A Multifaceted Examination of the 2020 United States Presidential Election*

Understanding and Comparing how Identity-Based, Socio-Economic, and Regional Variables Shaped the Outcome

Sima Shmuylovich

March 16, 2024

This study examines the factors that shaped support for the United States of America's Presidential Candidates, Joeseph Bidon and Donald J. Trump, during the 2020 federal election. Analyzing identity-based, socio-economic, and regional variables, we discovered that … age has minimal impact, higher education correlates with increased Liberal support, and higher income is linked to Conservative favor. … These insights offer a comprehensive understanding of the 2020 election dynamics, enabling more informed predictions for the upcoming 2024 election (potentially with the same candidates). This research contributes to a better grasp of American political affiliations, assisting citizens and policymakers in adapting to evolving electoral trends.

Table of contents

[*] Code and data are available at: https://github.com/sshmuylovich/us-election-2020.git

1 Introduction

The 2020 federal elections saw Democrat Joesph Biden defeat Republican Donald J. Trump, winning 306 of 538 electoral votes, to become the 46th President of the United States ((**citeTODO?**)). With the 47th federal election quickly approaching in 2024 and both Biden and Trump being the presumptive nominees for their respective political parties ((**citeTODO?**)), this paper conducts an analysis on support for these two politicians, specifically examining identity-based, socio-economic, and regional variables. The estimand of interest in this research is the average causal effect of these variables on the likelihood of individuals expressing support for either Democratic Candidate Biden or Republican Candidate Trump.

By delving into identity-based (gender, age, and race), socio-economic (education level, employment status, and income), and regional factors (region, urban status, and state), this paper seeks to illuminate the multifaceted influences that shaped political support in the 2020 US Election. The objective is to unravel the complex dynamics that dictated electoral outcomes, thereby providing a better understanding of the American political landscape and a better ability to forecast potential shifts in voter alignment in anticipation of future elections.

The subsequent sections follow a structured format. Section 2 outlines the source and variables central to our analysis. Section 3 details the construction and methodology of the statistical models used. Section 4 presents the key findings of our analysis, while Section 5 critically reviews the content, addresses the implications of the results, acknowledges model limitations, and suggests potential research directions.

2 Data

The data used in this paper was gathered from the 2020 Cooperative Election Study (CES) hosted on the Harvard Dataverse (Stephenson et al. 2022) and analyzed using R (R Core Team 2023) with help from `tidyverse` (Wickham et al. 2019), `rstanarm` (Goodrich et al. 2022), `modelsummary` (Arel-Bundock 2022), `testthat` (Wickham 2011), `here` (Müller 2020), `knitr` (Xie 2023), and `kableExtra` (Zhu 2021).

2.1 Cooperative Election Study 2020

The Cooperative Election Study (CES), previously known as the Cooperative Congressional Election Study, is an extensive research project aimed at understanding the intricacies of American electoral behavior, particularly focusing on congressional elections. Initiated in 2006, the CES has evolved to capture the nuances of how Americans perceive their representatives, their voting patterns, and the overall electoral experience, emphasizing the influence of political geography and social context. The methodology and scale of this study allow for an in-depth analysis of legislative constituencies across the United States, offering a unique perspective on the democratic process at both state and national levels (Stephenson et al. 2022).

The 2020 CES survey was conducted over the Internet by YouGov, a British international Internet-based market research and data analytics firm. Participants were interviewed from September 29 to November 2, 2020 (for pre-election data), and from November 8 to December 14, 2020 (for post-election data). A total of 61000 participants were interviewed for pre-election and 51551 returned for post-election interviews. The dataset includes 717 variables, many of which could have been included in the analysis but was narrowed down to 9 variables that could correlate with candidate support. These variables are gathered from the CPS portion of the survey with no open-ended answers and assigned numerical values with labels.The data was released on March 26, 2021 (Stephenson et al. 2022).

2.2 Identity-Based Variables

Gender was picked as one of the identity-based variables because gender can influence an individual's policy preferences. For instance, women might prioritize issues such as healthcare, reproductive rights, and gender equality more than men. Understanding gender differences in voting behavior can help in analyzing how these issues impact elections. From a preliminary look at Figure 6, we can see that gender does influence voting behavior, this paper aims to further analysis how.

Age was picked as one of the identity-based variables because different age groups often hold distinct values and perspectives shaped by their generational experiences. For instance, younger voters may prioritize climate change, education, and jobs, while older voters may focus on healthcare, social security, and national security. From a preliminary look at Figure 1, we can see that age influences voting behavior, this paper aims to further analysis how.

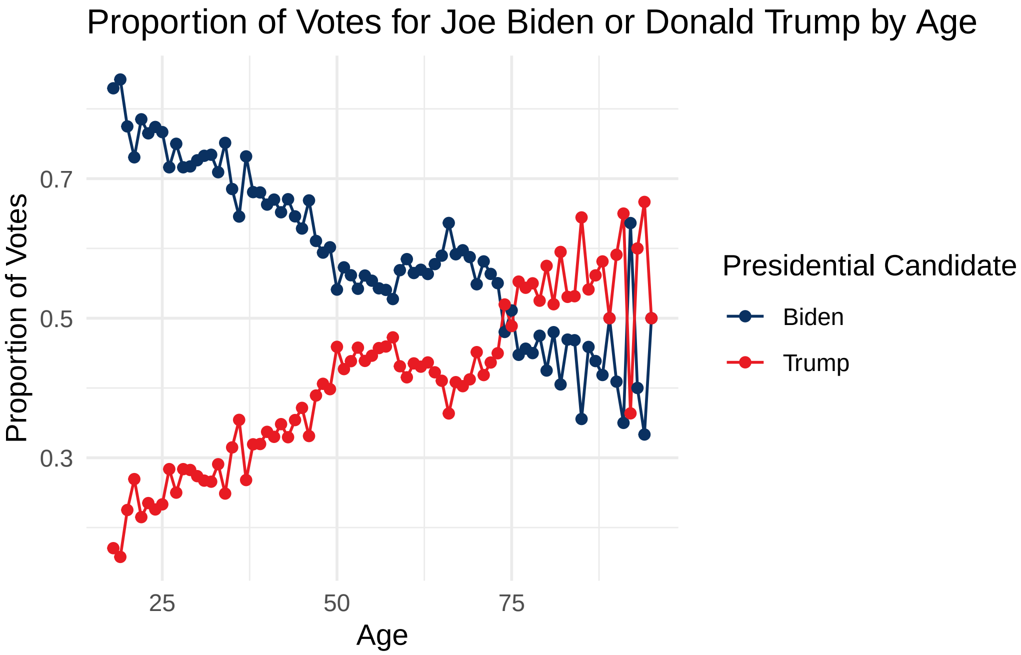

Figure 1: Proportion of Votes by Age in the 2020 Election

Race was picked as one of the identity-based variables because different racial and ethnic groups have unique historical and social experiences that influence their political views and behaviors. For example, policies on immigration, law enforcement, and affirmative action may be viewed differently by voters of different racial backgrounds. The Pew Research Center found that in the 2022 midterms, Black voters supported Democrats by overwhelming margins: 93% voted for Democrats while only 5% supported Republicans. This is similar to levels of support in 2020, 2018 and 2016 (**citePewRace?**). From a preliminary look at Figure 7, we can see that race influences voting behavior, this paper aims to further analysis how.

2.3 Socio-Economic Variables

Education was picked as one of the socio-economic variables because education level is closely linked to an individual's policy preferences and political awareness. Higher education levels often correlate with more liberal attitudes on social issues and a greater engagement in political processes. From a preliminary look at Figure 8, we can see that education influences voting behavior, this paper aims to further analysis how.

Employment status was picked as one of the socio-economic variables because employment status directly affects an individual's economic security and outlook, influencing their priorities at the polls. For example, unemployed or underemployed voters might prioritize job creation, economic recovery, and social safety nets more highly than those securely employed. From a preliminary look at Figure 9, we can see that employment status influences voting behavior, this paper aims to further analysis how.

Income was picked as one of the socio-economic variables because income levels influence voters' economic interests, with higher-income individuals potentially prioritizing tax policies and economic strategies that favor wealth preservation, while lower-income voters might focus on income redistribution, minimum wage increases, and access to affordable healthcare. From a preliminary look at Figure 10, we can see that income influences voting behavior, this paper aims to further analysis how.

2.4 Regional Variables

Region was picked as one of the regional variables because different regions in a country often have unique cultural and historical contexts that shape residents' values, attitudes, and political leanings. For example, historical voting patterns, regional industries, and local issues can significantly influence regional voting behaviors. Furthermore, specific issues may be more pressing in certain regions than others, such as environmental concerns in areas prone to climate-related disasters or economic policies in regions dominated by particular industries. From a preliminary look at Figure 11, we can see that region does influence voting behavior, this paper aims to further analysis how.

Urban status was picked as one of the regional variables because urban, suburban, and rural areas differ markedly in their socio-economic and demographic compositions. These differences can lead to distinct political priorities and voting behaviors, with urban areas often leaning more towards progressive policies and rural areas favoring conservative stances, influenced by factors like population density, diversity, and economic opportunities. The concentration of services and infrastructure in urban areas, as opposed to their scarcity in rural regions, can influence voter concerns and priorities, such as public transportation, education, and healthcare services. From a preliminary look at Figure 12, we can see that age influences voting behavior, this paper aims to further analysis how.

State was picked as one of the regional variables because state-level policies and governance can significantly affect residents' lives, influencing their political preferences. Issues such as education, healthcare, taxation, and environmental regulation can vary widely by state, affecting voting behavior. Some states also have a higher electoral significance due to their size, demographic composition, or status as swing states. Understanding the political dynamics at the state level is crucial for predicting and analyzing election outcomes, especially in systems like the United States' Electoral College. From a preliminary look at Figure 2, we can see that age influences voting behavior, this paper aims to further analysis how.

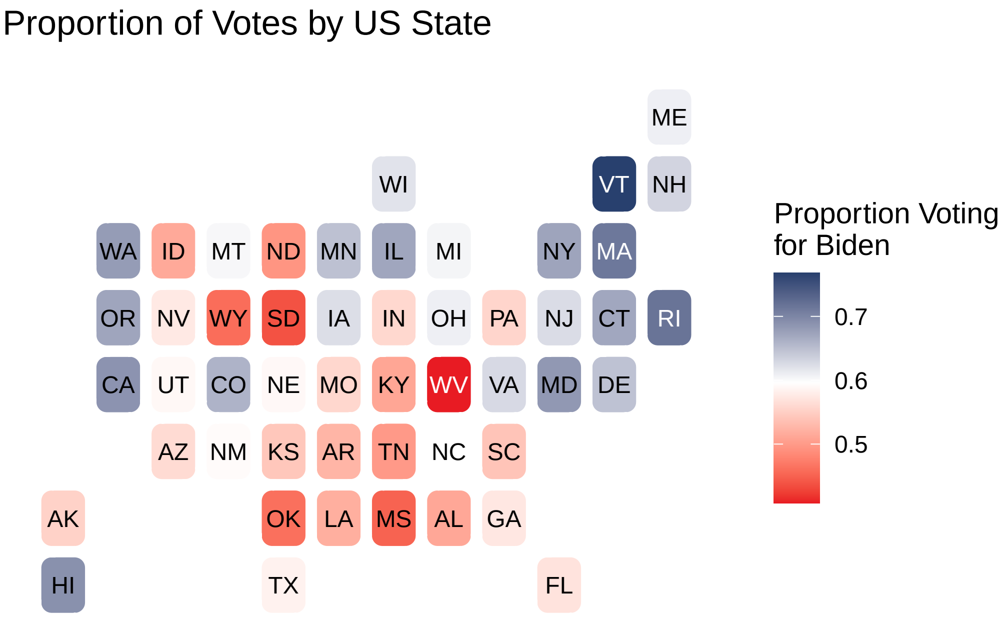

Figure 2: Proportion of Votes by State in the 2020 Election

2.5 Measurment

Age was calculated in years based on the year of birth the respondent inputted. Gender and Race were given a number with a corresponding label based on which radio button the respondent selected. When cleaning the data the following changes were made to the raw data: age was categorized as seen in Table 7, gender was limited to Male and Female, and race was mapped from a number to the categories seen in Table 8.

A preview of the identity-based variables used in this paper can be seen in Table 1.

Table 1: Sample of Identity-Based Data

gender	age_bucket	race
Male	45-64	White
Female	45-64	White
Male	45-64	White
Female	65+	White
Female	45-64	White

Education, employment status, and income were all given a number with a corresponding label based on which radio button the respondent selected. When cleaning the data the following changes were made to the raw data: education was mapped from a number to the categories seen in Table 9, employment status was mapped from a number to the categories seen in Table 10, and income was mapped from a number to the categories seen in Table 11.

A preview of the socio-economic variables used in this paper can be seen in Table 2.

Table 2: Sample of Socio-Economic Data

education	employment_status	income
Some college or assoc. degree	Not in the Workforce	Less than 30,000
College graduate	Not in the Workforce	100,000 - 199,999
Some college or assoc. degree	Employed	30,000 - 49,999
Some college or assoc. degree	Not in the Workforce	Less than 30,000
High school or less	Employed	50,000 - 99,999

Region, urban status, and state were all given a number with a corresponding label based on which radio button the respondent selected. When cleaning the data the following changes were made to the raw data: region was mapped to the categories seen in the CES survey (Northeast, Midwest, South, West), urban status was mapped from a number to the categories seen in the CES survey (City, Suburb, Town, Rural Area), and state was mapped from a number to the categories seen in the CES survey (all 50 states and the District of Columbia).

A preview of the regional variables used in this paper can be seen in Table 3.

Table 3: Sample of Regional Data

region	urban_status	state
Northeast	Suburb	Connecticut
Northeast	Rural Area	Massachusetts
Midwest	Suburb	Ohio
Midwest	Rural Area	South Dakota
Midwest	Rural Area	Ohio

It is important to note that this study is only interested in participants who were registered to vote in the 2020 election and who voted for either Biden or Trump, this was reflectcted in the data cleaning process.

3 Model

For a comprehensive analysis of how identity-based, socio-economic, and regional factors influenced voting behavior in the 2020 US Election, a generalized linear model (GLM) is an effective statistical approach. Given the nature of the dependent variable which is binary (voted for Joe Biden vs. Donald Trump), logistic regression is a suitable model within the GLM framework. This model predicts the log-odds of the outcome as a linear combination of the independent variables.

3.1 Identity-Based Model Setup

Define y_i as who the respondent voted for and equal to 1 if Joe Biden and 0 if Donald Trump. Then gender, age, and race are the respective answers of the respondent.

$$
\begin{aligned}
y_i|\pi_i &\sim \text{Bern}(\pi_i) && (1)\\
\text{logit}(\pi_i) &= \beta_0 + \beta_1 \times \text{gender}_i + \beta_2 \times \text{age}_i + \beta_3 \times \text{race}_i && (2)\\
\beta_0 &\sim \text{Normal}(0, 2.5) && (3)\\
\beta_1 &\sim \text{Normal}(0, 2.5) && (4)\\
\beta_2 &\sim \text{Normal}(0, 2.5) && (5)\\
\beta_3 &\sim \text{Normal}(0, 2.5) && (6)
\end{aligned}
$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of (**rstanarm?**). We use the default priors from `rstanarm`. The use of Normal(0, 2.5) priors is a conservative choice that imposes minimal prior beliefs on the magnitude of the coefficients. It is used in Bayesian logistic regression to reflect a lack of strong prior knowledge while still allowing the data to inform the final posterior distributions.

3.2 Socio-Economic Model Setup

Define y_j as who the respondent voted for and equal to 1 if Joe Biden and 0 if Donald Trump. Then education, employment status, and income are the respective answers of the respondent.

$$
\begin{align}
y_j|\pi_j &\sim \text{Bern}(\pi_j) \tag{7}\\
\text{logit}(\pi_j) &= \beta_0 + \beta_1 \times \text{education}_j + \beta_2 \times \text{eployment status}_j + \beta_3 \times \text{income}_j \tag{8}\\
\beta_0 &\sim \text{Normal}(0, 2.5) \tag{9}\\
\beta_1 &\sim \text{Normal}(0, 2.5) \tag{10}\\
\beta_2 &\sim \text{Normal}(0, 2.5) \tag{11}\\
\beta_3 &\sim \text{Normal}(0, 2.5) \tag{12}
\end{align}
$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of (**rstanarm?**). We use the default priors from `rstanarm`. The use of Normal(0, 2.5) priors is a conservative choice that imposes minimal prior beliefs on the magnitude of the coefficients. It is used in Bayesian logistic regression to reflect a lack of strong prior knowledge while still allowing the data to inform the final posterior distributions.

3.3 Regional Model Setup

Define y_k as who the respondent voted for and equal to 1 if Joe Biden and 0 if Donald Trump. Then region, urban status, and state are the respective answers of the respondent.

$$
\begin{align}
y_k|\pi_k &\sim \text{Bern}(\pi_k) \tag{13}\\
\text{logit}(\pi_k) &= \beta_0 + \beta_1 \times \text{region}_k + \beta_2 \times \text{urban status}_k + \beta_3 \times \text{state}_k \tag{14}\\
\beta_0 &\sim \text{Normal}(0, 2.5) \tag{15}\\
\beta_1 &\sim \text{Normal}(0, 2.5) \tag{16}\\
\beta_2 &\sim \text{Normal}(0, 2.5) \tag{17}\\
\beta_3 &\sim \text{Normal}(0, 2.5) \tag{18}
\end{align}
$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of (**rstanarm?**). We use the default priors from `rstanarm`. The use of Normal(0, 2.5) priors is a conservative choice that imposes minimal prior beliefs on the magnitude of the coefficients. It is used in Bayesian logistic regression to reflect a lack of strong prior knowledge while still allowing the data to inform the final posterior distributions.

4 Results

4.1 Identity-Based Results

Table 4: Explanatory model of support for the 2020 Presidential Candidates based on gender, age, and race

term	estimate	std.error	conf.low	conf.high
(Intercept)	1.89	0.15	1.63	2.15
genderMale	-0.27	0.04	-0.34	-0.20
age_bucket30-44	-0.37	0.09	-0.52	-0.22
age_bucket45-64	-0.94	0.09	-1.08	-0.80
age_bucket65+	-1.03	0.09	-1.18	-0.88
raceBlack	1.70	0.19	1.39	2.02
raceHispanic	-0.24	0.16	-0.52	0.02
raceOther	-1.09	0.26	-1.53	-0.66
raceWhite	-0.77	0.14	-1.01	-0.54

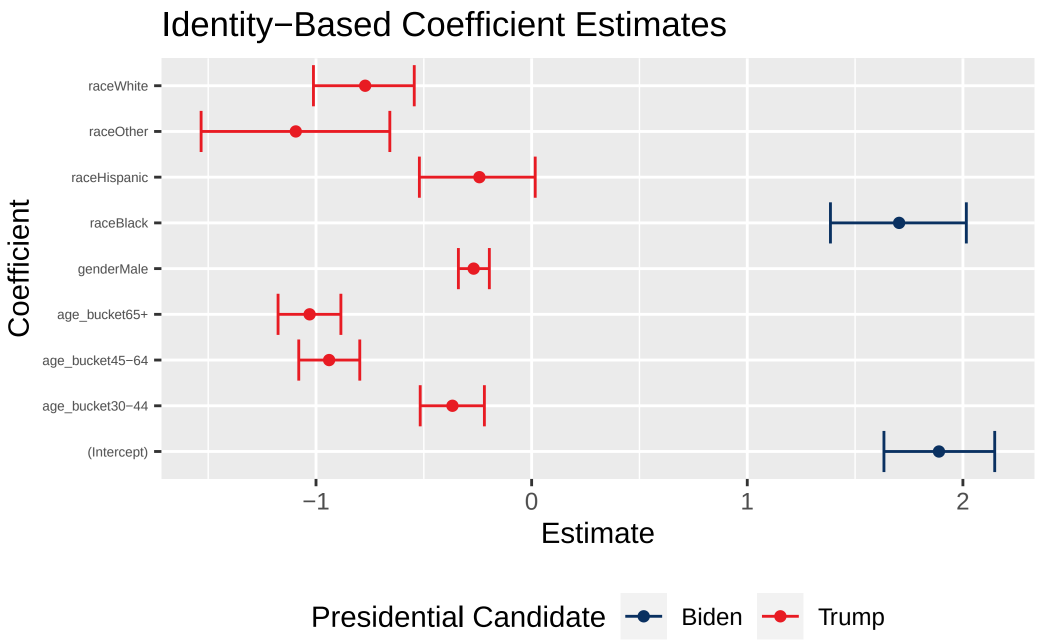

Figure 3: Explanatory graph of support for the 2020 Presidential Candidates based on gender, age, and race

4.2 Socio-Economic Results

Table 5: Explanatory model of support for the 2020 Presidential Candidates based on education, employment status, and income

term	estimate	std.error	conf.low	conf.high
(Intercept)	0.66	0.06	0.57	0.76
educationHigh school or less	-1.09	0.06	-1.20	-0.99
educationPostgraduate study	0.25	0.07	0.14	0.37
educationSome college or assoc. degree	-0.49	0.06	-0.58	-0.40
employment_statusNot in the Workforce	-0.09	0.05	-0.16	-0.01
employment_statusUnemployed	0.41	0.08	0.27	0.55
income200,000 or more	0.04	0.11	-0.13	0.23
income30,000 - 49,999	0.28	0.07	0.17	0.40
income50,000 - 99,999	0.11	0.06	0.01	0.21
incomeLess than 30,000	0.56	0.07	0.45	0.68

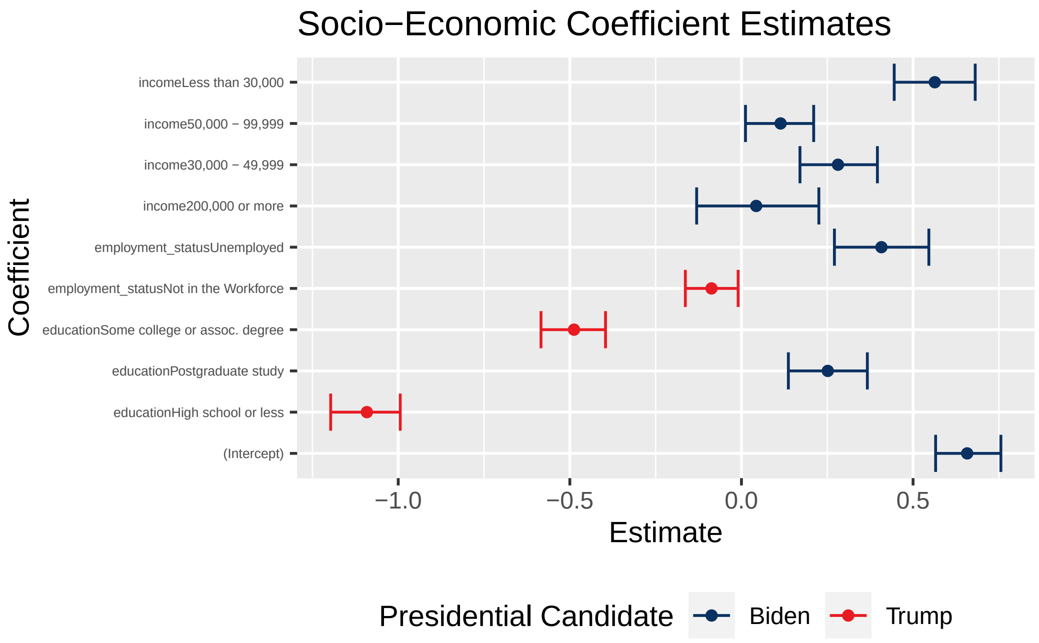

Figure 4: Explanatory graph of support for the 2020 Presidential Candidates based on education, employment status, and income

Table 6: Explanatory model of support for the 2020 Presidential Candidates based on region, urban status, and state

term	estimate	std.error	conf.low	conf.high
(Intercept)	1.11	2.96	-4.06	5.85
regionNortheast	0.04	4.46	-7.13	7.70
regionSouth	-0.41	2.97	-5.09	4.75
regionWest	0.05	4.50	-7.09	7.60
urban_statusRural Area	-1.36	0.07	-1.48	-1.25
urban_statusSuburb	-0.60	0.06	-0.70	-0.51
urban_statusTown	-0.85	0.07	-0.97	-0.73
stateAlaska	-0.65	4.40	-7.61	6.59
stateArizona	-0.21	4.31	-7.05	6.96
stateArkansas	0.12	0.29	-0.37	0.60
stateCalifornia	0.02	4.34	-6.84	7.18
stateColorado	-0.03	4.30	-6.86	7.15
stateConnecticut	0.28	4.47	-7.08	7.60
stateDelaware	0.44	0.34	-0.12	1.00
stateDistrict of Columbia	2.10	0.80	0.95	3.75
stateFlorida	0.25	0.19	-0.06	0.55
stateGeorgia	0.23	0.21	-0.11	0.57
stateHawaii	0.40	4.32	-6.65	7.65
stateIdaho	-0.39	4.31	-7.23	6.85
stateIllinois	0.14	2.98	-4.61	5.34
stateIndiana	-0.12	2.96	-4.86	5.03
stateIowa	0.18	3.03	-4.56	5.39
stateKansas	-0.80	2.99	-5.58	4.35
stateKentucky	-0.01	0.23	-0.39	0.36
stateLouisiana	0.09	0.26	-0.34	0.54
stateMaine	0.29	4.53	-6.95	7.60
stateMaryland	0.61	0.24	0.22	0.99
stateMassachusetts	0.42	4.51	-6.92	7.73
stateMichigan	0.17	2.96	-4.56	5.38
stateMinnesota	0.36	3.00	-4.37	5.56
stateMississippi	-0.01	0.31	-0.53	0.49
stateMissouri	-0.13	2.99	-4.84	4.98
stateMontana	0.31	4.32	-6.73	7.44
stateNebraska	0.14	3.03	-4.62	5.37
stateNevada	-0.37	4.31	-7.33	6.74
stateNew Hampshire	-0.11	4.50	-7.32	7.18
stateNew Jersey	0.05	4.50	-7.24	7.33
stateNew Mexico	-0.15	4.30	-7.07	7.05

term	estimate	std.error	conf.low	conf.high
stateNew York	0.21	4.52	-7.12	7.52
stateNorth Carolina	0.55	0.20	0.21	0.89
stateNorth Dakota	-0.23	3.09	-5.08	4.96
stateOhio	-0.03	3.03	-4.78	5.12
stateOklahoma	-0.28	0.30	-0.76	0.21
stateOregon	0.62	4.35	-6.17	7.81
statePennsylvania	-0.22	4.50	-7.51	7.06
stateRhode Island	0.10	4.54	-7.10	7.48
stateSouth Carolina	0.26	0.23	-0.12	0.62
stateSouth Dakota	0.03	3.02	-4.81	5.27
stateTennessee	-0.01	0.21	-0.37	0.33
stateTexas	0.28	0.19	-0.03	0.59
stateUtah	0.10	4.30	-6.86	7.33
stateVermont	1.32	4.60	-6.07	8.46
stateVirginia	0.37	0.21	0.03	0.71
stateWashington	0.27	4.33	-6.58	7.40
stateWest Virginia	0.01	0.30	-0.49	0.51
stateWisconsin	0.04	3.00	-4.68	5.18
stateWyoming	-0.45	4.34	-7.34	6.75

4.3 Regional Results

5 Discussion

6 Appendix

Table 7: Age bucket variable values

age_bucket
45-64
65+
30-44
18-29
NA

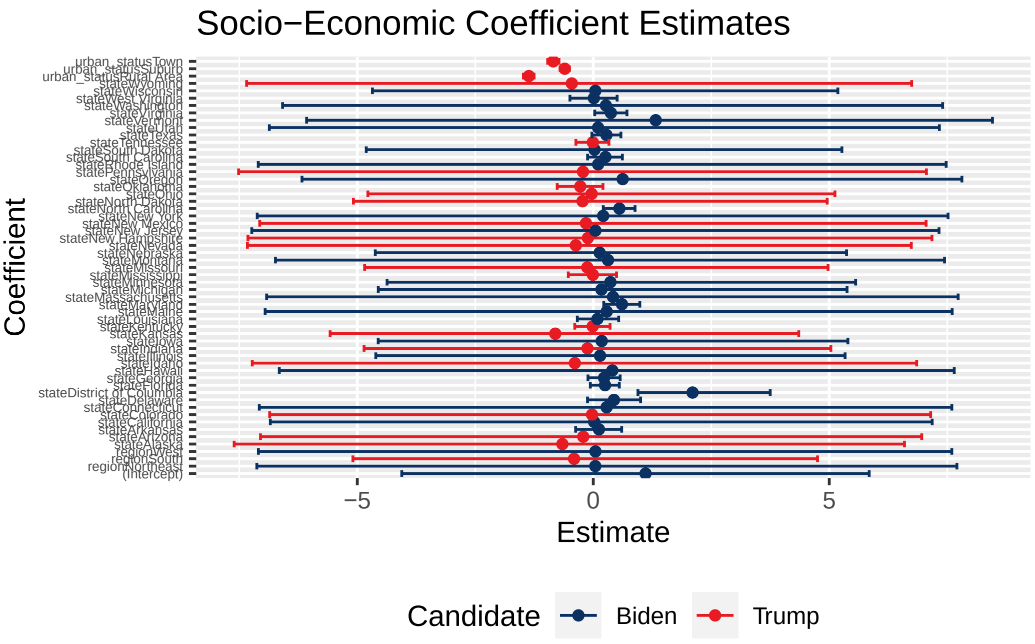

Figure 5: Explanatory graph of support for the 2020 Presidential Candidates based on region, urban status, and state

Table 8: Race variable values

race
White
Black
Hispanic
Asian
Other

Table 9: Education variable values

education
Some college or assoc. degree
College graduate
High school or less
Postgraduate study

Table 10: Eployment status variable values

emoloyment_status
Not in the Workforce
Employed
Unemployed

Table 11: Income variable values

income
Less than 30,000
100,000 - 199,999
30,000 - 49,999
50,000 - 99,999
200,000 or more

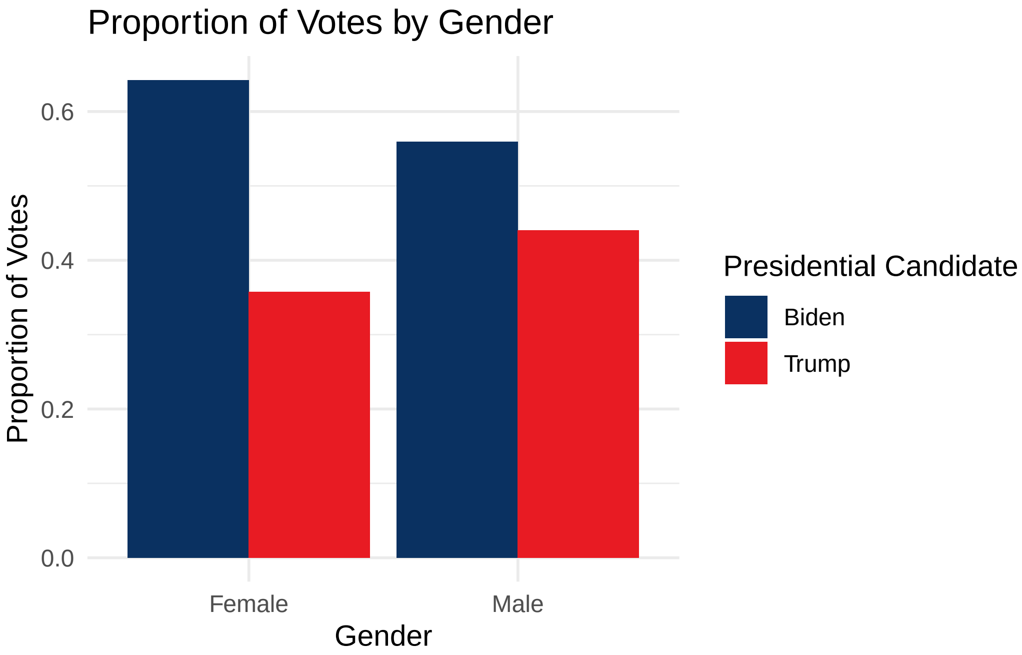

Figure 6: Proportion of Votes by Gender in the 2020 Election

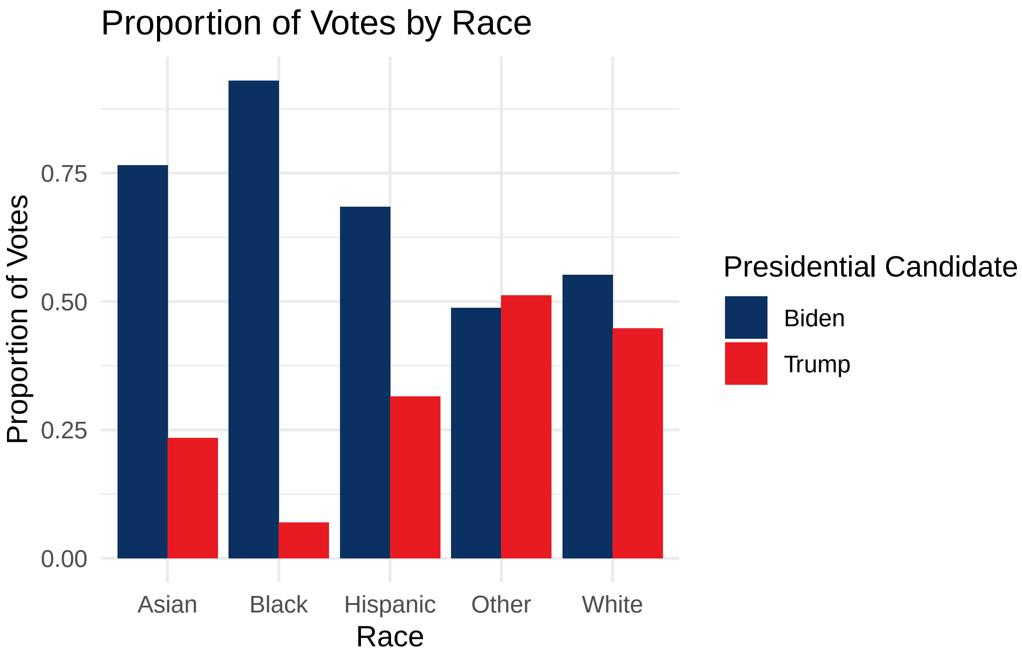

Figure 7: Proportion of Votes by Race in the 2020 Election

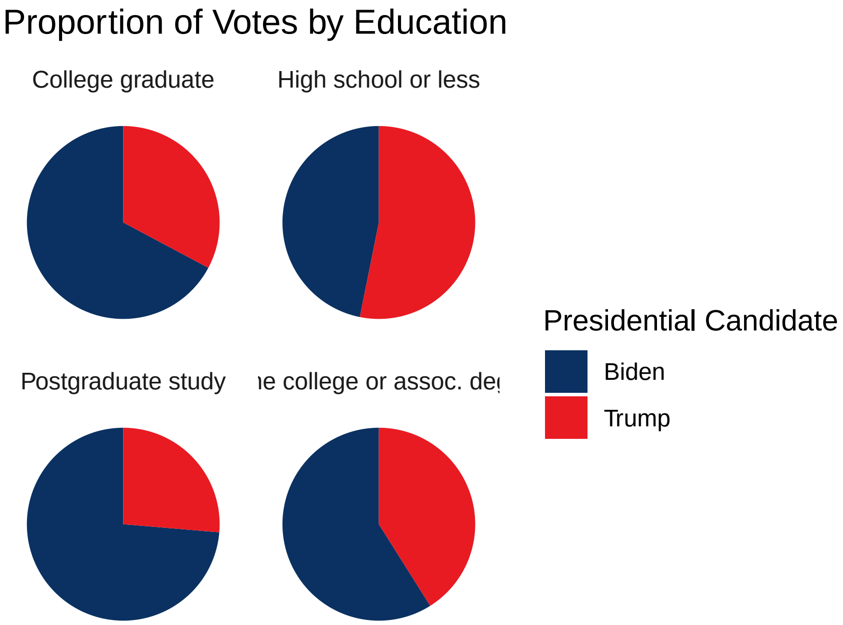

Figure 8: Proportion of Votes by Education in the 2020 Election

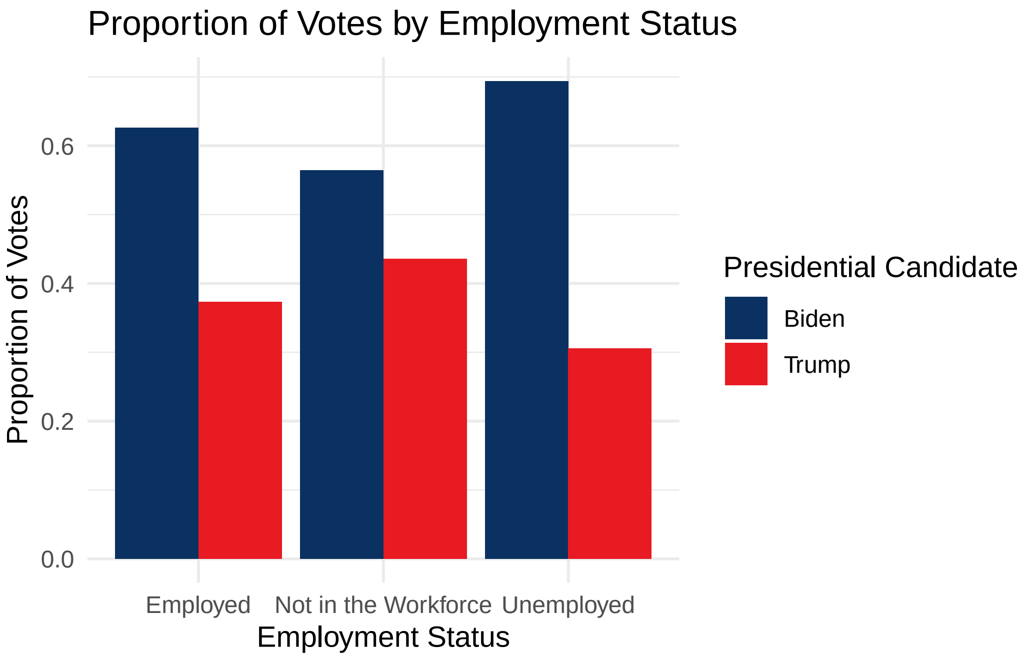

Figure 9: Proportion of Votes by Employment Status in the 2020 Election

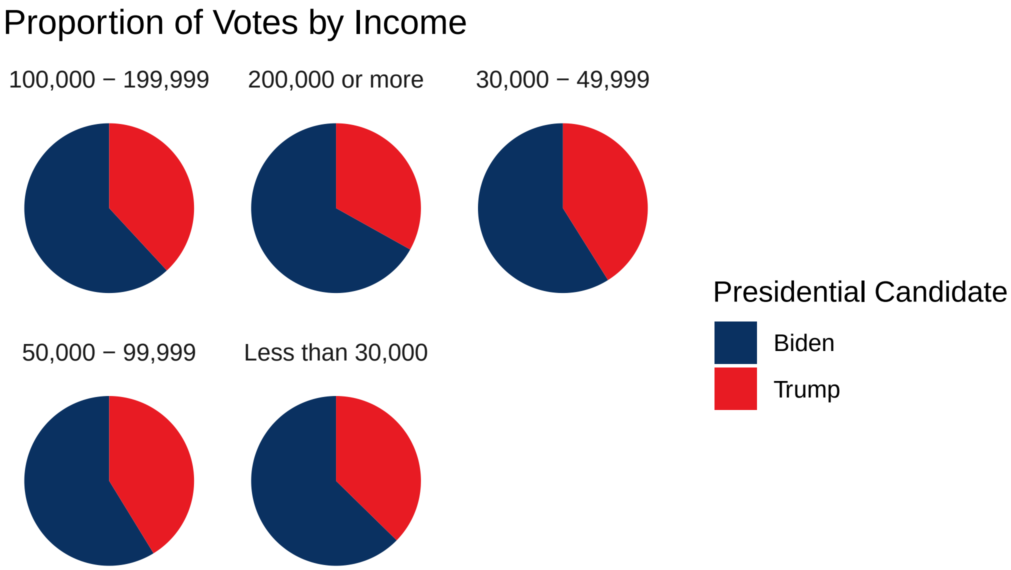

Figure 10: Proportion of Votes by Income in the 2020 Election

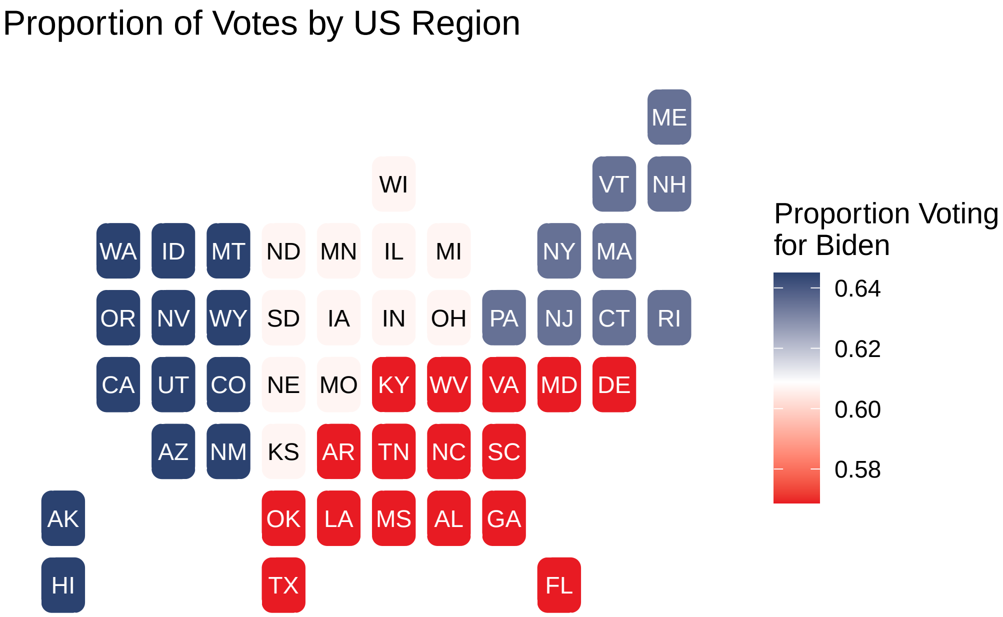

Figure 11: Proportion of Votes by Region in the 2020 Election

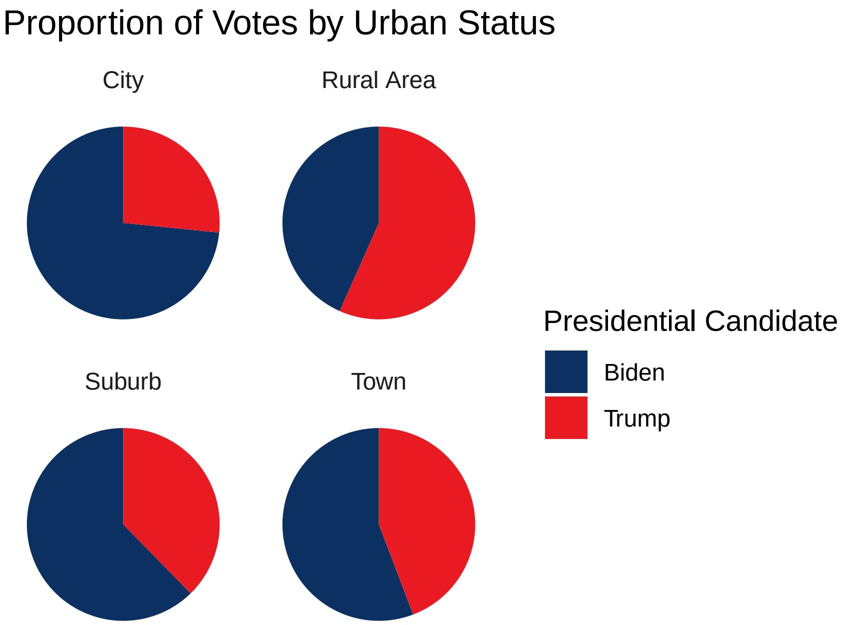

Figure 12: Proportion of Votes by Urban Status in the 2020 Election

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

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

Stephenson, Laura B, Allison Harell, Daniel Rubenson, and Peter John Loewen. 2022. "2021 Canadian Election Study (CES)." Harvard Dataverse. https://doi.org/10.7910/DVN/XBZHKC.

Wickham, Hadley. 2011. "Testthat: Get Started with Testing." *The R Journal* 3: 5–10. https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf.

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. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.