MACHINE LEARNING PROJECT REPORT

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.

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)

1. The dataset consists of 1525 observations and 9 variables.

2. 'Vote' and 'Gender' are the 2 categorical variables.

3. The rest of the variables are continuous numeric type variables.

	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

4. There is an unwanted column named "Unamed: 0" in the dataset that has been removed as shown from the above two images.

```
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
```

Information of the data types present in the dataset

	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge
count	1525.000000	1525.000000	1525.000000	1525.000000	1525.000000	1525.000000	1525.000000
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	24.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.000000
25%	41.000000	3.000000	3.000000	2.000000	2.000000	4.000000	0.000000
50%	53.000000	3.000000	3.000000	4.000000	2.000000	6.000000	2.000000
75%	67.000000	4.000000	4.000000	4.000000	4.000000	10.000000	2.000000
max	93.000000	5.000000	5.000000	5.000000	5.000000	11.000000	3.000000

Statistical summary table of the whole dataset

5. The minimum and maximum age of a voter is 24 years and 93 years respectively.

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

6. There are no null values present as shown in the above image.

7.There are 8 duplicate rows present in the dataset, and they have been removed from the dataset.

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

Number of Duplicate rows after treatment

```
Number of duplicate rows = 0
```

Number of Duplicate rows before treatment

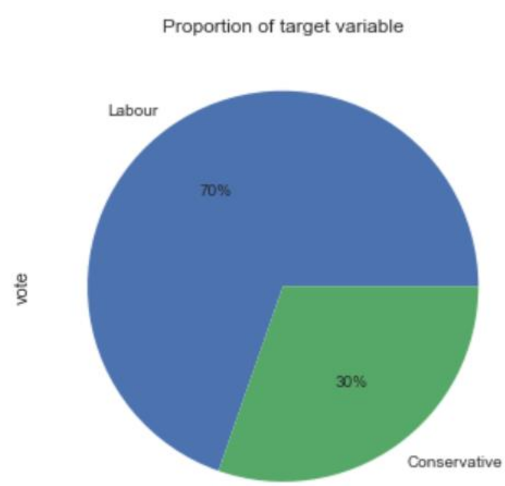
2. Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers. (7 Marks)

Exploratory Data Analysis

Proportion of Vote variable:

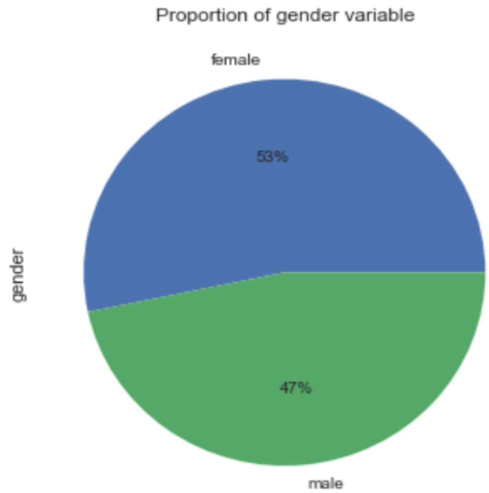
70% of voters voted for Labour party and 30% is for conservative party.

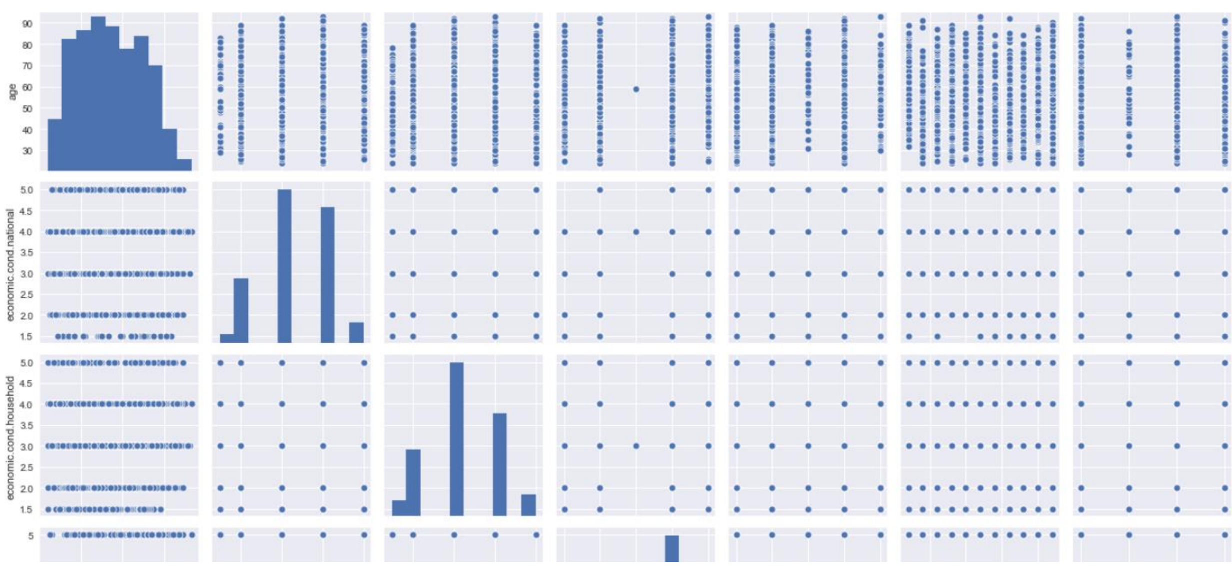
```
Labour          1051
Conservative     458
Name: vote, dtype: int64
```

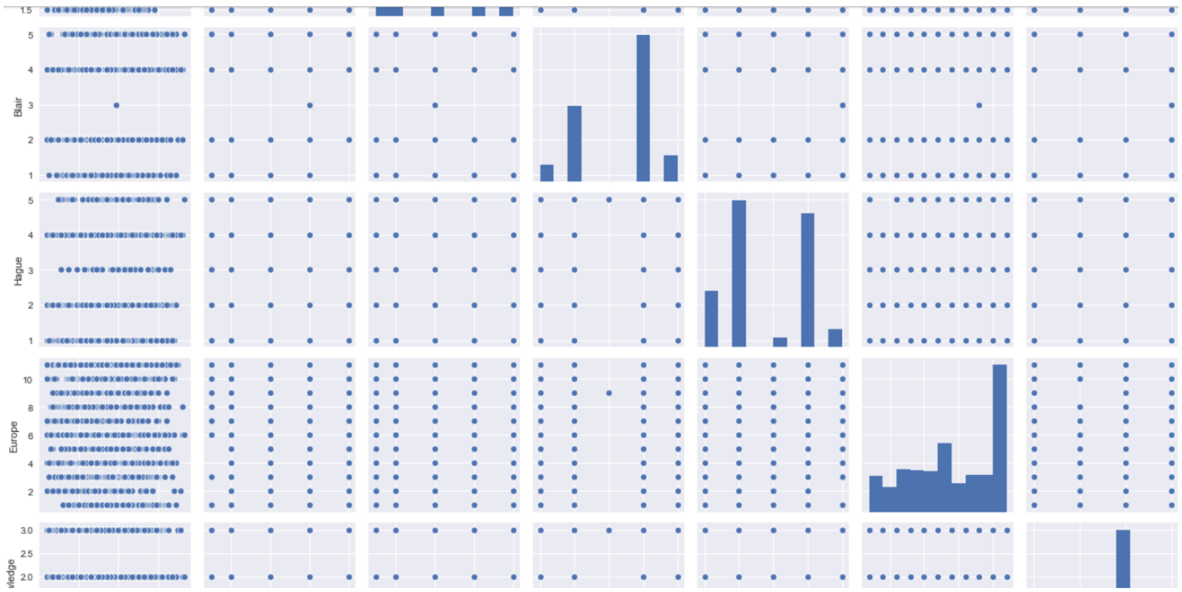


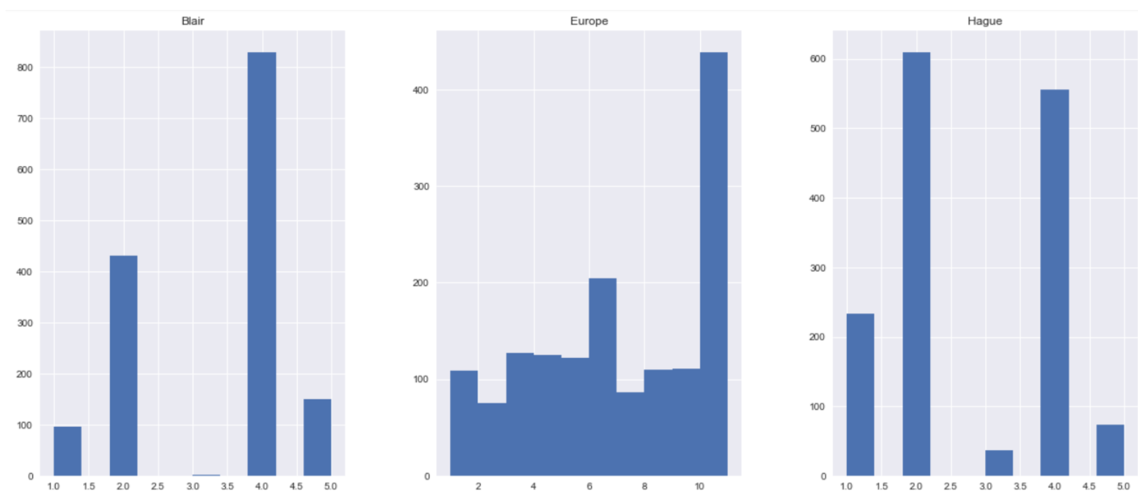
Proportion of Gender variable:

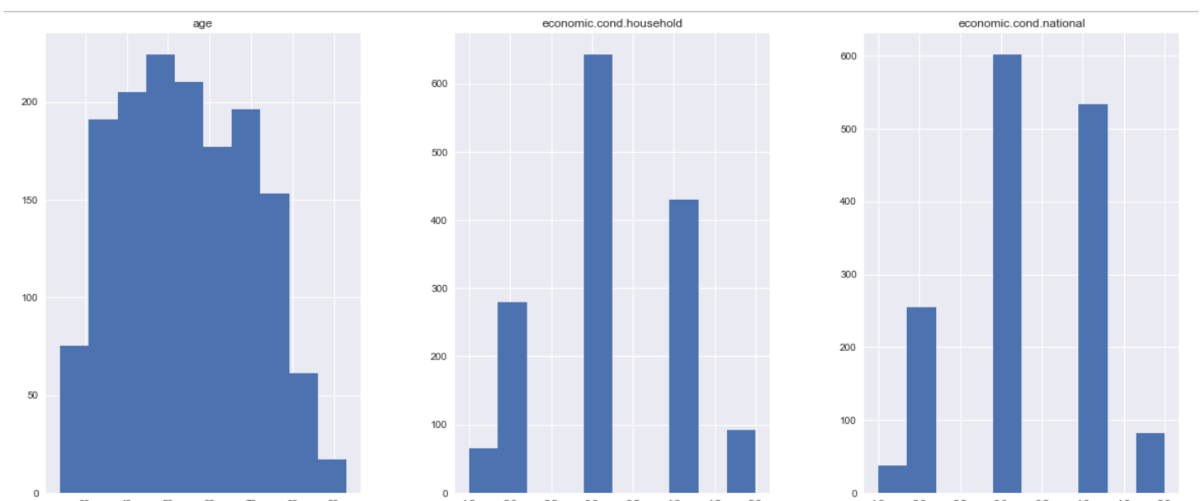
The proportion of female voters is 53% and whereas for male voters is 47%.

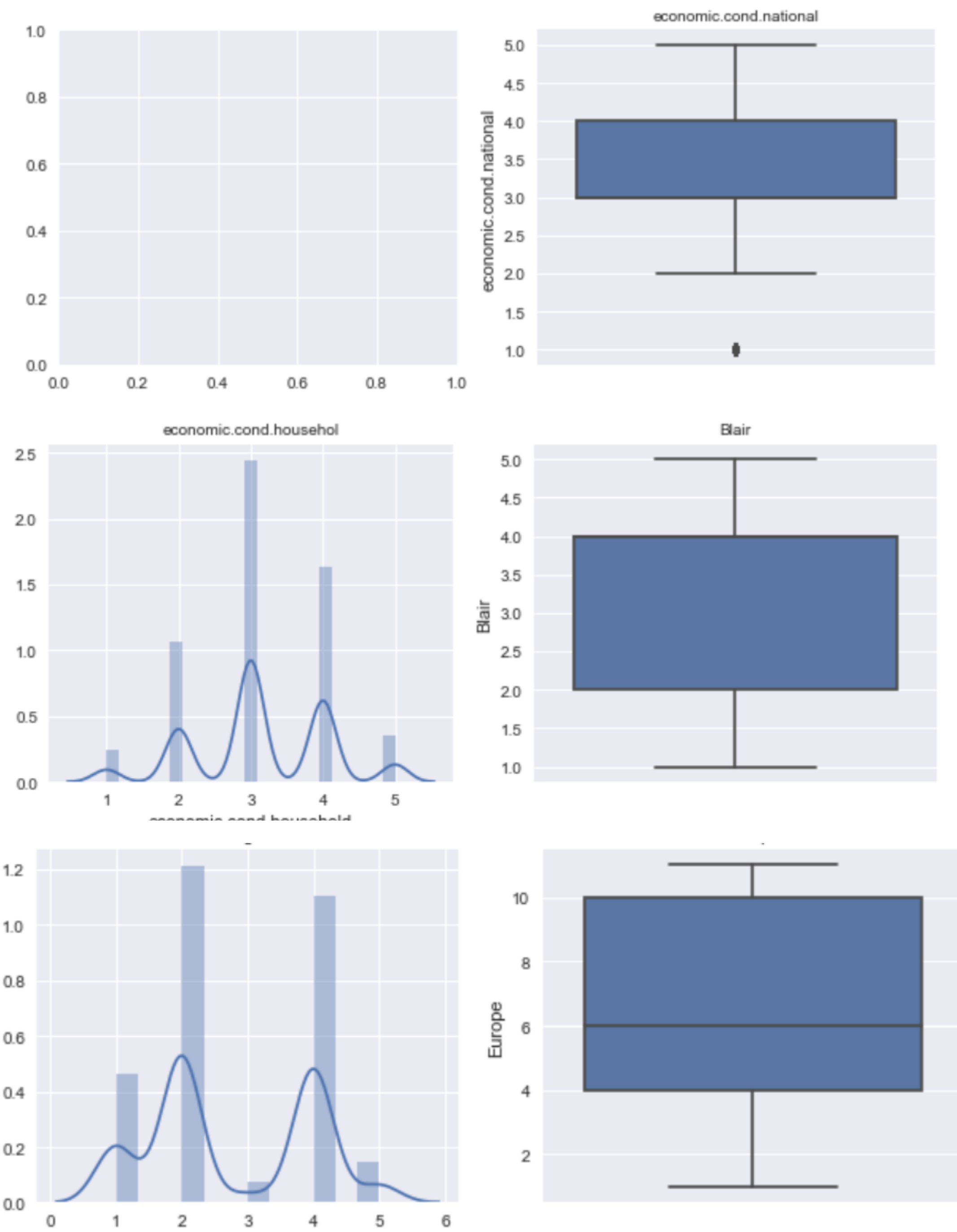


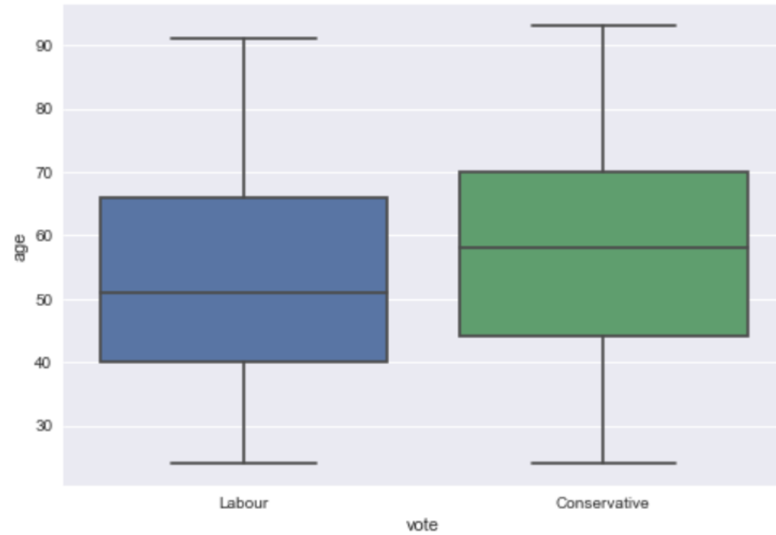



Blair
Hague
Europe
wledge

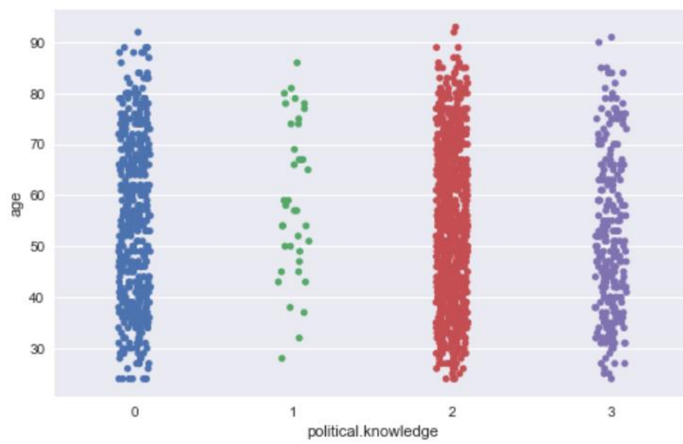

Blair
Europe
Hague

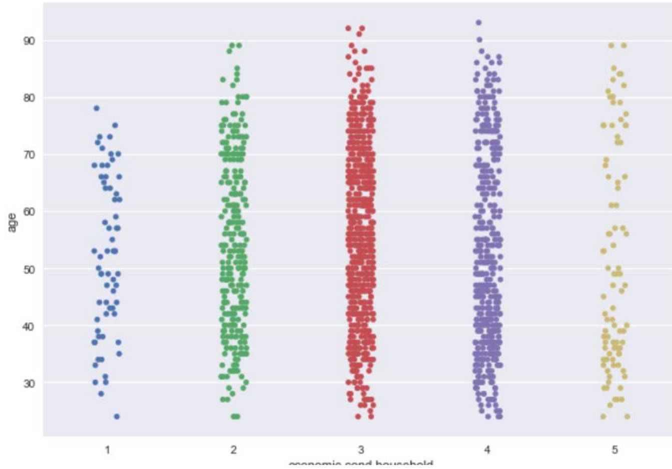

age
economic.cond.household
economic.cond.national


economic.cond.national
economic.cond.national
economic.cond.househol
Blair
Blair
Europe

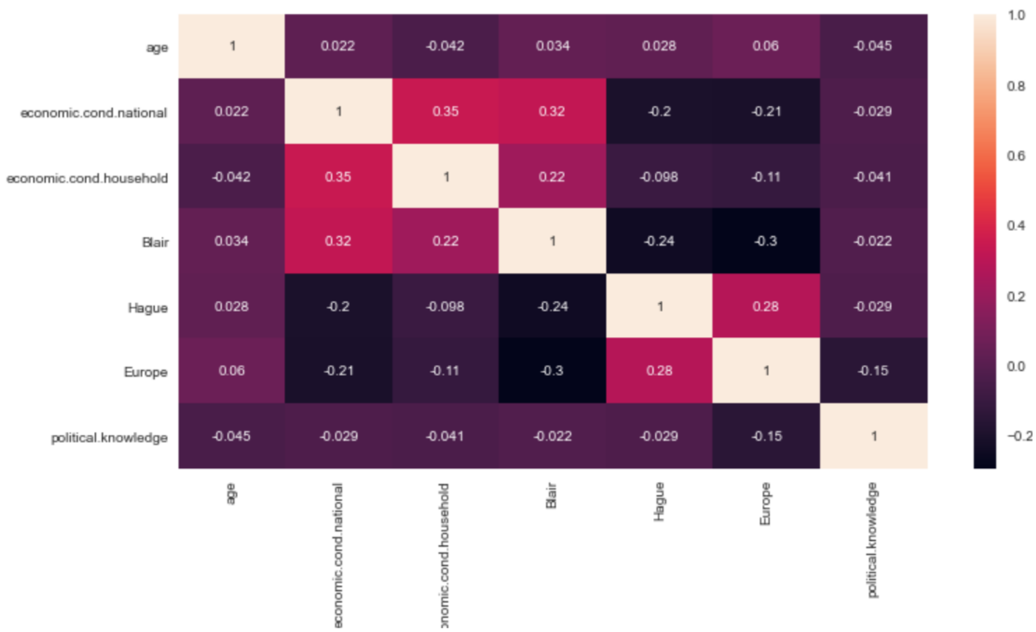


Older voter between 50 and 70 tend to vote for conservative party more.





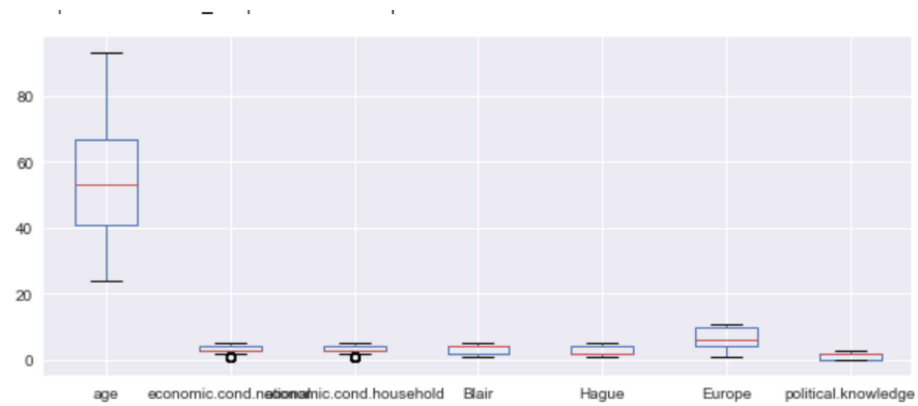
Pairplot:



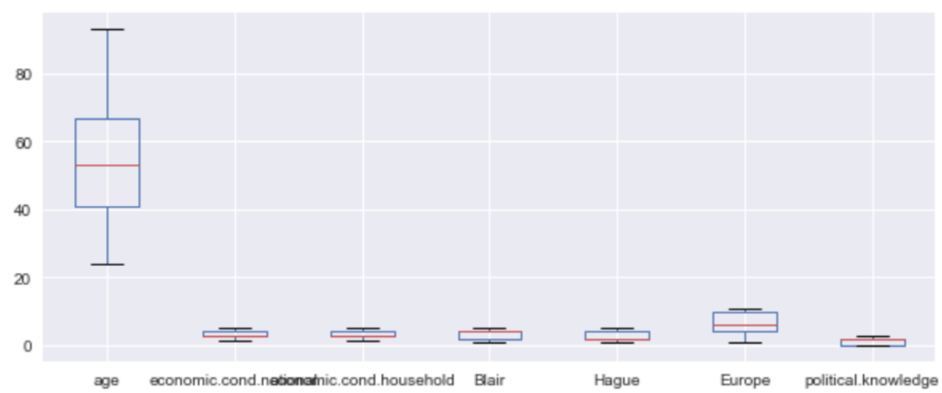
One of the inferences that can be drawn from the above pair plots and heat maps is that all the variables are not highly correlated to each other. There is no significant relationship between any variables.

Check for outliers:

There are outliers present in the dataset.



Outliers are to be treated to avoid influence of these variables on models.



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)

Data Encoding:

```
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   vote                     1509 non-null   int64
 1   age                      1509 non-null   float64
 2   economic.cond.national   1509 non-null   float64
 3   economic.cond.household  1509 non-null   float64
 4   Blair                    1509 non-null   float64
 5   Hague                    1509 non-null   float64
 6   Europe                   1509 non-null   float64
 7   political.knowledge      1509 non-null   float64
 8   gender                   1509 non-null   int64
dtypes: float64(7), int64(2)
memory usage: 117.9 KB
```

Vote and gender object data types are converted into integer data types for modelling as the algorithms do not accept string values.

Scaling:

As there are continuous variables present in the dataset, scaling has been done due to different weightage in the continuous variables.

Although, Scaling isn't necessary in this case, as it did not have any significant impact on the model results. Scaling would be effective if the algorithm use gradient descent method to converge faster. Scaling has been done on the dataset using z-scores, but there is no improvement in the results, therefore all the steps below have been done on an unscaled dataset.

The dataset is split into training set and testing set.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30 , random_state=1)
```

Modelling: 26 marks

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

Logistic Regression

Logistic regression model has been applied on split dataset.

```
LogisticRegression(max_iter=10000, n_jobs=2, penalty='none', verbose=True)
```

LDA

```
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
#Build LDA Model
clf = LinearDiscriminantAnalysis()
model=clf.fit(X_train,y_train)
```

Model	Training set accuracy	Testing set accuracy
Logistic Regression	83.42%	82.7%
LDA	83%	83%

The performance of the both models are similar. Their accuracy scores do not significantly vary. So we can choose any one of the models for solving the business case, as the results wouldn't be significantly different.

2. Apply KNN Model and Naïve Bayes Model. Interpret the results. (7 marks)

KNN Model

```
from sklearn.neighbors import KNeighborsClassifier

KNN_model=KNeighborsClassifier()
KNN_model.fit(X_train,y_train)
```

```
KNeighborsClassifier()
```

Naïve Bayes Model

```
from sklearn.naive_bayes import GaussianNB # using Gaussian algorithm from Naive Bayes

# creatw the model
naive_model = GaussianNB()

naive_model.fit(X_train, y_train.ravel())
```

```
GaussianNB()
```

Model	**Training set accuracy**	**Testing set accuracy**
Naïve Bayes	83%	83%
KNN	85.5%	77%

KNN has a higher training set accuracy than Naïve bayes. Although, the accuracy rate on an unseen data for naïve bayes is higher when compared to KNN.Therefore, Naïve bayes is the best model in comparison with KNN.

3. Model Tuning, Bagging (Random Forest should be applied for Bagging) and Boosting. (7 marks)

Model tuning has been done to 3 algorithm models:

Logistic Regression, KNN and Random forest.

This is done to maximize the recall rate and Gridsearch method is used in the process of model tuning.

Logistic Regression:

```
Accuracy Score : 0.8278145695364238
Precision Score : 0.7384615384615385
Recall Score : 0.6857142857142857
F1 Score : 0.7111111111111111
```

There has been an improvement in the recall rate of the training data. Initial recall rate was 0.65 and now it is 0.68

KNN:

```
0.8115363881401617
{'n_neighbors': 23}
```

Score and best parameter for KNN.

KNN model has being used with the best parameter and the accuracy scores and recall rates are shown in the below screenshots:

Training set:

```
Model Accuracy: 0.8258
```

	precision	recall	f1-score	support
0	0.84	0.93	0.88	738
1	0.79	0.58	0.67	318
accuracy			0.83	1056
macro avg	0.81	0.75	0.77	1056
weighted avg	0.82	0.83	0.82	1056

Testing set:

```
Model Accuracy: 0.8256
```

	precision	recall	f1-score	support
0	0.87	0.89	0.88	313
1	0.75	0.71	0.73	140
accuracy			0.84	453
macro avg	0.81	0.80	0.81	453
weighted avg	0.83	0.84	0.84	453

Recall rate has been maximized for the testing set from 0.60 to 0.71. Therefore, KNN model tuning worked absolutely fine.

Random Forest

```
from sklearn.ensemble import BaggingClassifier
from sklearn.ensemble import RandomForestClassifier
cart = RandomForestClassifier()
Bagging_model=BaggingClassifier(base_estimator=cart,n_estimators=100,random_state=1)
Bagging_model.fit(X_train, y_train)
```

Boosting

```python
from sklearn.ensemble import GradientBoostingClassifier
gbcl = GradientBoostingClassifier(random_state=1)
```

Model	Training set accuracy	Testing set accuracy
Bagging	97%	84%
Boosting	88%	83.6%

Bagging using Random forest classifier on training set has a very high accuracy rate, although on testing test, the score decreased to 84%.

The accuracies of both the models on the testing sets are almost the same. Hence, we have to look at the recall rate and other metrics to decide the best optimal model for the case study

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)

Let's look at the performance of all the models.

1. **Logistic regression:**

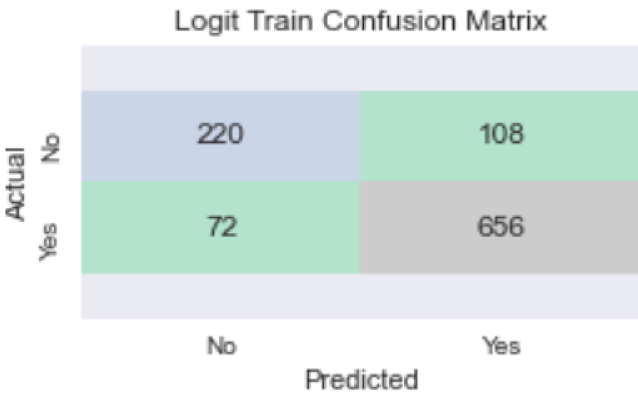
Training dataset

- Classification report

```
              precision    recall  f1-score   support

           0       0.86      0.91      0.89       738
           1       0.77      0.65      0.70       318

    accuracy                           0.83      1056
   macro avg       0.81      0.78      0.79      1056
weighted avg       0.83      0.83      0.83      1056
```

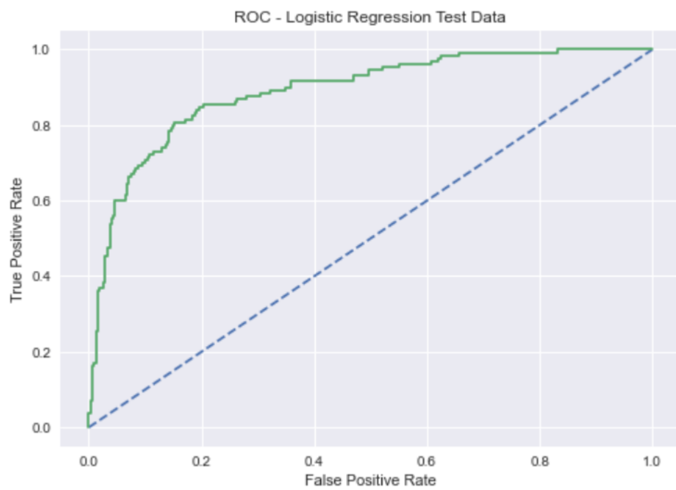
Training set accuracy: 83.42%

- Confusion matrix:



- ROC curve:

AUC: 0.890



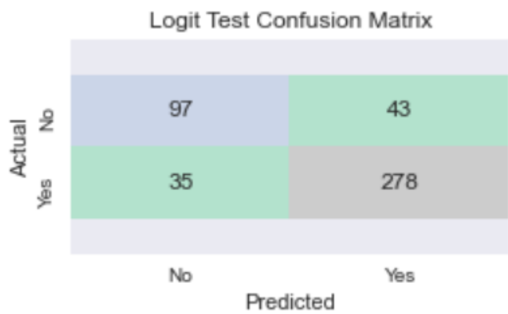
- AUC score:0.890

Testing dataset

- Classification report

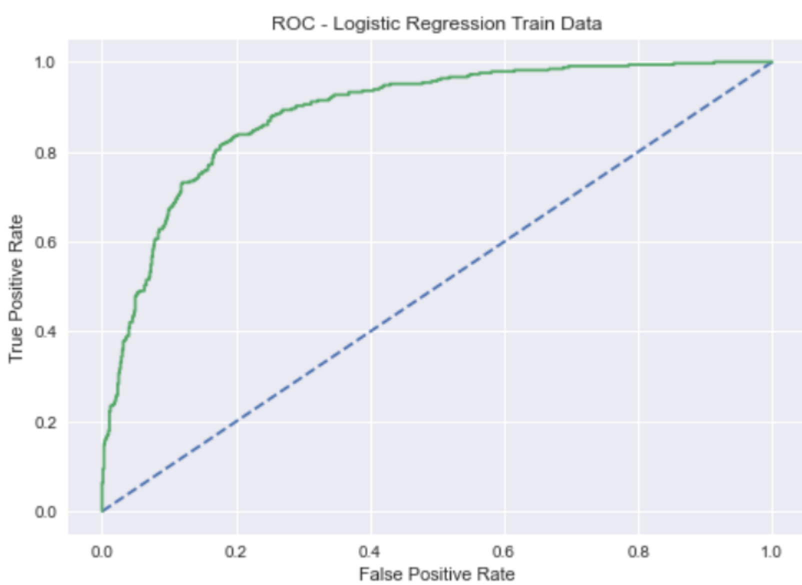
```
logit_test_precision  0.79
logit_test_recall  0.66
logit_test_f1  0.72
```

- Testing set accuracy: 82.7%
- Confusion matrix:



- ROC curve:

```
AUC: 0.887
```



- AUC score: 0.887

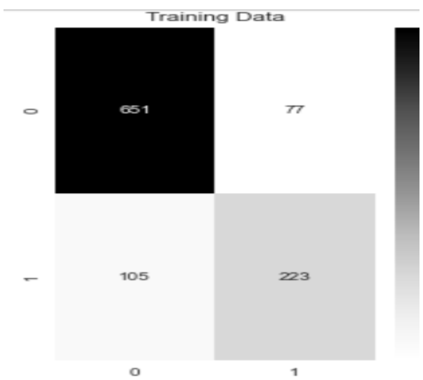
LDA:

Training dataset

- Classification report

```
Classification Report of the training data:

              precision    recall  f1-score   support

           0       0.86      0.89      0.88       728
           1       0.74      0.68      0.71       328

    accuracy                           0.83      1056
   macro avg       0.80      0.79      0.79      1056
weighted avg       0.82      0.83      0.83      1056
```

- Training set accuracy: 83%
- Confusion matrix:



- ROC curve:

```
AUC for the Training Data: 0.887
AUC for the Test Data: 0.891
```

- AUC score: 0.887

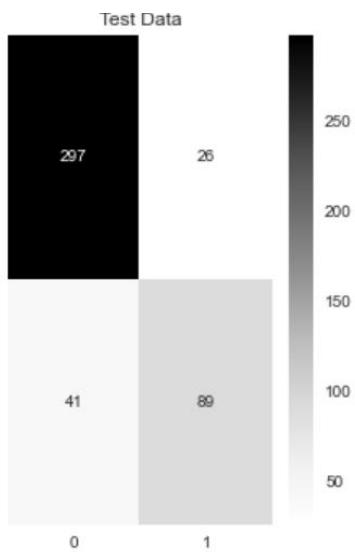
Testing dataset

- classification report

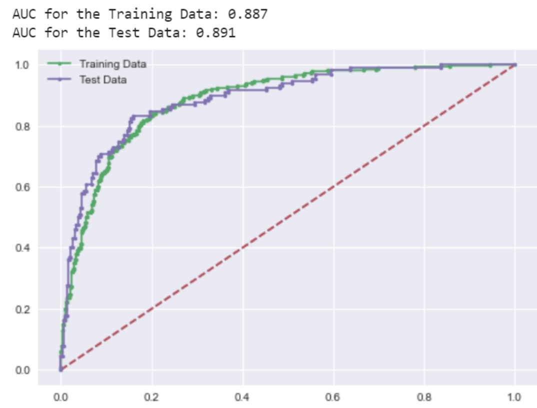
Classification Report of the test data:

	precision	recall	f1-score	support
0	0.87	0.88	0.88	313
1	0.73	0.70	0.72	140
accuracy			0.83	453
macro avg	0.80	0.79	0.80	453
weighted avg	0.83	0.83	0.83	453

- Testing set accuracy: 83%
- Confusion matrix:



- ROC curve:



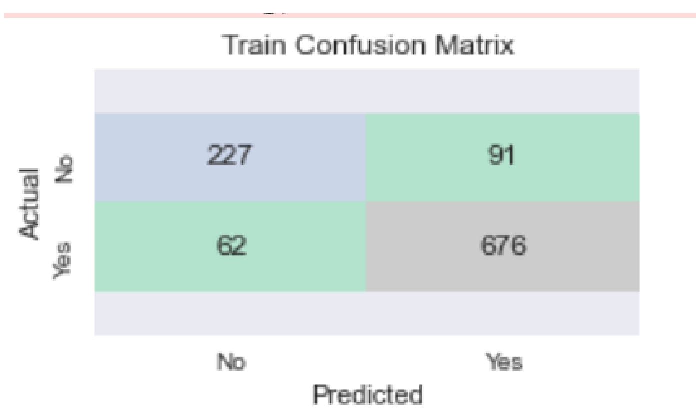
- AUC score: 0.891

KNN Model:

Training dataset

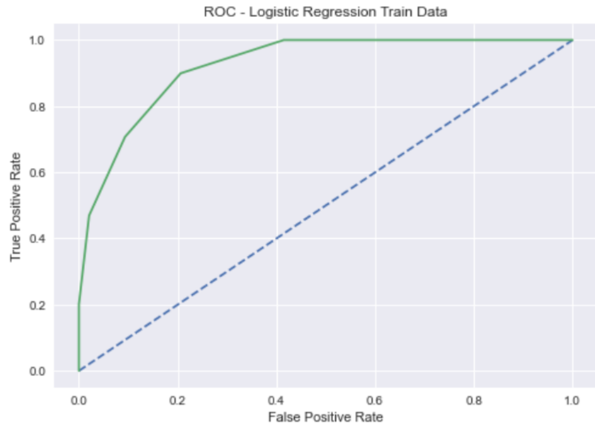
	precision	recall	f1-score	support
0	0.88	0.92	0.90	738
1	0.79	0.71	0.75	318
accuracy			0.86	1056
macro avg	0.83	0.81	0.82	1056
weighted avg	0.85	0.86	0.85	1056

- Training set accuracy: 85.5%
- Confusion matrix:



- ROC curve:

AUC: 0.924



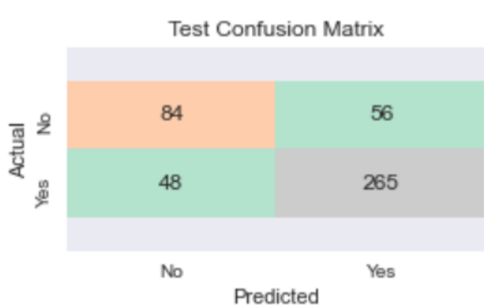
- AUC score: 0.924

Testing dataset

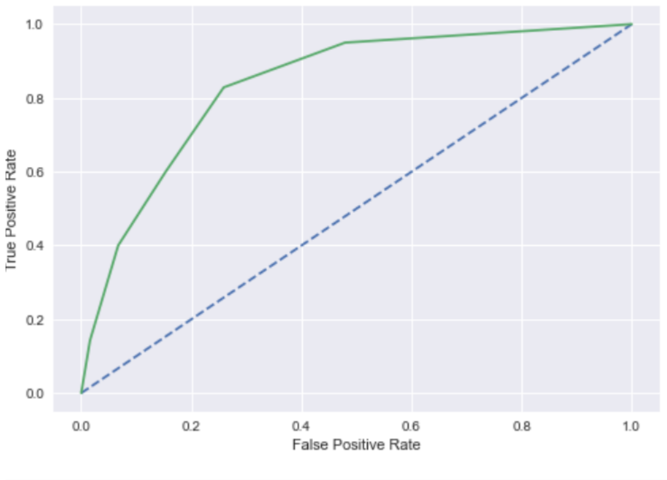
- classification report

	precision	recall	f1-score	support
0	0.83	0.85	0.84	313
1	0.64	0.60	0.62	140
accuracy			0.77	453
macro avg	0.73	0.72	0.73	453
weighted avg	0.77	0.77	0.77	453

- Testing set accuracy: 77%
- Confusion matrix:



- ROC curve:



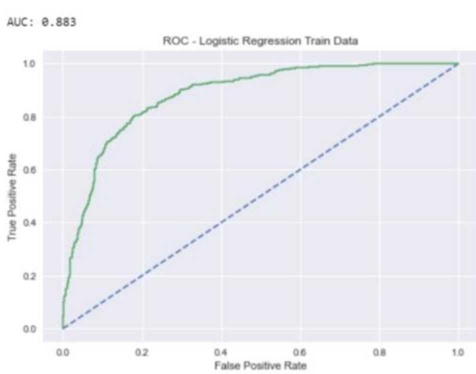
- AUC score: 0.837

Naïve Bayes Model:

Training dataset

```
Model Accuracy: 0.8286
```

- Training set accuracy: 82.86%
- ROC curve:



- AUC score:0.883

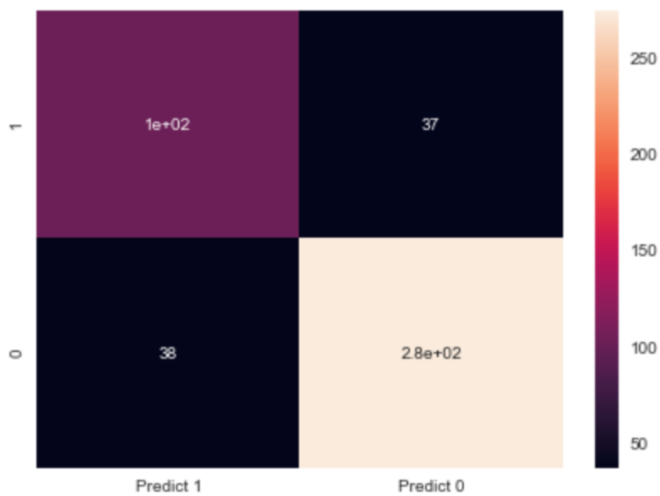
Testing dataset

- classification report

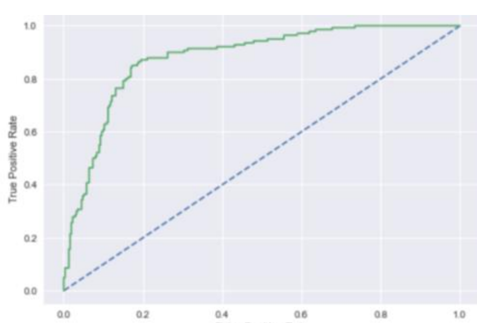
Classification Report

	precision	recall	f1-score	support
1	0.73	0.74	0.73	140
0	0.88	0.88	0.88	313
accuracy			0.83	453
macro avg	0.81	0.81	0.81	453
weighted avg	0.83	0.83	0.83	453

- Testing set accuracy: 83%
- Confusion matrix:



- ROC curve:



- AUC score:0.88

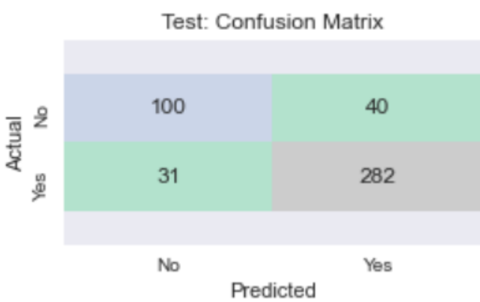
Bagging Model(Random Forest):

Training dataset

- classification report

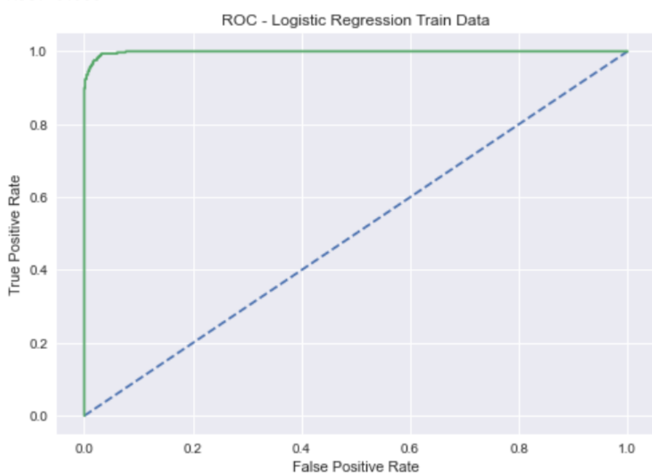
```
0.9715909090909091
[[730   8]
 [ 22 296]]
              precision    recall  f1-score   support

           0       0.97      0.99      0.98       738
           1       0.97      0.93      0.95       318

    accuracy                           0.97      1056
   macro avg       0.97      0.96      0.97      1056
weighted avg       0.97      0.97      0.97      1056
```

- Training set accuracy: 97%
- Confusion matrix:



- ROC curve:

```
AUC: 0.999
```



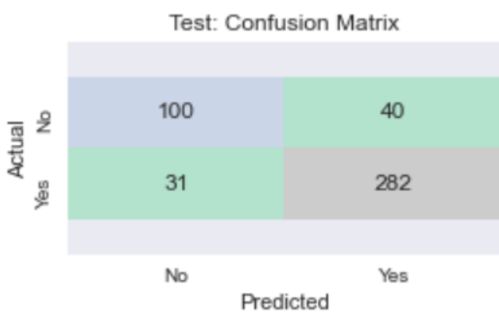
- AUC score: 0.99

Testing dataset

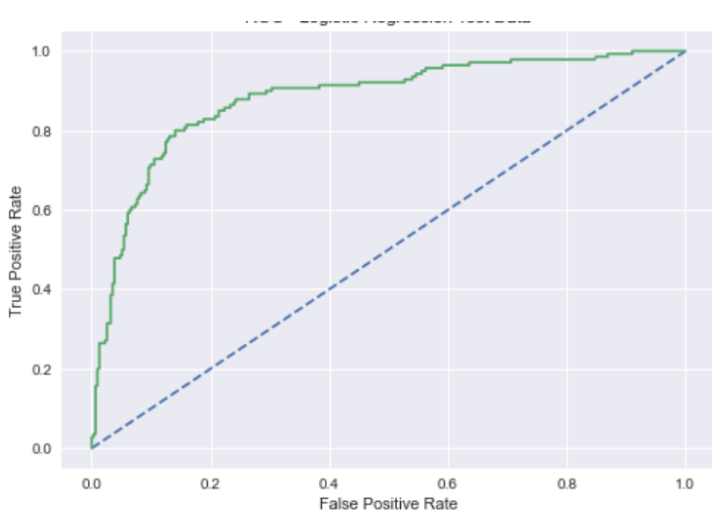
- classification report

	precision	recall	f1-score	support
0	0.88	0.90	0.89	313
1	0.76	0.71	0.74	140
accuracy			0.84	453
macro avg	0.82	0.81	0.81	453
weighted avg	0.84	0.84	0.84	453

- Testing set accuracy: 85%
- Confusion matrix:



- ROC curve:



- AUC score: 0.881

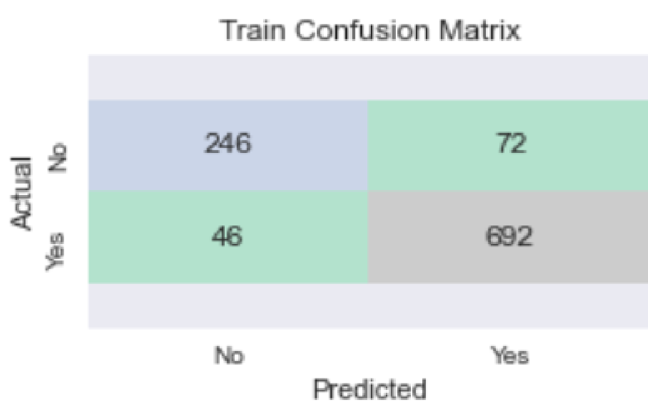
Boosting Model:

Training dataset

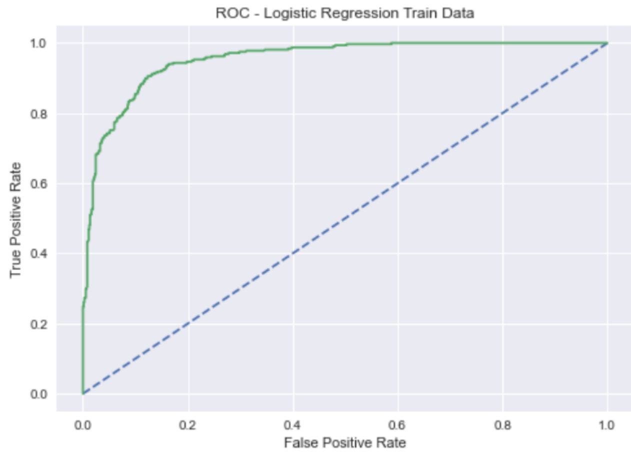
- classification report

```
0.8882575757575758
[[692  46]
 [ 72 246]]
              precision    recall  f1-score   support

           0       0.91      0.94      0.92       738
           1       0.84      0.77      0.81       318

    accuracy                           0.89      1056
   macro avg       0.87      0.86      0.86      1056
weighted avg       0.89      0.89      0.89      1056
```

- Training set accuracy: 88.8%
- Confusion matrix:



- ROC curve:



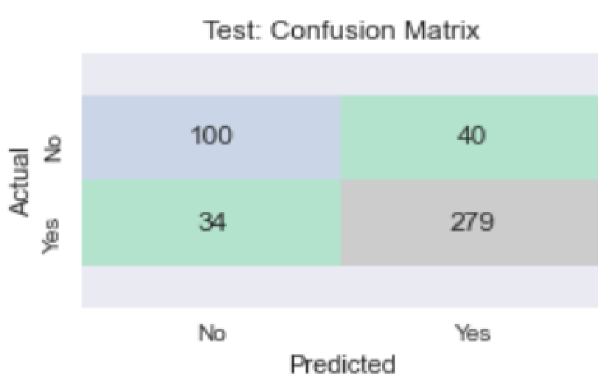
- AUC score: 0.954

Testing dataset

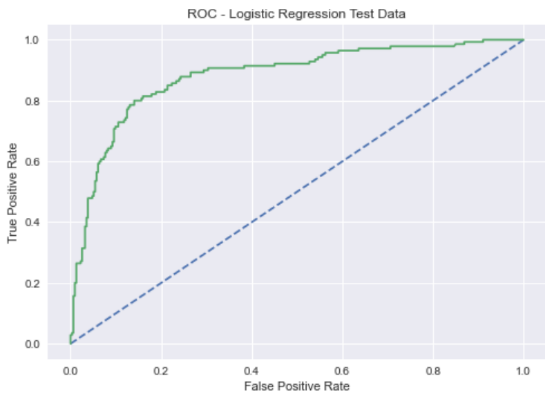
- classification report

```
0.8366445916114791
[[279  34]
 [ 40 100]]
              precision    recall  f1-score   support

           0       0.87      0.89      0.88       313
           1       0.75      0.71      0.73       140

    accuracy                           0.84       453
   macro avg       0.81      0.80      0.81       453
weighted avg       0.83      0.84      0.84       453
```

- Testing set accuracy: 83.6%
- Confusion matrix:



- ROC curve:



- AUC score: 0.881

Recall refers to the percentage of total relevant results correctly classified by the algorithm and hence we will compare Recall of class "1" for all models.

Based on the training and testing dataset performance metrics, recall for class "1" for the following models are as follows:

Models	**Recall - Train**	**Recall - Test**
Logistic Regression	0.65	0.69
LDA	0.68	0.70
KNN	0.71	0.60
Naïve Bayes	0.71	0.73
Bagging(Random Forest)	0.93	0.71
Boosting	0.77	0.71

So as per the train data, Worst performing models are – Logistic Regression, Linear Discriminant Analysis. Best Performing model among all the six models is Bagging (Random forest), however this model is clearly over fitted.

After looking at the recall rate of testing dataset, Naïve bayes has a good recall rate.

Models	Accuracy – Train	Accuracy- Test
Logistic Regression	83%	82.7%
LDA	83%	83%
KNN	85.5%	77%
Naïve Bayes	83%	83%
Bagging(Random Forest)	97%	84%
Boosting	88%	83.6%

All models have only slight difference in terms of accuracy rate and recall rate for class “1” and they have delivered satisfactory results.

But based on the scores, Bagging using random forest and Gradient boosting algorithms performed well in comparison to others.

Bagging using Random forest classifier is best suitable for this business case study as it’s accuracy and recall scores are 84% and 0.71 are higher on unseen data.

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

News channel can determine more positively correlated features to build a better model.

Factors such as marital status, diversity index, annual income etc; can also be included in the dataset.

Education, Place and information of who they voted for in the previous election etc; can be used as one of the variables to determine the voter’s choice.

Preferences change modestly and using data from previous elections will definitely help us in building a better and accurate model.

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:

1. **President Franklin D. Roosevelt in 1941**
2. **President John F. Kennedy in 1961**
3. **President Richard Nixon in 1973**

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

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

Case 1: President Franklin D. Roosevelt 1941

	president	text	char_count
1941-Roosevelt	Roosevelt - 1941	On each national day of inauguration since 178...	7571

Number of characters: 7571

]:

	president	text	char_count	word_count
1941-Roosevelt	Roosevelt - 1941	On each national day of inauguration since 178...	7571	1323

Number of words:1323

	president	text	char_count	word_count	sents_count
1941-Roosevelt	Roosevelt - 1941	On each national day of inauguration since 178...	7571	1323	68

Number of sentences:68

Case 2: President John F. Kennedy in 1961

1961-Kennedy	Kennedy - 1961	Vice President Johnson, Mr. Speaker, Mr. Chief...	7618

Number of characters: 7618

1961-Kennedy	Kennedy - 1961	Vice President Johnson, Mr. Speaker, Mr. Chief...	7618	1364

Number of words:1364

1961-Kennedy	Kennedy - 1961	Vice President Johnson, Mr. Speaker, Mr. Chief...	7618	1364	52

Number of sentences:52

Case 3: President Richard Nixon in 1973

1973-Nixon	Nixon - 1973	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	9991

Number of characters: 9991

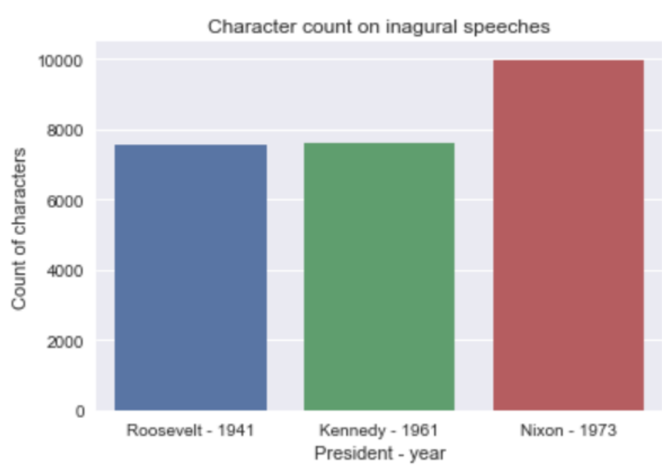
1973-Nixon	Nixon - 1973	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	9991	1769

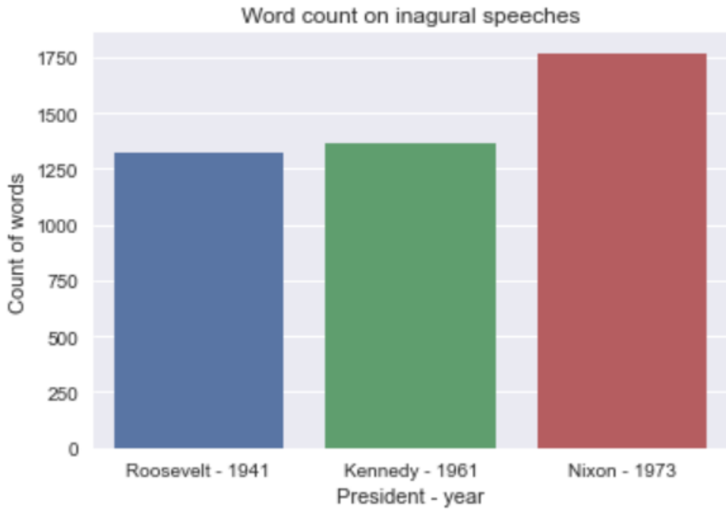
Number of words:1769

1973-Nixon	Nixon - 1973	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	9991	1769	68

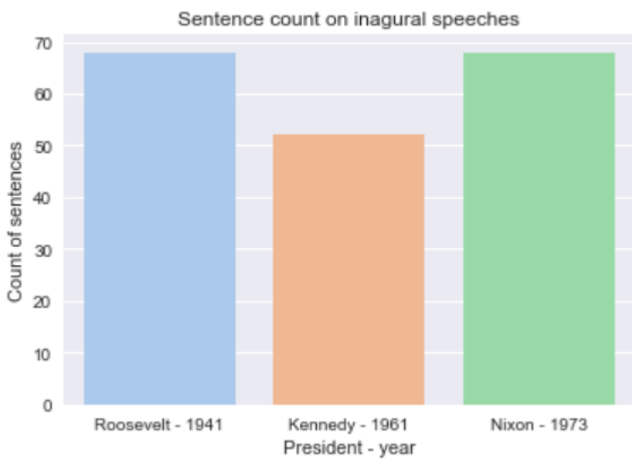
Number of sentences:68

Graphical representations:






```
Text(0, 0.5, 'Count of sentences')
```



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

Stop words are stored in a list in NLTK. Stop words can be removed using the following commands shown in the below picture. Modification has been done by adding the stop words that deemed to be necessary to the list.

```
from nltk.corpus import stopwords
stop_words = stopwords.words('english')
add_to_stop_words = ['mr','on','it','let','to','us','shall','the','in']
stop_words.extend(add_to_stop_words)
stop_words=set(stop_words)
```

```
inaugural_speech['text'] = inaugural_speech['text'].apply(lambda x: " ".join(x for x in x.split() if x not in stop_words))
inaugural_speech
```

	president	text	char_count	word_count	sents_count
1941-Roosevelt	Roosevelt - 1941	national day inauguration since 1789 people re...	7571	1323	68
1961-Kennedy	Kennedy - 1961	vice president johnson speaker chief justice p...	7618	1364	52
1973-Nixon	Nixon - 1973	vice president speaker chief justice senator c...	9991	1769	68

1941-Roosevelt Speech: Text after removing stop words.

```
list(inaugural_speech[inaugural_speech['president']=="Roosevelt - 1941"].text)
```

```
['national day inauguration since 1789 people renewed sense dedication united states washingtons day task people create weld toget
her nation lincolns day task people preserve nation disruption within day task people save nation institutions disruption without
come time midst swift happenings pause moment take stock recall place history rediscover may risk real peril inaction lives nation
s determined count years lifetime human spirit life man threescore years ten little little less life nation fullness measure live
men doubt men believe democracy form government frame life limited measured kind mystical artificial fate unexplained reason tyran
ny slavery become surging wave future freedom ebbing tide americans know true eight years ago life republic seemed frozen fatalist
ic terror proved true midst shock acted acted quickly boldly decisively later years living years fruitful years people democracy b
rought greater security hope better understanding lifes ideals measured material things vital present future experience democracy
successfully survived crisis home put away many evil things built new structures enduring lines maintained fact democracy action t
aken within threeway framework constitution united states coordinate branches government continue freely function bill rights rema
ins inviolate freedom elections wholly maintained prophets downfall american democracy seen dire predictions come naught democracy
dying know seen reviveand grow know cannot die built unhampered initiative individual men women joined together common enterprise
enterprise undertaken carried free expression free majority know democracy alone forms government enlists full force mens enlighte
ned know democracy alone constructed unlimited civilization capable infinite progress improvement human life know look surface sen
se still spreading every continent humane advanced end unconquerable forms human society nation like person bodya body must fed cl
othed housed invigorated rested manner measures objectives time nation like person mind mind must kept informed alert must know un
derstands hopes needs neighbors nations live within narrowing circle world nation like person something deeper something permanent
something larger sum parts something matters future calls forth sacred guarding present thing find difficult even impossible hit u
pon single simple word yet understand spirit faith america product centuries born multitudes came many lands high degree mostly pl
ain people sought early late find freedom freely democratic aspiration mere recent phase human history human history permeated anc
ient life early peoples blazed anew middle ages written magna charta americas impact irresistible america new world tongues people
s continent newfound land came believed could create upon continent new life life new freedom vitality written mayflower compact d
eclaration independence constitution united states gettysburg address first came carry longings spirit millions followed stock spr
ang moved forward constantly consistently toward ideal gained stature clarity generation hopes republic cannot forever tolerate ei
ther undeserved poverty selfserving wealth know still far go must greatly build security opportunity knowledge every citizen measu
```

1961-Kennedy Speech: Text after removing stop words.

```
list(inaugural_speech[inaugural_speech['president']=="Kennedy - 1961"].text)
```

```
['vice president johnson speaker chief justice president eisenhower vice president nixon president truman reverend clergy fellow c
itizens observe today victory party celebration freedom symbolizing end well beginning signifying renewal well change sworn almigh
ty god solemn oath forebears l prescribed nearly century three quarters ago world different man holds mortal hands power abolish f
orms human poverty forms human life yet revolutionary 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 time place friend foe alike torch passed new generat
ion americans born century tempered war disciplined hard bitter peace proud ancient heritage unwilling witness permit slow undoing
human rights nation always committed committed today home around world every nation know whether wishes well ill pay price bear bu
rden meet hardship support friend oppose foe order assure survival success liberty much pledge old allies whose cultural spiritual
origins share pledge loyalty faithful friends united little cannot host cooperative ventures divided little dare meet powerful cha
llenge odds split asunder new states welcome ranks free pledge word one form colonial control passed away merely replaced far iron
tyranny always expect find supporting view always hope find strongly supporting freedom remember past foolishly sought power ridin
g back tiger ended inside peoples huts villages across globe struggling break bonds mass misery pledge best efforts help help what
ever period required communists may seek votes right free society cannot help many poor cannot save rich sister republics south bo
rder offer special pledge convert good words good deeds new alliance progress assist free men free governments casting chains pove
rty peaceful revolution hope cannot become prey hostile powers neighbors know join oppose aggression subversion anywhere americas
every power know hemisphere intends remain master house world assembly sovereign states united nations last best hope age instrume
nts war far outpaced instruments peace renew pledge supportto prevent becoming merely forum invective strengthen shield new weak e
nlarge area writ may run finally nations would make adversary offer pledge request sides begin anew quest peace dark powers destru
ction unleashed science engulf humanity planned accidental selfdestruction dare tempt weakness arms sufficient beyond doubt certai
n beyond doubt never employed neither two great powerful groups nations take comfort present course sides overburdened cost modern
weapons rightly alarmed steady spread deadly atom yet racing alter uncertain balance terror stays hand mankinds final war begin an
ew remembering sides civility sign weakness sincerity always subject proof never negotiate fear never fear negotiate sides explore
problems unite instead belaboring problems divide sides first time formulate serious precise proposals inspection control arms bri
ng absolute power destroy nations absolute control nations sides seek invoke wonders science instead terrors together explore star
s conquer deserts eradicate disease tap ocean depths encourage arts commerce sides unite heed corners earth command isaiah undo he
```

1973-Nixon Speech: Text after removing stop words.

```
list(inaugural_speech[inaugural_speech['president']=="Nixon - 1973"].text)
```

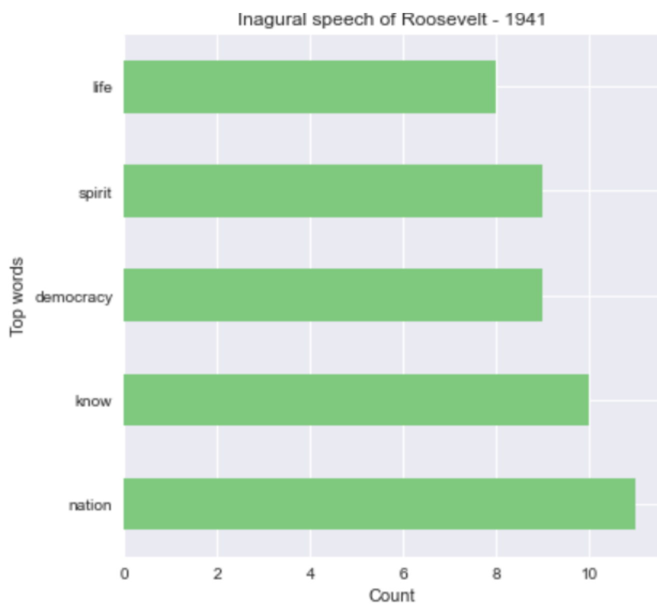
```
['vice president speaker chief justice senator cook mrs eisenhower fellow citizens great good country share together met four year
s ago america bleak spirit depressed prospect seemingly endless war abroad destructive conflict home meet today stand threshold ne
w era peace world central question use peace resolve era enter postwar periods often time retreat isolation leads stagnation home
invites new danger abroad resolve become time great responsibilities greatly borne renew spirit promise america enter third centur
y nation past year saw farreaching results new policies peace continuing revitalize traditional friendships missions peking moscow
able establish base new durable pattern relationships among nations world americas bold initiatives 1972 long remembered year grea
test progress since end world war ii toward lasting peace world peace seek world flimsy peace merely interlude wars peace endure g
enerations come important understand necessity limitations americas role maintaining peace unless america work preserve peace peac
e unless america work preserve freedom freedom clearly understand new nature americas role result new policies adopted past four y
ears respect treaty commitments support vigorously principle country right impose rule another force continue era negotiation work
limitation nuclear arms reduce danger confrontation great powers share defending peace freedom world expect others share time pass
ed america make every nations conflict make every nations future responsibility presume tell people nations manage affairs respect
right nation determine future also recognize responsibility nation secure future americas role indispensable preserving worlds pea
ce nations role indispensable preserving peace together rest world resolve move forward beginnings made continue bring walls hosti
lity divided world long build place bridges understanding despite profound differences systems government people world friends bui
ld structure peace world weak safe strong respects right live different system would influence others strength ideas force arms ac
cept high responsibility burden gladly gladly chance build peace noblest endeavor nation engage gladly also act greatly meeting re
sponsibilities abroad remain great nation remain great nation act greatly meeting challenges home chance today ever history make l
ife better america ensure better education better health better housing better transportation cleaner environment restore respect
law make communities livable insure godgiven right every american full equal opportunity range needs great reach opportunities gre
at bold determination meet needs new ways building structure peace abroad required turning away old policies failed building new e
ra progress home requires turning away old policies failed abroad shift old policies new retreat responsibilities better way peace
home shift old policies new retreat responsibilities better way progress abroad home key new responsibilities lies placing divisio
n responsibility lived long consequences attempting gather power responsibility washington abroad home time come turn away condesc
ending policies paternalism washington knows best person expected act responsibly responsibility human nature encourage individual
```

- 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

<u>1941 Inaugral speech:</u>

Top three words are nation, know and democracy.

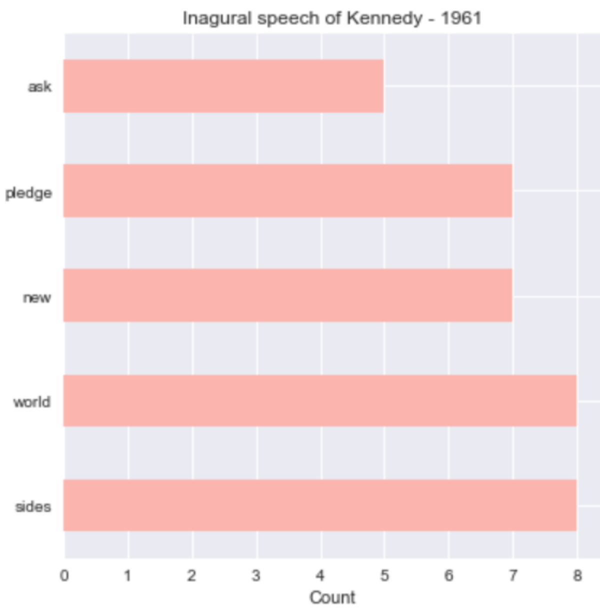
```
nation       11
know         10
democracy     9
```



1961 Inaugral speech:

Top three words are sides, world and new/pledge.

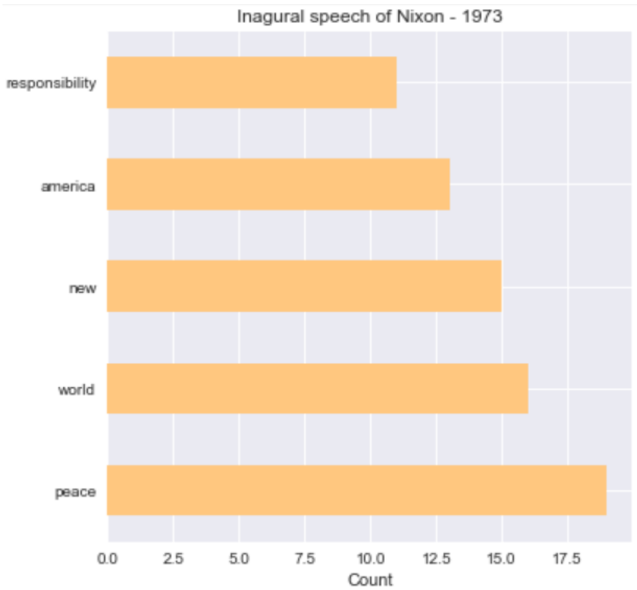
sides	8
world	8
new	7



1973 Inaugral speech:

Top three words are peace, world and new.

peace	19
world	16
new	15



When all the speeches are combined, the top words that keep occurring frequently are nation, sides and peace with 11,8 and 19 occurrences respectively.

- 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]

<u>1941 Inaugural Speech:</u> Nation, people, spirit and life are visualized bigger in comparison to the other words, therefore, these are the words whose frequency is higher in the speech.

<u>1961 Inaugural Speech :</u> Sides, world, nation and power's visualization is bigger than the other words present in the speech.

<u>1973 Inaugural Speech :</u> Peace, America and world are the words in this speech that keep occurring more frequently than the other terms.