**FINAL PROJECT**

**COUNTY-LEVEL PRESIDENTIAL ELECTION 2008 – 2016**

Logo

Description automatically generated

By:

**AKASH RAJ,**

**HARSHIT GAUR,**

**MIRAV PAREKH**

MASTER OF PROFESSIONAL STUDIES IN ANALYTICS

ALY 6015: INTERMEDIATE ANALYTICS

MARCH 19, 2022

To: **PROF. ROY WADA**

**TABLE OF CONTENT**

[Introduction 1](#_Toc97460363)

[Data Analysis 2](#_Toc97460364)

[Descriptive Statistics 3](#_Toc97460365)

[Exploratory Data Analysis 15](#_Toc97460366)

[Conclusion 18](#_Toc97460367)

[Bibliography 19](#_Toc97460368)

[Appendix 20](#_Toc97460369)

# INTRODUCTION

Every four years, the US general election is held to choose the president and the vice president of the United States. The two major political parties competing in the general elections are the Democrats, and the Republicans. On the election day, the US citizen vote for their presidential choice in their respective electoral college in each county. The electors with majority in turn cast direct votes for the president.

Before the general elections are held, the presidential nominee from each party campaign across the country to gain support of the people. The campaigns are massive in the swing states of Florida, Ohio, and Pennsylvania, also known as the Battleground states. These campaigns cost millions of dollars, and there are designated campaign teams to plan and raise funds.

The campaign teams from each political party leverage data analytics to understand and predict people’s stance on the political issues and candidates. The models and the insights help the political parties optimize their campaigns and target their outreach efforts.

**Project Proposal**

In the actual US presidential elections, data from different sources like national databases, consumer preferences, social media, etc. are used to gauge the sentiment of the voters, target floating voters, and define strategies for advertisement.

However, the scope of this project is to build classification model using historical election and demographic data to predict the winning political party in each county and understand the counties and swing states where the political parties should target their election campaign. The main questions that will be answered in this project are:

* ***Who will win the 2016 and 2020 general election?***
  + ***Who will win the election in each county?***
* ***Which swing states to focus on for the 2016 election campaign?***

**Models And Variables**

Since, predicting the winning political party in each county is a classification problem, a Logistics Regression model will be built. The following are the response and predictor variables for the logistic regression model:

* Outcome variable – Winner (Total Democratic Votes – Total Republican Votes)
* Predictor variables – Total votes of 2008 and 2012, Total republican votes of 2008 and 2012, Total democrat votes of 2008 and 2012, Unemployment rates 2011 – 2015, Gender distribution 2011 – 2015, Demographic distribution 2011 – 2015

The second question to obtain the swing states, a Linear Regression model will be built. The following are the response and predictor variables for the linear regression model:

* Outcome variable – Difference in votes for republican and democrats
* Predictor variables – Total votes of 2008 and 2012, Total republican votes of 2008 and 2012, Total democrat votes of 2008 and 2012, Unemployment rates 2011 – 2015, Gender distribution 2011 – 2015, Demographic distribution 2011 – 2015

Principal Component Analysis (PCA) will be applied to reduce the dimensionality of the data set.

**Understanding The Dataset:**

To perform the analysis and build the model, we will be using the county-level voter data from 2008, 2012, and 2016. The dataset also includes county-level socio economic factors and metrics like labor force participation, median household income, educational attainment, poverty, international and domestic migrations, population, race, gender, age, per capita income, and occupations. The data set has been collected from multiple sources:

* The employment and labor force data has been collected from
  + <https://www.bls.gov/lau/>
* Year-wise census information has been collected from
  + <https://data2.nhgis.org/main>
* County-level socioeconomic indicators have been collected from
  + <https://www.ers.usda.gov/data-products/county-level-data-sets/>

The data set contains ***3,143 observations*** and ***148 columns***, out of which 2 are categorical and 146 are quantitative.

Table

Description automatically generated

**Fig 1 – Data Description**

# DATA PREPROCESSING

## Missing Data Imputation

Before the analysis is performed, it’s important to identify and impute the missing values, as the missing values may bias the results. Therefore, the records with missing values were identified using the function ***complete.cases***, which outputs the observations with missing values in variable.

Since the number of records in the data are less, these records were not deleted from the data. These missing values were then imputed using the mean of variable.

missingRecords <- ElectionData %>% filter(!complete.cases(ElectionData))

# Retrieving the names of features with missing values.

missingValuesCols <- names(which(colSums(is.na(ElectionData)) > 0))

# Imputing missing values with their respective features' mean value

for(i in 1:ncol(ElectionData)) {

ifelse(is.numeric(ElectionData[,i]), ElectionData[is.na(ElectionData[,i]), i] <- mean(ElectionData[,i], na.rm = TRUE),

ifelse(is.character(ElectionData[,i]), "NULL", 0))}

**Table 2: Records with missing data**

| county | v2016 | vd2016 | vg2016 | pd2016 | pg2016 | diff2016 |
| --- | --- | --- | --- | --- | --- | --- |
| Aleutians East Borough |  |  |  |  |  |  |
| Aleutians West Census Area |  |  |  |  |  |  |
| Anchorage Municipality |  |  |  |  |  |  |
| Honolulu County | 285,683 | 175,634 | 90,296 | 0.61 | 0.32 | 85,338 |
| Oglala County | 2,896 | 2,504 | 241 | 0.86 | 0.08 | 2,263 |

**Observation:**

These missing values are due to unavailability of data for some counties or because elections were not conducted in these counties for some reason.

## Merging Additional Data

Although, the data set contains a exhaustive list of variables with all the variables related to the election results and the demographic attributes of the county, there are some key attributes of the county missing in the data, like population, etc. Furthermore, to extend the model and predict the results for 2020, we need the election results for 2020. Therefore, to improve the accuracy of the models and ensure that there are enough variables to explain the variance in the dependent variables, the following external data sources were augmented to the data.

* Population by year in each county
* County level election results in 2020

These data sets were sourced from the government portal, <www.census.gov>, and appended to the existing data set. The merge operation was done using the ***inner\_join*** function, from the ***dplyr*** package, on the county name and county code (County FIPS).

# Combining the external data sets with the original election data sets

ElectionData <- ElectionData %>% inner\_join(population, by = c("county", "state"))

election2020$county\_fips <- as.factor(election2020$county\_fips)

ElectionData$c\_fips <- as.factor(ElectionData$c\_fips)

ElectionData <- ElectionData %>% inner\_join(election2020, by = c("c\_fips" = "county\_fips"))

**Table 3: New variables added to the master data**

| Population 2017 | Population 2018 | Population 2019 | democrats | green | other | republican |
| --- | --- | --- | --- | --- | --- | --- |
| 55,390 | 55,533 | 55,869 | 35,595 | 160 | 1,869 | 99,981 |
| 212,521 | 217,855 | 223,234 | 110,442 | 1,033 | 8,413 | 377,557 |
| 25,157 | 24,872 | 24,686 | 31,316 | 46 | 449 | 33,487 |
| 22,550 | 22,367 | 22,394 | 13,160 | 52 | 565 | 36,402 |
| 57,787 | 57,771 | 57,826 | 20,203 | 154 | 1,800 | 118,769 |

**Observation:**

A total of 17 new variable were added which included the population of each county by year, and the total vote count for each party at each county in the 2020 election.

## Feature Engineering – Creating new variables

Apart from augmenting external data sets to enrich the data, some additional variables were created from the existing variables as a part of feature engineering. Two new variable, census region and geographics region were created from the state column. The region variables are categorical variables. The variable census region has values like ‘East South Central’, ‘Pacific’, ‘Mid Atlantic’, ‘South Atlantic’, etc. The variable geographic region has values like ‘South’, ‘West’, ‘Northeast’, ‘Midwest’, etc.

The following table indicates the sample of distinct values in the region variables created from state.

**Table 4: Sample of additional region variable created**

| state | State code | Census region | Geographic region |
| --- | --- | --- | --- |
| Alabama | AL | East South Central | South |
| Alaska | AK | Pacific | West |
| Arizona | AZ | Mountain | West |
| Arkansas | AR | West South Central | South |
| California | CA | Pacific | West |
|  | | | |

The next table is a snapshot of the sample data with external data added and new variables created.

**Table 5: New variables added to the master data**

| Population 2017 | Population 2018 | Population 2019 | Democrats 2020 | Republican 2020 | Census region | Geographic region |
| --- | --- | --- | --- | --- | --- | --- |
| 55,390 | 55,533 | 55,869 | 35,595 | 99,981 | East South Central | South |
| 212,521 | 217,855 | 223,234 | 110,442 | 377,557 | East South Central | South |
| 25,157 | 24,872 | 24,686 | 31,316 | 33,487 | East South Central | South |
| 22,550 | 22,367 | 22,394 | 13,160 | 36,402 | East South Central | South |
| 57,787 | 57,771 | 57,826 | 20,203 | 118,769 | East South Central | South |

# EXPLORATORY DATA ANALYSIS

## Descriptive Statistics

**Overall descriptive Statistics of all the variables**

The outcome variable of the regression model is percentage democratic and republican votes in a county. Before conducting the data analysis, in this report we have performed required steps to find out the descriptive statistics of the raw data to check the central tendency and dispersion of the numerical variables. The descriptive statistics of the overall data gives a sense of each of the variables, their central tendencies, and their dispersion. The information from the descriptive statistics table can be used to plan further exploration and data analysis. The following table gives the overall descriptive statistics of the data.

overallStats <- describeTable %>% describe(quant = c(.25, .75), IQR = TRUE)

overallStats <- round(overallStats, 2)

overallStats$vars <- rownames(overallStats)

**Table 6: Descriptive statistics of overall data**

| variable | n | mean | SD | median | min | max | skew | Q0.25 | Q0.75 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| pd2016 | 3,112 | 0.32 | 0.15 | 0.28 | 0.03 | 0.89 | 0.94 | 0.20 | 0.40 |
| pg2016 | 3,112 | 0.64 | 0.16 | 0.67 | 0.08 | 0.95 | -0.83 | 0.55 | 0.75 |
| ppd2016 | 3,112 | -0.32 | 0.31 | -0.38 | -0.92 | 0.81 | 0.89 | -0.55 | -0.15 |
| pd2012 | 3,112 | 0.38 | 0.15 | 0.37 | 0.03 | 0.93 | 0.49 | 0.28 | 0.48 |
| pg2012 | 3,112 | 0.60 | 0.15 | 0.61 | 0.06 | 0.96 | -0.46 | 0.51 | 0.70 |
| ppd2012 | 3,112 | -0.21 | 0.29 | -0.24 | -0.92 | 0.87 | 0.48 | -0.43 | -0.03 |
| pd2008 | 3,112 | 0.42 | 0.14 | 0.41 | 0.05 | 2.58 | 1.41 | 0.31 | 0.50 |
| pg2008 | 3,112 | 0.57 | 0.16 | 0.57 | 0.09 | 5.37 | 7.97 | 0.48 | 0.67 |
| Population 2016 | 3,112 | 103,405 | 331,501 | 25,924 | 117.00 | 10,105,708 | 13.76 | 11,173 | 68,375 |

**Observation:**

The percent democratic and percent republican votes (pd2016, pg2016, pd2012, pg2012, pd2008, pg2008) have quite similar mean and median indicating consistent data. The percentage variables are not skewed for 2016 and 2012, whereas they are skewed for 2008. The population variable also seems to be heavily skewed. A log transformation of the skewed variables may be required.

## Subgroup Analysis

**Descriptive Statistics of %Democratic votes by year (2008, 2012, 2016)**

Apart from looking at the overall statistics of the data, the subgroup analysis helps in understanding distribution of data across different subgroups. We’ve performed subgroup analysis on the variable ‘% Democratic votes’ grouped by year. The following table gives the descriptive statistics of the variable ‘% Democratic votes.’

percDemocraticVotesStats <- ElectionDataEDA %>% dplyr::select(pd2008, pd2012, pd2016) %>% describe(quant = c(.25, .75), IQR = TRUE) %>% mutate(year = c(2008, 2012, 2016)) %>% relocate(year)

**Table 7: Descriptive statistics of Democratic party %votes by year**

| year | n | mean | SD | median | min | max | skew | Q0.25 | Q0.75 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 2,008 | 3,112 | 0.41 | 0.14 | 0.41 | 0.05 | 2.58 | 1.40 | 0.31 | 0.50 |
| 2,012 | 3,112 | 0.38 | 0.14 | 0.37 | 0.03 | 0.93 | 0.49 | 0.28 | 0.48 |
| 2,016 | 3,112 | 0.31 | 0.15 | 0.28 | 0.03 | 0.89 | 0.93 | 0.20 | 0.40 |

**Observation:**

The % Democrat for 2008 seems to be heavily skewed whereas the variable for 2012 and 2016 does not seem to be skewed. The mean and median are similar for all the three years indicating that there are no outliers in the data.

**Descriptive Statistics of %Republican votes by year (2008, 2012, 2016)**

The next subgroup analysis is complimentary variable to the %Democratic votes, i.e., %Republican votes. The following code and table give the descriptive statistics for %Republican votes in each county by year.

percRepublicanVotesStats <- ElectionDataEDA %>% dplyr::select(pg2008, pg2012, pg2016) %>% describe(quant = c(.25, .75), IQR = TRUE) %>% mutate(year = c(2008, 2012, 2016)) %>% relocate(year)

**Table 7: Descriptive statistics of Republican party %votes by year**

| year | n | mean | SD | median | min | max | skew | Q0.25 | Q0.75 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 2,008 | 3,112 | 0.56 | 0.16 | 0.57 | 0.09 | 5.37 | 7.97 | 0.48 | 0.67 |
| 2,012 | 3,112 | 0.59 | 0.14 | 0.61 | 0.06 | 0.96 | -0.46 | 0.51 | 0.70 |
| 2,016 | 3,112 | 0.63 | 0.15 | 0.67 | 0.08 | 0.95 | -0.82 | 0.55 | 0.75 |

**Observation:**

The % Republican for 2008 seems to be heavily skewed whereas the variable for 2012 and 2016 does not seem to be skewed. The mean and median are similar for all the three years indicating that there are no outliers in the data.

**Descriptive Statistics of ‘% Republican Votes’ by Geographic Region**

Another set of subgroup analysis was performed for the % Republican votes by region.

**Table 7: Descriptive statistics of Republican party %votes by region**

| group | n | mean | SD | median | min | max | skew | Q0.25 | Q0.75 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| South | 1,420 | 0.65 | 0.16 | 0.69 | 0.08 | 0.95 | -0.98 | 0.57 | 0.77 |
| West | 421 | 0.57 | 0.17 | 0.60 | 0.09 | 0.91 | -0.45 | 0.45 | 0.72 |
| Northeast | 217 | 0.51 | 0.14 | 0.51 | 0.10 | 0.84 | -0.29 | 0.43 | 0.63 |
| Midwest | 1,054 | 0.66 | 0.12 | 0.67 | 0.08 | 0.93 | -0.77 | 0.59 | 0.75 |

**Observation:**

There are lesser number of counties in the West and Northeast, and there is no indication of outliers and skewness across regions.

**Descriptive Statistics of ‘Total Unemployment Rate’ in 2011, 2012, 2013, 2014, 2015**

One of the important variables that determines the % of votes for a political party is the unemployment rate in the county or the region. The next subgroup analysis is the

unemploymentRate <- ElectionData %>%   
 select(unemp11, unemp12, unemp13, unemp14, unemp15) %>%   
 describe(quant = c(.25, .75), IQR = TRUE) %>%   
 mutate(year = c(2011, 2012, 2013, 2014, 2015)) %>%   
 relocate(year)  
# Kable Classic Method  
round(unemploymentRate, 2) %>%   
 kbl(caption = "Table 4: Descriptive Statistics for Total Unemployment Rate") %>%   
 kable\_classic(html\_font = "Cambria")

Table 4: Descriptive Statistics for Total Unemployment Rate

Table

Description automatically generated

**Observation:**

The mean of the unemployment rate of the population in the US has decreased significantly from the year 2011 to 2015 from the value of 4,420.26 to 2,644.24. However, the year 2008 holds the maximum of the total votes registered by the republicans in the election poll. The range is big in the unemployment rate distribution belonging to all the 3 election years, but the standard deviation is correspondingly small which means the distribution is not dispersed to a significant level.

## Visual Analysis (Group & Sub-Group Analysis)

**Plot of outcome variable ‘% Democratic votes’ by region and year**

The outcome variable for the analysis is percent democratic votes for each county. We will plot the variable by region and year to understand the difference in votes across various regions and years.

ggplot(data = percDVotesL, mapping = aes(x = reorder(factor(geograhic.region), pdVotes, function(x) -1\*sum(x)), y = pdVotes, fill = year)) + geom\_bar(position = "dodge", stat = "identity") + labs(title = "Percentage Democrat Votes by Region & Year") + scale\_x\_discrete(name ="Region") + scale\_y\_continuous(name = "Percent Democrat Votes") + theme\_bw()

Chart, bar chart

Description automatically generated

**Observation:**

Since the republican party won in 2016 election, it’s evident from the plot that the democrats received the least percentage of votes in 2016. They received the least votes in the Midwest and South region. This could be because of two reason – either they do not have support in the southern and midwestern states, or the number of voters in these states are less.

**Plot of outcome variable ‘% Republican votes’ by region and year**

Another outcome variable for the analysis is percent republican votes for each county. The percent republican votes is complimentary to the percent democrat votes. We will plot the variable by region and year to understand the difference in votes across various regions and years.

ggplot(data = percRVotesL, mapping = aes(x = reorder(factor(geograhic.region), pgVotes, function(x) -1\*sum(x)), y = pgVotes, fill = year)) + geom\_bar(position = "dodge", stat = "identity") + labs(title = "Percent Republican Votes by Region & Year") + scale\_x\_discrete(name ="Region") + scale\_y\_continuous(name = "Percent Republican Votes") + theme\_bw()

Chart, bar chart

Description automatically generated

**Observation:**

Since the republican party won in 2016 election, it’s evident from the plot that the democrats received the least percentage of votes in 2016. They receive the least votes in the Midwest and South region. This could be because of two reasons – either they do not have support in the southern and midwestern states, or the number of voters in these states are less.

Since both republican and democrats received less votes in Southern and Midwestern region, it appears that the number of voters are less in these regions.

**Plot of variable ‘Total Votes’ by year**

totalVotesL <- ElectionData %>%   
 select(state, v2008, v2012, v2016) %>% group\_by(state) %>% summarise('2008' = sum(v2008, na.rm = TRUE), '2012' = sum(v2012, na.rm = TRUE), '2016' = sum(v2016, na.rm = TRUE)) %>% gather(year, tVotes, c('2008', '2012', '2016'))  
  
ggplot(data = totalVotesL, mapping = aes(x = reorder(factor(state), tVotes, function(x) -1\*sum(x)), y = tVotes, fill = year)) +  
 geom\_bar(position = "dodge", stat = "identity") + (title = "Total Votes by State & Year") +   
 scale\_x\_discrete(name ="States") + scale\_y\_continuous(name = "Total Votes", labels = label\_number(suffix = " M", scale = 1e-6)) + theme\_bw()

Chart, scatter chart

Description automatically generated

Chart, histogram

Description automatically generated

**Observation:**

The total number of votes registered by the states in all the 3 election years are almost same, but the states of California, Nebraska, North Dakota, Maine, Mississippi, Wyoming, Idaho registered very large number of total votes in 2008 than in other election years. The states of Nebraska, North Dakota, and Maine suffered a drastic reduction in the total votes registered from the year 2008 to 2012, and 2016. From the boxplot, we can figure out that there are some outliers present in the total votes’ distribution for all the 3 election years.

**Plot of variable ‘Total Democratic Votes’ by state and year**

totalDVotesL <- ElectionData %>%   
 select(state, vd2008, vd2012, vd2016) %>% group\_by(state) %>% summarise('2008' = sum(vd2008, na.rm = TRUE), '2012' = sum(vd2012, na.rm = TRUE), '2016' = sum(vd2016, na.rm = TRUE)) %>%  
 gather(year, tdVotes, c('2008', '2012', '2016'))  
  
ggplot(data = totalDVotesL, mapping = aes(x = reorder(factor(state), tdVotes, function(x) -1\*sum(x)), y = tdVotes, fill = year)) + geom\_bar(position = "dodge", stat = "identity") +  
 labs(title = "Total Democrat Votes by State & Year") + scale\_x\_discrete(name ="States") +   
 scale\_y\_continuous(name = "Total Democrat Votes", labels = label\_number(suffix = " M", scale = 1e-6)) + theme\_bw()

Chart, scatter chart

Description automatically generated

**Observation:**

From the boxplot, we can figure out that there are some outliers present in the total votes’ distribution for the democrat party for all the 3 election years. We would handle these outliers further in our analysis. Below is the bar plot for further insights.

Chart

Description automatically generated

**Observation:**

The total number of votes registered for the democrat party by the states in all the 3 election years are almost same, but the states of Nebraska, North Dakota, Maine, Mississippi, Wyoming, Idaho registered very large number of total votes in 2008 than in other election years. The states of Nebraska, North Dakota, and Maine suffered a drastic reduction in the total votes registered for the democrat party from the year 2008 to 2012, and 2016. The New York state experienced a significant increase in the outcome of total number of voters in support for democrats from 2008 to 2012, and 2016. Most of the states have experienced a reduction in total number of votes for the democrat party from 2008 to 2016.

**Plot of outcome variable ‘Total Republican Votes’ by state and year**

totalRVotesL <- ElectionData %>%   
 select(state, vg2008, vg2012, vg2016) %>% group\_by(state) %>% summarise('2008' = sum(vg2008, na.rm = TRUE), '2012' = sum(vg2012, na.rm = TRUE), '2016' = sum(vg2016, na.rm = TRUE)) %>%  
 gather(year, tgVotes, c('2008', '2012', '2016'))  
  
ggplot(data = totalRVotesL, mapping = aes(x = reorder(factor(state), tgVotes, function(x) -1\*sum(x)), y = tgVotes, fill = year)) + geom\_bar(position = "dodge", stat = "identity") +  
 labs(title = "Total Republican Votes by State & Year") + scale\_x\_discrete(name ="States") +   
 scale\_y\_continuous(name = "Total Republican Votes", labels = label\_number(suffix = " M", scale = 1e-6)) + theme\_bw()

Chart, scatter chart

Description automatically generated

Chart, histogram

Description automatically generated

**Observation:**

The total number of votes registered for the republican party by the states in all the 3 election years are almost same, but the states of Nebraska, North Dakota, Maine, registered very large number of total votes in 2008 than in other election years. These states of Nebraska, North Dakota, and Maine suffered a drastic reduction in the total votes registered from the year 2008 to 2012, and 2016. The New York, and North Carolina states experienced a significant increase in the outcome of total number of voters in support for republican party from 2008 to 2012, and 2016. Most of the states have experienced an increase in total number of votes for the republican party from 2008 to 2016 as opposed to reduction in votes for democrat party.

# ANALYTICAL METHODS

## Correlation of Numeric Variables

A correlation graph has been generated for the numeric & integer features of the dataset. It will help in figuring out the most correlated features and the least correlated features with our dependent feature - *pd2016* (Percentage of democratic votes in 2016).

*Chart, bubble chart

Description automatically generated*

*Figure 7: Correlation Plot of numerical features of the dataset.*

* The percentage of votes registered for the democrat party in 2016 is very highly correlated to the percentage of democrat votes registered in 2012 and a little less correlated to the percentage of votes registered for democratic party in 2008.
* As thought so, the population of the country would add significance to the correlation with the percentage of democrat votes in 2016.
* The correlation of the percentages of democrat and republican votes in 2008 have diminished a little.

#### **Splitting the Data into Train and Test Sets -**

We will create a partition of the data set randomly in the ratio of 80:20 where 80% of the random data will go into the training set and the rest 20% of the data will move into the testing set.

Dependent variable - ***pd2016***

The training and testing data sets have x and y variables which correspond to the predictor variables and predicting variable (responding variable).

#### **LASSO REGULARIZATION**

#### **Estimating the best value of Lambda using Cross Validation -**

The best value of lambda is calculated/estimated using the cv.glmnet() function and the training set (x and y variables). The plot of the mean squared error and Logarithmic of lambda is below with its observations.

Graphical user interface

Description automatically generated

*Figure 8: Plot to estimate the optimal Lambda value for Lasso Regularization*

#### **Observations:**

From the above plot of mean-squared error in relation with logarithm of lambda, we can figure out that

1. The minimum logarithmic value of Lambda: **-10.44232**
2. The logarithmic value of Lambda at one standard error: **-10.16322**
3. The mean squared error of the data set is very low for the dataset because we're considering the percentage of democratic votes in 2016. The difference between the percentages would be very low and squaring these terms would yield even a lower value. This is the reason behind the mean square error value being very low.
4. The mean-squared error at minimum and 1 standard error are almost the same because of the above reason.
5. For the mean-squared error at its minimum value, the number of features which can be considered is around 39. For the mean-squared error at 1 standard error, the number of features which can be considered is around 36.

Measure: Mean-Squared Error

Lambda Index Measure SE Nonzero

min 2.917e-05 93 2.657e-05 9.421e-07 39

1se 3.856e-05 90 2.727e-05 1.077e-06 36

1. The minimum value of Lambda: **0.00002917152**
2. The value of Lambda at one standard error: **0.00003856303**

**Table 9: Lasso Regression on Training Data set using Lambda at 1 Standard Error**

| **Features** | **Coefficients** |
| --- | --- |
| (Intercept) | 0.0441 |
| countya | 0.0000 |
| statea | 0.0000 |
| pg2016 | 0.7676 |
| diff2016 | 0.0000 |
| ppd2016 | 0.8691 |
| pd2012 | 0.0847 |
| pg2012 | 0.0575 |
| diff2012 | 0.0000 |
| ppd2012 | 0.0000 |

#### **Observations:**

Firstly, since the total number of variables are 159, we haven't taken all of them out in the above table. We have added the first 12 features in the table.

From the table of Lasso Regularization on the training data set, we can figure out that the features with coefficient values as 0 have been penalized by the lasso regularization model. This means that these features are either very less significant or almost insignificant to the model and the data set.

Below are some of the variables which have been penalized by the regularization model and can be considered very less significant or almost insignificant to the model.

1. countya 0.0000000
2. statea 0.0000000
3. diff2016 0.0000000
4. diff2012 0.0000000
5. ppd2012 0.0000000

The variables with high value of coefficients have not been penalized heavily by the lasso regularization model. This means they are significant to the model and the data set and should be included in the predicting the *pd2016* feature.

One example of the significant features is ***ppd2016***: 0.8691

#### **Making Prediction on the Training Data Set and Determining Performance -**

The fit model was used to predict the *pd2016* feature of the training data set. The performance of the model at predicting the training data feature can be interpreted using the value of root mean square error.

The RMSE value of the regularization model is **0.005086561**

The interpretation of this root mean square error value is that it is good enough for our initial model for this prediction. The residuals of the data points in the training data are not spread very much and the data is concentrated to a good degree around the line of best fit as well.

#### **Making Prediction on the Testing Data Set and Determining Performance -**

The fit model was used to predict the *pd2016* feature of the testing data set. The performance of the model at predicting the testing data feature can be interpreted using the value of root mean square error.

The RMSE value of the regularization model is **0.005404997**

The RMSE value of this model on the testing data is also good. The value is almost close and not very significantly deviating from the RMSE value of the training data set regression. The residuals of the data points in the testing data are not spread very much and the data is concentrated to a good degree around the line of best fit as well.

**Is the Model Overfitting?**

Since the values of Root Mean Squared Error (RMSE) for both the training and testing data set are almost same, we can say that the Lasso Regularization Model is not overfitting on the data set.

#### **LINEAR REGRESSION**

We will apply linear regression on the data set with *pd2016* feature as predictor variable using all the features available in it. The summary of the model is below

## Table 10: Linear Regression Model on Training Data set using features from Stepwise Selection

| Features | Estimate | Std. Error | t value | Pr(>|t|) |
| --- | --- | --- | --- | --- |
| (Intercept) | -0.1830 | 0.1574 | -1.1625 | 0.2451 |
| countya | 0.0000 | 0.0000 | 0.5674 | 0.5705 |
| statea | 0.0000 | 0.0000 | -0.4326 | 0.6653 |
| vd2016 | 0.0000 | 0.0000 | 3.2668 | 0.0011 |
| vg2016 | 0.0000 | 0.0000 | 2.3228 | 0.0203 |
| pg2016 | 0.8665 | 0.0095 | 91.5749 | 0.0000 |
| diff2016 | 0.0000 | 0.0000 | -0.3780 | 0.7055 |
| ppd2016 | 0.9245 | 0.0053 | 173.6980 | 0.0000 |
| vd2012 | 0.0000 | 0.0000 | -2.1236 | 0.0338 |
| vg2012 | 0.0000 | 0.0000 | -1.6404 | 0.1010 |
| pd2012 | 0.0735 | 0.0208 | 3.5253 | 0.0004 |
| pg2012 | 0.0667 | 0.0207 | 3.2293 | 0.0013 |
| ppd2012 | 0.0044 | 0.0174 | 0.2534 | 0.8000 |
| vg2008 | 0.0000 | 0.0000 | -0.0581 | 0.9537 |

**Adjusted R-Squared value: 0.99**

**Root Mean Square Error value: 0.004**

**Observations:**

The Adjusted R-Squared value of Linear Regression model is significantly very high and upon it, our initial impressions are that the model has been fitted properly. The root mean square error (RMSE) value is also significantly very low which shows that the data points are concentrated in a single region and are very close to each other. This shows that the model has been fitted properly in our initial impressions.

# CONCLUSION

In this assignment, we performed the initial analysis of the data to uncover the hidden patterns in the data and understand the distribution of the data across different groups. We started with the preprocessing of the data to identify the missing values and impute them using the mean of the variables. Another important step of the data preprocessing process was merging additional data and feature engineering to create additional variables. It is important to add new variables to explain the variance of the dependent variable, which might not be explained by the existing variables.

The next step to was to explore and analyze the outcome and predictor variables using plot to understand the overall distribution and the distribution across various subgroups. The subgroup analysis was performed on the variable year and region. The percentage of

In the next weeks, we will perform further analysis on the data set to be able to come to the point where we can answer the questions, we are trying to find answers of. We will build classification model using the demographic data and historical election to predict the winning political party in each county. We will also try to understand the counties and the swing states to help the political parties about where their election campaign should focus more on. Upon performing the analysis, we will be able to answer who will win the 2016 general election and even predict the 2020 general election's winner.

The variables which we have figured out as the outcome variables for our models is 'Winner' which we would compute using the difference of Total Democratic Votes and Total Republican Votes. The predictor variables which we have investigated to use for our models and analysis belong to the Total votes of 2008 and 2012, Total republican votes of 2008 and 2012, Total democrat votes of 2008 and 2012, Unemployment rates from the year 2011 till 2015, Gender distribution from 2011 to 2015, and the Demographic distribution from the year 2011 till 2015.

# BIBLIOGRAPHY

*Downloadable Data – DASIL*. (2017). County-Level Presidential Election Data 2008 – 2016. <https://dasil.sites.grinnell.edu/downloadable-data/>

*Bluman, A.* (2017). Elementary Statistics: A Step By Step Approach (10th ed.). McGraw-Hill Education.

*Kabacoff, R., I.* (2022). R in Action, Third Edition. Manning.

*F.* (2021, April 2). tidyverse in r – Complete Tutorial. R-Bloggers. Retrieved October 30, 2021, from <https://www.r-bloggers.com/2021/04/tidyverse-in-r-complete-tutorial/>

# APPENDIX

