|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Decorative | | | | |
|  |  |  | |  |
| MACHINE LEARNING  Business report | | |
|  | | Sivani Varma |

# Problem 1:

## You are hired by one of the leading news channels 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](https://olympus.greatlearning.in/courses/16934/files/2478796/download?verifier=UbU5R6L52H9Ur9mDr7QX8xF63ubYUb6Cy6cKFmy7&wrap=1)

## Data Ingestion: 11 marks

1.1 Read the dataset. Do the descriptive statistics and do the null value condition check. Write an inference on it. (4 Marks)  
1.2 Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers. (7 Marks)

Data Preparation: 4 marks

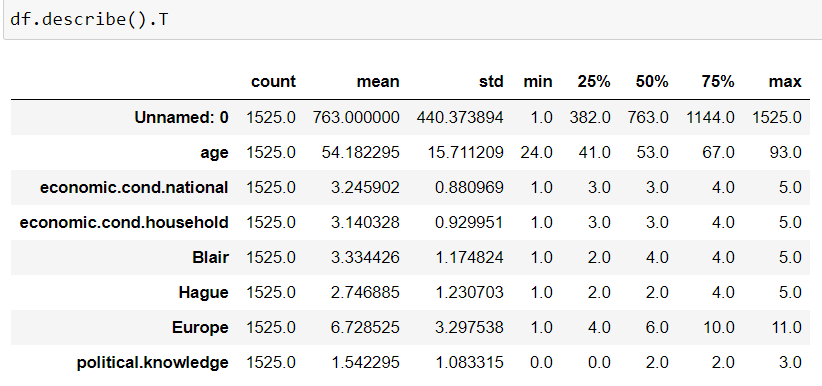
1.3 Encode the data (having string values) for Modelling. Is Scaling necessary here or not? Data Split: Split the data into train and test (70:30). (4 Marks)

Modeling: 22 marks

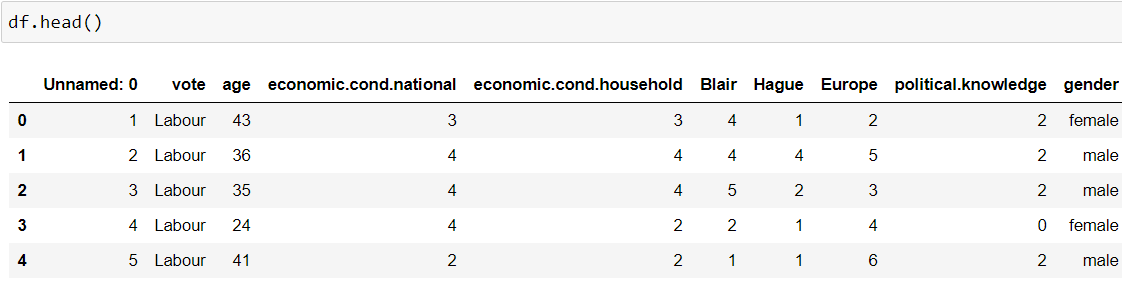
1.4 Apply Logistic Regression and LDA (linear discriminant analysis). (4 marks)  
1.5 Apply KNN Model and Naïve Bayes Model. Interpret the results. (4 marks)  
1.6 Model Tuning, Bagging (Random Forest should be applied for Bagging), and Boosting. (7 marks)  
1.7 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model. Final Model: Compare the models and write inference which model is best/optimized. (7 marks)

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

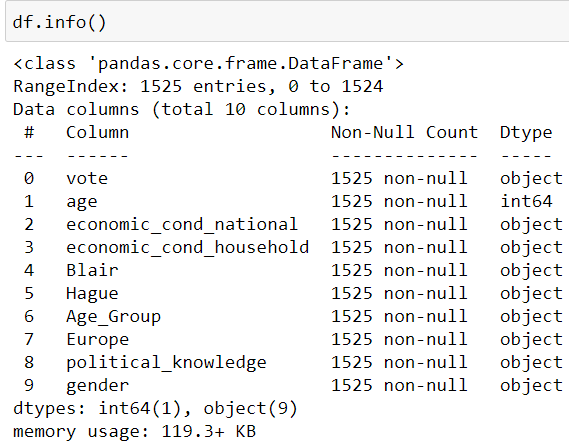
**1.1) Read the dataset. Describe the data briefly. Interpret the inferences for each. Initial steps like head () .info (), Data Types, etc. Null value check, Summary stats, Skewness must be discussed**

****

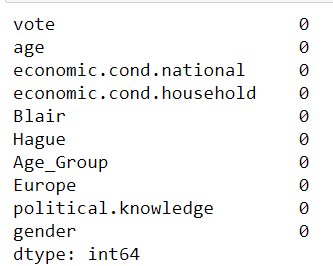
* The election date set has 1525 rows of observations from an election poll, across 9 dimensional of integer and object data type.
* The min age of voters starts with 24 and max age is 93.
* From the descriptive summary of the data set, it can be inferred that the average age of the voters is 54 years and median is 53 years, which indicates the normal distribution of the age.
* 50% of the voters are between the age of 41 years and 67 years.
* The average voter’s assessment of the national house hold economic condition is around 3.
* The national economic conditions, household economic conditions, Blair and Hague gives the range of the assessment, which is from 1 to 5.
* Whereas, knowledge of parties’ positions on European integration starts from 0 to 3.
* Only “age” variable is found to be both numerical and continuous in nature and rest are categorical in nature.
* ‘vote’ and ‘age’ represents the categories which they voted and their gender.



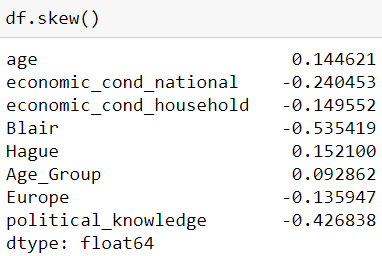
* Head function helps in visualizing weather the data from the sheet has been imported or not.
* It reflects the top 5 rows of the election data sheet.
* Where the column 1 will be removed in further analysis, as it will not be used or will not add any meaningful value in data learning.



* From the Info summary, one can interpret that the data frame has total 1525 entries, range indexing from 0 to 1524.
* Having total 10 data columns (Ideally data frame has 9 columns but, 1 extra dimension (‘Age Group’) has been created from continuous variable ‘age’.
* After converting the continuous variables into object type, now the data frame has total 9 dimension of object data type and 1 dimension of integer type.
* Like, Age is of integer datatype while other dimensions are of object type.
* Also, the dimension name of political knowledge, national and household economic condition has been changed for the ease.



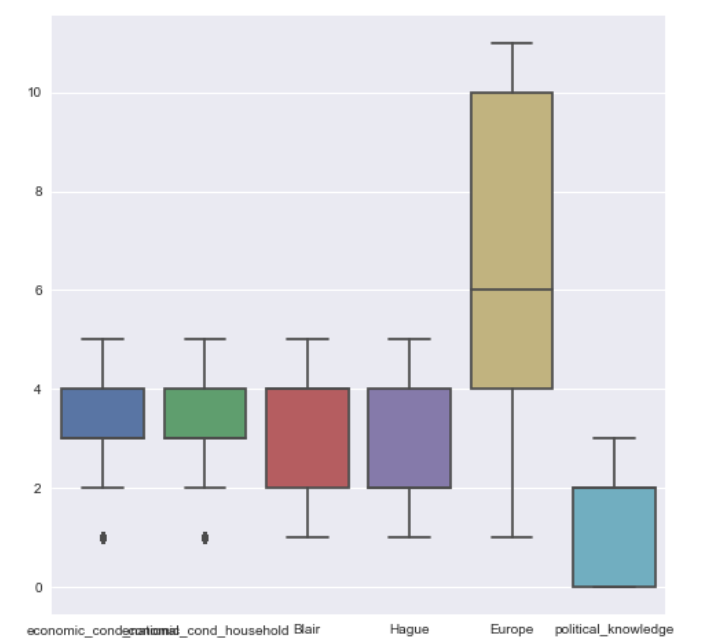
* No variables were found to be having ‘null’ values.
* The categorical variables in the integer type were converted to objective type for the ease of analysis and visualization.
* There is total 8 duplicate rows in the data set which either can be deleted or not.



* Age, Hague and age group skewness lies in 0 to 0.15, which is neither negative nor greater than 1 and can be said as the mentioned dimension as perfectly symmetrical.
* Where are, Blair, Europe, political knowledge national, national and household economic condition dimensions are negatively skewed.
* It means that the left tail is longer.

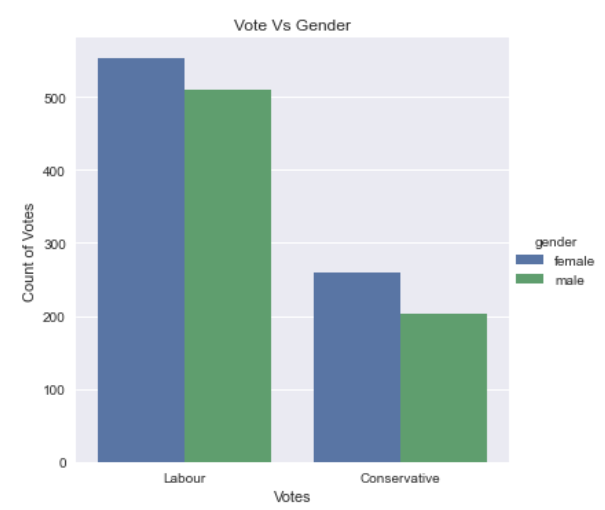
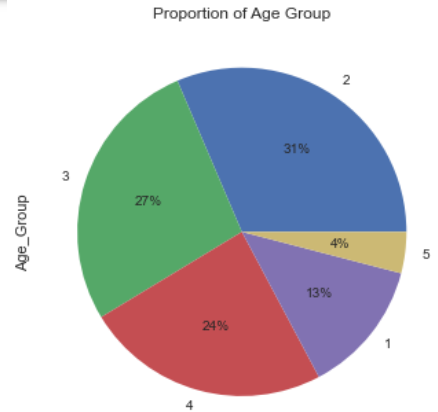
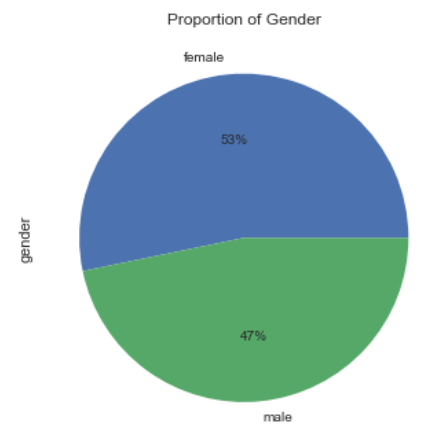
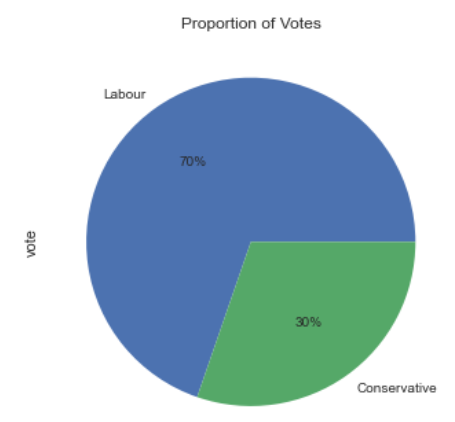
**1.2) Perform EDA (Check the null values, Data types, shape, Univariate, bivariate analysis). Also check for outliers (4 pts). Interpret the inferences for each (3 pts) Distribution plots(histogram) or similar plots for the continuous columns. Box plots, Correlation plots. Appropriate plots for categorical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

NOTE- Null Values, data type and shape of the data frame has been discussed above already.

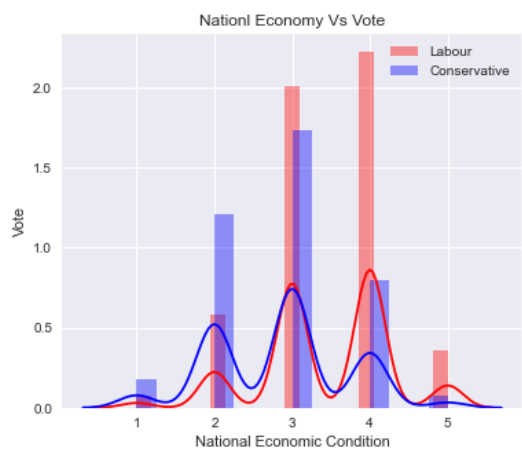
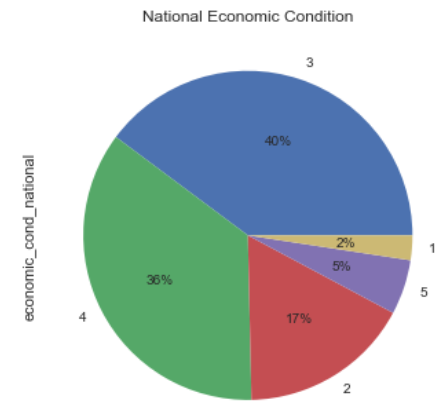
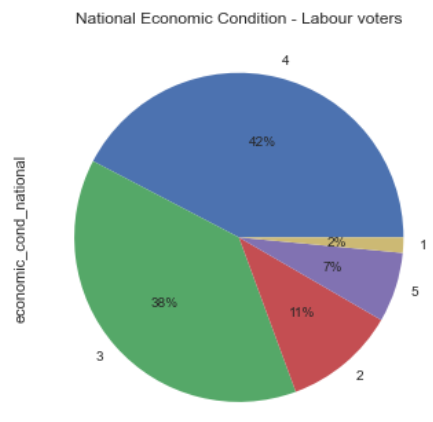
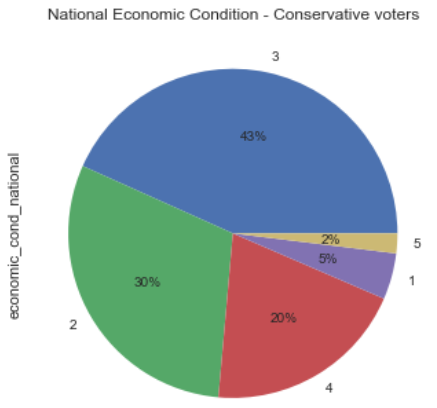
****

* We have outliers present for national and household economic condition data, which can be avoided as it doesn’t affect much in the analysis.
* The age distribution of voters follows a normal distribution with the mean of 54 years and median of 53 years. There are no outliers.
* Having said that, rest of the dimension doesn’t have any outliers.
* Hence outlier treatment is not required.

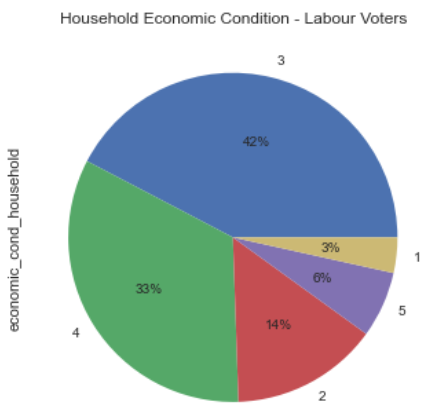
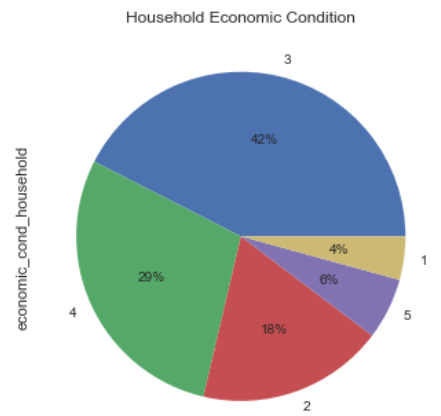
DEMOGRAPHIC FACTORS:

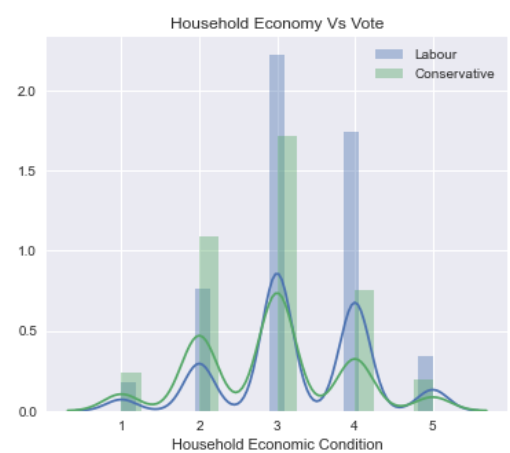
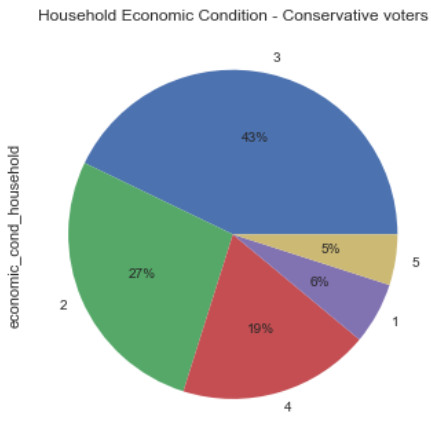


* The poll data (proportion of votes) clearly shows that the winning party is the labour party with 70% of votes in their favor. While 30% is with conservative party.
* From the data, we can refer that 53% of the poll respondents are female and 47% of them are male.
* The age group distribution of voters provides interesting sights on the voting patterns.
* 57% of the voters are found to be above 55 years of age.
* 31% of them are in the age range of 36 to 50 years.
* The group between age 51 and 65 years make 27% of the votes.
* On analyzing the votes polled, it can be observed that the labour party got more votes from all the age groups than the conservatives.
* The conservative party is observed to be clearly favored by the elderly votes than the younger ones.

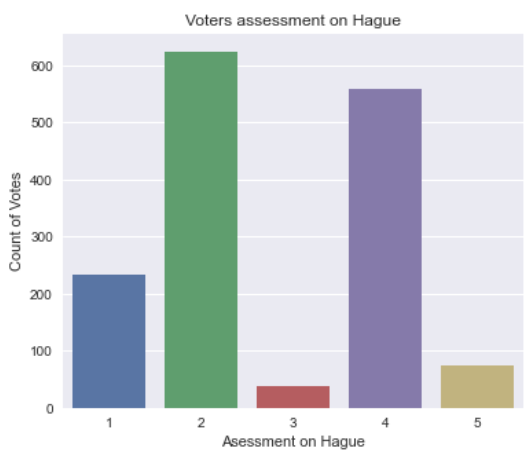
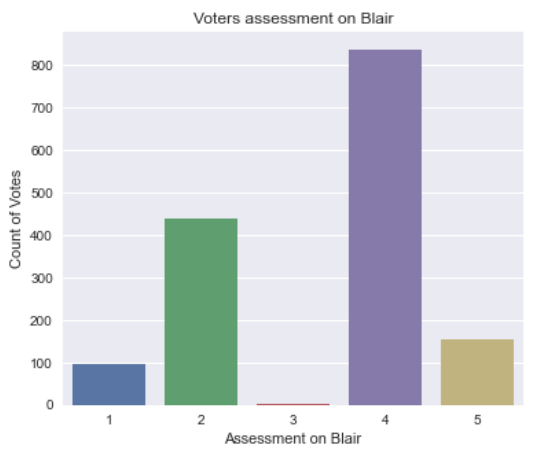


* The features representing every voters’ assessment on the national and household economic conditions gives us an insight into the prevailing economic situation.
* In the national economic condition, 40% of the labour have fallen under 3 ranking.
* From the distribution of the economic rankings, it can be said that those who ranked both the national and household conditions high i.e, 4 or 5 are more likely to vote for labour.
* 81% of the labour votes have ranked the household economic condition on or above 3, on a scale of 1 to 5.
* And those ranked the national and household economic condition to be lower than 3 are more likely to vote for conservatives.

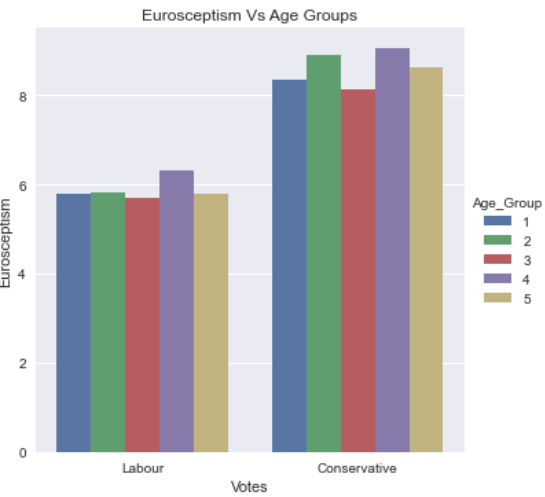
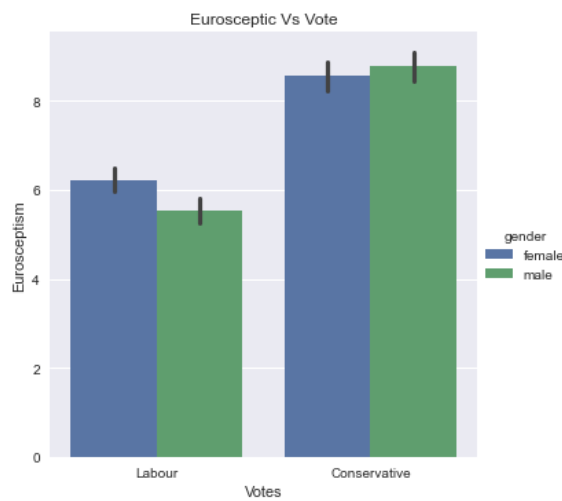
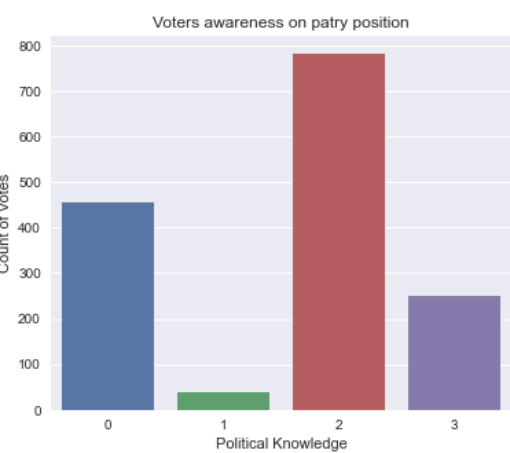




* 81% and 77% of all the voters have ranked the national and household economic cnditions respectively on or above 3, which indicates that the prevailing economic condition of the period to be great.



* The feature ‘Blair’ and ‘Hague’ represents voters’ assessment on the labour and conservative party leaders.
* From the data it can be inferred that the irrespective of party, 65% of the voters have rated Blair ranking of 3.
* 40% of the conservative voters and 78% of the labour voters rated Blair above 3.
* Blair is rated high among youth and those between 75 and 85 years.
* Hague is rated lowest among 40 and 50 years and 80 to 85 years.

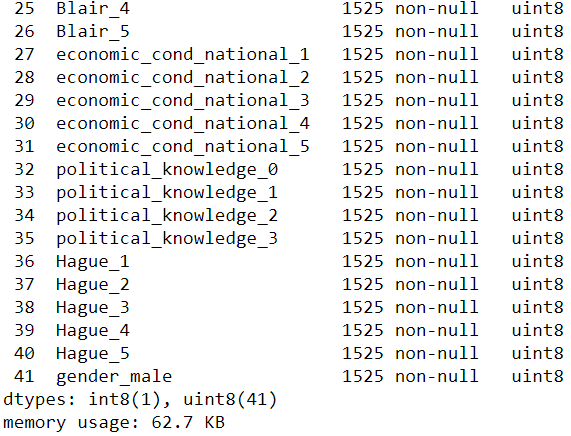
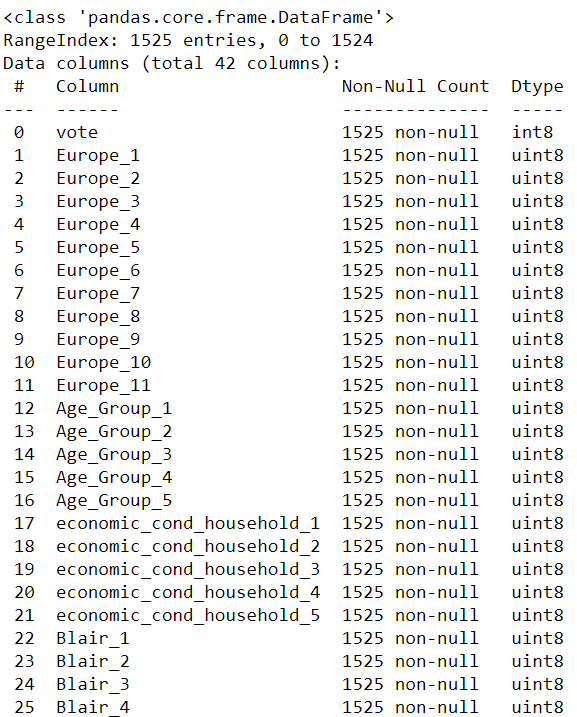


* The factor ‘Eurpoe’ gauged the level of Euroscepticism of the voter on a scale of 0 to 11.
* The female labour voter are observed to be more Eurosceptic than the male voters where as both male and female conservative voters share same level of Euroscepticism.
* Highly Eurosceptic voters are more likely to vote for conservative, where as labour voter is more likely to be less Eurosceptic.

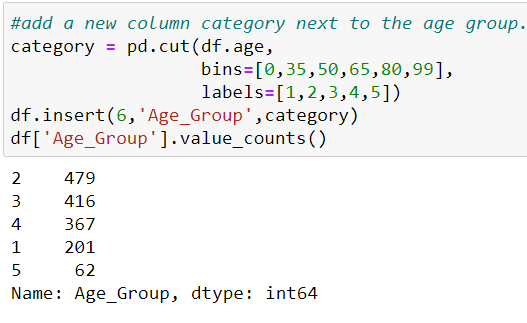
**1.3) Encode the data (having string values) for Modelling. Is Scaling necessary here or not?( 2 pts), Data Split: Split the data into train and test (70:30) (2 pts). The learner is expected to check and comment about the difference in scale of different features on the bases of appropriate measure for example std dev, variance, etc. Should justify whether there is a necessity for scaling. Object data should be converted into categorical/numerical data to fit in the models. (pd.categorical().codes(), pd.get\_dummies(drop\_first=True)) Data split, ratio defined for the split, train-test split should be discussed.**

**Data Encoding-**

* One hot coding methods is used to encode the categorical variables in the data set in order to improve the explain ability of the models.
* 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 explained in the feature set.
* The final dataset included the target class variable of 42 features.



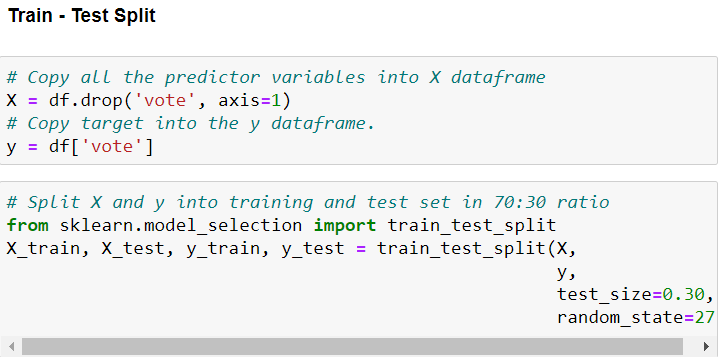
* The only numerical variable in the dataset ‘age’ is binned to create the categorical variable ‘age group’ to classify the voters into 5 different groups in increasing order of their group.
* The variable age is dropped after binning the value into a categoric variable, as keeping both the variables will bring collinearity and redundancy between these variables.



**SCALING-**

* Scaling is not required in this case as there are no numerical variable in different scales present in the dataset after dropping age.
* Binning and encoding have ensured that all factors are uniform scales.

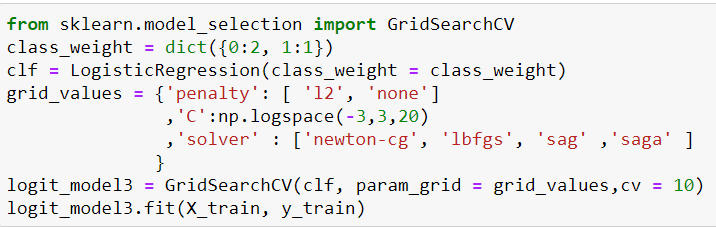
**Data Splitting-**

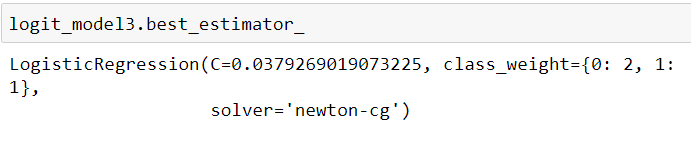


* Data has been split into train and test sets on 70:30 Ratio, with random state=27.

**1.4) Apply Logistic Regression and LDA (Linear Discriminant Analysis) (2 pts). Interpret the inferences of both model s (2 pts). Successful implementation of each model. Logical reason behind the selection of different values for the parameters involved in each model. Calculate Train and Test Accuracies for each model. Comment on the validness of models (over fitting or under fitting)**

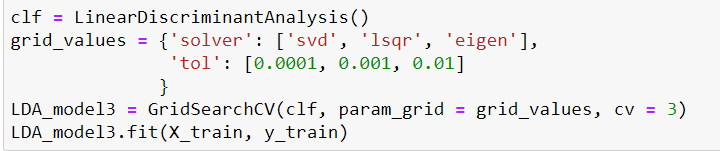
Logistic Regression-

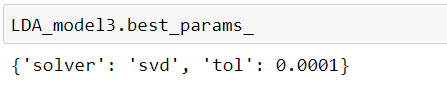




* The final model selected after all the iterations of model tuning as follows.
* The conservative class and the majority class(labour) is on a ratio of 1:2:3, which is used to set the class weight parameter in logistic regression.
* The grid search cv() returned hyperparameters for penalty and solver algorithm l2 and newton cg respectively.
* The model found to be underfit.

LDA-



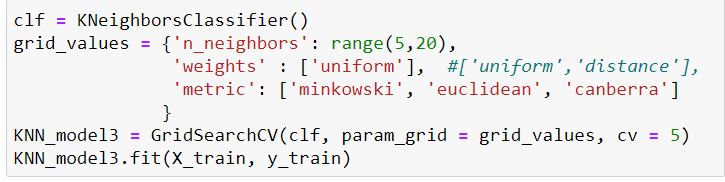


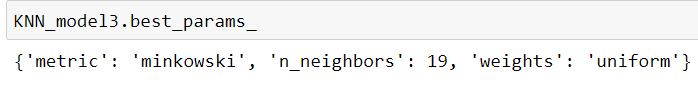
* Grid search was applied on the LDA model to identify the right solver algorithm and tolerance value.
* The default alogirithm SVD and the default tolerance value was returned as the ideal hyperparameter.
* The model scores are found to be the right fit, but the recall of minority is lower.
* The model found to underfit.

NOTE- Train and Test accuracies are calculated in code book for each LR, LDA, KNN and Naïve bayes model.

**1.5) Apply KNN Model and Naïve Bayes Model (2pts). Interpret the inferences of each model (2 pts). Successful implementation of each model. Logical reason behind the selection of different values for the parameters involved in each model. Calculate Train and Test Accuracies for each model. Comment on the validness of models (over fitting or under fitting)**

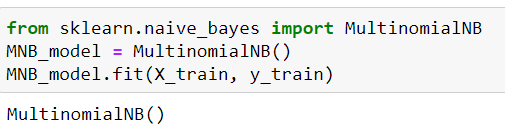
KNN Model-





* For tuning the hyperparameters of the KNN model grid search was applied to find the 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 caputuring almost 100% of the variance in train.
* Hence uniform weight is used to create a right fit model and default minkowski metric for distance measure.
* The model was found to be the right fit, but the recall of labour was lowest.

Naïve Bayes model (Multinomial)-



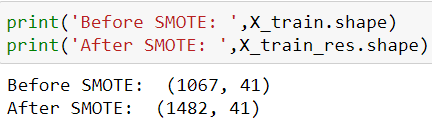
* For building a naïve bayes model, multinomial naïve bayes algorithm is used. The reason being that the data set contain only nominal factors and it meets the basic assumption.
* Neither Gridsearch nor any further tuning was applied on the model as the function does not have too many hyper parameters to be optimized.
* The resulting model was found to be the right fit with acceptable recall rate of both classes.

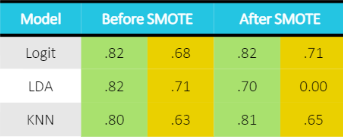
NOTE- Train and Test accuracies are calculated in code book for each LR, LDA, KNN and Naïve bayes model.

**1.6) Model Tuning (4 pts) , Bagging ( 1.5 pts) and Boosting (1.5 pts). Apply grid search on each model (include all models) and make models on best\_params. Define a logic behind choosing particular values for different hyper-parameters for grid search. Compare and comment on performances of all. Comment on feature importance if applicable. Successful implementation of both algorithms along with inferences and comments on the model performances.**

**Model Tuning-**

* The approach taken for model tuning and selection is to first balance the precision and recall of the conservative class with that of the labour class in train and test data base sets before accuracy and area under curve of the model.
* The labour is 70% and conservative is 30% of the total target class 2:3:1.



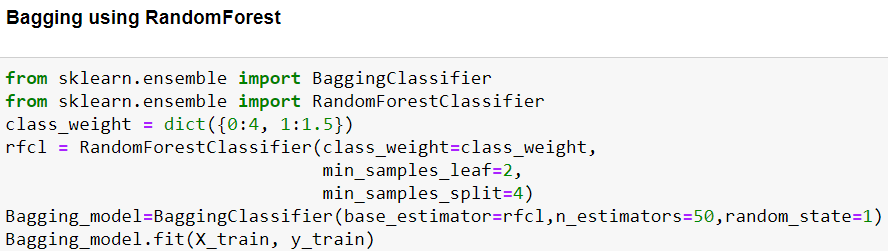




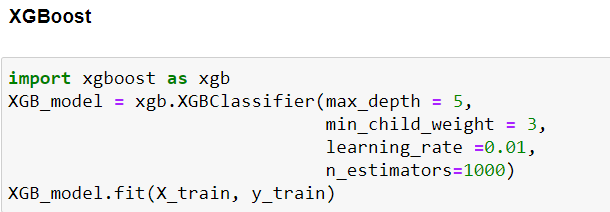
**Grid search-**

* Except for multinomial Naïve Bayes and bagging classifier, grid search was applied to optimize the hyper parameters of all the models.
* Logspace was used in the logistic regression to find the optimal value of the regularization parameter C in log scale.

**Bagging and Boosting-**

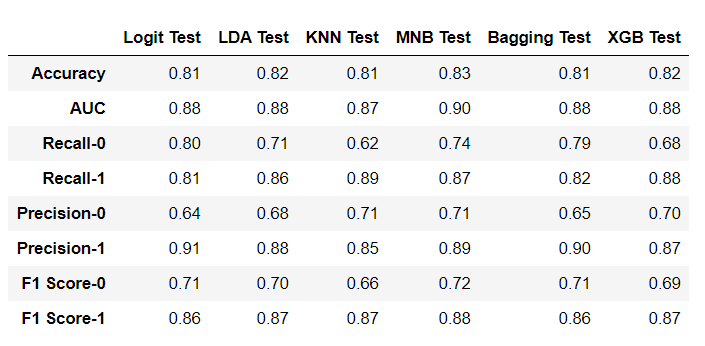
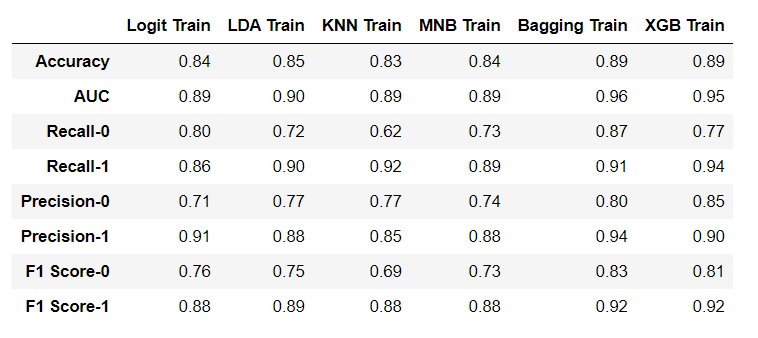


* The values for min sample leaf and min samples split was also tuned iteratively based on the accuracy score and recall.
* The random forest model was used as base estimator for the bagging classifier and n\_estimator was optimized so that model wont overfit.
* The final bagging classifier model was found to be the slightly overfit but acceptable as it returned equally high recall.



* Extreme gradient boosting or known as XGboost which is used to build a boosting model.
* The hyperparameters are tuned iteratively so that model is right fit.
* Boosting Is observed to be more appropriate for very large data sets than smaller ones as in the given business case.

**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, classification report (4 pts) Final Model - Compare and comment on all models on the basis of the performance metrics in a structured tabular manner. Describe on which model is best/optimized, After comparison which model suits the best for the problem in hand on the basis of different measures. Comment on the final model.(3 pts)**

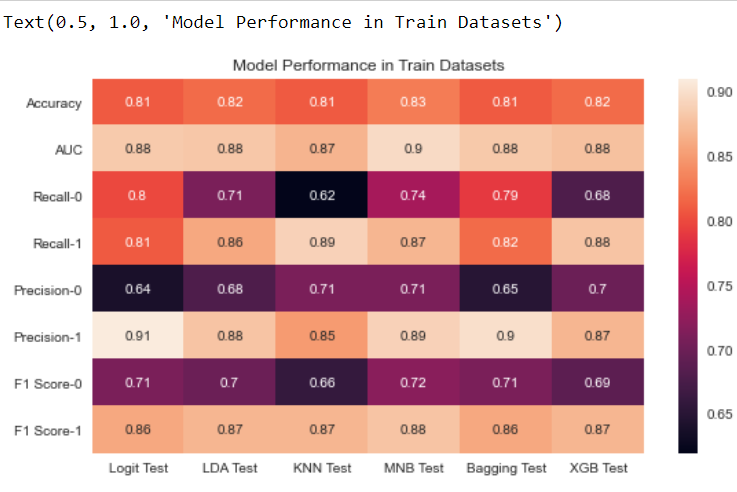
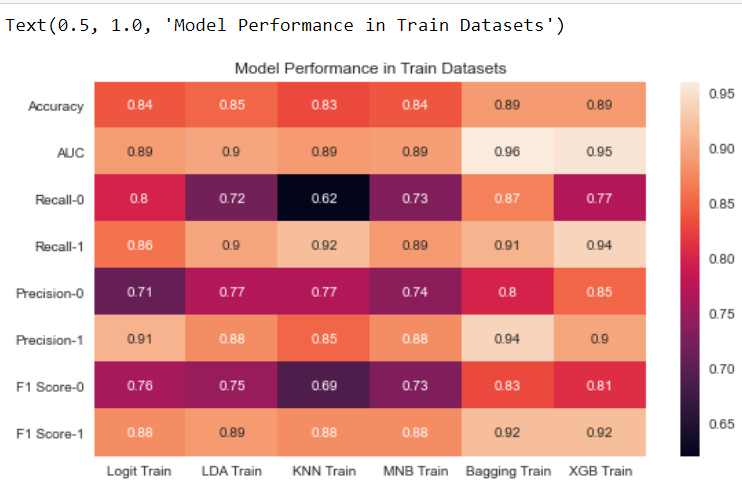


Accuracy-

* While bagging and XGboost classifiers delivered the highest accuracy of 89% in train model
* Logistic regression and Naïve bayes gives the highest accuracy of 83% in test.

Recall-

* For labour, XG boost got 94% recall, followed by KNN at 91% in test.
* Bagging gave most balanced recall of both the classes than other models.



F1 Score-

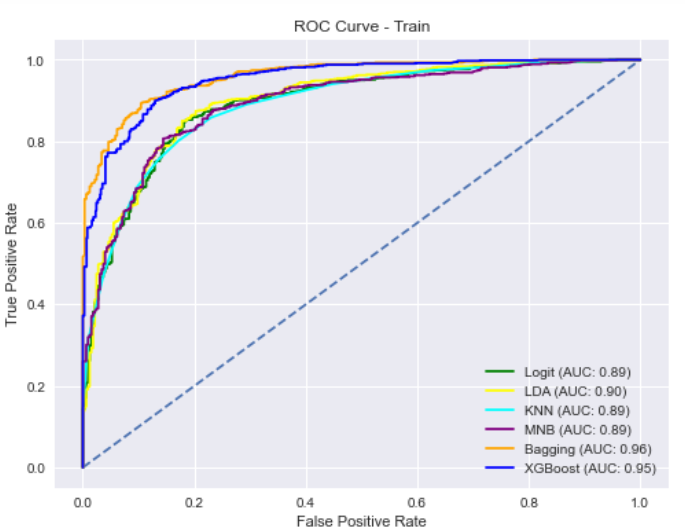
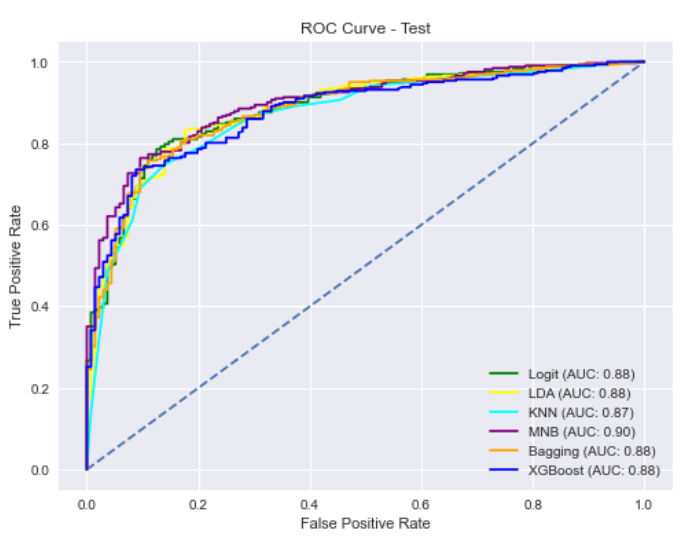
* The mean of recall and precision in train is high for bagging and XG boost at 83% and 81% for conservative class and 92% for labour class.
* In test, logit is highest for conservative class at 73% and 87% by logit, LDA, KNN and XGboost for majority.

AUC Score-

* Bagging and XGboost have the highest score on train, 96 and 95% respectively, but on test they gave 88% around each.
* On test, Naïve bayes delivered the highest AUC score at 90%, followed by Logit at 89%.
* All the models have consistency given an AUC score above 89% in train and abve 88% in test,
* Except for KNN, KNN gave the least AUC score in with 87%

Cross Validation-

* In test, except bagging all other model gave accuracy of 81%.
* LDA model got highest inconsistency with 20% std.



Conclusion-

* There are inconsistency in class predictions and computation performance in bagging and XGboost.
* Naïve and Logit showed better performance.
* MNB has delivered highest AUC score in test and right fit in train and test without any complex model tuning.
* The recall rate of the target class is also on par with logit.
* Considering the robustness of model on cross validation, recall across both the labour classes and consistent accuracy on train and test dataset, Logistic regression and MNB could be chosen as the final model.
* Logit model has produced very consistent and balanced performance in terms of conservative class predictions and over all 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.

**1.8) Based on your analysis and working on the business problem, detail out appropriate insights and recommendations to help the management solve the business objective. There should be at least 3-4 Recommendations and insights in total. Recommendations should be easily understandable and business specific, students should not give any technical suggestions. Full marks should only be allotted if the recommendations are correct and business specific.**

* The most important political factor of the poll is observed to be the European integration of UK.
* The national and household economic condition prevailed during the period is found to be an non issue during the poll, which is an indicator of the fledging economic situation of the period.
* Gender of the voter doesn’t appear to have a significant influence on the outcome of the election
* The age of the voter appears to be a decisive factor in predicting votes.
* Age group from 36 to 45 years and lesser 35 and below appears to have favored labour party.
* And those from 66 years and above are more likely to vote for conservative party.
* The opinion of voters on the labour and conservative leaders, Blair and Hague is the most decisive factor in the poll than any other factor, making the election a mandate on these political leaders.

## 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 1973

## (Hint: use .words(), .raw(), .sent() for extracting counts)

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

2.2 Remove all the stop words from all three speeches. – 3 Marks

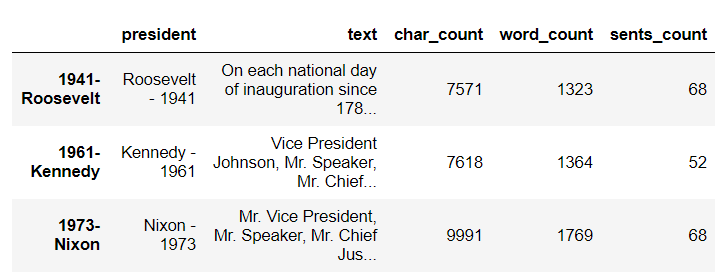
2.3 Which word occurs the most number of times in his inaugural address for each president? Mention the top three words. (after removing the stopwords) – 3 Marks

2.4 Plot the word cloud of each of the speeches of the variable. (after removing the stopwords) – 3 Marks [ refer to the End-to-End Case Study done in the Mentored Learning Session ]

Code Snippet to extract the three 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')  
"

**2.1) Find the number of characters, words and sentences for the mentioned documents. (Hint: use .words(), .raw(), .sent() for extracting counts)**

****

Character count-

* President Roosevelts inangular speech in 1941 consists of 7571 characters.
* In 1961, president kennedy gave similarly long inangular speech of 7618 character long.
* President Nioxn in 1973, appears to gave the longest of the inangular speech of the three which is of 9991 characters long.

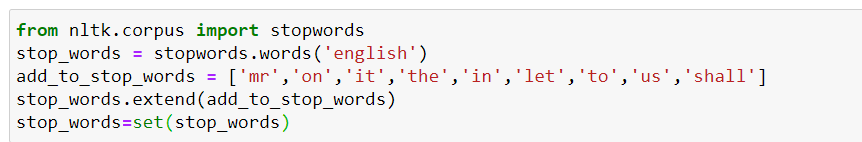
Word count-

* President Roosevelts inangular speech in 1941 consists of 1323 words.
* In 1961, president kennedy gave similarly long inangular speech of 1364 words long.
* President Nioxn in 1973, consist of 1769 words.

Sentence count-

* It looks like the president Nixon gave longer sentences while delivering his speeches, because with larger word base he delivered equal number of sentences as other presidents.
* Both Roosevelts and Nixon inangular speech contains 68 sentences.
* Where as president Kennedy with 52 sentences delivers the speech.

**2.2) Remove all the stopwords from the three speeches.**

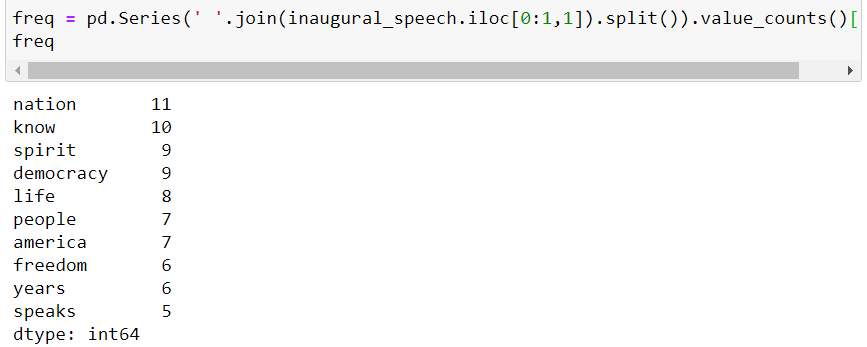


* Before removing the stopwords all the alphabets in the inangular sppecg texts were converted to lower case.
* After lower case conversion the punctuations including all special characters were removed from the speeches.
* A set of words were added next to the corpus of stop words.
* The stopwords were removed from the speeched..

(Note- speech , after removing the stopwords can be seen in the codebook.)

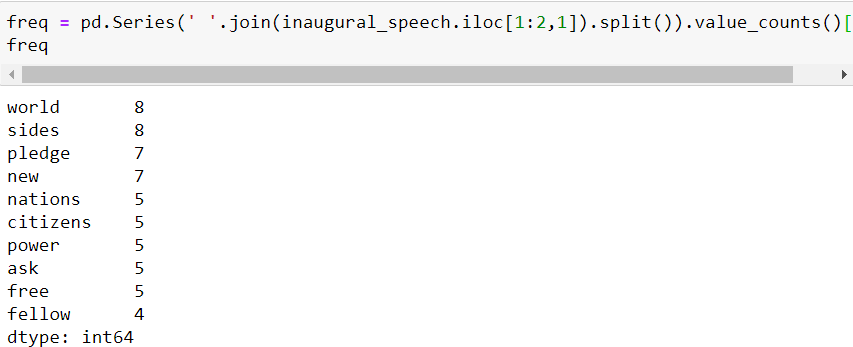
**2.3) Which word occurs the most number of times in his inaugural address for each president? Mention the top three words. (after removing the stopwords)**

ROOSEVELT 1941-



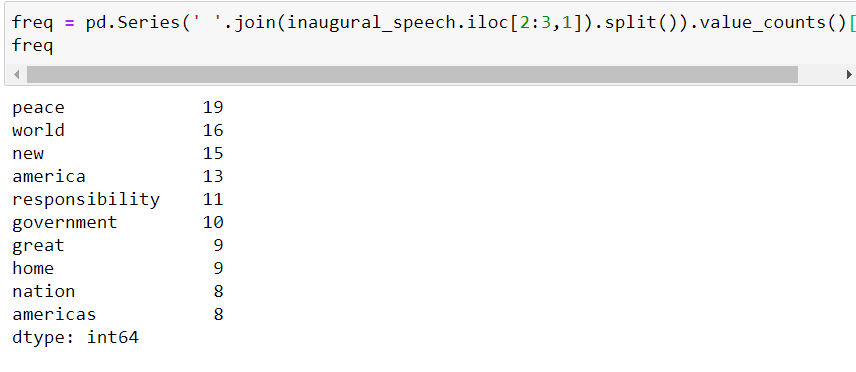
* In president Roosevelts inangural speech in 1941, ‘nation’ is the most often used word- 11 times, followed by ‘Know’ occurring 10 times.
* The word spirit and democracy came under third most used or occurring words in the speech i.e, 9 times.

KENNEDY 1961-



* President kennedy in his 1961 inangular speech- ‘World’ and ’sides’ were used most often, which is 8 times.
* And, followed by ‘new’ and ‘Pledge’ used most of the times, which is 7 times.

NIXON 1973-



* In his 1973 inangular speech, president Nixon used ‘peace’ word 19 times.
* Followed by ‘world’ 16 times.
* And ‘new’ & ‘America’ stand 3rd and 4th most used words in the speech. i.e, 15 and 13 times.

**2.4) Plot the word cloud of each of the three speeches. (after removing the stopwords)**

ROOSEVELT 1941-



KENNEDY 1961-



NIXON 1973-

