**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 21, 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 5: 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:

## 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 2: 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 |

Observations:

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,

# DATA ANALYSIS

## Descriptive Statistics

The outcome variable of the regression model is Total Votes, Total Democrat Votes, and Total Republican Votes. Before conducting the exploratory 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.

**Descriptive Statistics of ‘Total Votes’ in 2008, 2012, 2016**

totalVotesStats <- ElectionData %>%   
 select(v2008, v2012, v2016)  
  
# Kable Classic Method  
totalVotesStats <- totalVotesStats %>%   
 describe(quant = c(.25, .75), IQR = TRUE) %>%   
 mutate(year = c(2008, 2012, 2016)) %>%   
 relocate(year)  
round(totalVotesStats, 2) %>%   
 kbl(caption = "Table 1: Descriptive Statistics for Total Votes") %>%   
 kable\_classic(html\_font = "Cambria")

Table 1: Descriptive Statistics for Total Votes

Table

Description automatically generated

**Observation:**

* The mean of the total votes registered for the election decreased in the year 2012 when compared to 2008 but increased in the year 2016 when compared to 2012. However, the highest mean of total votes registered is in the year 2008 only.
* The year 2008 also holds the maximum of the total votes registered in the election poll.
* The means of total votes registered in all the 3 years election polls deviate significantly in the north region from their respective medians.

**Descriptive Statistics of ‘Total Democratic Votes’ in 2008, 2012, 2016**

totalDemocraticVotesStats <- ElectionData %>%   
 select(vd2008, vd2012, vd2016) %>%   
 describe(quant = c(.25, .75), IQR = TRUE) %>%   
 mutate(year = c(2008, 2012, 2016)) %>%   
 relocate(year)  
# Kable Classic Method  
round(totalDemocraticVotesStats, 2) %>%   
 kbl(caption = "Table 2: Descriptive Statistics for Total Democratic Votes") %>%   
 kable\_classic(html\_font = "Cambria")

Table 2: Descriptive Statistics for Total Democratic Votes

Table

Description automatically generated

**Observation:**

* The mean of the total democratic votes registered for the election decreased in the year 2012 when compared to 2008 but increased in the year 2016 when compared to 2012. However, the highest mean of total votes registered by them is in the year 2008 only.
* The year 2008 also holds the maximum of the total votes registered by the democrats in the election poll.
* The means of total votes registered in all the 3 years election polls are very large from their respective medians.
* The minimum votes registered by the democrats in a county in all the 3 years has decreased up to just 4 votes in the year 2016 from 8 in 2008.

**Descriptive Statistics of ‘Total Republican Votes’ in 2008, 2012, 2016**

totalRepublicanVotes <- ElectionData %>%   
 select(vg2008, vg2012, vg2016) %>%   
 describe(quant = c(.25, .75), IQR = TRUE) %>%   
 mutate(year = c(2008, 2012, 2016)) %>%   
 relocate(year)  
# Kable Classic Method  
round(totalRepublicanVotes, 2) %>%   
 kbl(caption = "Table 3: Descriptive Statistics for Total Republican Votes") %>%   
 kable\_classic(html\_font = "Cambria")

Table 3: Descriptive Statistics for Total Republican Votes

Table

Description automatically generated

**Observation:**

* The mean of the total republican votes registered for the election decreased in the year 2012 but increased 2016 and the highest mean of total votes registered by them is in the year 2016 only.
* However, the year 2008 holds the maximum of the total votes registered by the republicans in the election poll.
* The means of total votes registered in all the 3 years election polls are large from their respective medians.
* The minimum votes registered by the republicans in a county in all the 3 years has decreased up to 57 votes in 2016 from their highest of 67 in 2008.

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

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 4420.26 to 2644.24.
* 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.

## Exploratory Data Analysis (Group & Sub-Group Analysis)

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

* 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 outcome 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

* 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

* 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

* 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 variables

The total

## Lasso Regression

The total

## Linear Regression

The total

## Stepwise Regression

The total

# CONCLUSION

In the coming 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

