# Biden Projected to Win in the 2024 US Presidential Election*

Monica Sainani Kaavya Kalani

March 15, 2024

This study investigates the dynamics of voting behaviour of americans leveraging demographic, socioeconomic, and geographic factors, such as age, family income, education, state of residency, and sex. We use a Bayesian Logistic Regression model with Post-Stratification to predict the outcome between Biden and Trump with the mentioned factors as predictors. Our study indictates that Biden is likely to be re-elected as the president with a 10% lead. This research contributes to a better understanding of the factors influencing political engagement in American democracy by offering insightful information that might guide future electoral strategies and policy discussion.

## Table of contents



*Code and data supporting this analysis is available at: https://github.com/sainanim/Predictions-for-the-US-2024-Elections

# 1 Introduction

A key turning point in the political environment of the United States occurred with the 2020 presidential election, which was marked by extreme conflict and serious economic and social challenges. We analyze data from the American Community Survey (ACS) 2022 sample, which offers a thorough understanding of demographic, socioeconomic, and geographic aspects across the nation, in order to comprehend the patterns underlying voting choices during this 2024 election period.

Our estimand is the relationship between demographic, socio-economic, and geographic variables (such as age, income, education, state of residence, and gender) and voting behavior during the 2022 presidential election. Using this extensive dataset, our goal is to identify trends and factors that influence voting behaviour, clarifying the complex relationship between personal traits and local factors that influences election results.

Our research adds to a greater understanding of the factors influencing voting decisions through comprehensive statistical analysis, yielding insightful data that may influence future election campaigns and policy discussions. Through highlighting the subtleties of voting patterns in the 2022 election, our research provides a critical insight on the condition of American democracy and the factors motivating involvement in politics and engagement.

The paper is further organised into three sections: Data, Model, Results, and Discussion. In the Data section, we discuss how the datasets to be used for the analysis were obtained and pre-processed. We will explain the variables of interest in the datasets for the analysis. The Model section describes the model being used for the analysis. The Results section will then highlight and discuss the trends and associations found during the analysis. Lastly, the Discussion section will talk about some interesting trends found in Results in depth, link it to the real world and also highlight the weaknesses of our analysis.

# 2 Data

For this analysis, we have used two datasets. The datasets were cleaned and analysed using the statistical programming software R (R Core Team 2023) along with the help tidyverse (Wickham et al. 2019), knitr (Xie 2014), ggplot2 (Wickham 2016), here (Müller 2020), ADD MORE

## 2.1 Survey Data

Our survey data is from the Ipsos Poll [Roper #31120637]. It was collated from a web-based survey conducted on November 3 - 4, 2023 using KnowledgePanel. The survey is conducted by Ipsos and the sponsor is ABC News. This poll is based on a nationally representative probability sample of 949 adults age 18 or older with oversamples among Black and Hispanic respondents. The poll is a collection of demographic information and political preferences.

A person becomes an entry in this dataset if they were 18 years or older on November 3-4, 2023, an American national and participated in this poll. They are recruited through a scientifically developed addressed-based sampling methodology using the latest Delivery Sequence File of the USPS – a database with full coverage of all delivery points in the US. Households invited to join the panel are randomly selected from all available households in the US. People in the sampled households are invited to join and participate in the panel. Those selected who do not already have internet access are provided a tablet and internet connection at no cost to the panel member. Those who join the panel and who are selected to participate in a survey are sent a unique password-protected log-in used to complete surveys online. Panelists receive a unique login to the survey and are only able to complete it one time. No reminder emails were sent for this study.

The results from this dataset are used for our model. Among the range of variables available, we selected age, income, education, state and sex as our independent variables.

- Age is the age group the panelist falls into. The poll had the ages of the panelists at the time of the poll which we then classified into one of the four age groups: 18-29, 30-44, 45-59, 60+
- Income is the annual household income reported by the panelist. They were given the option of seven categories: Less than $10,000, $10,000 to $24,999, $25,000 to $49,999, $50,000 to $74,999, $75,000 to $99,999, $100,000 to $149,999 and $150,000 or more.
- Education is the highest level of education attained by the panelist. They were given the option between: Less than high school, High school, Some college and Bachelor's degree or higher. We then classified these into three categories: Less than high school, High school, More than high school
- State is the state of residence of the panelist.

- Sex is the sex reported by the panelist and it is either male or female

For the variable we are predicting, we created a binary variable which is 1 if the panelist favors Biden and 0 if the panelist favors Trump. There was no question in the poll which asked this direct question so we used "Overall, do you have a favorable or unfavorable impression of...Joe Biden?". We acknowledge that not favoring Joe Biden does not necessarily imply favoring Donald Trump, but since they are the two major potential candidates, we classify the panelists who chose "Unfavorable" as supporting Trump, i.e not supporting Biden.

### 2.2 Post-stratification Data

Our post-stratification data is from the most recent American Community Survey (ACS), the 2022 Sample. We use this to represent the voter population and adjust the weight of the survey data. The population for this sample was the entire US population, categorized into households as subgroups, and the responses of the sampled subgroups were collected. ACS 2022 dataset consists of 3,373,378 observations. Using a 1-in-100 national random sample, the ACS 2022 sample provides a thorough overview of the demographic, socioeconomic, and geographic features of the US population. Most importantly, this dataset includes people who live in both home and group quarters, offering a comprehensive view of the American population. A person becomes an entry in this dataset if they are a part of the sampled subgroups and have responded to the survey.

There are a large number of different features available and the variables we chose are age, income, education, state and sex to match with the variables from our survey. These variables were cleaned and the dataset was preprocessed so the values and categories match with the survey dataset.

## 3 Model

We used a Bayesian Logistic Regression model with Post-Stratification to predict the 2024 US Presidential Elections. Logistic regression is a method used for binary classification to predict the probability of a categorical dependent variable.

For our analysis, a logistic regression model will be first used to model the proportion of the voters who will vote for Biden. The model will be based on five independent demographic variables: state, sex, age, education and income.

The logistic regression model we will be using is:

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \beta_0 + \beta_1 \times \text{state} + \beta_2 \times \text{sex} + \beta_3 \times \text{age} + \beta_4 \times \text{education} + \beta_5 \times \text{income} \quad (1)$$

where,

- $\hat{p}$ represents the probability that someone will vote for Biden
- $\beta_0$ represents the intercept term of this logistical regression. It is the probability that someone will vote for Biden if the predictors' values are zero
- $\beta_1$ is the coefficient corresponding to state
- $\beta_2$ is the coefficient corresponding to sex
- $\beta_3$ is the coefficients corresponding to age
- $\beta_4$ is the coefficients corresponding to education
- $\beta_5$ is the coefficients corresponding to income

In our model, normal priors with a mean of 0 and a standard deviation of 2.5 are used for both the coeffients and the intercept. Setting the mean of the priors to 0 implies that there is no expectation of a particular direction or magnitude for the coefficients or intercept. We chose this as we have no expectation of the same. The standard deviation of 2.5 reflects the uncertainty or variability in the prior beliefs. We chose a moderately wide prior to allow for a reasonable amount of uncertainty.

The chosen priors allow the data to largely determine the posterior distribution as they are relatively non-informative. They don't heavily influence the results unless the data provide strong evidence to the contrary.

The use of moderately wide priors can also help regularize the model, preventing overfitting and providing more stable estimates, particularly when dealing with limited data.

Table 1: Coefficients of the model

| term | estimate | std.error | conf.low | conf.high |
|---|---|---|---|---|
| (Intercept) | -0.54 | 0.73 | -1.79 | 0.68 |
| sexmale | -0.24 | 0.17 | -0.50 | 0.03 |
| age30-44 | -0.20 | 0.27 | -0.65 | 0.26 |
| age45-59 | -0.31 | 0.27 | -0.77 | 0.14 |
| age60+ | 0.28 | 0.26 | -0.14 | 0.69 |
| stateAlaska | 48.60 | 43.20 | 5.39 | 137.72 |
| stateArizona | -0.02 | 0.88 | -1.52 | 1.47 |
| stateArkansas | -0.85 | 1.10 | -2.76 | 0.81 |
| stateCalifornia | 0.58 | 0.68 | -0.49 | 1.70 |
| stateColorado | -0.55 | 0.86 | -1.95 | 0.86 |
| stateConnecticut | 0.48 | 0.95 | -1.12 | 2.10 |
| stateDelaware | 1.05 | 1.18 | -0.85 | 3.11 |
| stateDistrict of Columbia | 47.03 | 39.82 | 4.86 | 136.07 |
| stateFlorida | 0.21 | 0.71 | -0.91 | 1.37 |
| stateGeorgia | 1.19 | 0.88 | -0.22 | 2.56 |
| stateHawaii | 0.37 | 1.81 | -2.89 | 3.49 |
| stateIdaho | 0.59 | 1.26 | -1.56 | 2.64 |
| stateIllinois | -0.06 | 0.81 | -1.35 | 1.26 |

| term | estimate | std.error | conf.low | conf.high |
|---|---|---|---|---|
| stateIndiana | -0.25 | 0.90 | -1.80 | 1.24 |
| stateIowa | -0.98 | 1.05 | -2.88 | 0.67 |
| stateKansas | 1.19 | 1.10 | -0.55 | 3.10 |
| stateKentucky | -0.77 | 0.96 | -2.38 | 0.78 |
| stateLouisiana | 0.79 | 0.87 | -0.68 | 2.23 |
| stateMaine | 0.60 | 1.27 | -1.48 | 2.78 |
| stateMaryland | 0.74 | 0.89 | -0.64 | 2.23 |
| stateMassachusetts | 0.61 | 0.85 | -0.73 | 2.02 |
| stateMichigan | 1.01 | 0.78 | -0.21 | 2.31 |
| stateMinnesota | 0.16 | 0.76 | -1.10 | 1.46 |
| stateMississippi | 1.46 | 0.97 | -0.11 | 3.17 |
| stateMissouri | -0.88 | 0.97 | -2.50 | 0.69 |
| stateNebraska | 0.29 | 1.06 | -1.49 | 2.11 |
| stateNevada | -1.00 | 1.09 | -2.87 | 0.73 |
| stateNew Hampshire | 24.84 | 20.31 | 4.26 | 70.33 |
| stateNew Jersey | 0.51 | 0.80 | -0.75 | 1.81 |
| stateNew Mexico | 2.08 | 1.39 | 0.02 | 4.82 |
| stateNew York | 0.92 | 0.69 | -0.21 | 2.08 |
| stateNorth Carolina | 0.27 | 0.72 | -0.90 | 1.46 |
| stateNorth Dakota | 1.45 | 1.55 | -0.96 | 4.30 |
| stateOhio | 0.07 | 0.73 | -1.07 | 1.29 |
| stateOklahoma | -1.25 | 1.06 | -3.15 | 0.37 |
| stateOregon | 0.60 | 0.95 | -0.99 | 2.21 |
| statePennsylvania | 0.24 | 0.70 | -0.91 | 1.41 |
| stateRhode Island | 0.96 | 1.71 | -2.10 | 4.17 |
| stateSouth Carolina | -0.97 | 1.05 | -2.91 | 0.71 |
| stateSouth Dakota | -0.33 | 1.60 | -3.25 | 2.17 |
| stateTennessee | 0.52 | 0.95 | -0.97 | 1.98 |
| stateTexas | 0.28 | 0.68 | -0.81 | 1.41 |
| stateUtah | -17.54 | 13.51 | -48.72 | -3.44 |
| stateVirginia | 0.69 | 0.77 | -0.54 | 1.95 |
| stateWashington | 0.78 | 0.79 | -0.52 | 2.12 |
| stateWest Virginia | 0.81 | 1.21 | -1.14 | 2.87 |
| stateWisconsin | 0.57 | 0.82 | -0.74 | 1.95 |
| stateWyoming | -33.43 | 29.58 | -93.91 | -3.22 |
| educationLess than high school | 0.54 | 0.37 | -0.05 | 1.12 |
| educationMore than high school | 0.60 | 0.21 | 0.27 | 0.97 |
| income$100,000 to $149,999 | -0.27 | 0.38 | -0.89 | 0.38 |
| income$150,000 or more | -0.24 | 0.38 | -0.84 | 0.40 |
| income$25,000 to $49,999 | -0.37 | 0.40 | -1.01 | 0.27 |
| income$50,000 to $74,999 | -0.45 | 0.40 | -1.10 | 0.21 |

| term | estimate | std.error | conf.low | conf.high |
|---|---|---|---|---|
| income$75,000 to $99,999 | -0.35 | 0.40 | -0.99 | 0.31 |
| incomeLess than $10,000 | -2.06 | 0.77 | -3.44 | -0.87 |

Table 1 shows the coefficients for our model that can be used in the equation along with the standard error and the 95% confidence interval. The standard error (SE) is a measure of the precision with which a sample statistic estimates a population parameter. It quantifies the variability of sample statistics around the population parameter. A 95% confidence interval means that if you were to take many samples and construct confidence intervals in the same way, approximately 95% of those intervals would contain the true population mean.

After we created the model, we then post-stratify these proportions and to conclude who we predict to win the 2024 US Presidential Elections. Post-stratification is a statistical technique employed to refine survey estimates by aligning them with known population characteristics. It helps us ensure that our findings accurately represent the diversity of the entire population.

In this process, we first the organize the survey data based on the demographic factors we decided to investigate. This involved counting individuals in each group to understand the sample distribution and computing proportions to depict the relative contribution of each demographic group to the overall survey sample.

Predictive model is then employed to generate probabilities for each demographic group, capturing the nuanced political preferences within distinct strata. These probabilities were summarized by aggregating them within each demographic group.

Applying this post-stratification process results in enhanced predictions that align with the known population distribution within each stratum. These adjusted values contribute to a more accurate representation of party votes within specific demographic categories, ensuring our findings are reflective of the broader population.

## 4 Results

Figure 1 In Figure 1 we see the percentage of the votes for Biden/Democratic party. This graph uses the data we collected after the post-stratification. Here, the red dots on the graph represent the voting percentage and the red lines are the error bars. The black vertical line with dashes indicates the line of proportion. Here, it is the equal split between support for Biden and support for the opposing candidate, Trump.

The states that have points to the left of the black vertical line are the states voting for Trump, and the states on the right are voting for Biden. We see that the majority of the votes for the democratic party are found from the Northeast,Midwest and the West Coast states. For

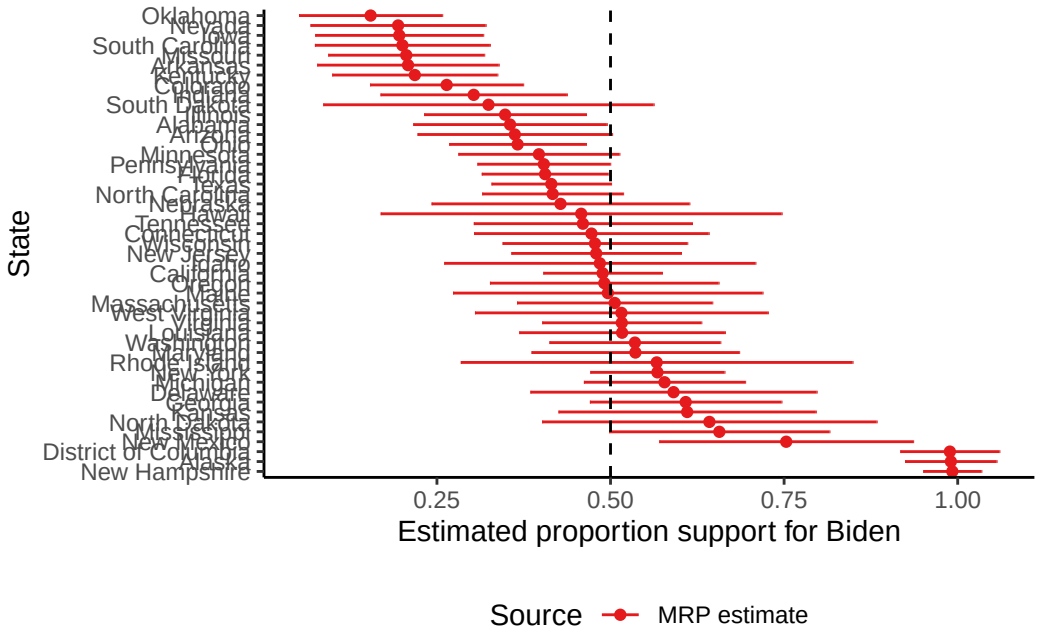


Figure 1: Estimated proportion support for Biden by State

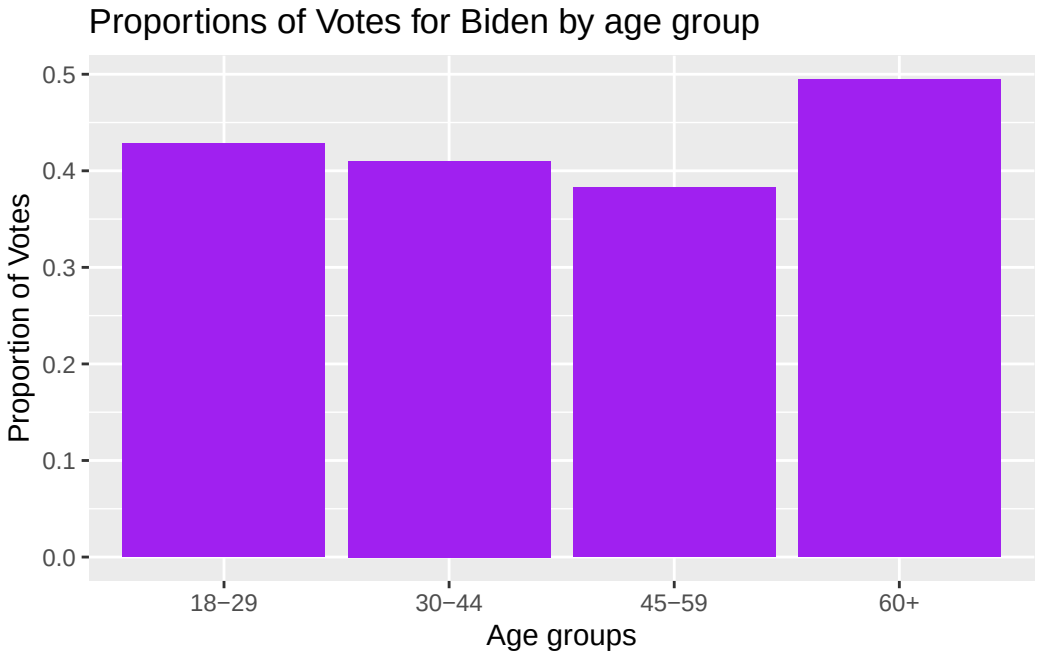


Figure 2: Estimated proportion support for Biden by Age Groups

example, the difference in votes for Biden between Oklahoma and New Mexico is more than 50%, but states like Oregon and Maine are neutral in terms of voting for either parties.

Figure 2 In Figure 2, we used the data to find the proportion of the voters categorized by age, who would vote for the Democratic party. We found a pattern that the group of youngest voters and the group of the senior citizen votes are more likely to vote for the Democratic Party(Biden).

We see that 45% of the individuals from age group 18-29 are voting for biden as well as 50% of the individuals from the age group of 60+ years. Meanwhile, only 35% of people in the age group of 45-59 are voting for Biden. Hence, in this graph we see the trend that people of ages from 30-59 are less likely to vote for the democratic party.

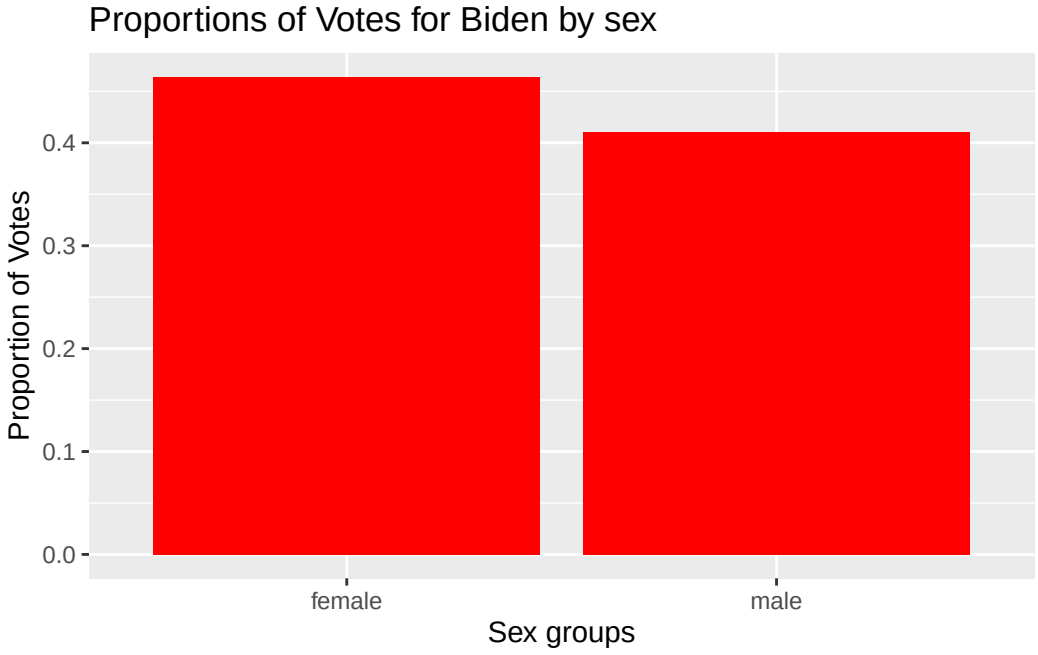


Figure 3: Estimated proportion support for Biden by Sex

Figure 3 Figure 3 shows the proportion of Female voters that would vote for the Democratic Party as compared to the Male voters. We see a 10% difference in the proportion. Here, we see that Women are 10% more likely to vote

Figure 4

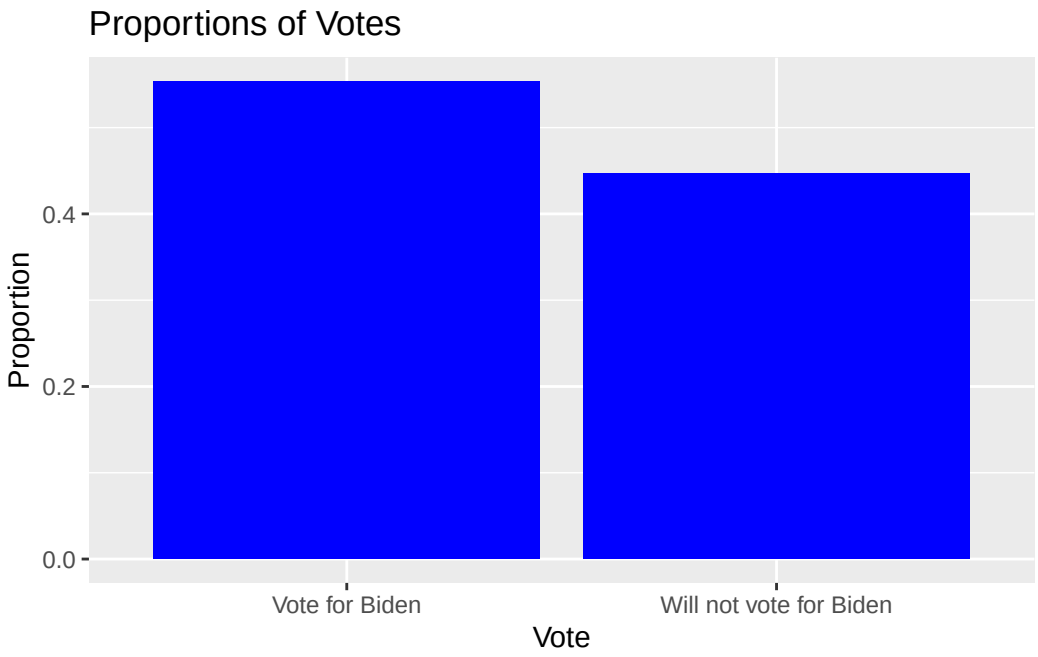


Figure 4: Estimated proportion support for Biden

# 5 Discussion

## 5.1 Insights Gained:

Our analysis, based on sample data from the 2022 survey, followed by post-stratification focusing on key variables, Age, State, and Sex, yielded valuable insights into voter behavior. Notably, our findings indicated an inclination within the population towards voting for Biden or the Democratic Party.

In the graph showcasing different regions and their voting for biden, we noticed that This is similar to the findings in https://www.brookings.edu/articles/forecasting-the-presidential-election-what-can-we-learn-from-the-models/ that address the same prediction for 2024. Our model and the article predict that states that support Trump are from the middle and Midwest parts of the US, whereas coastal states and the Northeast tend to vote for Biden. Historically, rural, conservative-leaning towns with strong cultural ties to traditional values and industries like manufacturing and agriculture have defined the middle and Midwest regions of the United States. These areas frequently support gun rights, minimal government intrusion, and religious liberty—issues that are strongly aligned with the agenda of the Republican Party, which Trump advocated. In contrast, the Northeast and coastline states, which are renowned for their progressive views on culture, urban facilities, and diverse populations, have historically

tended toward more liberal policies and social ideals that are more in line with the Democratic Party platform, which Biden advocated.

We also created a graph highlighting different age groups and what their voting opinions are. We noticed that the youngest group of voters (age 18-24). The emphasis on social justice and LGBTQ+ rights among the younger generation may be the cause of this difference. Since there are undoubtedly more young people who identify as sexual minorities, there is a trend for them to support Biden and the Democratic Party. We also noticed that the older generation i.e individuals of age group higher than 60 years old were inclined to voting for Biden as well. National security and conservative principles, which are frequently connected to the Republican Party, may have an impact on the elder generation.

Lastly, our third variable of study was Sex. We noticed a slight bias towards the democratic party from the females than the male gender. Women are predicted to vote for the democratic party 0.6 times more than men. Studies have consistently demonstrated that there is a gender difference in political ideology, with women generally favoring liberal or progressive policies over those of males. https://www.pewresearch.org/politics/2023/07/12/voting-patterns-in-the-2022-elections/#:~:text=In%20the%202022%20midterms%2C%2054,2022%20while%2051%25%20favored%
This discrepancy is sometimes linked to the fact that women prioritize issues more than men do, with women giving social welfare, healthcare, education, and reproductive rights—all of which are more in line with the Democratic Party program. With its focus on women's rights, healthcare reform, and inclusivity, Biden's campaign may have connected more deeply with female voters, resulting in higher levels of support.

## 5.2 Comparative Analysis:

Our findings align with existing literature on electoral behavior, which underscores the significance of demographic variables in shaping voter preferences. While our study offers insights into the inclination towards voting for Biden, it is important to contextualize our findings within the broader landscape of political science research. https://projects.economist.com/us-2020-forecast/president/how-this-works

Our analysis contributes to a growing body of evidence highlighting the complex interplay between socio-economic factors and political attitudes, providing valuable insights for future research and decision-making in the political arena.

## 5.3 Contributions to Knowledge: (Need to edit this)

By focusing our analysis on specific demographic variables, we contributed to a deeper understanding of the underlying determinants shaping electoral outcomes. Our study corroborates existing literature suggesting that demographic factors such as age, income, education, and geographic location play pivotal roles in influencing voter preferences. (https://projects.economist.com/us-2020-forecast/president/how-this-works) Moreover, our

utilization of post-stratification techniques allowed for a more granular examination of voting behavior, offering insights into the heterogeneity of voter sentiment across different demographic groups.

## 5.4 Weaknesses and Limitations: (Need to edit this)

While our analysis provides valuable insights into voter behavior, it is important to acknowledge several limitations. Firstly, our reliance on sample data from the 2022 survey may introduce biases and inaccuracies in our predictions. Our reliance on sample data from the 2022 survey may introduce a sampling bias, as the sample might not fully represent the diversity of the electorate. Consequently, this could lead to inaccuracies in our predictions and limit the generalizability of our findings to the broader population.

Additionally, while we focused on key demographic variables, other influential factors such as political ideology and campaign messaging were not fully accounted for in our model. This means that our research has the inaccuracy of Incomplete Variable Consideration.

Furthermore, uncertainties in the post-stratification process may impact the reliability of our estimates, warranting caution in interpreting the observed inclination towards voting for Biden.

## 5.5 Future Directions:

Moving forward, further research is warranted to address the limitations identified in our study and advance our understanding of electoral dynamics. Future studies could explore the integration of additional data sources, such as social media sentiment analysis and polling data, to enhance the accuracy of election predictions. Additionally, efforts to improve transparency and reproducibility in data collection and analysis are essential for enhancing the reliability and validity of electoral forecasts.

## 5.6 Data Source:(Add where the second dataset is from)

The data utilized in this study were sourced from the 2022 survey, supplemented by post-stratification data obtained from the American Community Surveys (ACS). These datasets provided comprehensive information on voter demographics, socio-economic characteristics, and geographic location, enabling a robust analysis of voting behavior in the 2020 US presidential election.

# 6 Appendix

## 6.1 Cleaning

## 6.2 Survey dataset

Here is a glimpse of the survey data set used

```
head(survey_data)
```

```
# A tibble: 6 x 6
  age   education             income               state     sex    vote_biden
  <chr> <chr>                 <chr>                <chr>     <chr>       <dbl>
1 60+   High school           $25,000 to $49,999   Wisconsin male            1
2 60+   More than high school $150,000 or more     Florida   male            1
3 60+   More than high school $50,000 to $74,999   Maine     male            1
4 60+   High school           $10,000 to $24,999   Virginia  female          0
5 60+   More than high school $100,000 to $149,999 Colorado  male            0
6 30-44 More than high school $25,000 to $49,999   New York  male            1
```

## 6.3 Poststratification dataset

Here is a glimpse of the poststratification data set used

```
head(poststrat_data)
```

```
# A tibble: 6 x 5
  age   education             state   income           sex
  <chr> <chr>                 <chr>   <chr>            <chr>
1 60+   More than high school Alabama $150,000 or more female
2 45-59 High school           Alabama $150,000 or more male
3 30-44 Less than high school Alabama $150,000 or more female
4 60+   Less than high school Alabama $150,000 or more male
5 45-59 More than high school Alabama $150,000 or more male
6 30-44 High school           Alabama $150,000 or more male
```

## References


Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://here.r-lib.org/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.