**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 03, 2022

To: **PROF. ROY WADA**

**Table of Contents**

[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 the 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**

To predict the winning political party in each county, a Logistics Regression model will be built:

* 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

**Understanding the dataset:**

To perform the analysis and build the model, we will be using the county-level voter data from the 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

# Data Analysis

**Importing the libraries required for the analysis**

# Declaring the names of packages to be imported  
packageList <- c("tidyverse", "vtable", "RColorBrewer", "corrplot", "car", "psych", "stargazer", "scales", "DT")  
  
for (package in packageList) {  
 if (!package %in% rownames(installed.packages()))   
 { install.packages(package) }  
   
 # Import the package  
 library(package, character.only = TRUE)}

## Descriptive Statistics

**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

**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

**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

**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

## Exploratory Data 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, histogram

Description automatically generated

**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

Description automatically generated

**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, histogram

Description automatically generated

# Conclusion

The data set of car sales by CarDekho.com has provided various insights about the types of cars sold in the car industry and the patterns between them. The data set contains 8128 data points along with 13 features related to car details, engine details, and sale details.

# Bibliography

Vehicle dataset. (2020, October 24). Kaggle. <https://www.kaggle.com/nehalbirla/vehicle-dataset-from-cardekho>

Xie, Y. C. D. (2021, October 7). R Markdown Cookbook. R Markdown. Retrieved October 30, 2021, from <https://bookdown.org/yihui/rmarkdown-cookbook/>

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.

CarDekho. (2021). About Us | CarDekho.com. <https://www.cardekho.com/info/about_us>

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/>

D. (2021a, March 22). Data Analytics for Car Dealers. Automated Metrics. <https://www.automatedmetrics.io/data-analytics-for-car-dealers/>

# Appendix

The RMD file of the analysis is included with the analysis report.