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 develop a statistical model to predict which party a state will support in future elections based on 2024 presidential election data and general demographic data. Using publicly available data for all 50 states, we compiled measures of gun ownership, median household income, unemployment, COVID-19 mortality, transportation patterns, racial composition, and more. We then linked these variables to each state's 2024 presidential outcome. After initial exploration with correlation plots and pairs plots, we fit a series of logistic regression models and used LASSO regularization for variable selection to isolate the most informative predictors. We then estimated a Bayesian logistic regression model using the selected variables and evaluated its performance on specific states. The final model achieved roughly 90% accuracy, identifying gun ownership, median household income, and Age-Adjusted COVID-19 death rate as the key predictors. These results suggest that, although past voting behavior remains the strongest single predictor in practice, a combination of demographic indicators can also identify state level support, while highlighting how cultural, economic, and health-related factors jointly shape the modern U.S. voting map.

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:

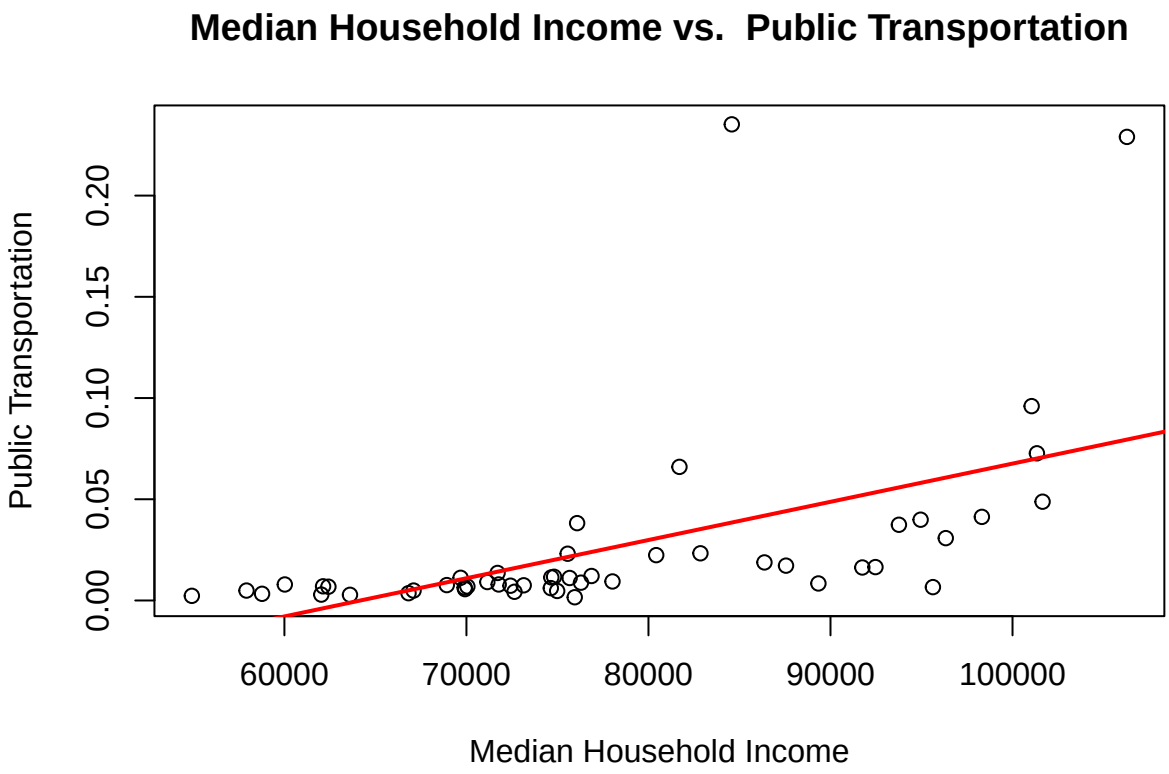
[1] "Location" "White"
[3] "Black" "Hispanic"
[5] "Asian" "American Indian or Alaska Native"
[7] "Native Hawaiian or Pacific Islander" "Multiple Races"
[9] "Total" "Bicycle"
[11] "Walked" "Taxi, motorcycle, or other"
[13] "Public transportation" "Worked at home"
[15] "Carpool" "Drove alone"
[17] "PercentageOfHouseholdsThatOwnGuns"

[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"

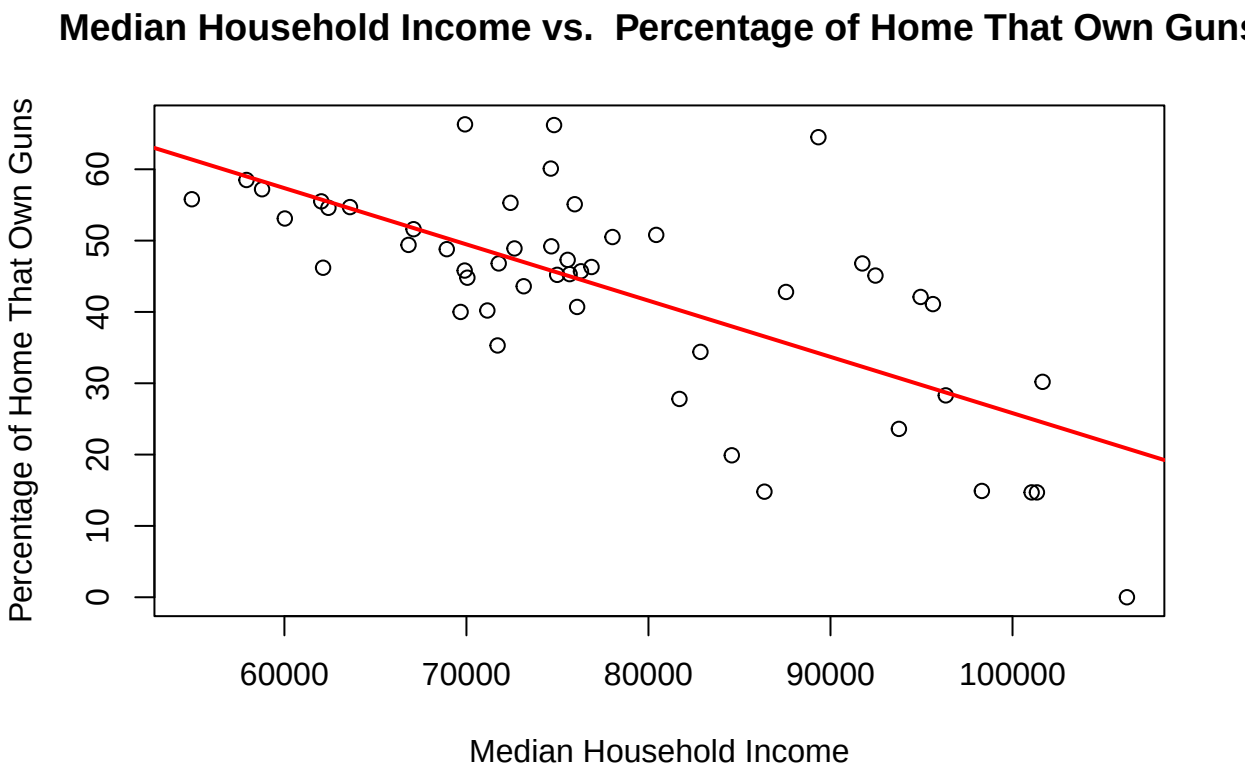
Data cleaning consisted of removing columns that were not necessary and could not be used for our future model. This included dates, fully empty columns, etc. We then replaced NA values in numerical columns with 0, as it will allow for the data set to be more manageable when it comes to implementing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

Visualization:

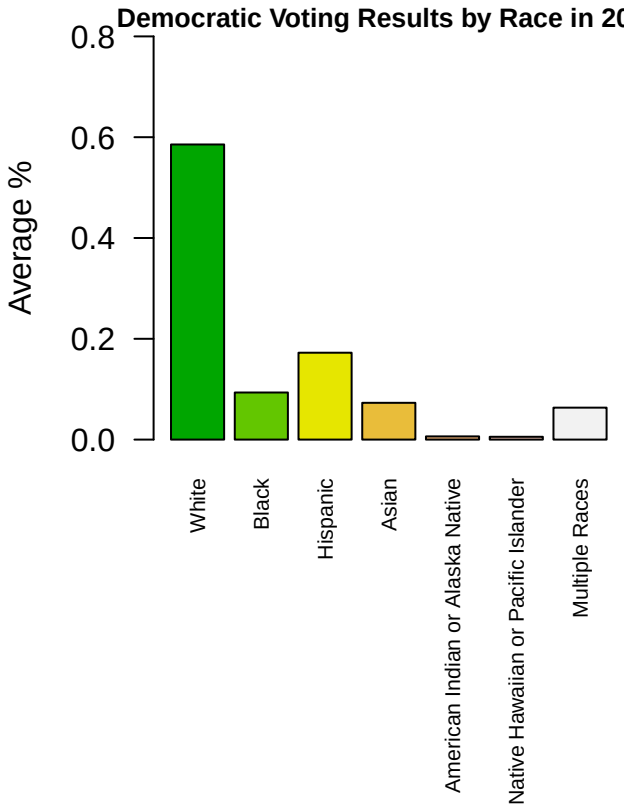
We want to explore the relationships between some of our variables, and see if they ar correlated. If there are correlations between different variables in our data then we can start to gain a clearer idea of how they might influence voting.

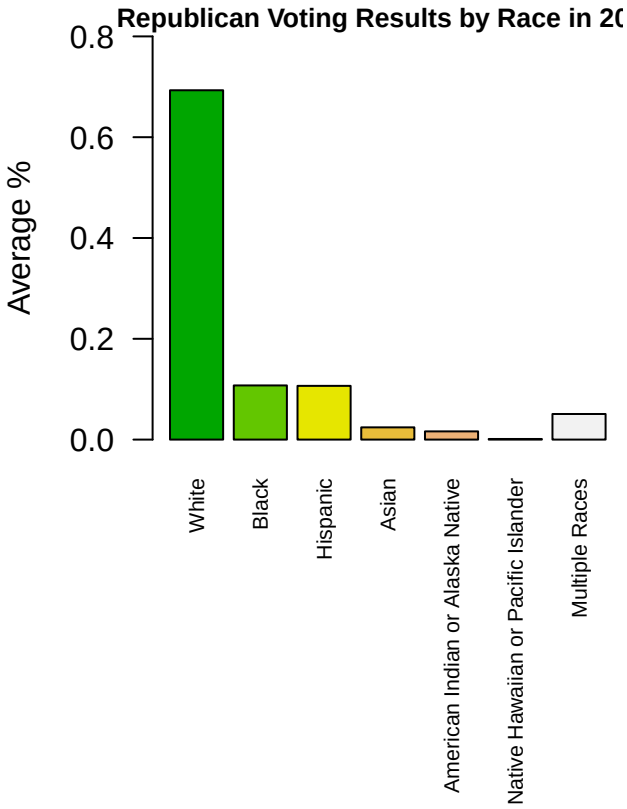


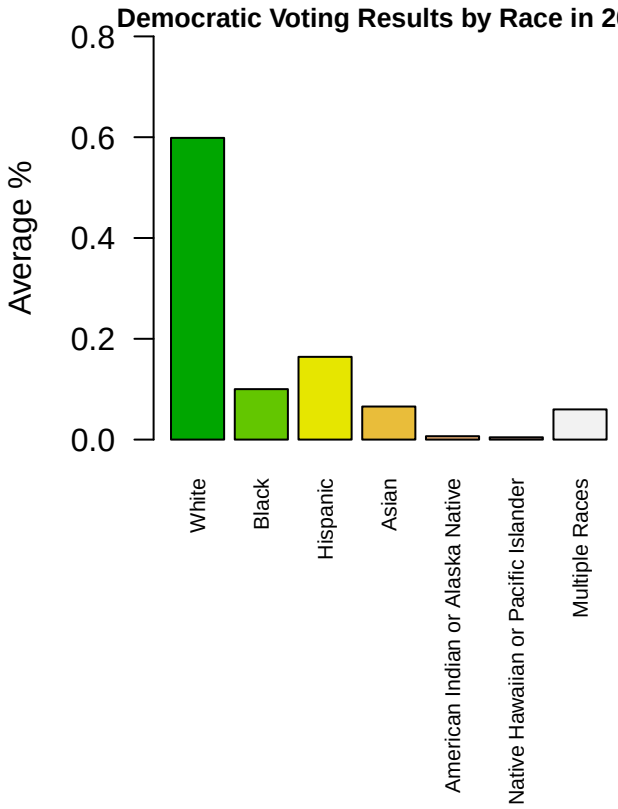
From the plot above we can see the relationship between the median household income and the rate of using public transportation to get to work. From the line of best fit we can tell that there is a relationship between the two variables. The states with higher median household income see higher rates 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.

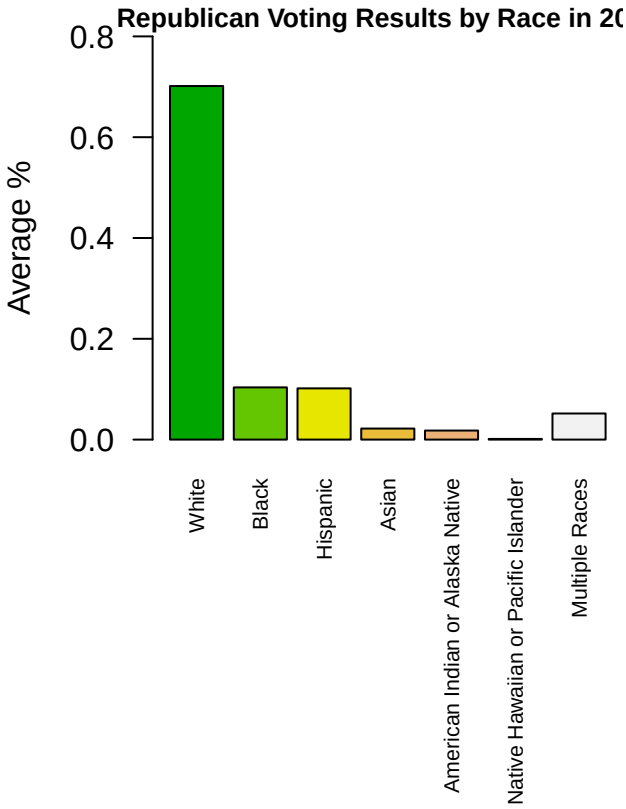


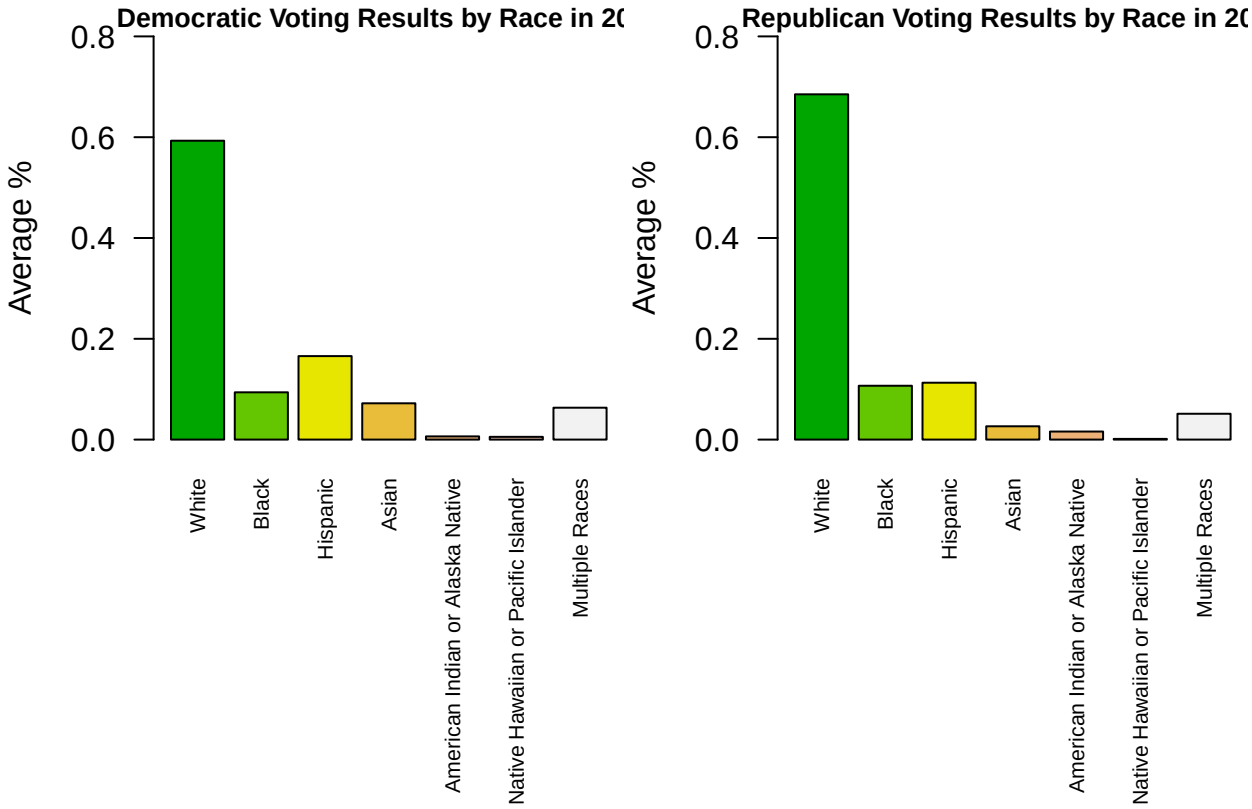
When comparing the median household income of a state to the percentage of homes that own guns, we see a very clear trend that indicates that states with a higher median income own less guns, and vice versa.


Democratic Voting Results by Race in 2(
Average %
0.0
0.2
0.4
0.6
0.8
White
Black
Hispanic
Asian
American Indian or Alaska Native
Native Hawaiian or Pacific Islander
Multiple Races

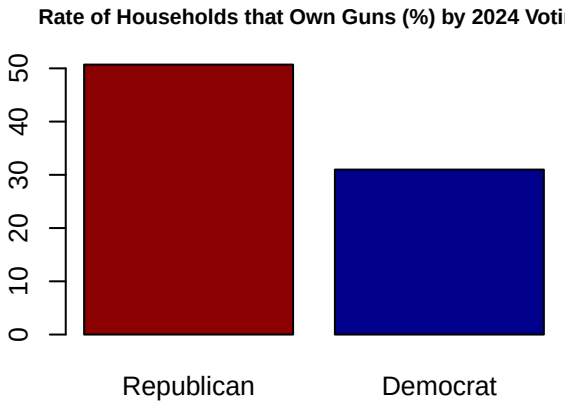

Republican Voting Results by Race in 2(
Average %
0.0
0.2
0.4
0.6
0.8
White
Black
Hispanic
Asian
American Indian or Alaska Native
Native Hawaiian or Pacific Islander
Multiple Races

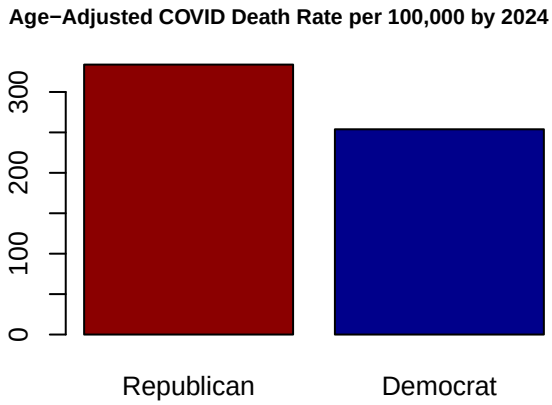

Democratic Voting Results by Race in 2(
Average %
0.0
0.2
0.4
0.6
0.8
White
Black
Hispanic
Asian
American Indian or Alaska Native
Native Hawaiian or Pacific Islander
Multiple Races

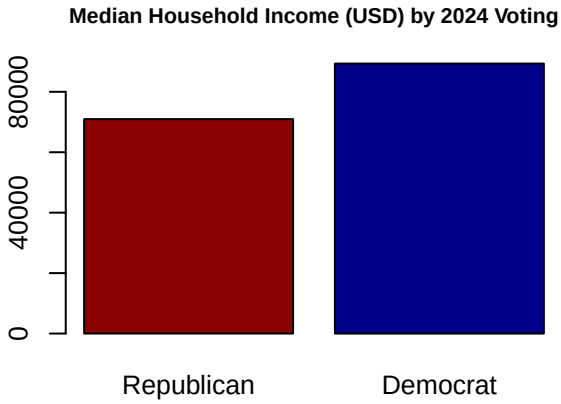

Republican Voting Results by Race in 2(
Average %
0.0
0.2
0.4
0.6
0.8
White
Black
Hispanic
Asian
American Indian or Alaska Native
Native Hawaiian or Pacific Islander
Multiple Races



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.

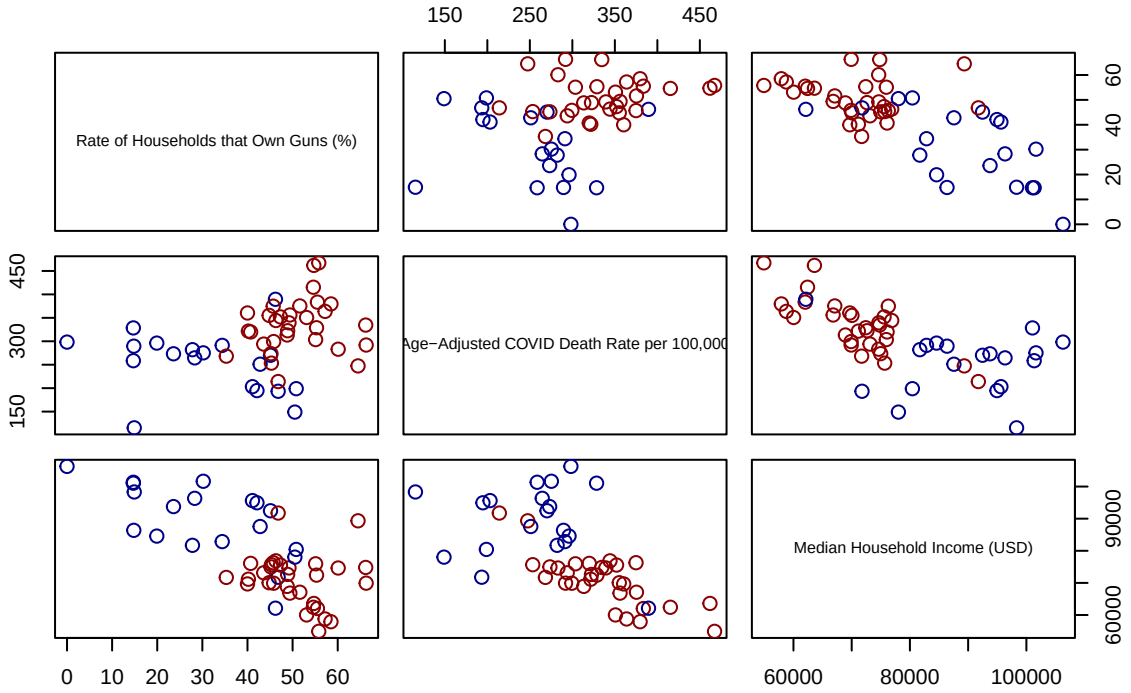


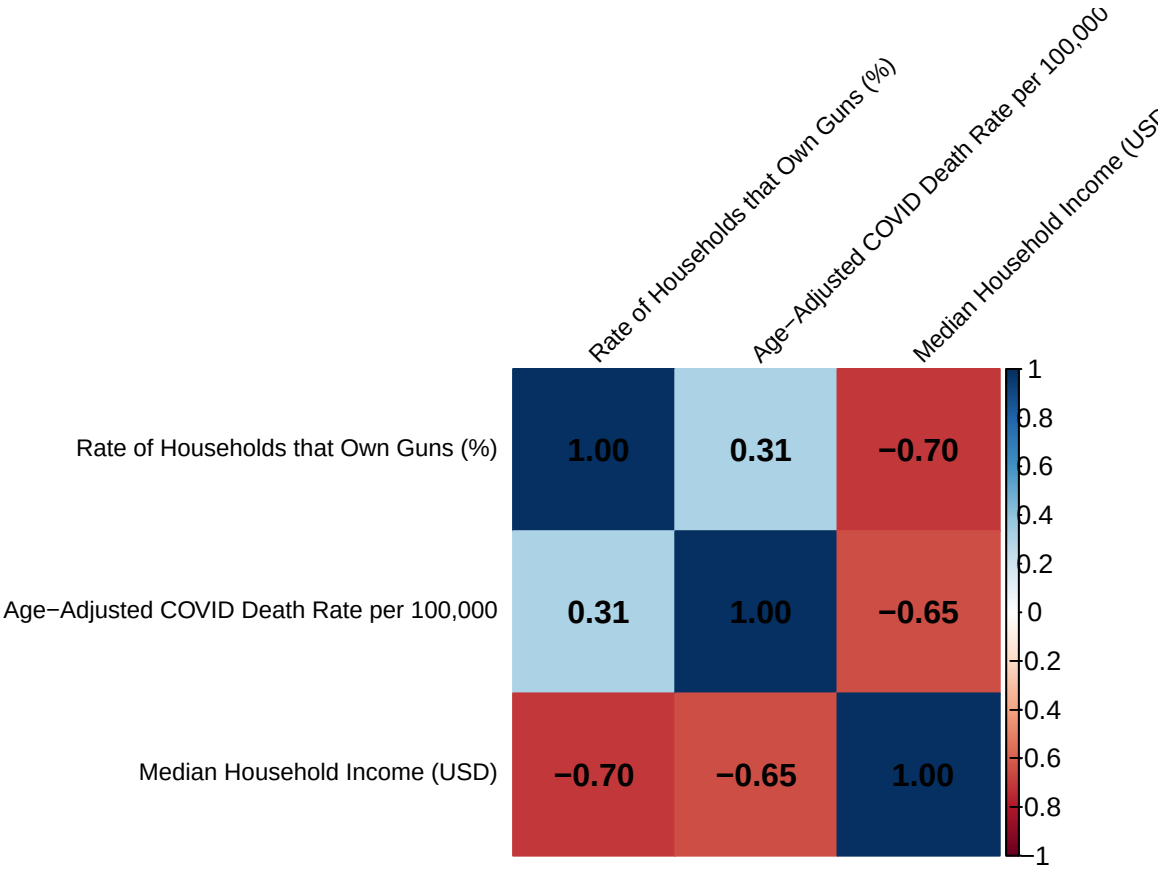




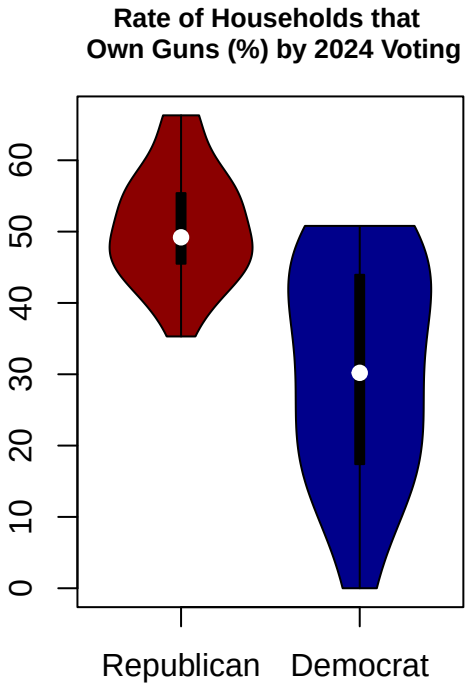
We can compare the average gun ownership percentage, pandemic COVID-19 rate, and median household income for states that went blue and red in the 2024 election. We can see that states that voted Republican in 2024 have a much higher average gun ownership rate and average COVID-19 rate during the pandemic,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

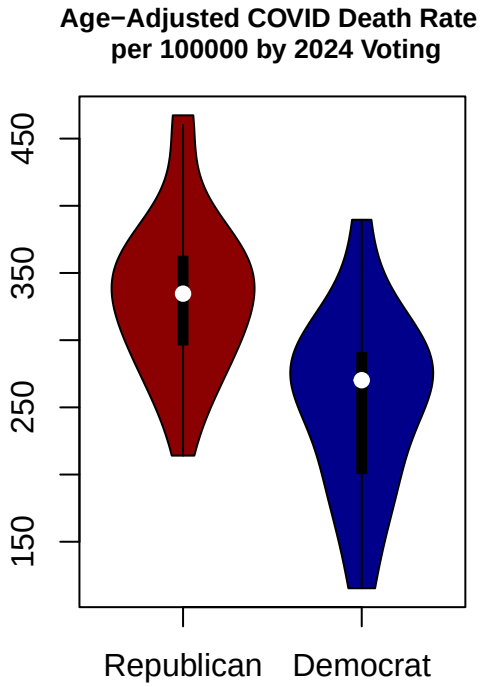
Pairs Plot of Assessed Variables Colored by 2024 Election Result

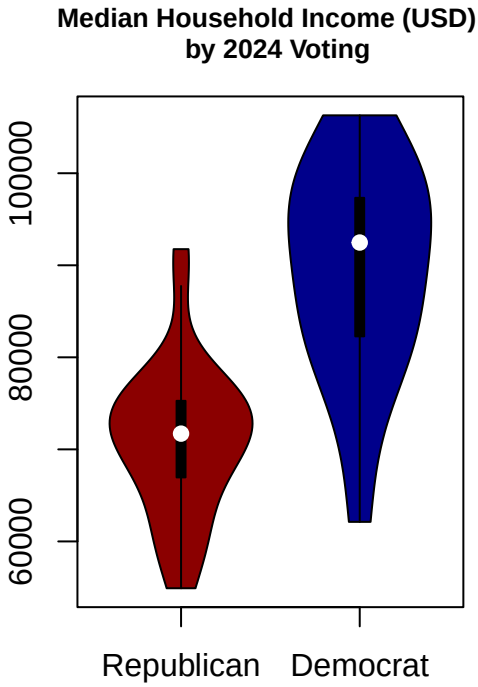


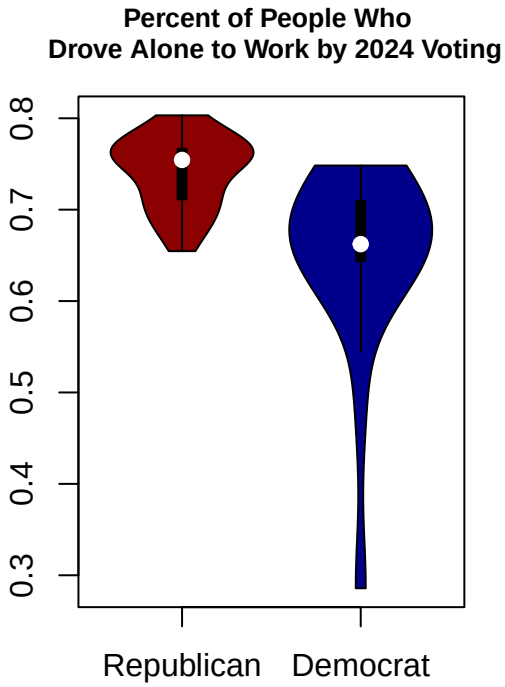


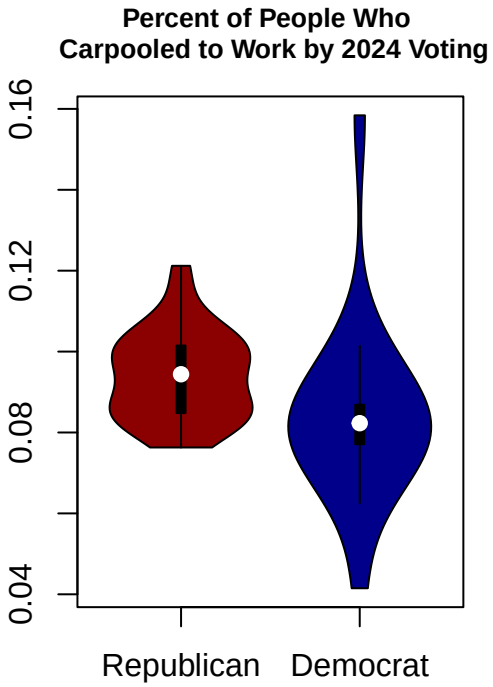
We can see from this pairs plot and related correlation heat map how each assessed variable relates to one-another. Each states Rate of Households that Own Guns (%), Age-Adjusted COVID Rate per 100000 people, and Median Household Income are compared to each other on scatter plots, and then assessed for an r correlation value, displayed on the heat map. The higher the correlation, the higher the absolute value of the r value, with positive values representing positive linear correlations and negative values representing negative linear correlations. We can see that every variable has a maximum correlation of 1.0 when being compared to itself, which is obvious, but the two highest correlations I want to talk about are the correlation between Median household income and gun ownership and the correlation between median household income and COVID rate. Both of these values of -0.70 and -0.65 respectively show a strong, negative correlation between the two variables.


Rate of Households that
Own Guns (%) by 2024 Voting
0
10
20
30
40
50
60
Republican
Democrat


Age−Adjusted COVID Death Rate
per 100000 by 2024 Voting
150
250
350
450
Republican
Democrat


Median Household Income (USD)
by 2024 Voting
100000
80000
60000
Republican
Democrat


Percent of People Who
Drove Alone to Work by 2024 Voting
0.8
0.7
0.6
0.5
0.4
0.3
Republican
Democrat



Lastly, we can create violin plots to understand the distribution of the analyzed variables in the data set 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.

The three variables LASSO chose for our regression model,Rate of Households that Own Guns (%), Age-Adjusted COVID Rate per 100000 people, and Median Household Income all showed relatively similar ranges with one political parties distribution clearly above the others. However, for the driving data that our forward selecting model chose, The ranges varied a lot more. For the percent of people who drove alone, our violin plot shows that although a majority of people drove alone to work across both parties, there were a few democratic states that showed a significantly lower rate of driving alone. Interestingly, When looking at the rate of people who carpooled to work, even though a similar rate can be seen across states from both political parties, a few democratic states features significantly lower rates of carpooling and a few democratic states features significantly higher rates of carpooling, giving the distribution a much higher range over republican states.

Analysis:

Our modeling strategy focused on constructing a logistic regression model that maximized predictive accuracy while maintaining interpretability. To validate our results, we employed a cross-validation technique by randomly splitting our 50-state dataset into a training set (75%) and a testing set (25%). This separation ensured that the model was trained on roughly 37 states and then evaluated on 13 "unseen" states, preventing the model from simply memorizing the data and ensuring it could generalize to new observations.

Variable Selection (LASSO Regression)

Our initial dataset contained over 25 potential predictor variables, ranging from transportation methods to racial demographics. To determine which variables were most relevant without succumbing to overfitting,

we implemented *LASSO (Least Absolute Shrinkage and Selection Operator) Regression*. LASSO is a regularization technique that improves model accuracy by applying a penalty to the regression coefficients. It mathematically shrinks the coefficients of less important variables to exactly zero, effectively removing them from the model and leaving only the most impactful predictors.

Crucially, before running the selection process, we removed the 2016 and 2020 election results from the dataset. While past voting behavior is the strongest predictor of future voting, including it would have overshadowed the demographic trends we aimed to study. The goal of our project is to understand what variables are most important in the way a state will vote; if we were to have used 2016 and 2020 election results, we could have very possibly arrived at a model with 100% accuracy. However, it would not have given us any insight unto what demographic variables are most important

The LASSO algorithm identified three distinct variables as the strongest drivers of the 2024 vote:

1. *Percentage of Households that Own Guns* (Cultural)

- This variable seems to represent a more cultural aspect of the population. We hypothesized that a higher percentage of households that own guns will result in a higher likelihood that the state will vote Republican.

2. *Median Household Income* (Economic)

- This variable represent the economic aspect of the state's population. We hypothesized that a higher median household income will result in higher likelihood of the state voting Democrat. (States with higher Median Household Income such as California, New York and Illinois seem to favor democrats.)

3. *Age Adjusted COVID Death Rate* (Health)

- From the CDC "Rates are based on deaths occurring in the specified week/month and are age-adjusted to the 2000 standard population using the direct method (see https://www.cdc.gov/nchs/data/nvsr/nvsr70/nvsr70-08-508.pdf).". This variable seems to encapsulate the health stantadards of each state. We hypothesize that states with higher Age Adjusted COVID rates will lean more towards republicans. *Model Refinement (Bayesian Logistic Regression)*

We used these three selected variables to fit our final *Bayesian Logistic Regression* model. We chose the Bayesian approach (`bayesglm`) over standard logistic regression due to the small size of our dataset ($n = 50$). Standard logistic regression often becomes unstable with small samples, especially when predictors are very strong, leading to infinite coefficients and unreliable standard errors. Bayesian logistic regression solves this by applying a prior distribution that acts as a stabilizer, gently pulling coefficients towards realistic values and ensuring the math remains robust.

Results and Interpretation

The Model achieved a **92.3% accuracy** on the test data, correctly classifying the voting outcome for **12 out of the 13 states** in the testing set.

Table 1: Final Model Results (Bayesian Logistic Regression)

Predictor Variable	Coefficient	P-Value	Significance
(Intercept)	3.845	0.575	
Gun Ownership (%)	−0.119	**0.027**	*
Median Household Income	$5.12e^{-5}$	0.356	
AA COVID Death Rate	−0.012	0.160	
Test Set Accuracy	*92.3%*	**(12/13 Correct)*	

As seen in Table 1, Gun Ownership was the primary driver of the model and the only statistically significant variable at the $\alpha = 0.05$ level ($p = 0.027$). The coefficient for gun ownership is negative (−0.119), indicating a strong inverse relationship with voting Democrat. Specifically, the odds ratio suggests that for every 1% increase in a state's gun ownership, the odds of that state voting Democrat decrease by approximately 11.2%. This magnitude identifies cultural identity as a massive divider in the current political landscape.

While Median Household Income ($p = 0.356$) and AA COVID Death Rate ($p = 0.160$) yielded p-values above the traditional 0.05 threshold, their inclusion was essential for the model's predictive power. In small datasets, variables that overlap (multicollinearity) often split the statistical "credit," causing their individual p-values to rise. However, without these variables, the model lacks the nuance to identify "swing" states. Income helps distinguish wealthy suburban areas (which may lean Democrat despite moderate gun ownership), while COVID rates likely help identify specific regional voting patterns in the South.

Diagnostics

The binned residual plot confirms that the final model is well-calibrated and fits the data effectively. In an ideal model, roughly 95% of the residuals should fall within the ±2 standard error bounds (represented by the grey lines). As shown in the figure, the residuals are predominantly contained within these error bands across the entire range of predicted probabilities. The points in the middle probability range (0.4 to 0.6) which represent the swing states—fall within the calculated bounds. This indicates that the model successfully captures the complex dynamics of battleground elections without significant systematic bias or underfitting.

Comparison with Alternative Models

To confirm the necessity of our three-variable approach, we tested simpler models using traditional Forward Selection methods based on p-values ($p < 0.05$). Interestingly, this greedy algorithm did not select Income or Gun Ownership; instead, it identified **Commuting Patterns** (specifically driving alone or carpooling) as the statistically significant separators of the training data.

Table 2: Forward Selection Model Results (Commuting Patterns)

Predictor Variable	Coefficient	P-Value	Accuracy
(Intercept)	53.12	0.004	
Drove Alone (Rate)	−52.88	0.010	
Carpool (Rate)	−179.73	0.013	
Test Set Accuracy	**76.9%**	**(10/13 Correct)**	

As shown in Table 2, this model relies on proxies for urban density (where driving alone is less common) rather than direct cultural or economic indicators. While the p-values appear significant ($p < 0.05$), the

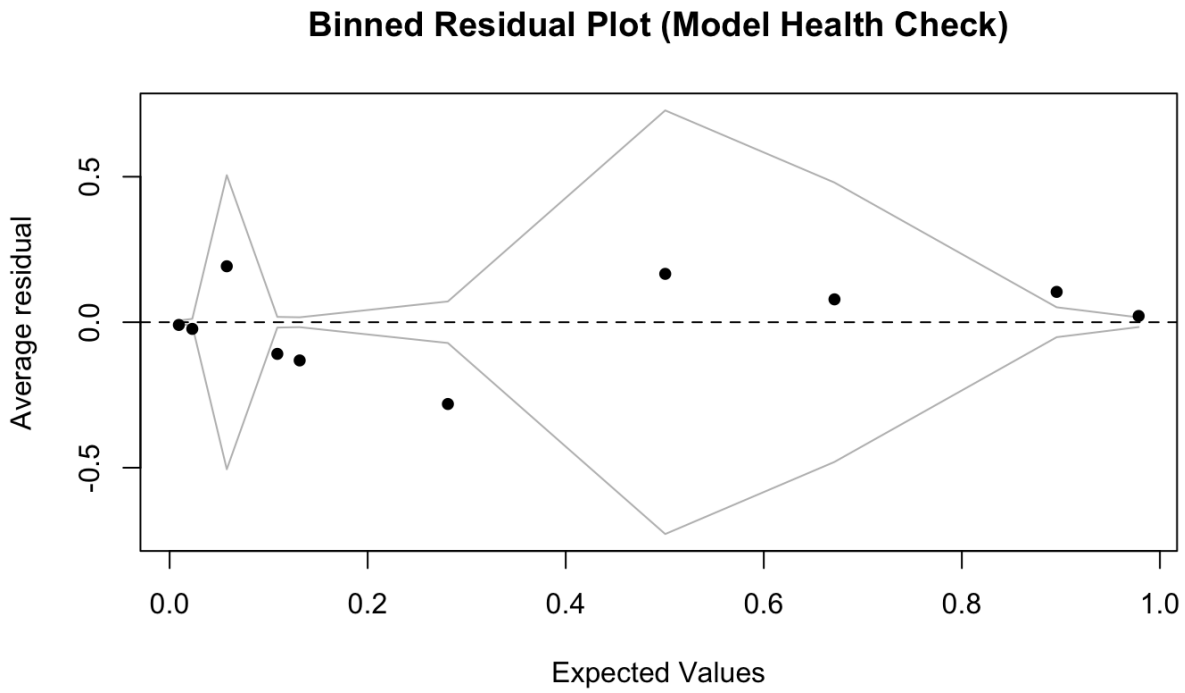


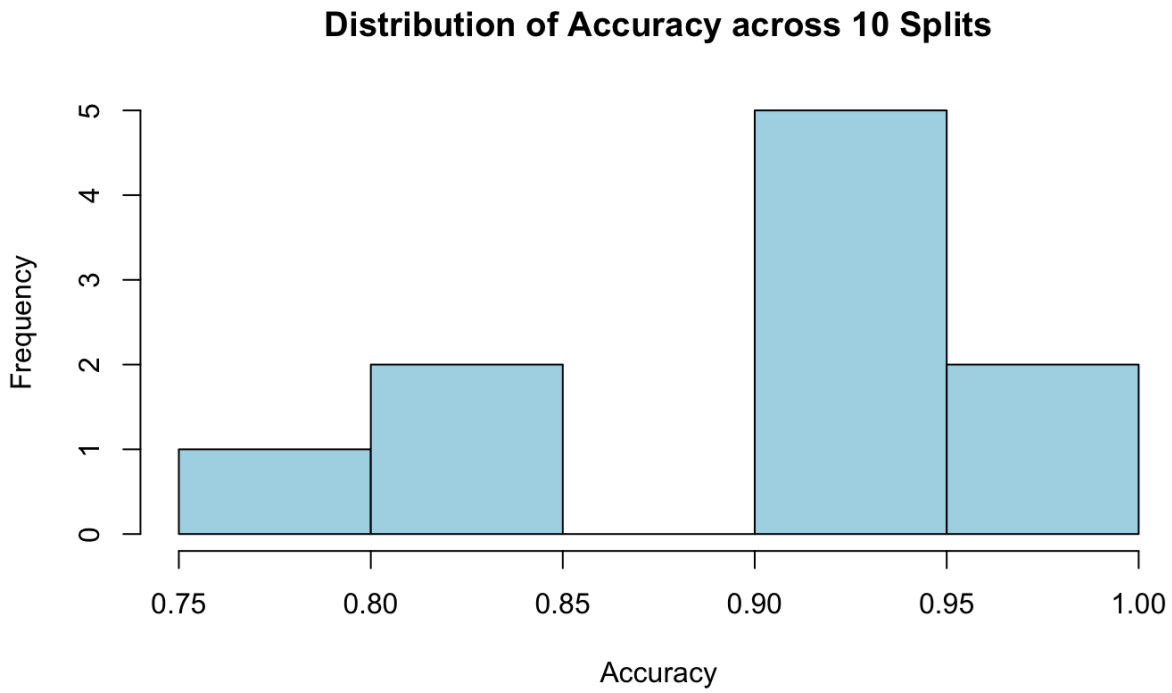
Figure 1: alt text here

model shows signs of instability, with massive coefficients (-179.7) and warnings of perfect separation in the training phase. Furthermore, this model achieved an accuracy of only **76.9%** on the test data. The drop in accuracy from the Champion Model (92.3%) to this benchmark demonstrates that while urban/rural proxies like commuting can identify the "obvious" states, they lack the nuance of the Culture (Guns) + Economy (Income) approach needed to correctly predict the more difficult swing states.

Table 3: Testing our model on different test/train splits

In this section, we tested our model using the same variable predictors on different train and test splits. We did this a total of ten times, using a different seed each time. Our results are as follows.

Run / Metric	Seed	Accuracy
Run 1	101	92.31 %
Run 2	202	100 %
Run 3	303	76.92 %
Run 4	404	84.62 %
Run 5	505	92.31 %
Run 6	606	100 %
Run 7	707	92.31 %
Run 8	808	92.31 %
Run 9	909	92.31 %
Run 10	1010	84.62 %
Mean Accuracy		**90.77 %**
Std Dev		**7.0687**
Range (Min - Max)		**76.92% - 100%**



With this iterative approach to testing our model, we are able to solidify the validity of our model, proving that the training and test split has a low effect on the accuracy of our model; as it manages to remain at around 90 percent. This leads us to believe that although we arrived at statistically insignificant predictive values ($p > 0.05$), we still managed to arrive at a model that tests with a high accuracy.

Conclusion:

Overall, the final Bayesian logistic regression model demonstrated that state-level demographic and contextual variables can predict 2024 presidential voting outcomes with roughly 92% accuracy, correctly classifying roughly 12 of 13 states on average. The LASSO technique identified gun ownership, median household income, and Age-Adjusted COVID-19 death rate as the most influential predictors, with gun ownership being the only statistically significant variable. Income and COVID mortality, while weaker individually, improved model stability and helped differentiate competitive "swing" states, highlighting the roles of culture, economics, and regional health in shaping voting results. While these findings suggest that past voting remains the strongest predictor in practice, they also show that carefully chosen demographic indicators can meaningfully anticipate state-level outcomes. Because this model includes only a single election cycle, these findings should be viewed as exploratory, and future work could incorporate data from multiple election years to test whether the model's predictive patterns remain consistent over time.

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. Implemented 31_LogisticRegressionModel. Added analysis, diagnostics and testing in report.

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 race bar graphs, and three related bar graphs for our LASSO data, a pairs-plot for the three, a heatmap of correlation between the three, and five violin plots for the distribution of the three LASSO variables and the two forwardly selected variables.

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