Machine Learnig

Project

Business Report

Contents:

Problem:1

1.1 Read the dataset. Do the descriptive statistics and do the null value condition check. Write an inference on it……..3
1.2 Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers…….4
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)……..7
1.4 Apply Logistic Regression and LDA (linear discriminant analysis). ……..8
1.5 Apply KNN Model and Naïve Bayes Model. Interpret the results……8
1.6 Model Tuning, Bagging (Random Forest should be applied for Bagging), and Boosting….8
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.. ……9
1.8 Based on these predictions, what are the insights?.....9

Problem 2

2.1 Find the number of characters, words, and sentences for the mentioned documents…10

2.2 Remove all the stopwords from all three speeches…….10

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) ……10

2.4 Plot the word cloud of each of the speeches of the variable…….11

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.

From the above problem statement we can understand that the data is collected from survey.

Data Dictionary:

1. **Vote :** party choice : conservative or labour
2. **Age :** Age of the voter
3. **economic.cond.national :** Assessment of current national economic conditions1 to 5 with cardinality of 5
4. **economic.cond.household :** Assessment of current household economic conditions 1 to 5 with cardinality of 5
5. **Blair :** assessment of the labour leader 1 to 5 with cardinality of 5
6. **Hague :** Assessment of Conservative leader 1 to 5 with cardinality of 5
7. **Europe :** An 11-point scale that measures respondents attitudes toward European integration.High scores represent Eurosceptic sentiment with cardinality of 11.
8. **political.knowledge :** Knowledge of parties positions on European integration,0 to 3 with cardinality of 4
9. **gender :** Female or Male

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

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

From the dataset we can see that there is an extra column with labeling which is not required .so, droping the variable **Unnamed :0** .After dropping ,the dataset is as follows.

	vote	age	economic.cond.national	economic.cond.household	Blair	Hague	Europe	political.knowledge	gender
0	Labour	43	3	3	4	1	2	2	female
1	Labour	36	4	4	4	4	5	2	male
2	Labour	35	4	4	5	2	3	2	male
3	Labour	24	4	2	2	1	4	0	female
4	Labour	41	2	2	1	1	6	2	male

	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

- There are 5 Unique values in variables economic.cond.national, economic.cond.household, Blair, Hague which represents the assessment levels of voters ranging from 1 to5.
- There are 11 Unique values which represents Assessment levels for Europe ranging from 1 to 11.
- There are 4 Unique values in variable political.knowledge which represents the Knowledge level of voters ranging from 0 to 3.
- The minimum age of voter present is 24yrs and maximum age of the voter in the data is 93 yrs which says that there is a high probability the data collected is genuine.
- There are no null values
- 8 Duplicate rows are present in the data.

1.2 Perform Univariate and Bivariate Analysis. Do exploratory data analysis. Check for Outliers.

Univariate Analysis

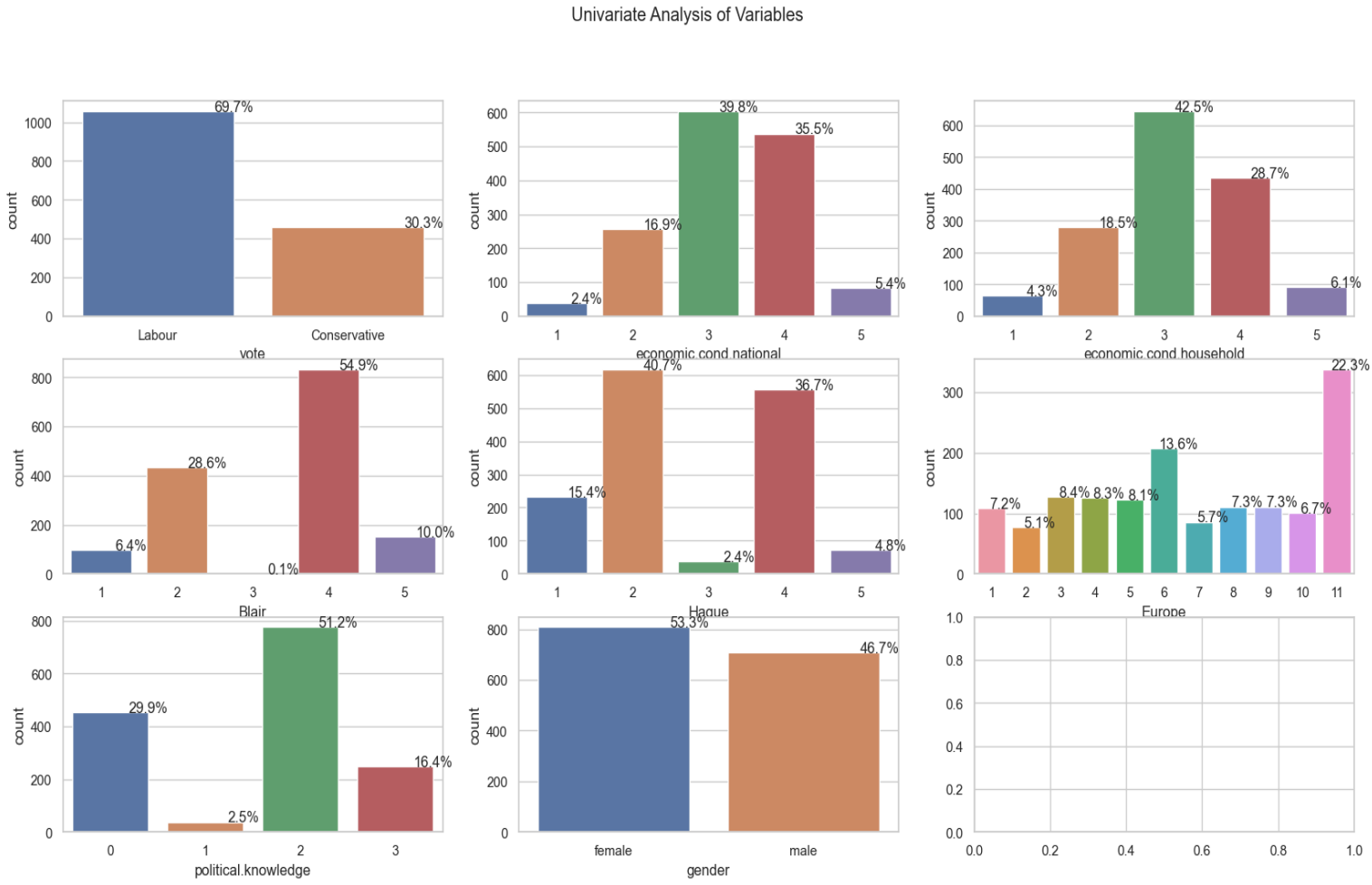


Figure :1

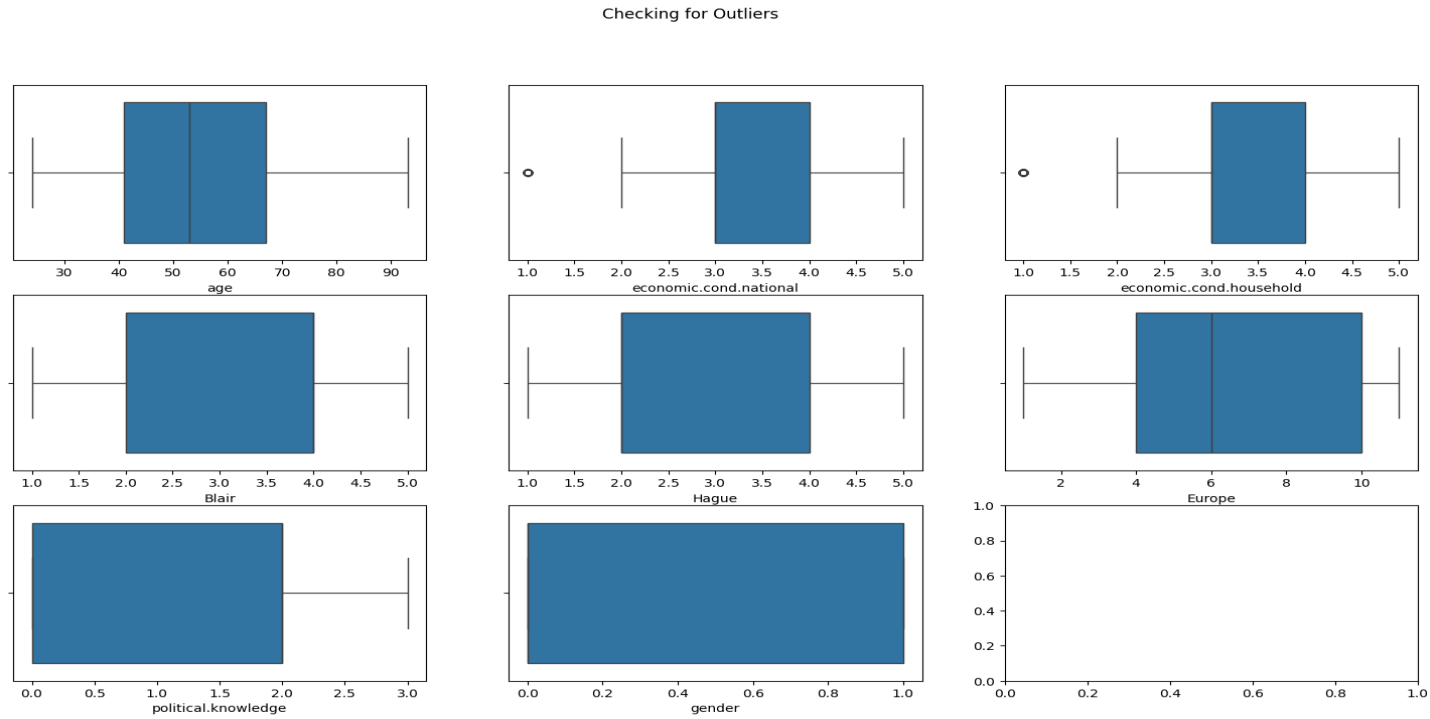


Figure :2

Bivariate Analysis

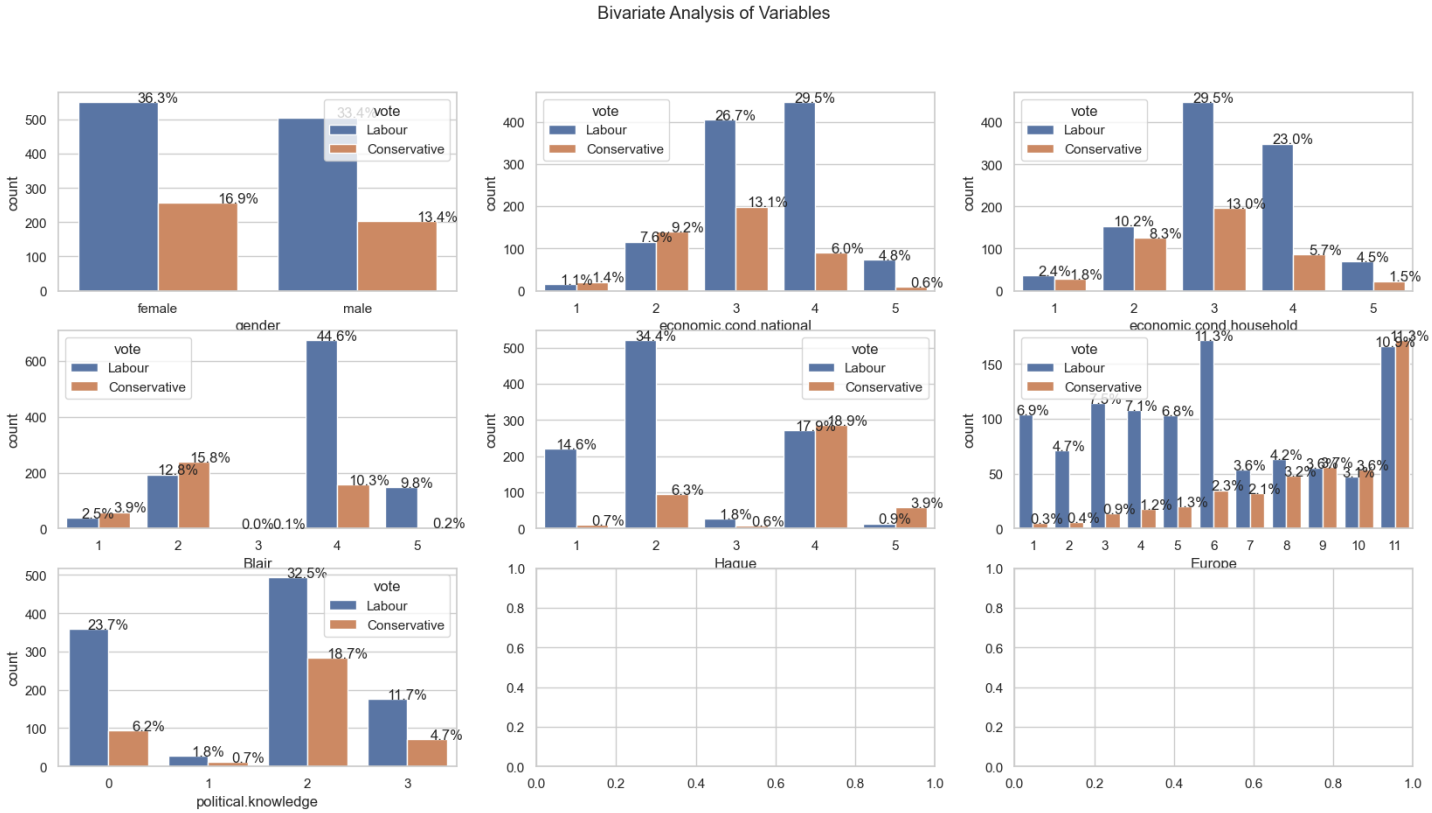


Figure :3

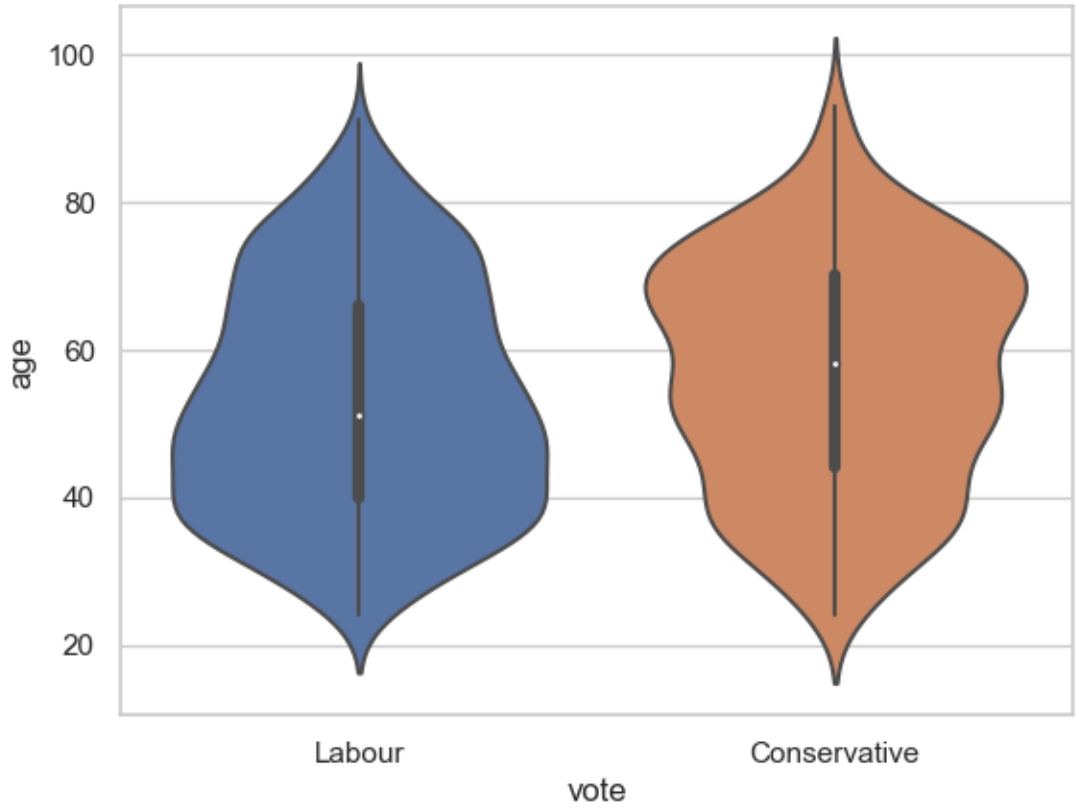


Figure :4

INFERENCES

- **69.7%** of voters voted for **Labour party** and **30.3%** of voters voted for **Conservative Party**.
- **39.8%** and **42.5%** of voter have average Assesment of **current national economic conditions** and **current household economic conditions**.
- **22.3%** (Highest) of voters are **Eurosceptic**
- In Figure: 2 We can see outliers for current national economic conditions and current household economic conditions.But that indicates very few people have Less Knowledge.
- In figure:3 we can see people who are not much Eurosceptic voted for Labour Party and people who are more Eurosceptic voted for Conservative Party.
- In Figure:4 we can see that Maximum people who voted for Labour Party are in the age group of 40-50 and Maximum people who voted for Conservative Party are in the age group of 70.

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)

- Perfomed **label Encoding** for two Categorical variable **vote** and **gender** in order to fit into Modelling.
- Performed MinMax Scaling Before Spliting and Fiting to the data so that variables that are measured at different scales might not create a bias and Preserves the shape of the original distribution and Values contribute equally to the analysis
- Splited the data variables into two.Target variable **vote** in one and all other dependent variables into other and than splitted both the dataset into Train and Test with 70% of data into Train data and 30% of Data into Test Data

1.4 Apply Logistic Regression and LDA (linear discriminant analysis).

- Fitted the train data into **Logistic Regression** model and than predicted with test data .The model score for both data is as follows

Model score for Train Data 0.8350612629594723
Model score for Test Data 0.8245614035087719
Since the difference of Train and Test Accuracies is approx 1.1% it does not comes under Overfitting or underfitting.

- Fitted the train data into **LDA (Linear Discriminant Analysis)** model and than predicted with test data .The model score for both data is as follows

Model score for Train Data 0.8341187558906692
Model score for Test Data 0.8333333333333334
The Accuracies of both Train and Test data are 83%.LDA is doing better compared to Logistic Regression.

1.5 Apply KNN Model and Naïve Bayes Model. Interpret the results.

- Fitted the train data into **KNN Model model** and than predicted with test data .The model score for both data is as follows

Model score for Train Data 0.8557964184731386
Model score for Test Data 0.8245614035087719
Here,its neither overfitting nor underfitting but there is 3% difference between both train and test which is not a bad difference but more compared to Logistic regression model and LDA.

- Fitted the train data into **Naive Bayes model** and than predicted with test data .The model score for both data is as follows

Model score for Train Data 0.8350612629594723
Model score for Test Data 0.8223684210526315
1.3% Accuracy difference is there between Train and Test data which is not a bad one but LDA and Logistic regression are performing better than Naive bayes Model.

1.6 Model Tuning, Bagging (Random Forest should be applied for Bagging), and Boosting.

- Fitted the train data into **Random Forest** and than predicted with test data .The model score for both data is as follows

Model score for Train Data 0.9905749293119699
Model score for Train Data 0.8201754385964912

- Fitted the train data into **AdaBoostClassifier** and than predicted with test data .The model score for both data is as follows

Accuracy on training set : 0.8463713477851084
Accuracy on test set : 0.8135964912280702
Recall on training set : 0.9124668435013262
Recall on test set : 0.8778877887788779
Precision on training set : 0.8764331210191083
Precision on test set : 0.8471337579617835

- Fitted the train data into **GradiantBoostingClassifier** and than predicted with test data .The model score for both data is as follows

Accuracy on training set : 0.8925541941564562
Accuracy on test set : 0.8355263157894737
Recall on training set : 0.9389920424403183
Recall on test set : 0.9108910891089109
Precision on training set : 0.9123711340206185

Precision on test set : 0.8518518518518519

- Fitted the train data into **XGBClassifier** and than predicted with test data .The model score for both data is as following

Accuracy on training set : 0.9915174363807728

Accuracy on test set : 0.831140350877193
Recall on training set : 0.9973474801061007
Recall on test set : 0.8910891089108911
Precision on training set : 0.9907773386034255
Precision on test set : 0.8598726114649682

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.

Based on analysis of all parameters From different models performed LDA is the best model

1.8 Based on these predictions, what are the insights?

Based on the prediction Labour Party has more chances to win and LDA Model is working better for the prediction

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**

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

No of words that are present in each Presidents Speech is as bellow.

	Name	Speech	totalwords	char_count
0	Roosevelt	On each national day of inauguration since 178...	1323	7651
1	Kennedy	Vice President Johnson, Mr. Speaker, Mr. Chief...	1364	7673
2	Nixon Mr.	Vice President, Mr. Speaker, Mr. Chief Jus...	1769	10106

No of sentences in President Roosevelt speech 69
No of sentences in President Kennedy speech 56
No of sentences in President Nixon speech 70

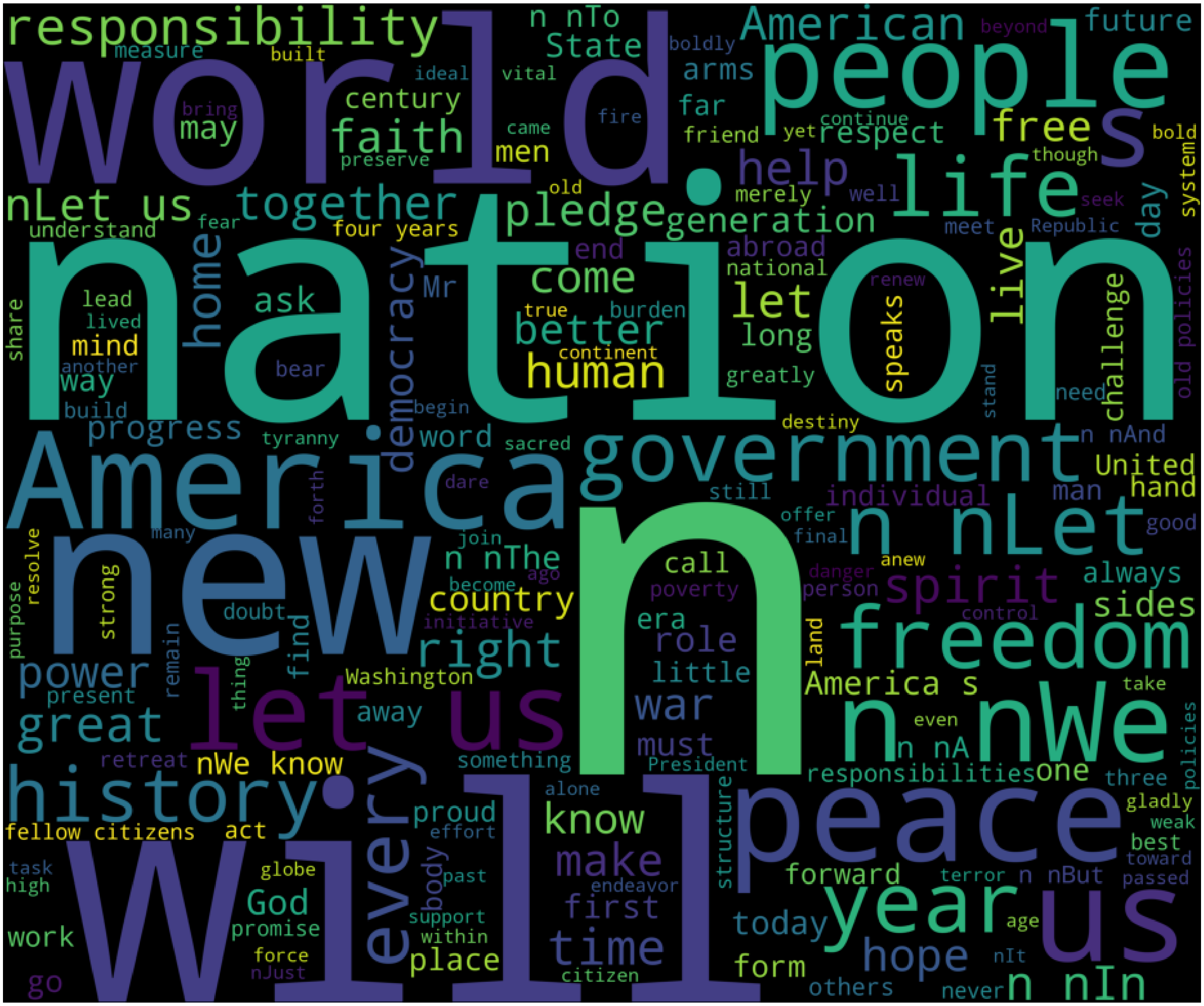
2.2 Remove all the stopwords from all three speeches

Stopwords are removed from all 3 speeches.

No of words Before and After removing stopwords:

	Name	Speech	totalwords	char_count	all clean words
0	Roosevelt	On each national day of inauguration since 178...	1323	7651	632
1	Kennedy	Vice President Johnson, Mr. Speaker, Mr. Chief...	1364	7673	618
2	Nixon	Mr. Vice President, Mr. Speaker, Mr. Chief Jus...	1769	10106	899

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


world
nation
n
will
new
America
people
peace
government
freedom
let us
history
us
year
life
responsibility
American
n nWe
n nLet
faith
together
nLet us
pledge
generation
free
help
every
right
power
great
human
country
spirit
democracy
home
time
hope
make
know
first
come
better
let
live
war
role
little
must
speaks
challenge
God
proud
place
today
form
forward
sides
always
United
hand
man
individual
America s
n nThe
call
era
find
away
nWe know
fellow citizens
act
work
progress
word
century
may
men
State
arms
far
friend
respect
future
beyond
boldly
measure
built
ideal
vital
bring
fire
came
continue
yet
bold
though
system
preserve
old
merely
well
seek
day
understand
fear
four years
end
abroad
national
meet
Republic
old policies
renew
lead
lived
mind
true
burden
continent
another
way
bear
greatly
share
build
begin
stand
need
destiny
tyranny
sacred
n nAnd
dare
still
forth
offer
good
many
join
final
resolve
purpose
strong
ago
danger
anew
become
poverty
person
control
doubt
initiative
remain
thing
Washington
present
retreat
something
President
alone
structure
responsibilities
one
three
gladly
weak
effort
body
best
toward
task
high
globe
past
endeavor
terror
n nBut
passed
support
within
promise
force
nJust
go
citizen
age
nIt
others
never
n nIn
take
even
n nA
policies
n nTo
land