FinalReport

Jaime Valencia Lopez, Evan Silzle, and Sumukh Chanda

2025-11-04

R Markdown

Title: A Predictive Model of USA Presidential Election by States.

Link to the Github: https://github.com/JaimeVal24/STAT107_Final_Project

Abstract:

In this project, we aim to create a predictive model based on State data that aims to accurately predict the party that a state will vote for in the presidential electionbased on demographic data and past voting history. Our question is "Are we able to somewhat accurately predict a states voting tendency based on demographic data using logistic regression? If so, which variables are most important to the model?". **Conclusion to be done**

Introduction:

The purpose of this analysis is to find certain how certain characteristics in individuals, and how their demographic background impacts their voting tendencies. By creating a successful predictive model, one is able to act on these predictions, whether it is to change them or to ensure that they happen. Knowing which states have a chance of flipping to either political party is also extremely important, as it is in these states that presidential candidates have the hardest battles. Not only can this benefit politicians, but also the voters, as the model highlights what factors are more important to the way in which they vote, and will therefore force these politicians to address these factors. If we were to find that unemployment rate seems to be heavily associated with a voter voting for the Democratic party, the democratic party should aim to address their constituents and lower unemployment.

Data:

The data available is mostly demographic data relating to the 50 states, meaning that there are 50 observations (plus territories and overall USA data that will not be used). The data sets are as follows:

Demographic Data: (https://www.kff.org/state-health-policy-data/state-indicator/distribution-by-raceethnicity/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D)

Income Data: (https://hdpulse.nimhd.nih.gov/data-portal/social/table?age=001&age_options=ageall_1&demo=00011&demo_options=income_3&race=00&race_options=race_7&sex=0&sex_options=sexboth_1&socialtopic=030&socialtopic_options=social_6&statefips=00&statefips_options=area_states)

Unemployment Rates: (https://www.bls.gov/web/laus/laumstrk.htm)

Commuter Mode: (https://www.bts.gov/browse-statistical-products-and-data/state-transportation-statistics/commute-mode)

Gun Ownership: (https://worldpopulationreview.com/state-rankings/gun-ownership-by-state)

Covid Deaths: (https://catalog.data.gov/dataset/provisional-covid-19-death-counts-rates-and-percent-of-total-deaths-by-jurisdiction-of-res)

In order to clean and aggregate the data, we performed a few transformations. Most importantly, for the commuter mode data set, the rows were formatted as (State_Name_Commuter_Mode), where it was necessary to modify it so that the state became the row on its own and the commuter mode into a column for every state. All conversions were made so that we could have the data for each state.

The following are the variables for each of our observations:

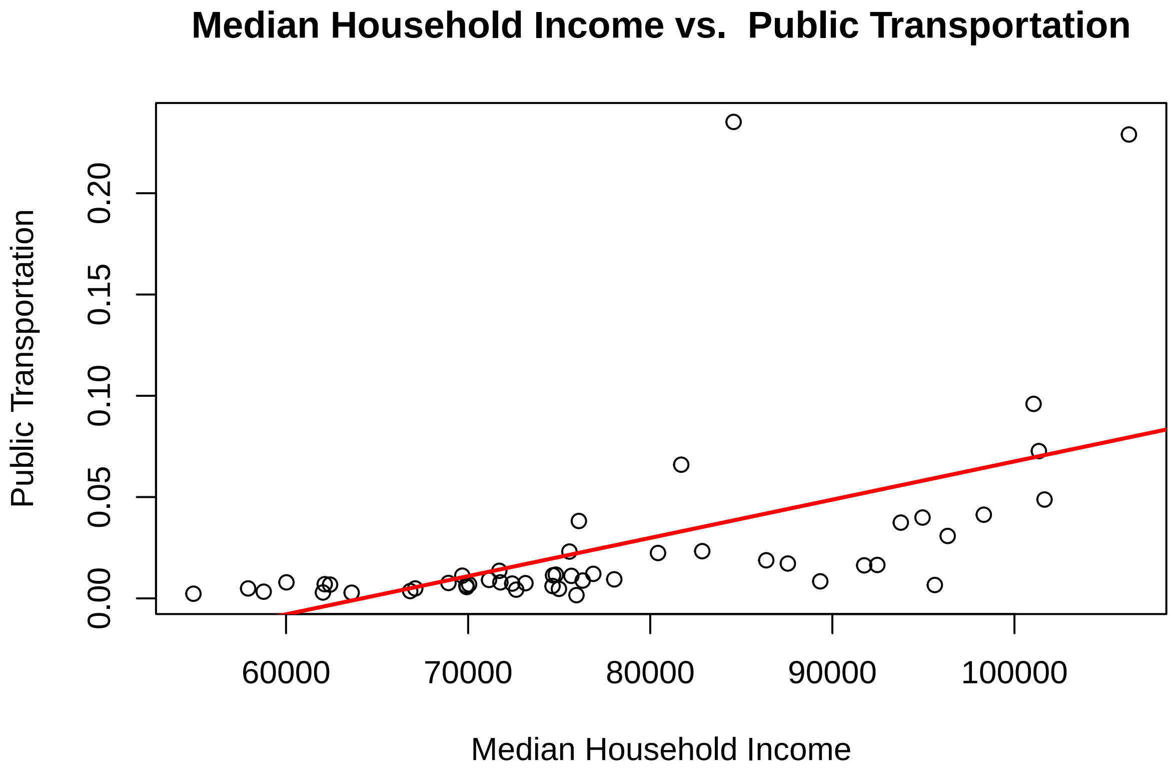
```
[1] “Location” “White”
[3] “Black” “Hispanic”
[5] “Asian” “American Indian or Alaska Native”
[7] “Native Hawaiian or Pacific Islander” “Multiple Races”
[9] “Total” “Bicycle”
[11] “Walked” “Taxi, motorcycle, or other”
[13] “Public transportation” “Worked at home”
[15] “Carpool” “Drove alone”
[17] “PercentageOfHouseholdsThatOwnGuns”

[18] “GunOwnership_NumOfGunLicenses_num_2022” “data_as_of”
[20] “Group” “data_period_start”
[22] “data_period_end” “COVID_deaths”
[24] “COVID_pct_of_total” “pct_change_wk”
[26] “pct_diff_wk” “crude_COVID_rate”
[28] “aa_COVID_rate” “crude_COVID_rate_ann”
[30] “aa_COVID_rate_ann” “footnote”
[32] “end_date” “Unemployment.Rate.August.2025”
[34] “Median.Household.Income” “Voting.Results.2016”
[36] “Voting.Results.2020” “Voting.Results.2024”
[38] “Electoral.College.Votes”
```

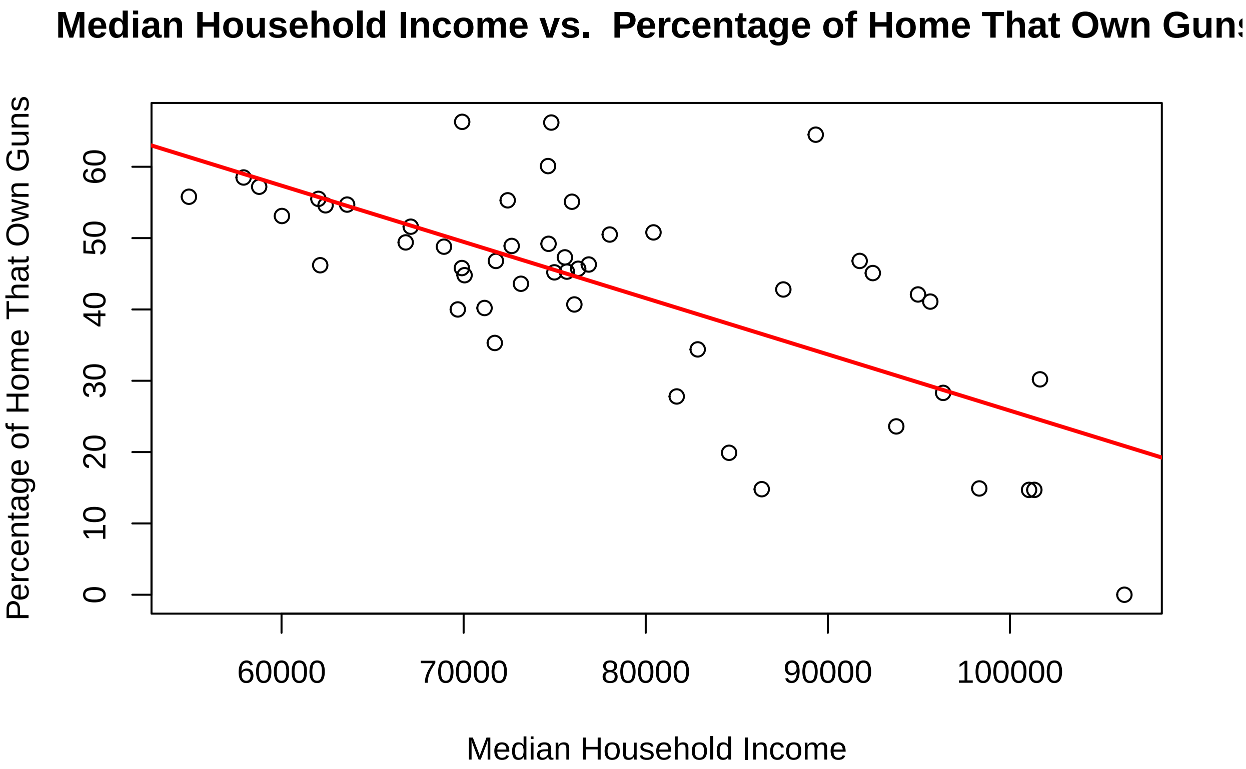
Data cleaning consisted of removing columns that were not necessary and could not be used for our future model. This included dates, fully empty columns, etc. We then replaced NA values in numerical columns with 0, as it will allow for the data set to be more manageable when it comes to implementing our model. Lastly, we turned the character columns of voting results from 2016, 2020 and 2024 into factor values.

Visualization:

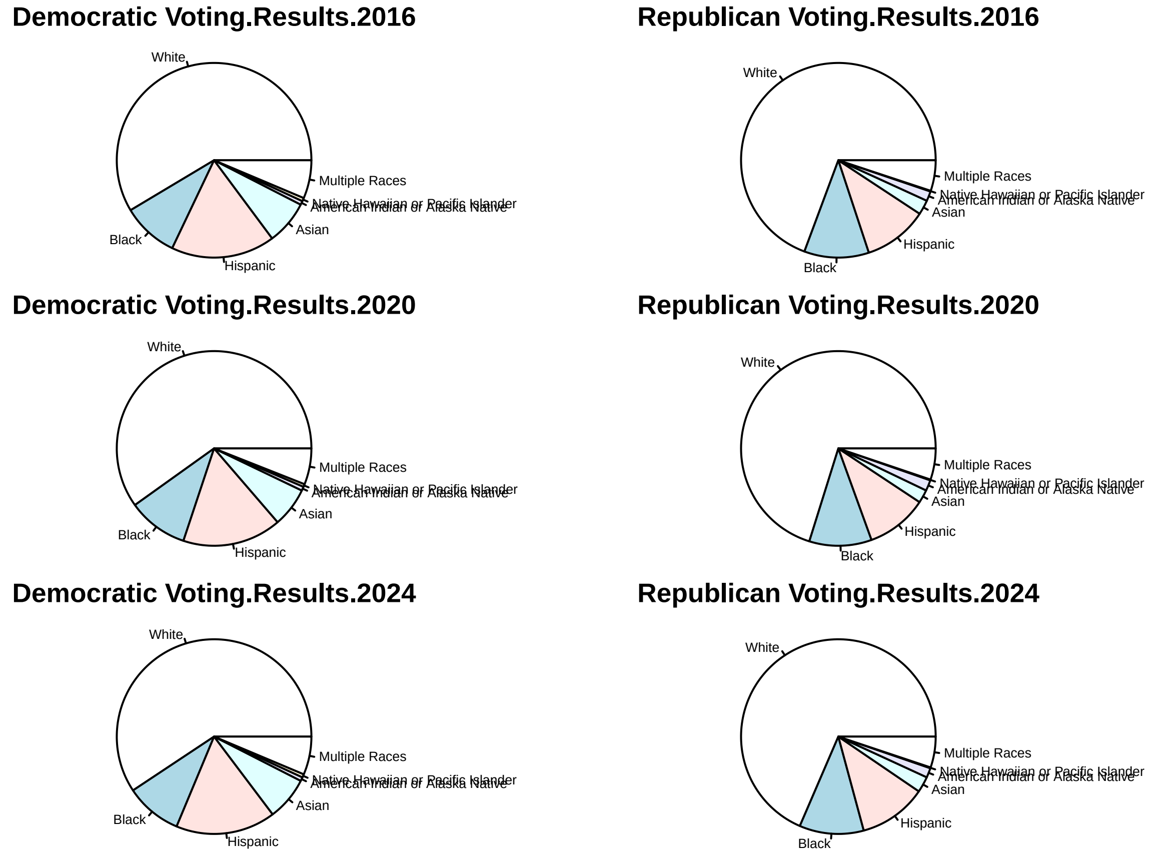
We want to explore the relationships between some of our variables, and see if they ar correlated. If there are correlations between different variables in our data then we can start to gain a clearer idea of how they might influence voting.



From the plot above we can see the relationship between the median household income and the rate of using public transportation to get to work. From the line of best fit we can tell that there is a relationship between the two variables. The states with higher median household income see higher rates of their workforce using public transportation to work. One thing that sticks out are the two high outliers at the top. New York has a median household income of 84,578 and a public transportation rate of 0.2352, and D.C. has a median household income of 106,287 and a public transportation rate of 0.2290.



When comparing the median household income of a state to the percentage of homes that own guns, we see a very clear trend that indicates that states with a higher median income own less guns, and vice versa.


Democratic Voting.Results.2016

Republican Voting.Results.2016

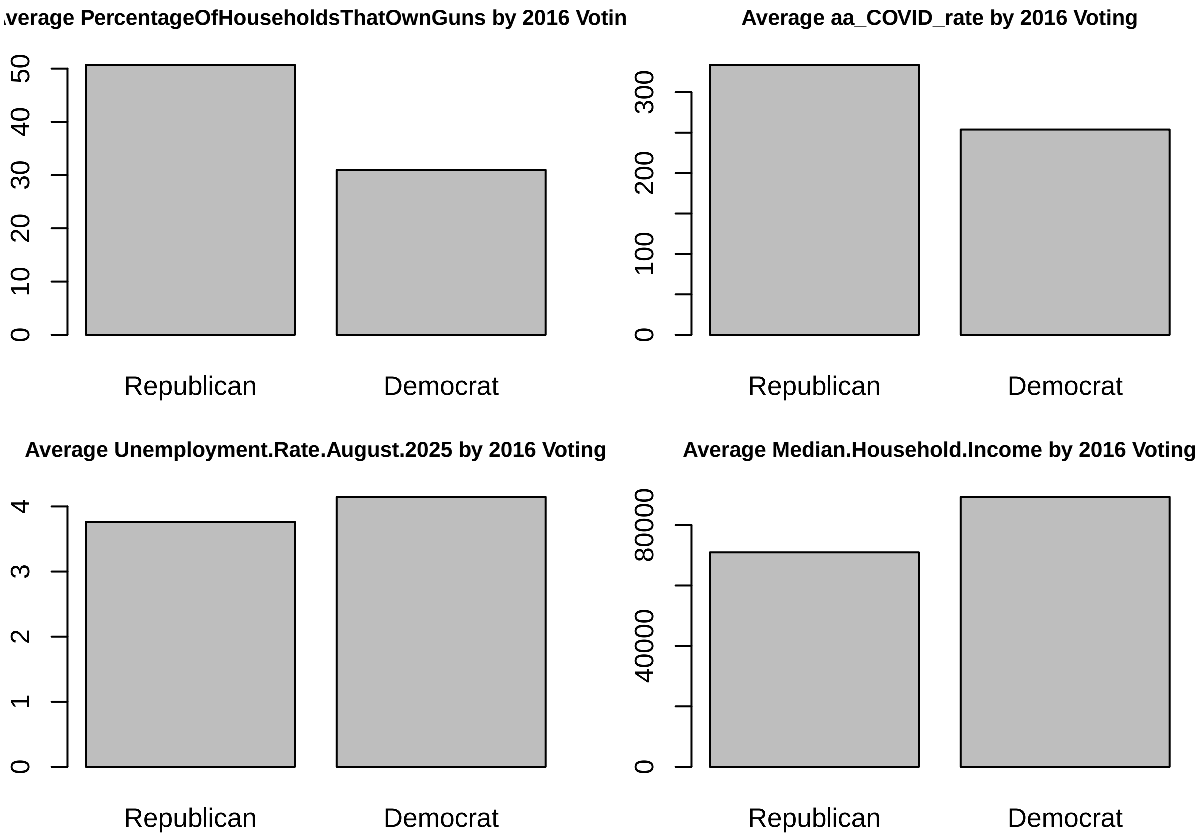
Democratic Voting.Results.2020

Republican Voting.Results.2020

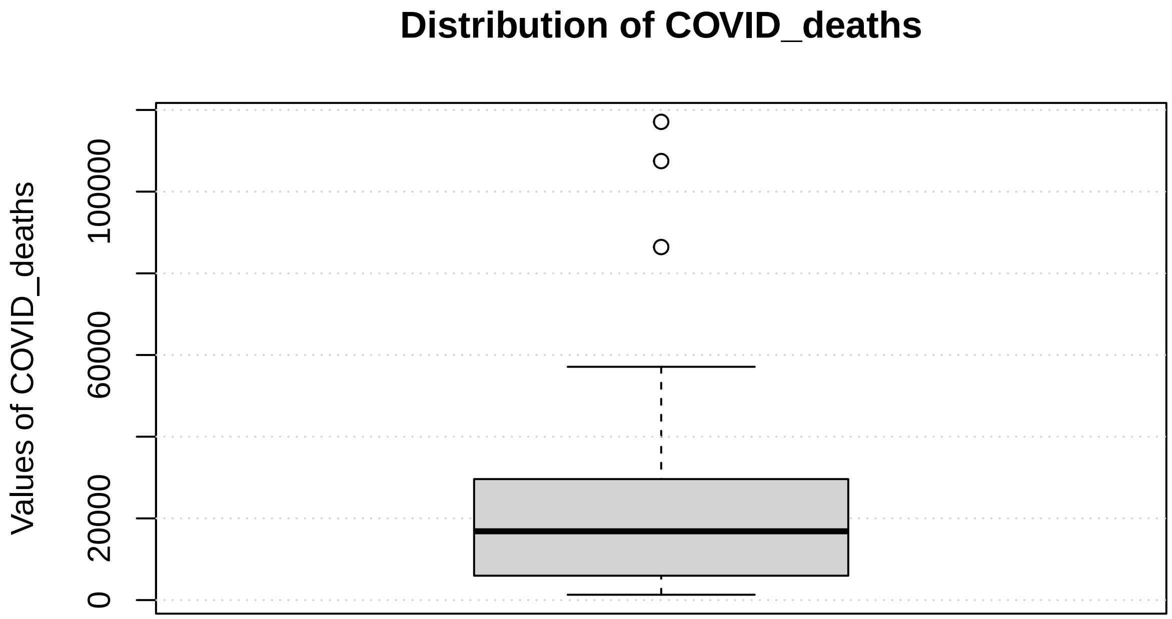
Democratic Voting.Results.2024

Republican Voting.Results.2024

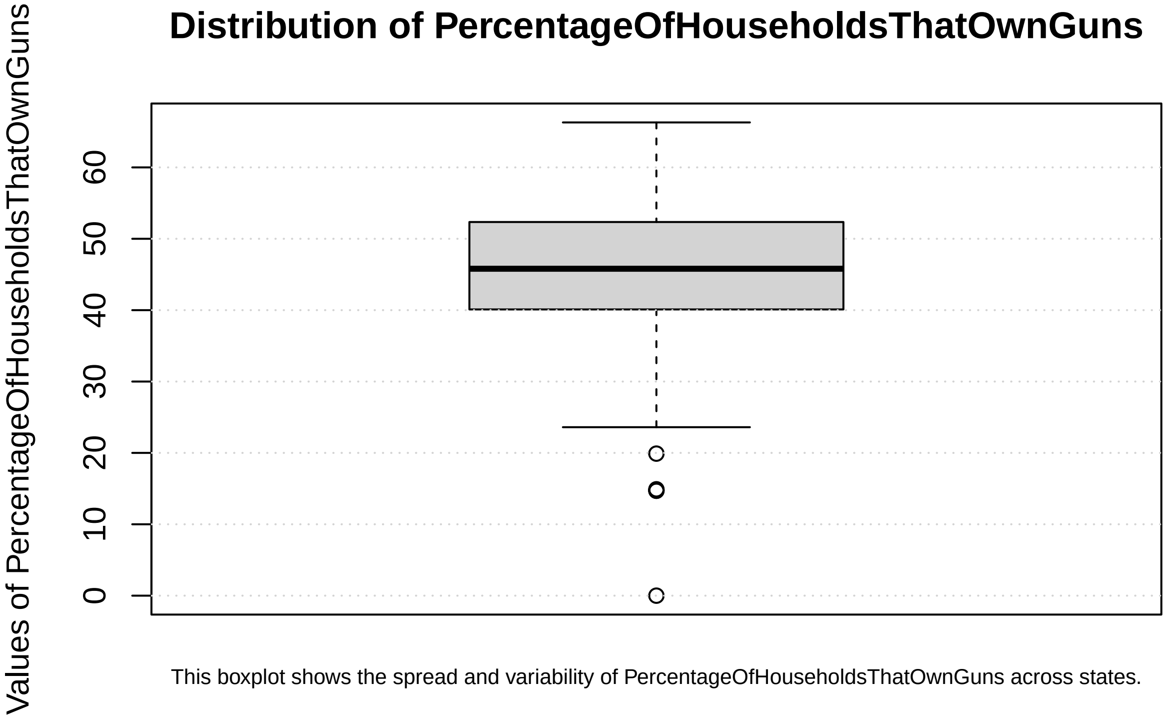
Next, we can look at the ethnic breakdown of each election, split by states that went red in each election and states that went blue. We can see that for every election the Democrats had less of a proportion of their votes from white voters, and more of a proportion of their votes from minority races. Additionally, although the ethnic breakdowns of each election differ greatly for democratic states vs republican states, the breakdowns barely differ when comparing different years.



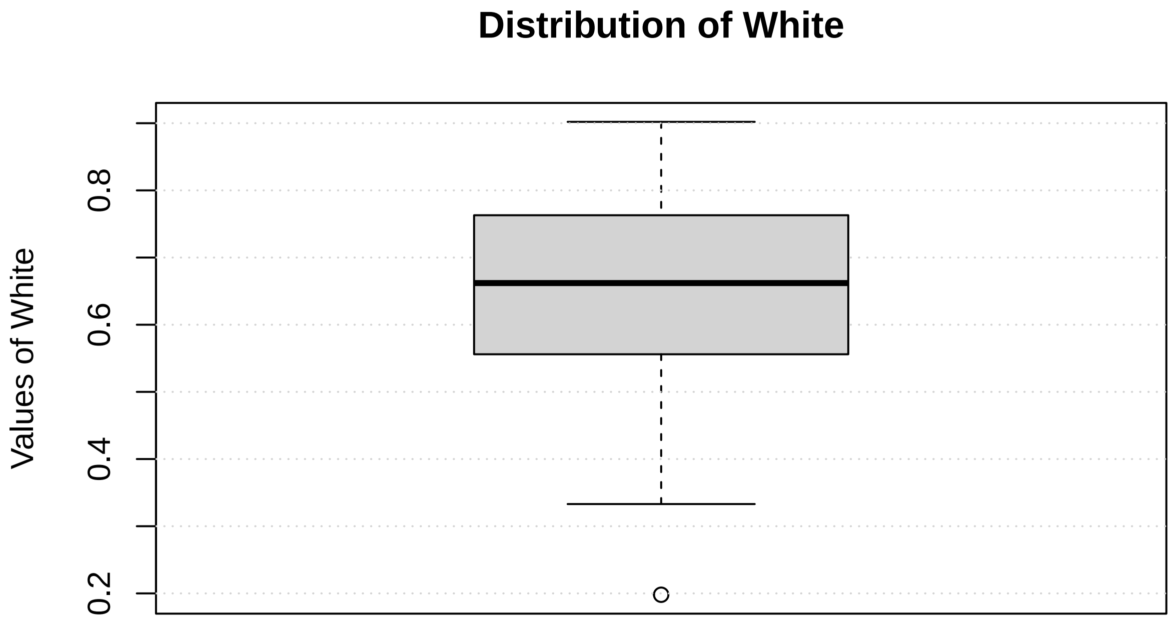
We can compare the average gun ownership percentage, pandemic COVID-19 rate, unemployment rate, and median household income for states that went blue and red in the 2024 election. We can see that states that voted Republican in 2024 have a much higher average gun ownership rate and average COVID-19 rate during the pandemic, and democratic states have a slightly higher average unemployment rate and average median household income.



This boxplot shows the spread and variability of COVID_deaths across states.



This boxplot shows the spread and variability of PercentageOfHouseholdsThatOwnGuns across states.



This boxplot shows the spread and variability of White across states.

Lastly, we can create box plots to understand the distribution of the data sets through their median, spread, and overall range. They highlight the central tendency, variability, and possible outliers and will give us more insight into our specific data and the validity of it as we are able to identify any skewed distributions. These specific ones are just a few of all of the box plots we can create for each variable introduced.

Analysis:

We plan to implement a logistic regression model, while also possibly implementing a K-Nearest Neighbors model for comparison of performance. We plan to use backwards selection in order to implement and discern only the most influential variables to our model. This way, not only can we build the most accurate model, but we can find the issues and demographics that are most important to how a state might vote. We expect to find that demographic variables such as race and median household income will play a significant role in the way a state votes, it is often that class and ethnic backgrounds demonstrate an affinity for either political party.

Once we construct and test our model, we can construct more data visualizations to demonstrate which variables have the largest effect on a states voting patterns, beyond basic correlations.

Individual Contributions are as follows:

Contributions:

Jaime Valencia Lopez: Wrote 21_DataProcessing.Rmd (merged the first four datasets and added the descriptions), 22_DataCleaning.Rmd (all code and descriptions), 01_funct_DataCleaning.R, FinalReport (excluding vizualizations) and README.md.

Evan Silzle: Wrote 02_func_PLots.r, Helped contribute to data cleaning process by merging a set of simple data sets manually, loading and cleaning them, then merging them with the main data set, as well as some cleaning of the merged data frame. Created and analyzed comparative data visualizations using loops and functions: two scatterplots, six related pie charts, and four related bar graphs.

Sumukh Chanda: Helped contribute to the idea and worked on the data visualization models.