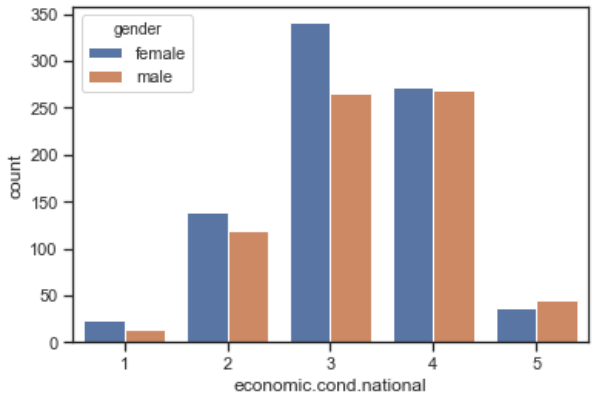
5. Maximum males and females have done Assessment of the Labour leader for 2 and 4 and zero for 3.

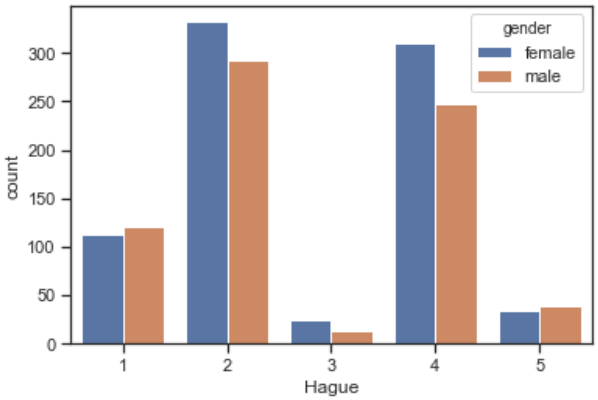
6.Proportion of female candidates is more than male candidates.

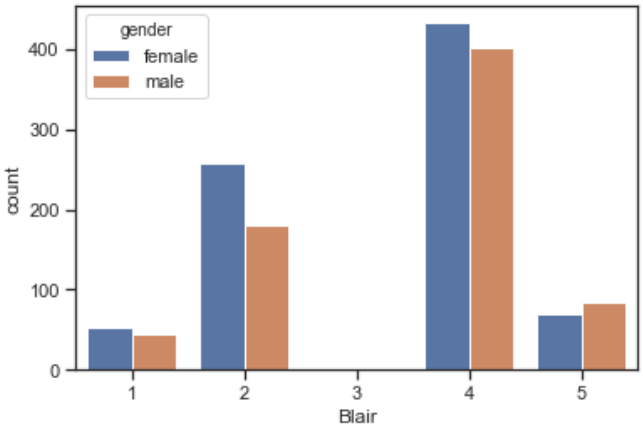
7. Maximum people have given 0 and 2 Knowledge level of parties' positions on European integration.

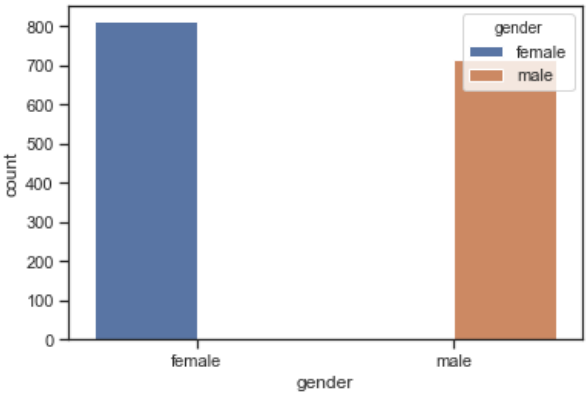


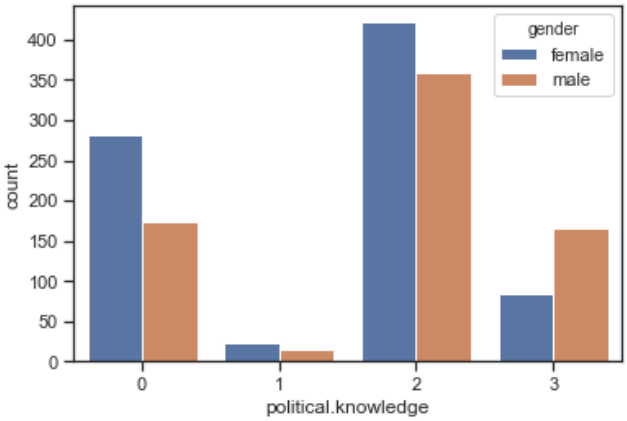




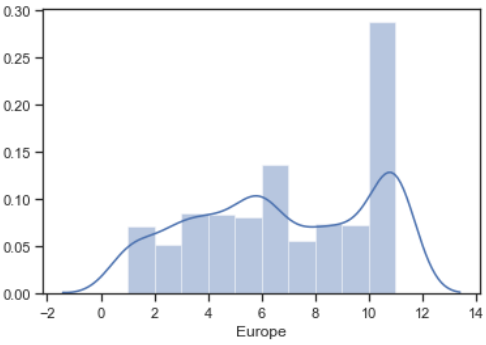


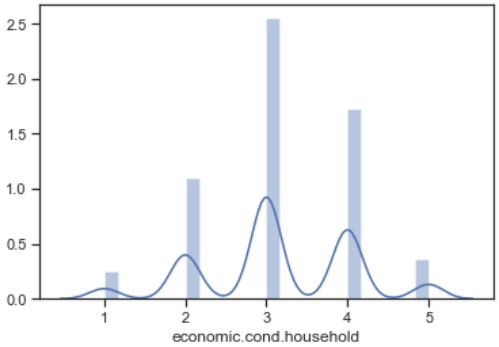


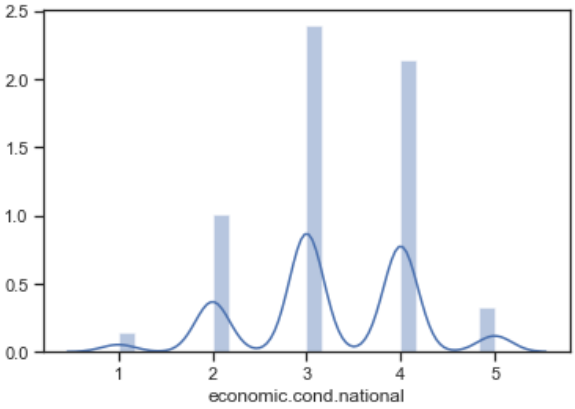


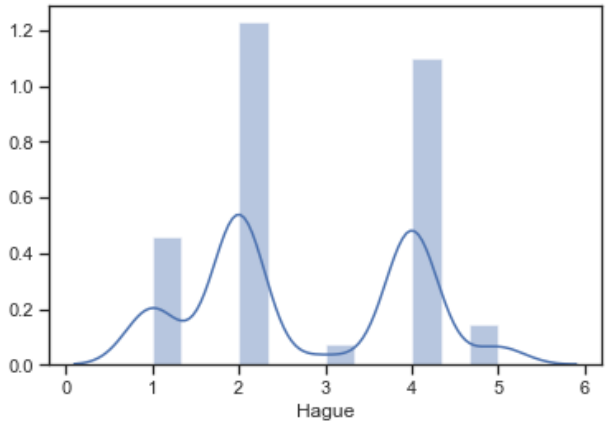


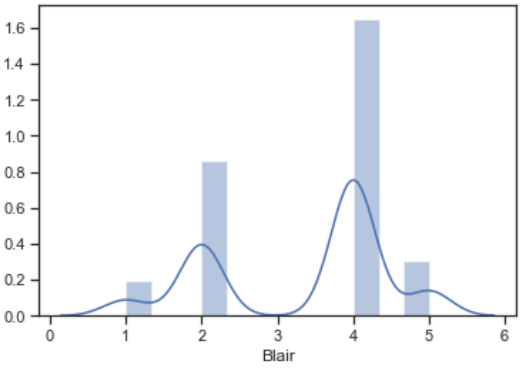
3.Frequency distribution with kde:

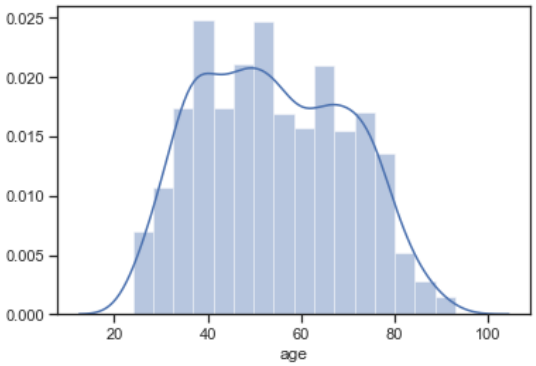


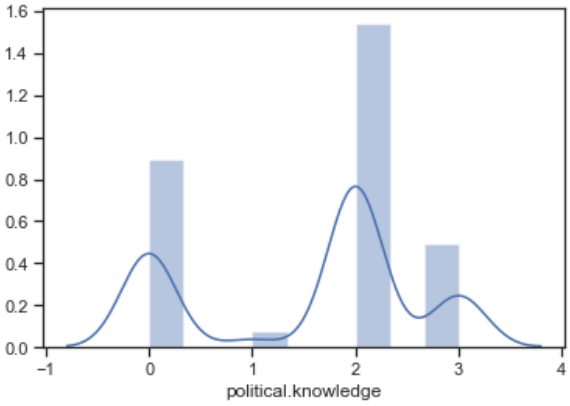












1.Europe is little multimodal with maximum value of 11

2.Maximum household is multimodal and not normally distributed.

3.National condition is multimodal and with maximum is 3

4.Hauge is again multimodal with 2 value maximum

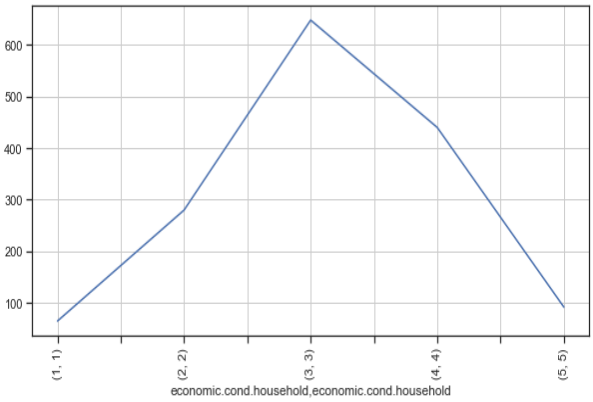
5.Blair is also multimodal with 4 as max value

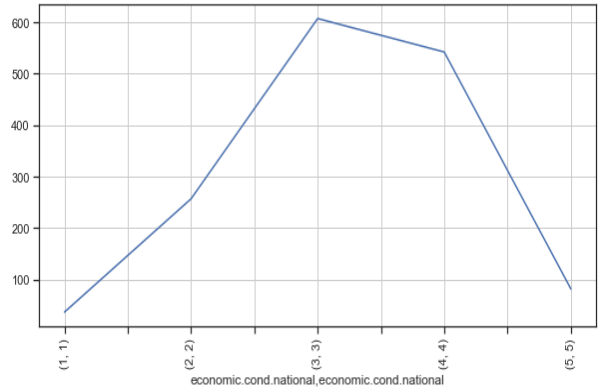
7.Age is a continuos data so gives approximately normal distribution

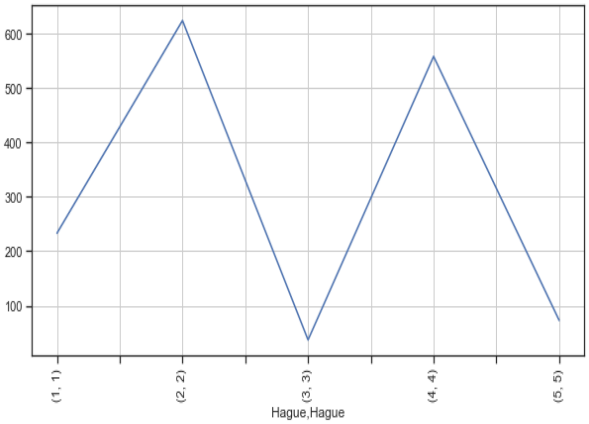
7.Political knowledge is multimodal with value 2 as maximum.

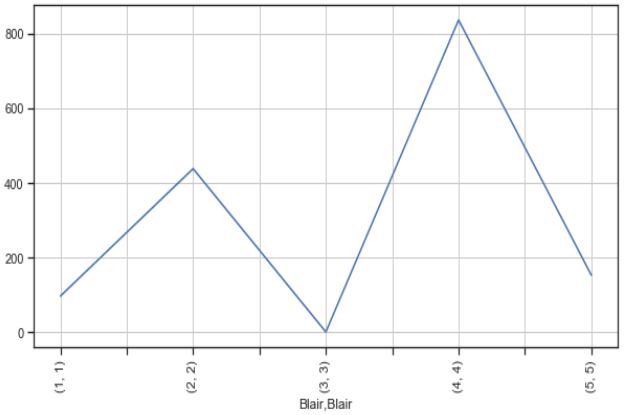
4.Line plot:

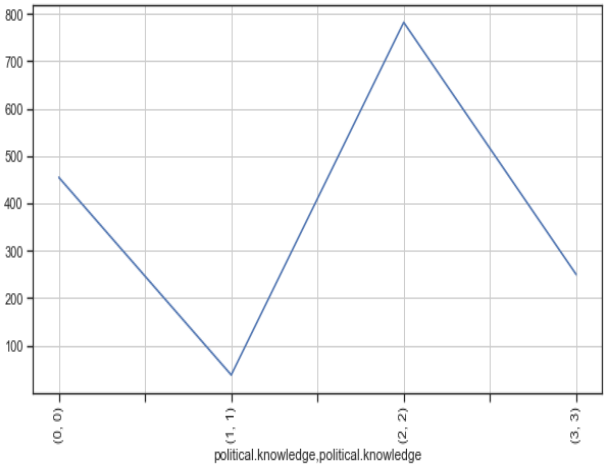
It gives count of each category in variable. Economic Household has maximum class 3.

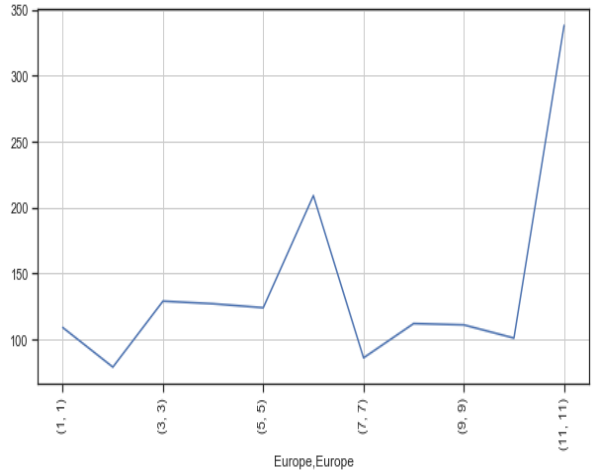

600
500
400
300
200
100
(1, 1)
(2, 2)
(3, 3)
(4, 4)
(5, 5)
economic.cond.household,economic.cond.household


600
500
400
300
200
100
(1, 1)
(2, 2)
(3, 3)
(4, 4)
(5, 5)
economic.cond.national,economic.cond.national

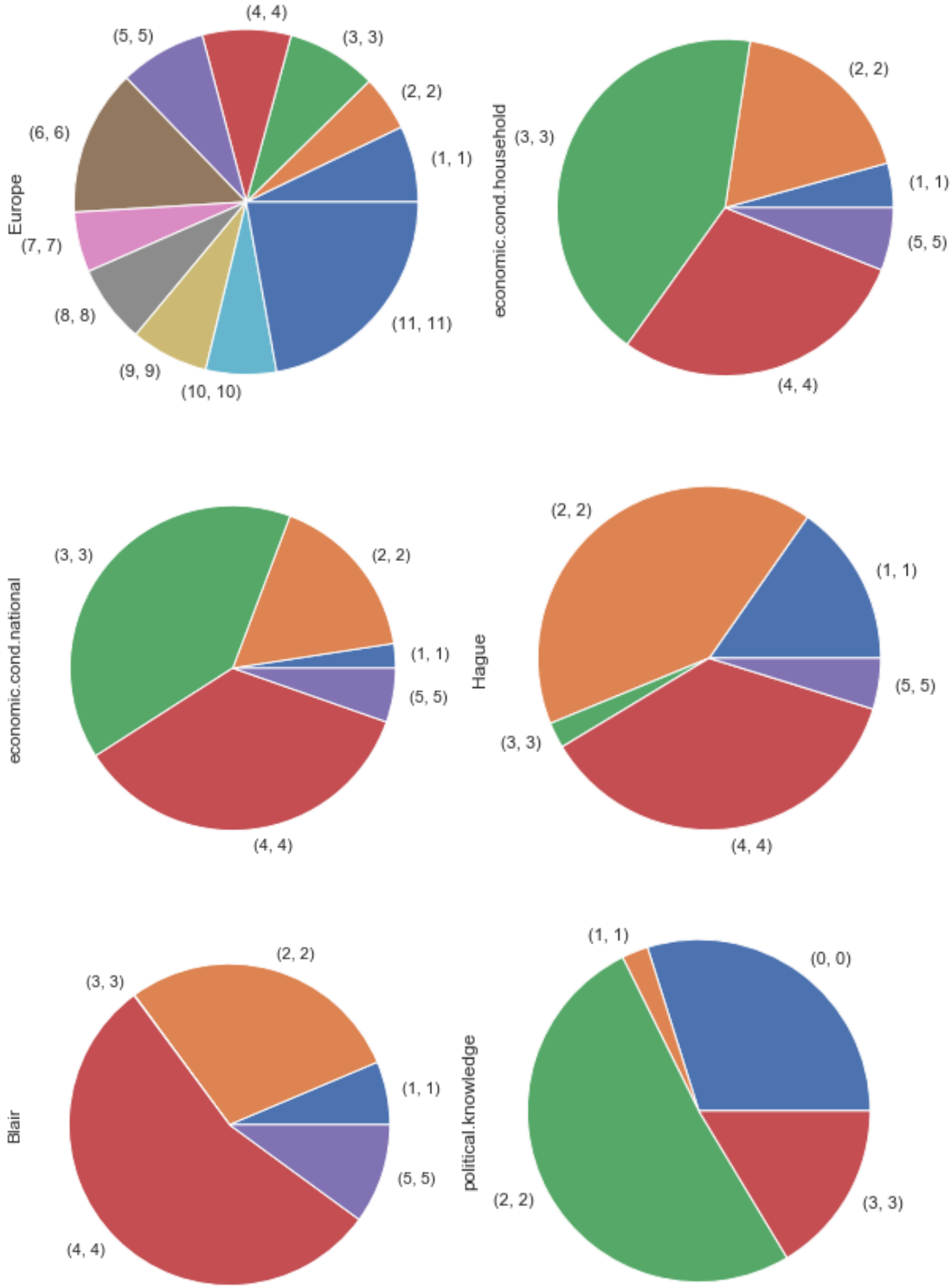

600
500
400
300
200
100
(1, 1)
(2, 2)
(3, 3)
(4, 4)
(5, 5)
Hague,Hague


800
600
400
200
0
(1, 1)
(2, 2)
(3, 3)
(4, 4)
(5, 5)
Blair,Blair


800
700
600
500
400
300
200
100
(0, 0)
(1, 1)
(2, 2)
(3, 3)
political.knowledge,political.knowledge


350
300
250
200
150
100
(1, 1)
(3, 3)
(5, 5)
(7, 7)
(9, 9)
(11, 11)
Europe,Europe

5.Pie plot: Shows the concentration of each variable of feature. More the area more is concentration/weightage.



2.Bivariate Analysis:

1.Pairplot:

As data is mostly ordinal so no any correlation is seen in any of the variable and it is multimodal.

Diagonal shows Europe and age follows approximate normal distribution.

2.Line plot:

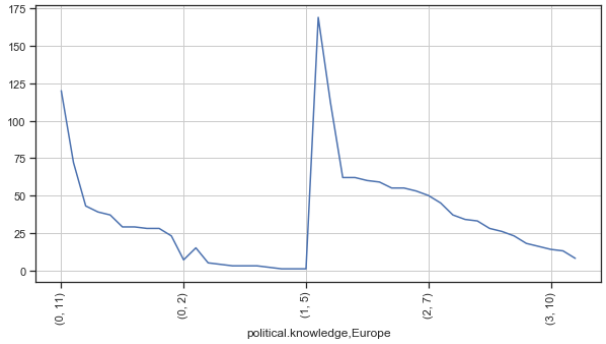
These graph gives the count of pair in each variable.

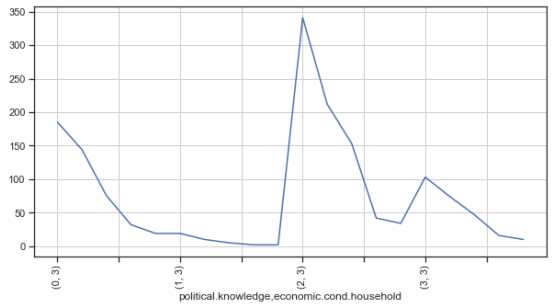
Example:In first graph of (1,5)

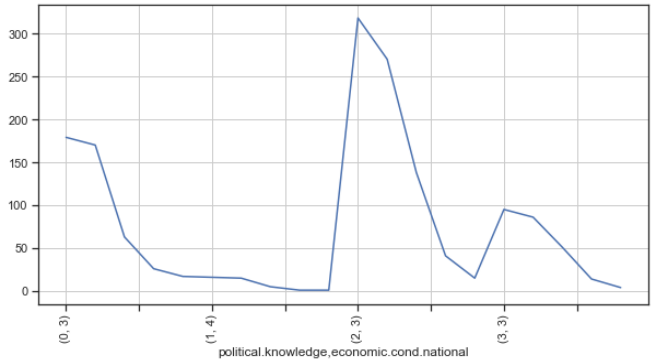
1: Political knowledge

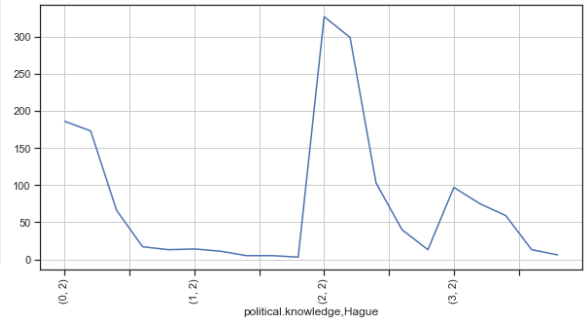
5: Europe

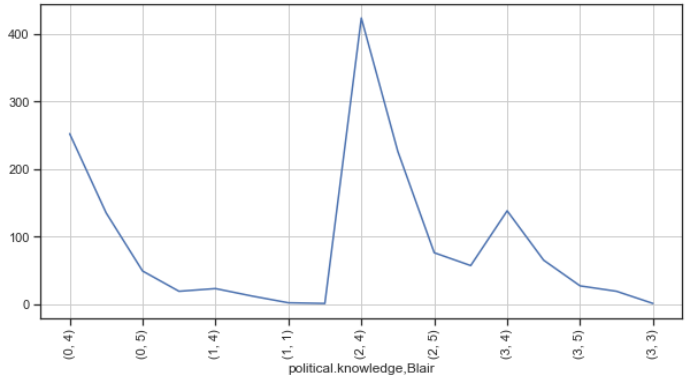
More than 150 peoples have pair of (1,5)

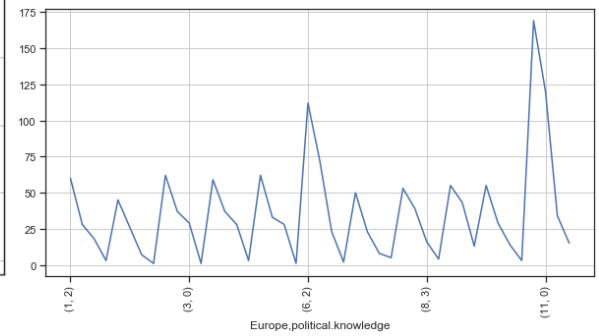


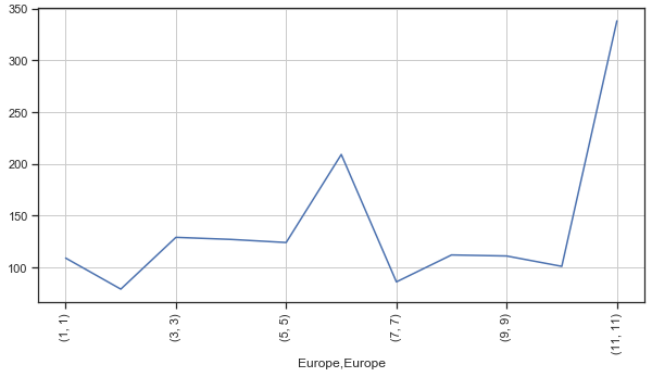


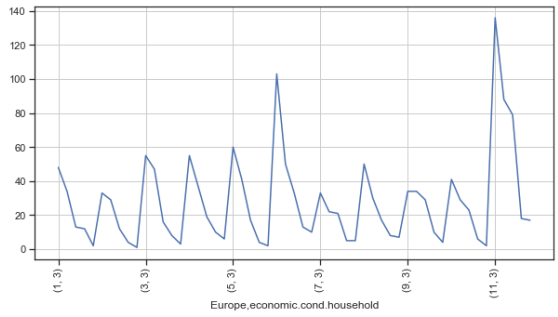


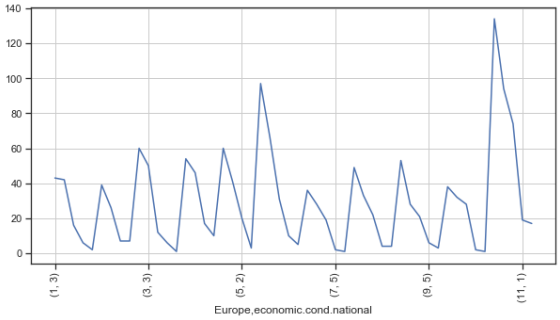


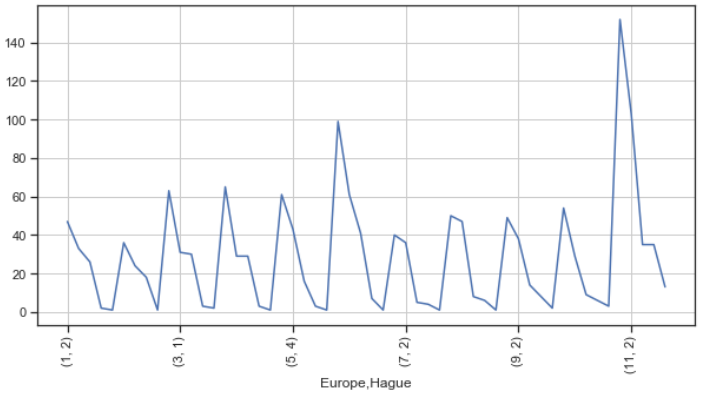


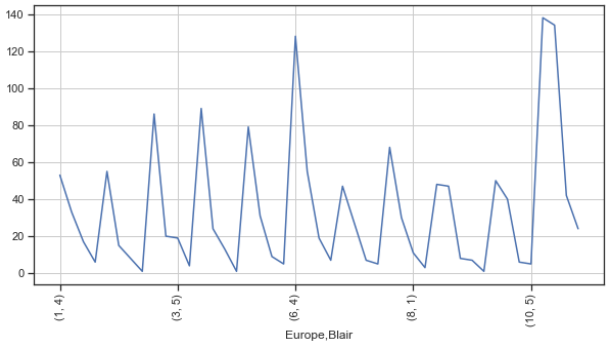


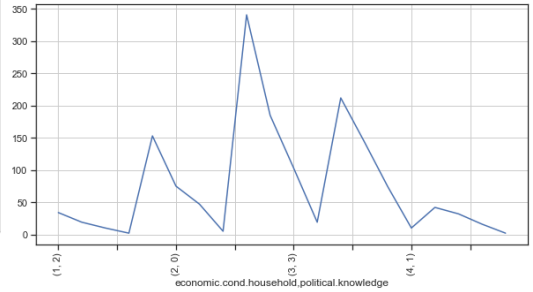


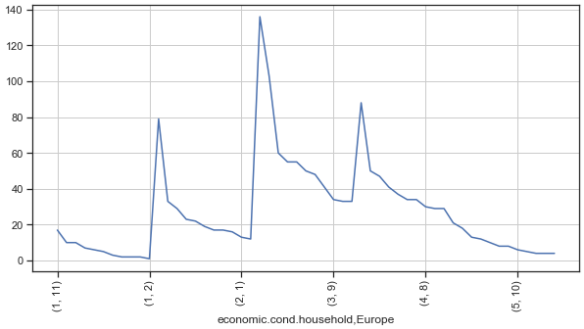


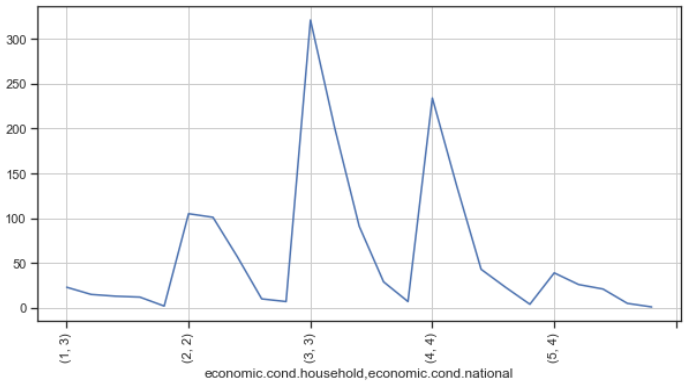

Europe,economic.cond.national

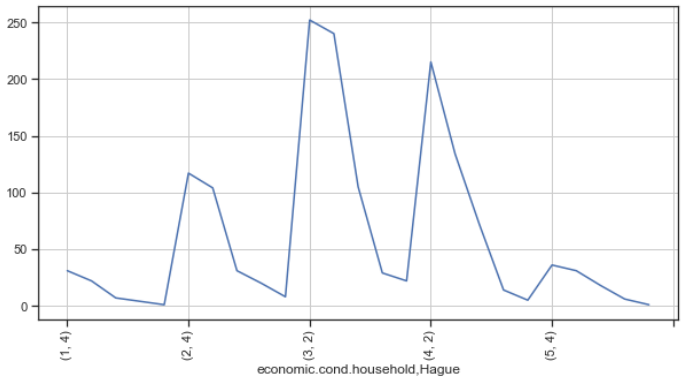

Europe,Hague

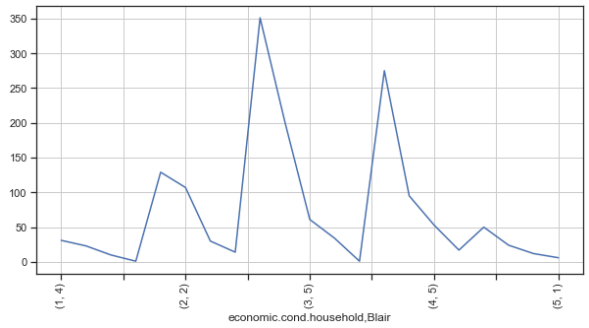

Europe,Blair

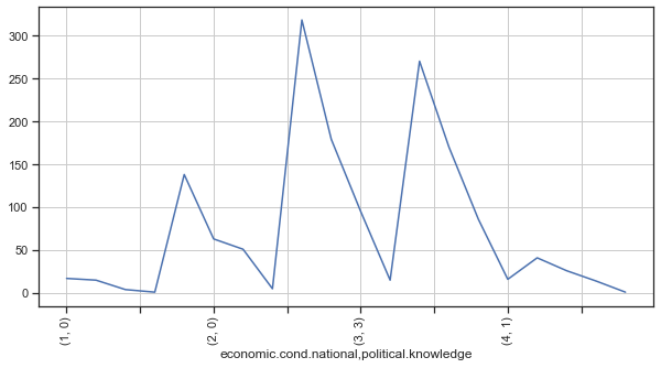

economic.cond.household,political.knowledge

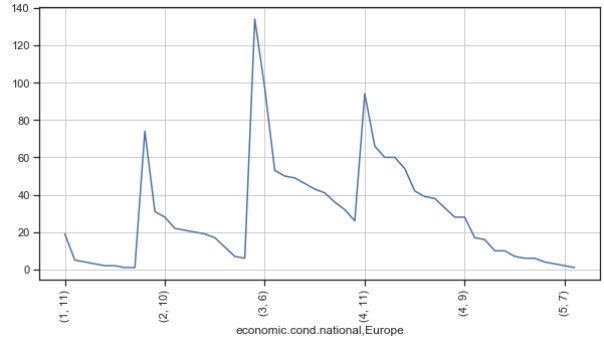

economic.cond.household,Europe

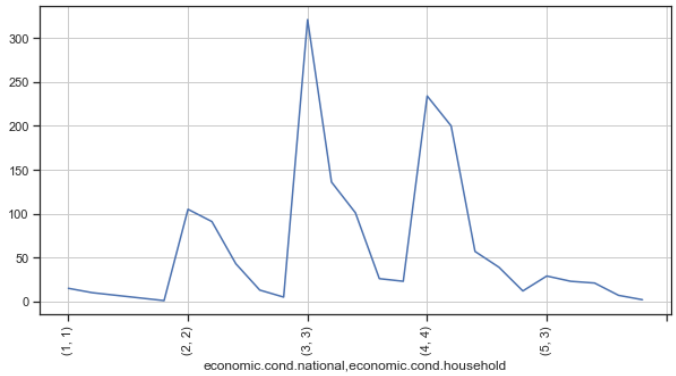

economic.cond.household,economic.cond.national

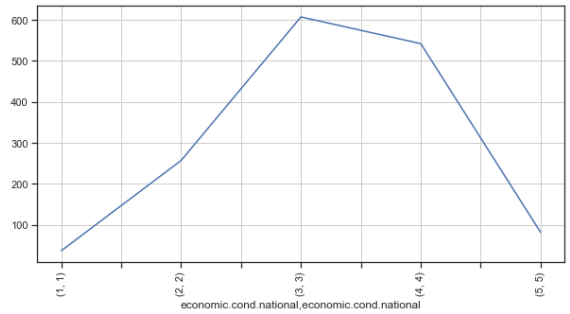

economic.cond.household,Hague

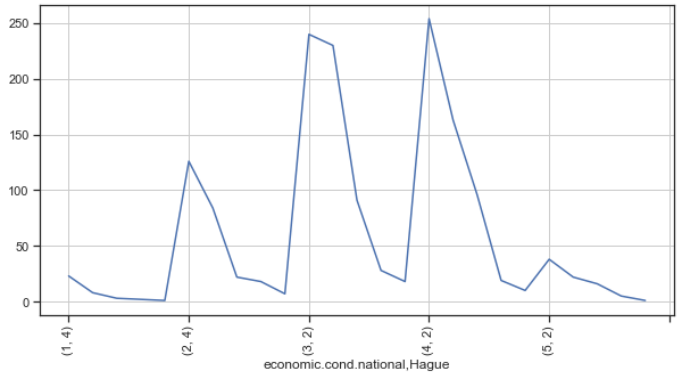

economic.cond.household,Blair

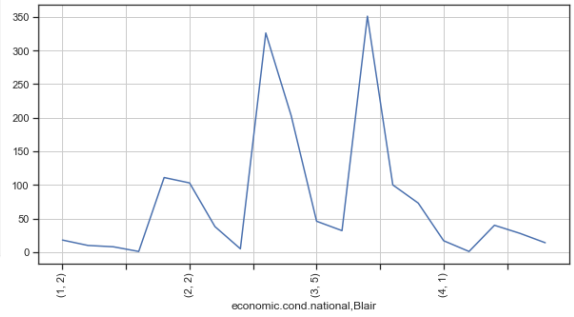

300
250
200
150
100
50
0
(1, 0)
(2, 0)
(3, 3)
(4, 1)
economic.cond.national,political.knowledge

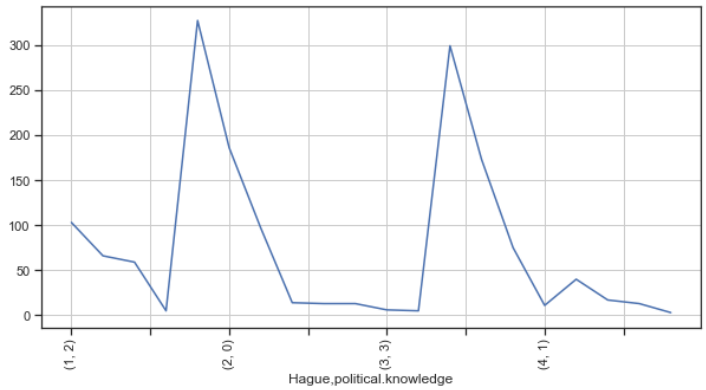

140
120
100
80
60
40
20
0
(1, 11)
(2, 10)
(3, 6)
(4, 11)
(4, 9)
(5, 7)
economic.cond.national,Europe

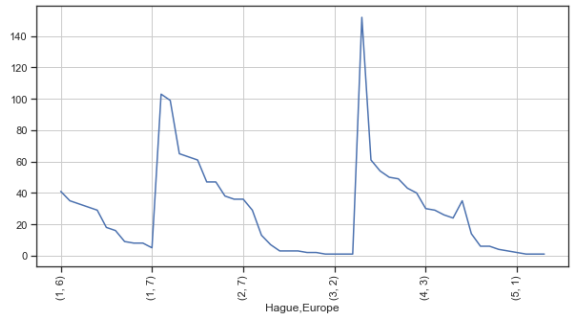

300
250
200
150
100
50
0
(1, 1)
(2, 2)
(3, 3)
(4, 4)
(5, 3)
economic.cond.national,economic.cond.household

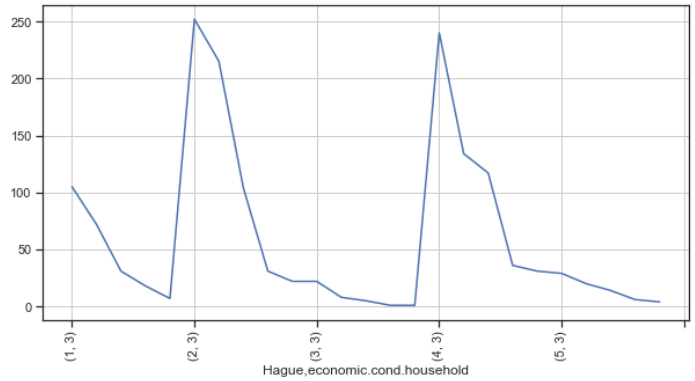

600
500
400
300
200
100
(1, 1)
(2, 2)
(3, 3)
(4, 4)
(5, 5)
economic.cond.national,economic.cond.national

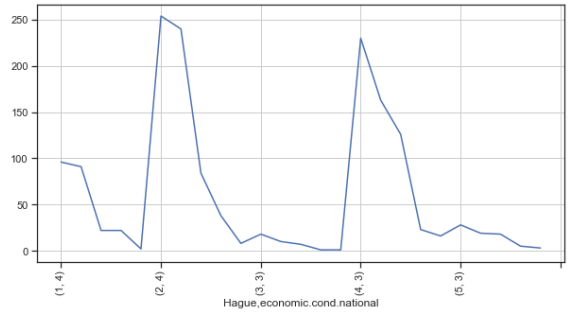

250
200
150
100
50
0
(1, 4)
(2, 4)
(3, 2)
(4, 2)
(5, 2)
economic.cond.national,Hague

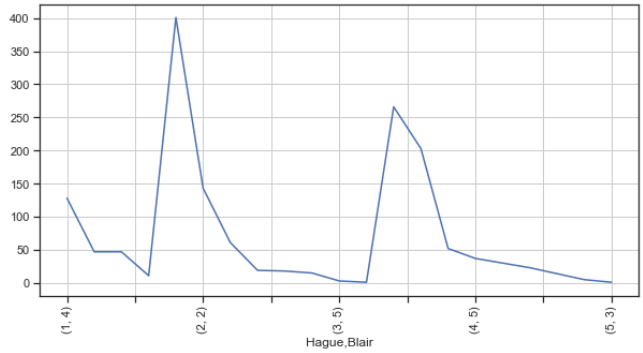

350
300
250
200
150
100
50
0
(1, 2)
(2, 2)
(3, 5)
(4, 1)
economic.cond.national,Blair

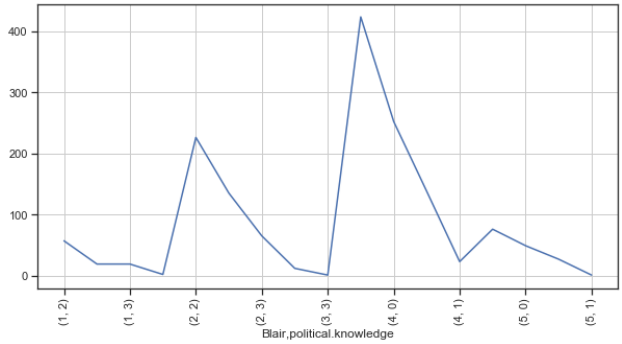

300
250
200
150
100
50
0
(1, 2)
(2, 0)
(3, 3)
(4, 1)
Hague,political.knowledge

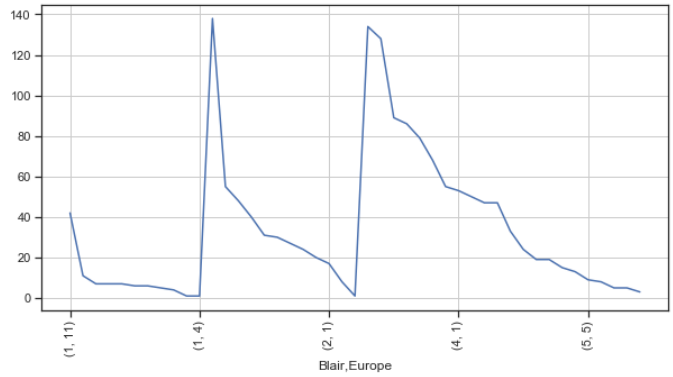

140
120
100
80
60
40
20
0
(1, 6)
(1, 7)
(2, 7)
(3, 2)
(4, 3)
(5, 1)
Hague,Europe

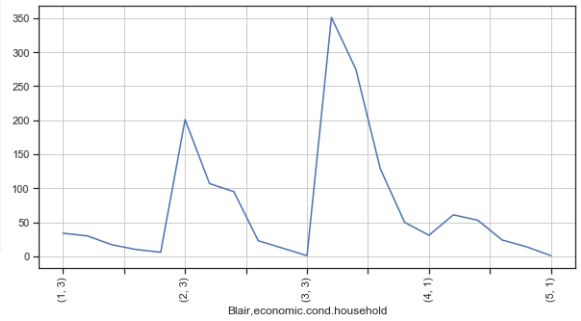


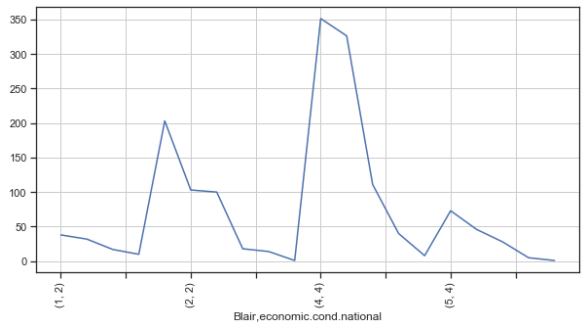


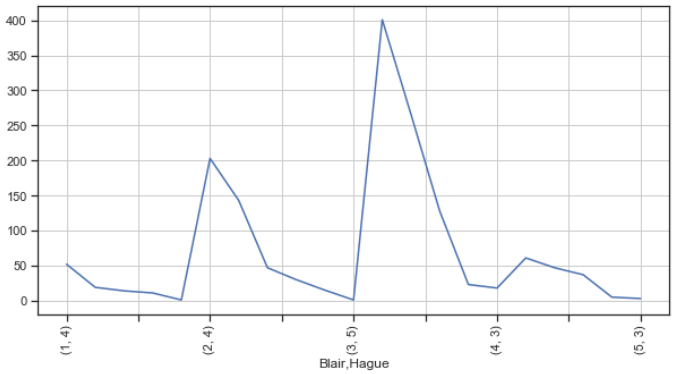






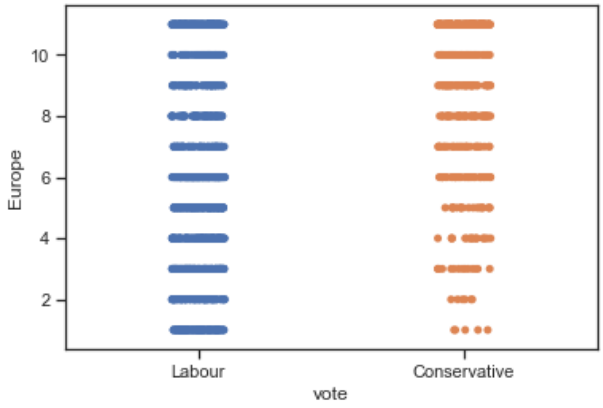


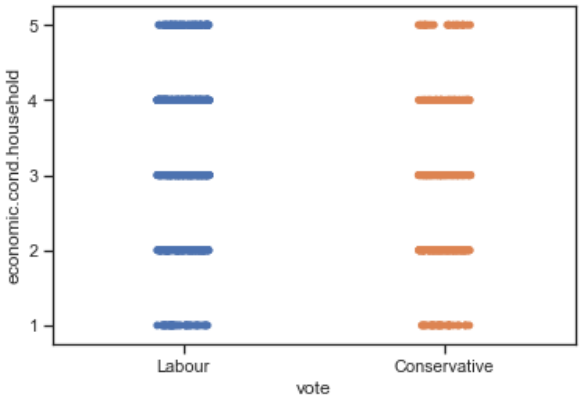


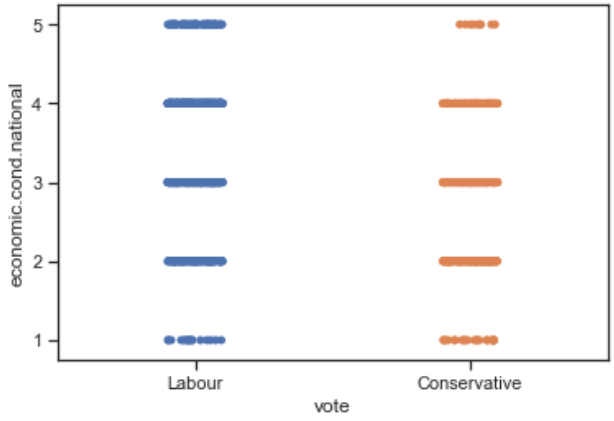


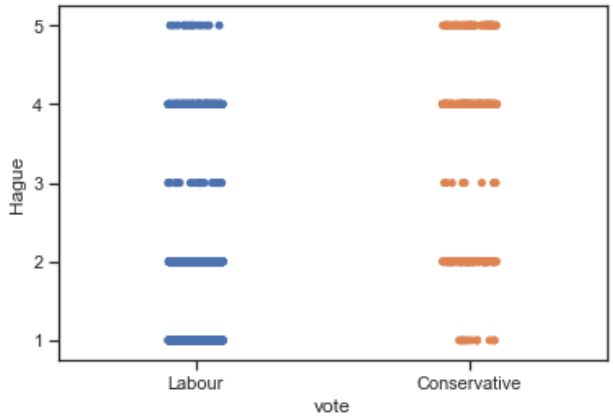
3.Stripplot:

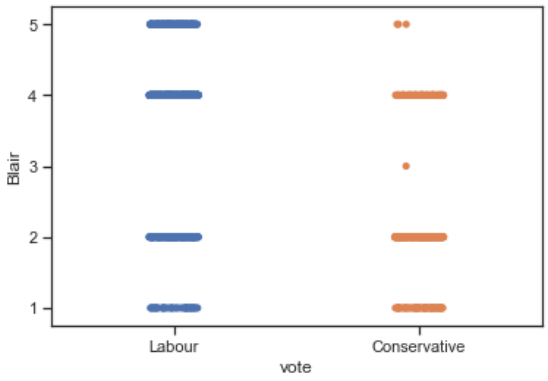
Stripplot gives all other variables vs vote.So for Europe has 11 categories.Those are distributed in vote.So,it can be concluded that,more points from Europe is for Labour.So category 1,2,3,4 and 5 has less count in Conservative.

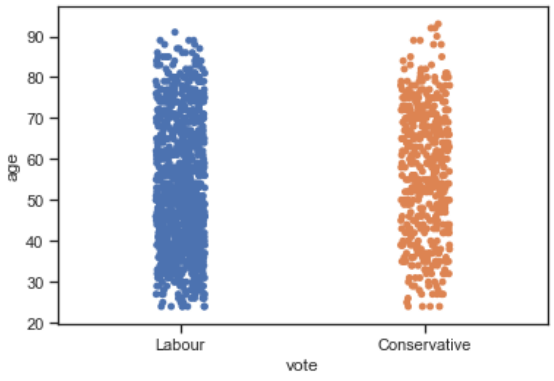

Europe
10
8
6
4
2
Labour
Conservative
vote

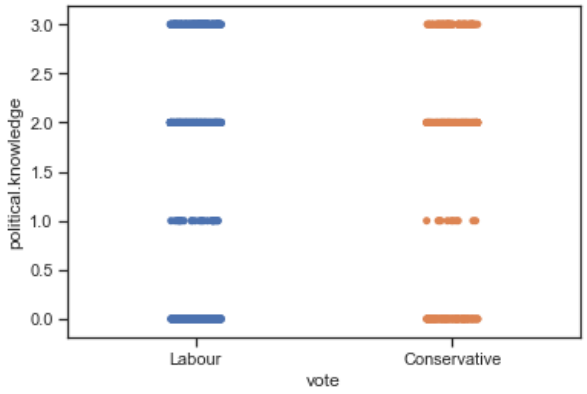

economic.cond.household
5
4
3
2
1
Labour
Conservative
vote


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


Hague
5
4
3
2
1
Labour
Conservative
vote

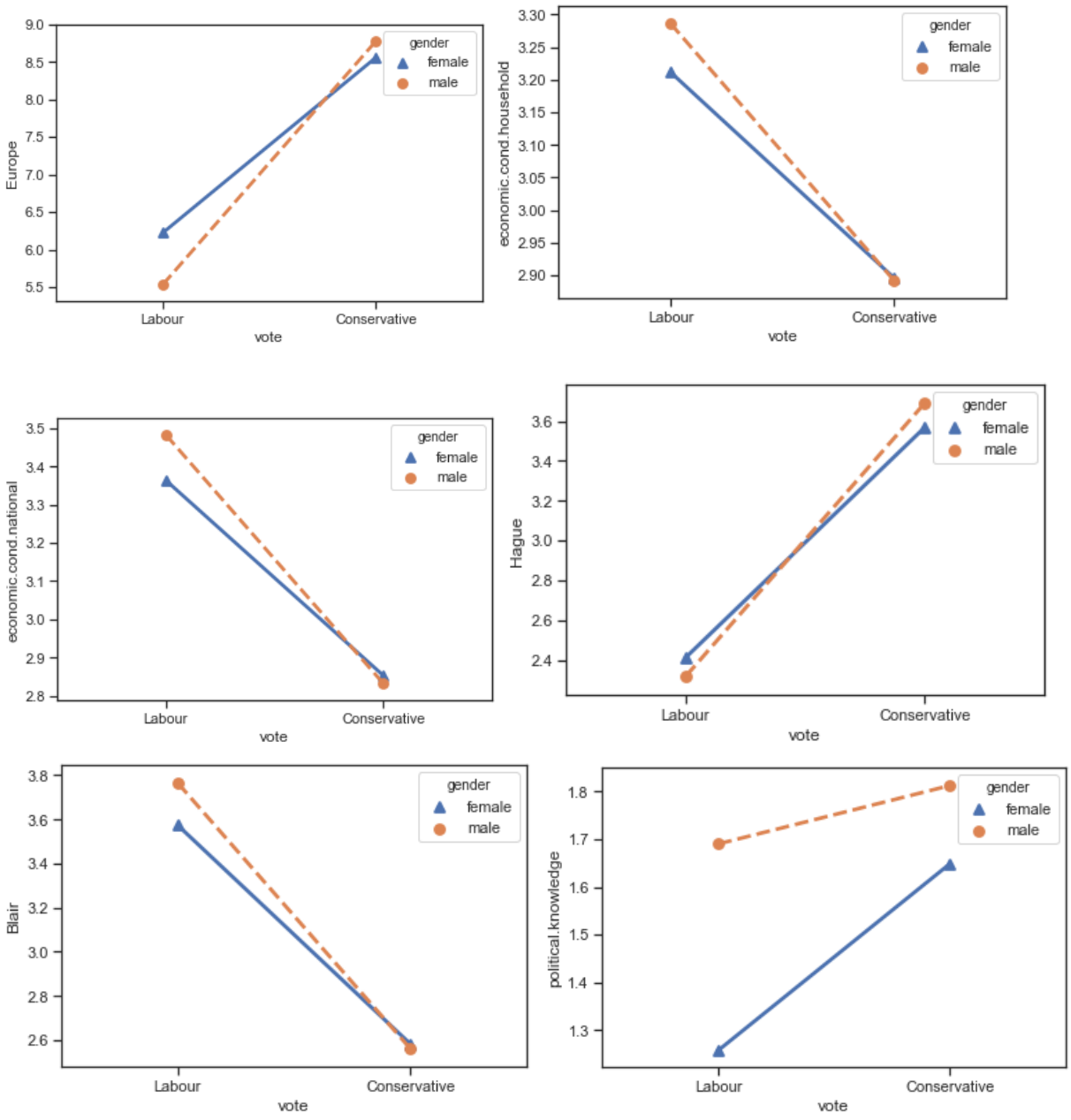

Blair
5
4
3
2
1
Labour
Conservative
vote


age
90
80
70
60
50
40
30
20
Labour
Conservative
vote


political.knowledge
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Labour
Conservative
vote

If we see age vs vote, then Labor is having more points than conservative. So all age categories gave good votes to labours.

4.Pointplot of each variable vs Vote:



First point plot shows labour party is less Eurosceptic' sentiment. Similar is for Hauge and political knowledge.

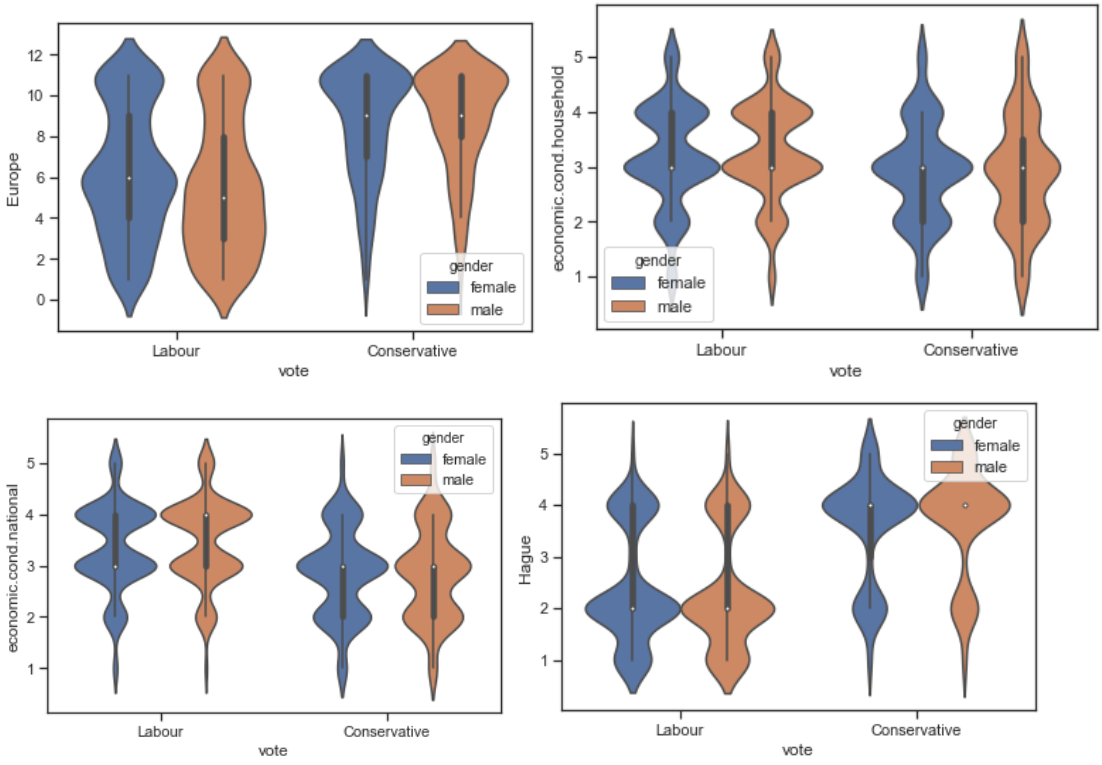
In second plot shows less economic household condition. Similar is for Blair and economic national condition.

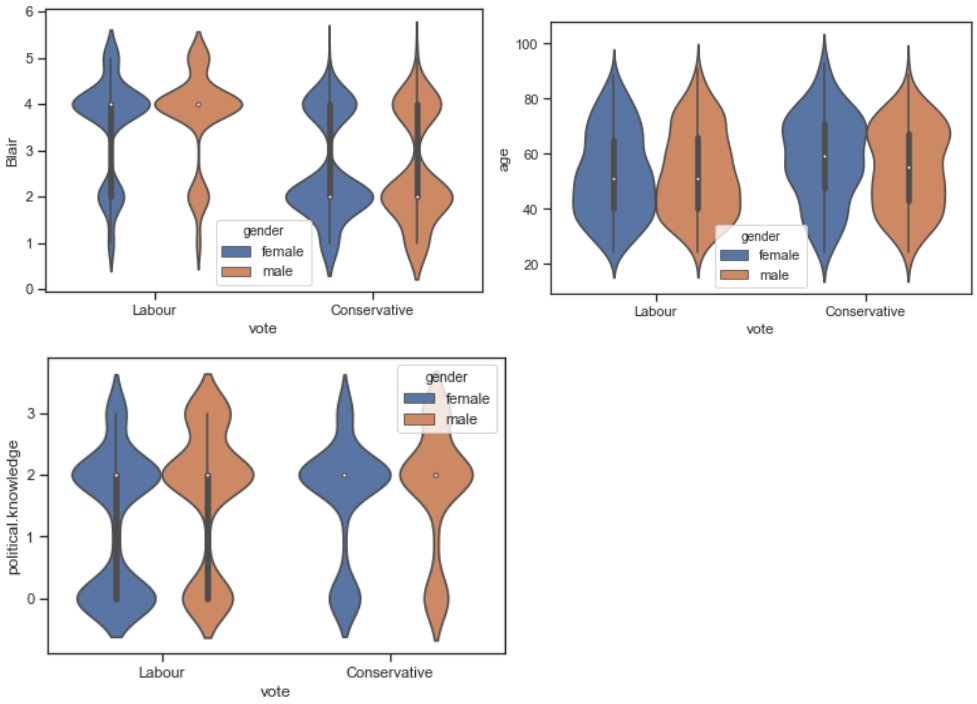
5.Violinplot:

Violin plot shows frequency distribution on vertical axis with vote.

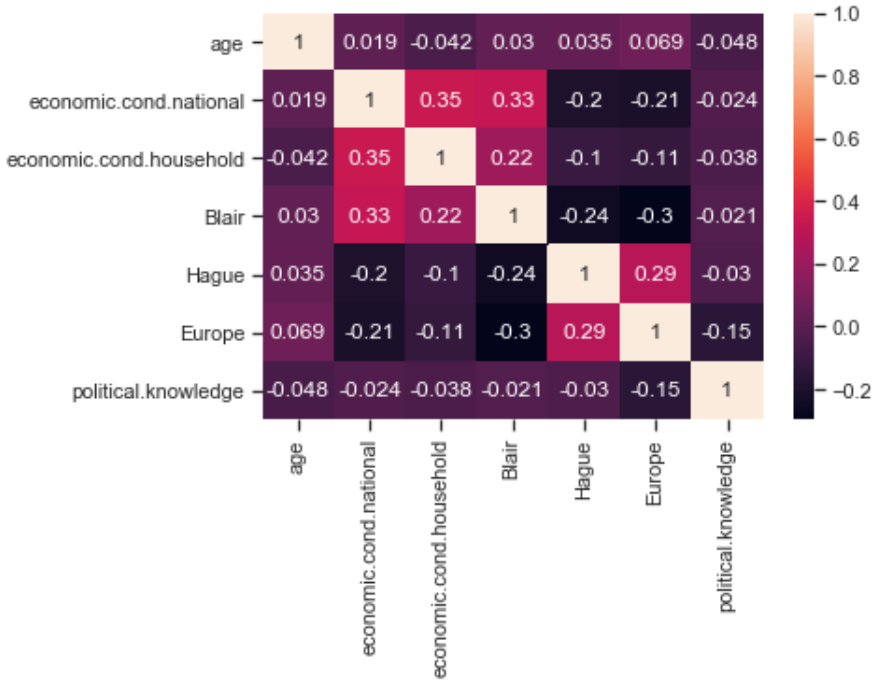
Ex.First graph shows distribution of male and female on Europe score.On labour side maximum score is near to 6 and for conservative max score for male and female is near to 11.

For age, distribution is approximately normal for labour.





6.Correlation heat map



No correlation amongst variable.

7.bivariate counts:

	gender	female	male	All
vote	Blair			

Conservative	**1**	28	31	59
	2	142	100	242
	3	0	1	1
	4	88	69	157
	5	1	2	3
Labour	**1**	24	14	38
	2	116	80	196
	4	345	334	679
	5	68	82	150
All		812	713	1525

	gender	**female**	**male**	**All**
vote	**Hague**			
Conservative	**1**	6	5	11
	2	58	38	96
	3	6	3	9
	4	161	126	287
	5	28	31	59
Labour	**1**	106	116	222
	2	274	254	528
	3	18	10	28
	4	149	122	271
	5	6	8	14
All		812	713	1525

	gender	**female**	**male**	**All**
vote	**economic.cond.national**			
Conservative	**1**	12	9	21
	2	74	66	140
	3	117	83	200
	4	52	40	92
	5	4	5	9
Labour	**1**	11	5	16
	2	64	53	117
	3	224	183	407
	4	221	229	450
	5	33	40	73
All		812	713	1525

	gender	**female**	**male**	**All**
vote	**economic.cond.household**			
Conservative	**1**	14	14	28
	2	68	58	126
	3	118	80	198
	4	49	38	87
	5	10	13	23
Labour	**1**	23	14	37
	2	85	69	154
	3	232	218	450
	4	178	175	353
	5	35	34	69
All		812	713	1525

	gender	**female**	**male**	**All**
vote	**Europe**			
Conservative	**1**	4	1	5
	2	4	2	6
	3	7	7	14
	4	13	5	18
	5	8	12	20
	6	24	12	36
	7	21	11	32
	8	25	24	49
	9	25	31	56
	10	32	22	54
	11	96	76	172
Labour	**1**	46	58	104
	2	28	45	73
	3	51	64	115
	4	51	58	109
	5	54	50	104
	6	106	67	173
	7	32	22	54
	8	33	30	63
	9	28	27	55
	10	28	19	47
	11	96	70	166
All		812	713	1525

	gender	**female**	**male**	**All**
vote	**political.knowledge**			
Conservative	**0**	57	38	95
	1	5	6	11
	2	169	115	284
	3	28	44	72
Labour	**0**	225	135	360
	1	18	9	27
	2	253	245	498
	3	57	121	178
All		812	713	1525

2.Data Preparation: 5 marks

2.1 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). (5 Marks)

Ans:

1.Encoding:

Gender has string values so encoded as dummy variable.

X_scaled=pd.get_dummies(X_scaled)

2.Scaling:

Here scaling is necessary because because age is continuous variable and has different unit.

Other variables are having scaling from 0 to 3,1 to 5 and 1 to 11.

So bring it to same unit we are using minmax scaler.

X=df.iloc[:,1:8]

y=df['vote']

X_scaled = X.apply(lambda x:(x-x.min()) / (x.max()-x.min()))

	age	national_ ecnomic_cond	household_ economic_ cond	Blair	Hague	Europe	political_ knowledge	gender_ female	Gend er_male
0	0.275362	0.5	0.5	0.75	0	0.1	0.666667	1	0
1	0.173913	0.75	0.75	0.75	0.75	0.4	0.666667	0	1
2	0.15942	0.75	0.75	1	0.25	0.2	0.666667	0	1
3	0	0.75	0.25	0.25	0	0.3	0	1	0
4	0.246377	0.25	0.25	0	0	0.5	0.666667	0	1

3.Splitting data:

from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.3 , random_state=1)

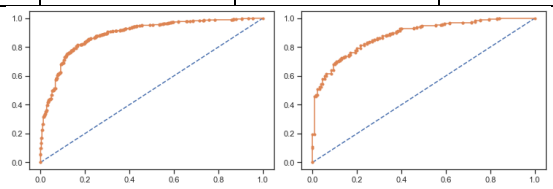
2.2Modelling:

Models and their performance is shown in following table;

Codes are available in same sequence in ipynb file attached.

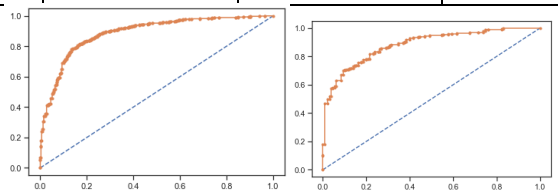
1.LOGISTIC REGRESSION

	Acc	Conf matrix		precision	recall	f1	AUC
Train	0.836926	[[225 107]	conservative	0.77	0.68	0.72	0.89
		[67 668]]	labour	0.86	0.91	0.88	
Test	0.82096	[[84 46]	conservative	0.7	0.65	0.67	0.883
		[36 292]]	labour	0.86	0.89	0.88	



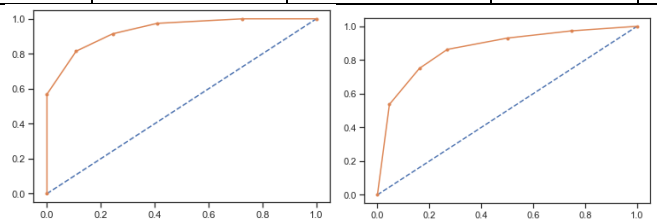
2.LDA

	Acc	Conf matrix		precision	recall	f1	AUC
Train	0.836926	[[233 99]	conservative	0.76	0.7	0.73	0.889
		[75 660]]	labour	0.87	0.9	0.88	
Test	0.81877	[[86 44]	conservative	0.69	0.66	0.67	0.884
		[39 289]]	labour	0.87	0.88	0.87	



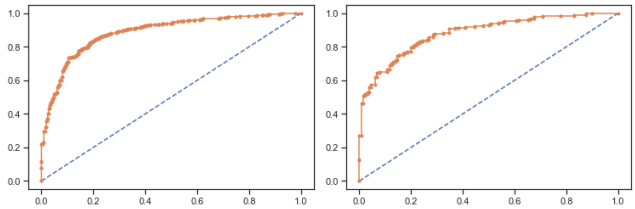
3.KNN

	Acc	Conf matrix		precision	recall	f1	AUC
Train	0.865042	[[251 81]	conservative	0.8	0.76	0.78	0.935
		[63 672]]	labour	0.89	0.91	0.9	
Test	0.82532	[[95 35]	conservative	0.68	0.73	0.7	0.865
		[45 283]]	labour	0.89	0.86	0.88	



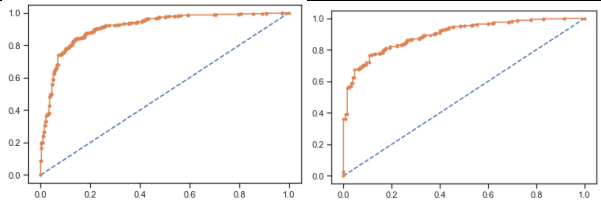
4.Naïve Bayes Model

	Acc	Conf matrix		precision	recall	f1	AUC
Train	0.829428	[[238 94]	conservative	0.73	0.72	0.72	0.886
		[88 647]]	labour	0.87	0.88	0.88	
Test	0.816594	[[93 37]	conservative	0.66	0.72	0.69	0.882
		[47 281]]	labour	0.88	0.86	0.87	



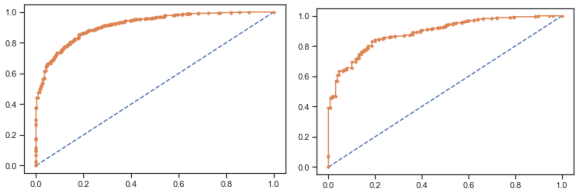
5.Support vector machine (SVM) model

	Acc	Conf matrix		precision	recall	f1	AUC
Train	0.852858	[[229 103]	conservative	0.81	0.69	0.74	0.913
		[54 681]]	labour	0.87	0.93	0.9	
Test	0.820961	[[83 47]	conservative	0.7	0.64	0.67	0.899
		[35 293]]	labour	0.86	0.89	0.88	

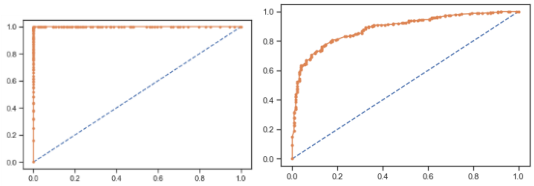


6.Random Forest with model tuning/GridSearch

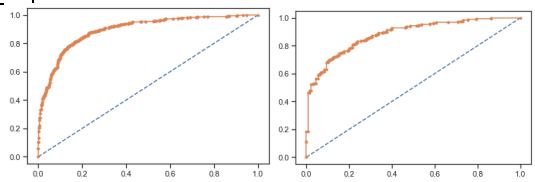
	Acc	Conf matrix		precision	recall	f1	AUC
Train	0.845361	[[229 103]	conservative	0.79	0.69	0.74	0.914
		[62 673]]	labour	0.87	0.92	0.89	
Test	0.818777	[[83 47]	conservative	0.7	0.64	0.67	0.893
		[36 292]]	labour	0.86	0.89	0.88	



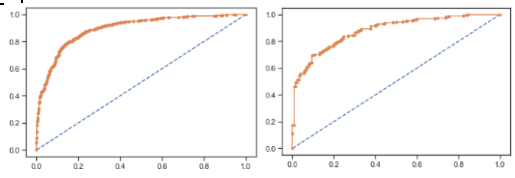
7.Random forest without tuning							
	Acc	Conf matrix		precision	recall	f1	AUC
Train	0.999063	[[331 1]	conservative	1	1	1	1
		[0 735]]	labour	1	1	1	
Test	0.82314	[[89 41]	conservative	0.69	0.68	0.69	0.886
		[40 288]]	labour	0.88	0.88	0.88	



8.Logistic regression with SMOTE							
	Acc	Conf matrix		precision	recall	f1	AUC
Train	0.817687	[[594 141]	conservative	0.82	0.81	0.82	0.889
		[127 608]]	labour	0.81	0.83	0.82	
Test	0.79694	[[102 28]	conservative	0.61	0.78	0.69	0.882
		[65 263]]	labour	0.9	0.8	0.85	
Score(max)	0.85034						
Score(min)	0.789116						
diff	0.061224						

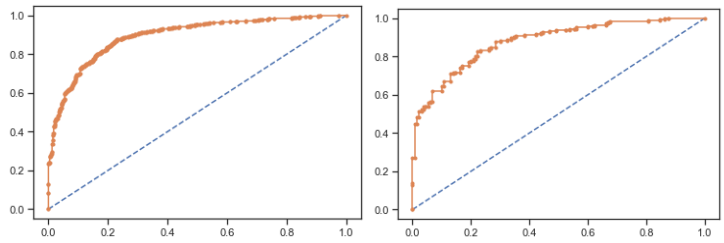


9.LDA with SMOTE							
	Acc	Conf matrix		precision	recall	f1	AUC
Train	0.817	[[593 142]	conservative	0.82	0.81	0.82	0.895
		[127 608]]	labour	0.81	0.83	0.82	
Test	0.7925	[[102 28]	conservative	0.6	0.78	0.68	0.882
		[67 261]]	labour	0.9	0.8	0.85	
Score(max)	0.843537						
Score(min)	0.789116						
diff	0.054422						



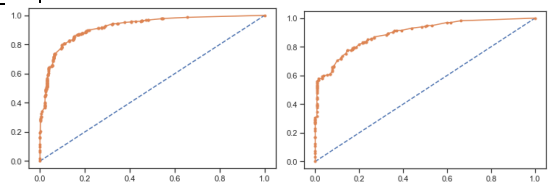
10.NB With SMOTE

	Acc	Conf matrix		precision	recall	f1	AUC
Train	0.815	[[594 141]	conservative	0.82	0.81	0.81	0.894
		[131 604]]	labour	0.81	0.82	0.82	
Test	0.7817	[[104 26]	conservative	0.58	0.8	0.68	0.879
		[74 254]]	labour	0.91	0.77	0.84	
Score(max)	0.843537						
Score(min)	0.768707						
diff	0.07483						



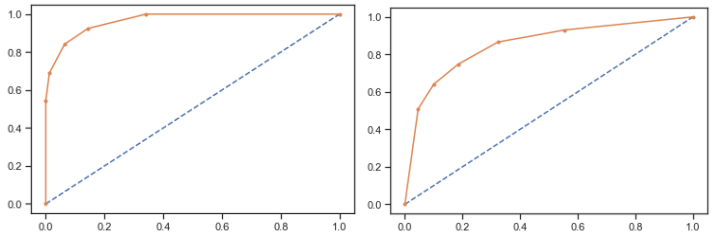
11.XGBOOST WITH SMOTE

	Acc	Conf matrix		precision	recall	f1	AUC
Train	0.851	[[624 111]	conservative	0.85	0.85	0.85	0.921
		[108 627]]	labour	0.85	0.85	0.85	
Test	0.8122	[[104 26]	conservative	0.63	0.8	0.71	0.895
		[60 268]]	labour	0.91	0.82	0.86	
Score(max)	0.863946						
Score(min)	0.782313						
diff	0.081633						

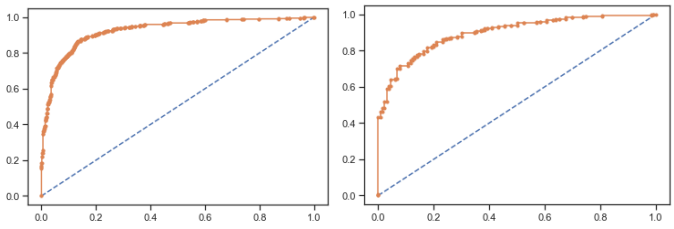


12.KNN With SMOTE

	Acc	Conf matrix		precision	recall	f1	AUC
Train	0.8884	[[688 47]	conservative	0.85	0.94	0.89	0.966
		[117 618]]	labour	0.93	0.84	0.88	
Test	0.7664	[[106 24]	conservative	0.56	0.82	0.66	0.851
		[83 245]]	labour	0.91	0.75	0.82	
Score(max)	0.891156						
Score(min)	0.789116						
diff	0.102041						



13.SVM With SMOTE							
	Acc	Conf matrix		precision	recall	f1	AUC
Train	0.8619	[[641 94]	conservative	0.85	0.87	0.86	0.926
		[109 626]]	labour	0.87	0.85	0.86	
Test	0.8079	[[109 21]	conservative	0.62	0.84	0.71	0.899
		[67 261]]	labour	0.93	0.8	0.86	
Score(max)	0.870748						
Score(min)	0.816327						
diff	0.054422						



4.Inference: 5 marks

1. Based on these predictions, what are the insights? (5 marks)

The basic application of all these model is to predict correct result.

Ideal case is that my model should identify all predictions correctly in training and when the same is fitted on testing data it should identify all predictions in testing also.

But it is not the real case. Model performs good in training but doesn't perform well on testing data.

So first we will see which model performs well.

4.1. Interpretation of model without tuning

Model 1 to 7:

1.Accuracy: All models are having good accuracy (more than 0.82) on both train and test data. But accuracy is not the main criteria. Criteria is that whether model identified correct class or not.

Accuracy gives us whether model is over fitted or under fitted. When accuracy for training and testing data if same then model is not over fitted/under fitted.

But in actual 10% variation is allowed.

In our models Random forest model(no.7) is over fitted because it shows 99.92% accuracy in training but 82% for test data.

Remaining all models are more or less good fitted.

2.Confusion Matrix:

This matrix gives us the count of Actual Conservative/Predicted Conservative and Actual Labour/Predicted labour.

In this matrix ideally non diagonal elements should be zero means actual and predicted are same. But its not the case here.

Only Random forest model(no.7) training confusion matrix shows it has identified correct class.

3.Classification report:

Acc.=(TP+TN)/(TP+TN+FP+FN)

Recall=TP/(TP+FN)

Precision=TP/(TP+FP)

All models performed well for training data but could not perform well in testing data.

As per the classification report SVM without SMOTE is poor model which performed poor in prediction of conservative class.

From above table KNN model is better than others

4.2. Interpretation of model with SMOTE

Model 8 to13

Model 8 to 13 are tuned using SMOTE (Synthetic Minority Oversampling Technique).

As we observe that models performed poor in conservative class in test data.

The reason being that the proportion of classes in target variable is not proper.

Labour 69.7049%
Conservative 30.2951%

Already conservative has less proportions and after splitting it get reduced. So for model to learn conservative class some more data is required. So in SMOTE minority class is resampled and increased in proportion.

So all above models are tuned by resampling.
After resampling,it can be observed that recall is increased significantly but could not improve precision signicantly.
This means that,after SMOTE model 8 to 13 ,prediction of conservative have increased but decreased the labor prediction.thus model has increased recall but reduced precision.

1.LOGISTIC REGRESSION

	Acc	Conf matrix		precision	recall	f1	AUC
Train	0.836926	[[225 107]	conservative	0.77	0.68	0.72	0.89
		[67 668]]	labour	0.86	0.91	0.88	
Test	0.82096	[[84 46]	conservative	0.7	0.65	0.67	0.883
		[36 292]]	labour	0.86	0.89	0.88	

8.Logistic regression with SMOTE

	Acc	Conf matrix		precision	recall	f1	AUC
Train	0.817687	[[594 141]	conservative	0.82	0.81	0.82	0.889
		[127 608]]	labour	0.81	0.83	0.82	
Test	0.79694	[[102 28]	conservative	0.61	0.78	0.69	0.882
		[65 263]]	labour	0.9	0.8	0.85	

4.3 Inference

It's difficult to say optimized model but as per performance SVM with SMOTE is good model.

Moreover, proportions of both classes shold not be biased.This means biased data does not produce good model even after SMOTE.

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

President John F. Kennedy in 1961

President Richard Nixon in 1963

Find the number of characters, words and sentences for the mentioned documents. – 3 Marks

Remove all the stopwords from all the three speeches. – 3 Marks

Which word occurs the most number of times in his inaugural address for each president? Mention the top three words. (after removing the stopwords) – 3 Marks
Plot the word cloud of each of the speeches of the variable. (after removing the stopwords) – 3 Marks

1.Find the number of characters, words and sentences for the mentioned documents
Ans:

Imported liabraries and text speeches.

```
import nltk

nltk.download('inaugural')

from nltk.corpus import inaugural

inaugural.fileids()

inaugural.raw('1941-Roosevelt.txt')

inaugural.raw('1961-Kennedy.txt')

inaugural.raw('1973-Nixon.txt') "
```

Speech	Sentence	Total words	Total Characters
Nixon	67*	1820	9922
Roosevelt	69	1361	7503
Kennedy	56	1390	7563

As split is done through ".", so first 4 columns in df_Nixon is merged in one sentence so it becomes 70-3=67

2.Remove all the stopwords from all the three speeches.

Ans:

*Before removing stop words - do consider to remove punctuation and digits.

After removing stopwords:

Speech	Sentence	Total words	Total Characters
Nixon	67*	833	5882
Roosevelt	69	627	4523
Kennedy	56	693	4719

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)

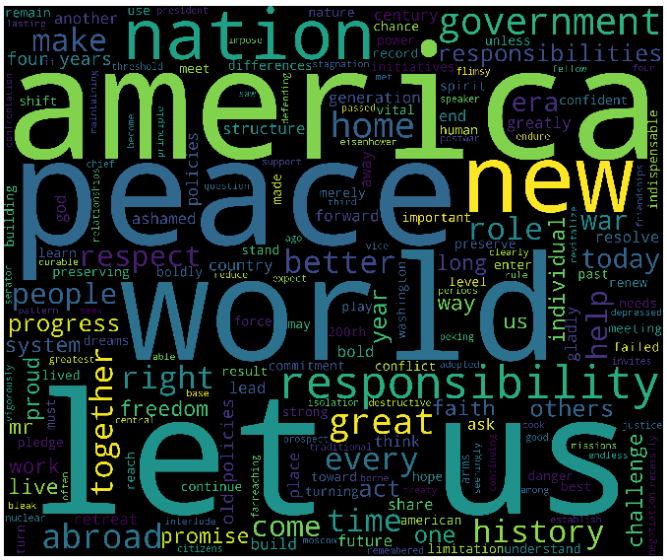
Ans:

```
In Nixon speech frequency of words is:
us        26
let       22
peace     19
dtype: int64
In Roosevelt speech frequency of words is:
nation    11
know      10
spirit     9
dtype: int64
In Kennedy speech frequency of words is:
let       16
us        12
world      8
dtype: int64
```

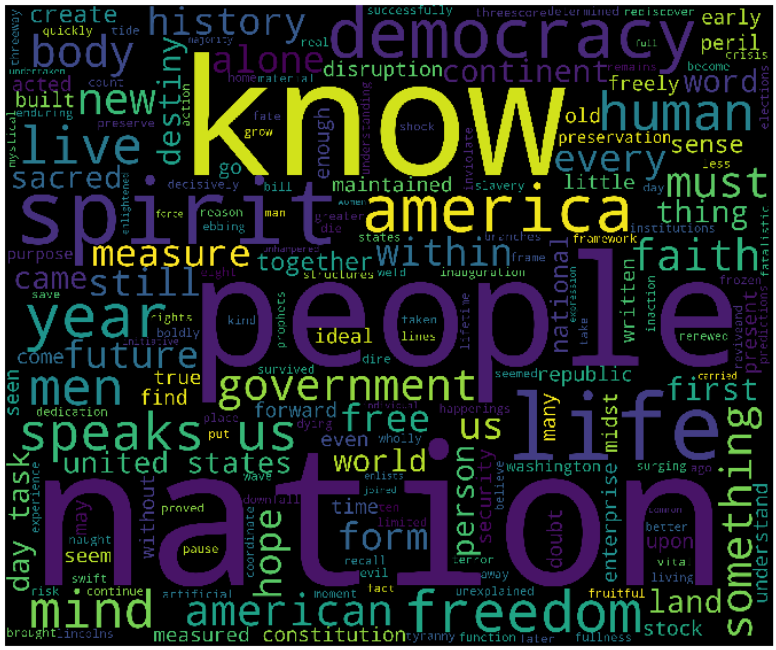
4. Plot the word cloud of each of the speeches of the variable. (after removing the stopwords)

Ans:

World cloud for Nixon Speech:



World cloud for Roosevelt Speech:



World cloud for Kennedy Speech: