Lower tax brackets and employment levels, and people living in populous areas are more likely to vote for Biden*

Victor Ma

December 3, 2024

In this study, I investigate demographic patterns that influence political support during the 2020 U.S. Presidential Election, focusing on whether socio-economic factors like employment status, income level, and living environment influence voters' likelihood to support Joe Biden or Donald Trump. I employ a logistic regression model to analyze these predictors and identify trends within each demographic category. The results show that urban residents, lower-income groups, and individuals in specific employment categories (e.g., students and retirees) are more likely to support Joe Biden. These findings align with broader trends in U.S. politics, where socio-economic conditions and regional contexts play pivotal roles in shaping political preferences.

1 Introduction

The 2020 United States presidential election highlighted key societal shifts amidst crises. Beyond the COVID-19 pandemic, which reshaped healthcare and exposed systemic vulnerabilities, the election underscored debates about economic inequality as unemployment surged (Blustein 2020; Labor Statistics 2020). Climate change became a priority for many voters, driven by the growing frequency of natural disasters (Climate Change 2020).

Demographic patterns also shaped the election. Younger, more diverse voters emphasized generational and racial divides in priorities (Frey 2020). Suburban voting patterns, particularly among college-educated women, had shifts in key swing areas (Cook 2020). The widespread adoption of mail-in voting during the pandemic increased accessibility but fueled political and legal debates over election integrity (State Legislatures 2020). The election also tested

*Code and data are available at: https://github.com/bestmustard/political_support

democratic institutions against misinformation and polarization (Ben-Gurion 2020; Sunstein 2020).

Living environments strongly influenced political alignment. Urban residents predominantly supported Joe Biden, reflecting progressive policy preferences often found in cities (Bishop 2019). Conversely, rural voters favored Donald Trump, consistent with support for conservative values (Cramer 2016a).

Employment status was another factor. Data from the Cooperative Congressional Election Study (CCES, 2020) indicated that students and retirees leaned toward Biden, likely due to policy interests in education and healthcare. Full-time workers exhibited mixed preferences influenced by industry and region (B. F. Schaffner, Ansolabehere, and Luks 2020).

Income disparities also defined political divides. Lower-income groups largely supported Biden, aligning with Democratic policies on healthcare and social safety nets, while higher-income voters leaned Republican, especially in rural and suburban settings, reflecting preferences for tax policies benefiting their financial interests (Center 2020a).

This study uses a logistic regression model to predict voting preferences based on living environment, employment status, and income levels. Logistic regression is suited for binary outcomes, providing probabilities for voter alignment with each candidate.

The primary dataset is the 2020 Cooperative Congressional Election Study (CCES), a stratified survey conducted by YouGov and sourced from the Harvard Dataverse (B. F. Schaffner, Ansolabehere, and Luks 2020). Supporting data includes Pew Research Center's 2020 voter analysis and the Census Bureau's Current Population Survey Voting and Registration Supplement, which validate and contextualize findings (Center 2020b; Bureau 2020).

The report includes five sections and an appendix. After the introduction, the first section describes the CCES dataset, highlighting demographic attributes and distributions. The second section explains the logistic regression model and presents preliminary findings. The third section analyzes the influence of living environment, employment status, and income on election outcomes using visualizations and numerical results. The fourth section discusses the implications, limitations, and directions for future research.

The appendix explores survey methodologies, focusing on sampling and observational data in the CCES dataset. It compares methodologies, discusses sampling variability and bias, and uses simulations to expand on points in the Data section.

This analysis was done in R, with the packages tidyverse for data manipulation, dataverse for dataset access, knitr for report generation, modelsummary for model interpretation, and rstanarm for Bayesian regression (R Core Team 2021; Leeper 2021; Xie 2021; Arel-Bundock 2021; Goodrich et al. 2022; Kay 2021; Wickham et al. 2021). Plots and graphs were created with ggplot, and portions of the discussion, data, and appendix were assisted by ChatGPT4 (OpenAI 2023).

2 Data

The dataset for this analysis is the 2020 Cooperative Election Study (CCES) (B. Schaffner, Ansolabehere, and Luks 2021), accessed via the Harvard Dataverse. The 2020 iteration surveyed 61,000 respondents, capturing a broad range of U.S. political opinions. The release includes detailed documentation and questionnaires, ensuring transparency and replicability. Vote validation was conducted by Catalist, a large-scale organization maintaining data on over 240 million U.S. individuals (Catalist 2017).

2.1 Strengths

2.1.1 Sample Size

Sources for Figure 1 below: (American National Election Studies 2021),(Center 2020c),(Research 2020), (B. Schaffner, Ansolabehere, and Luks 2021).

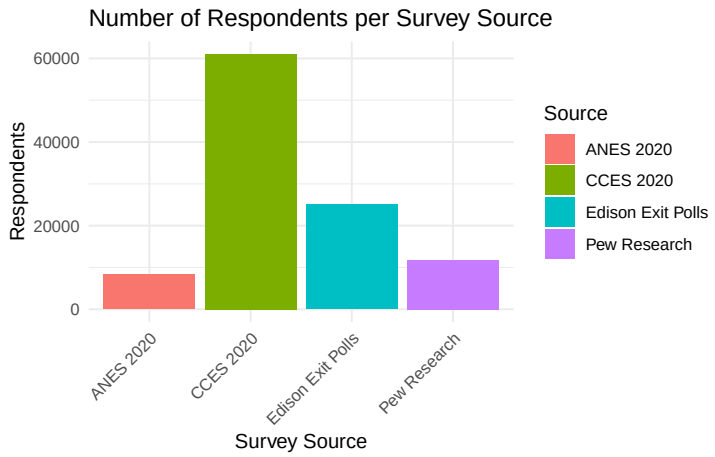

Figure 1: Number of respondents for various 2020 election survey sources

Figure 1 shows that CCES poll features the largest respondent pool compared to other surveys. This extensive sample size enables granular analyses, ensuring the representation of smaller subgroups that may be underrepresented in smaller datasets (Ansolabehere and Schaffner 2020).

2.1.2 Data Collection Phases

The CCES collected data in two waves, immediately before and after the 2020 election. This helps capture the shifts in voter attitudes and preferences during the final stages of the election

campaign (B. Schaffner, Ansolabehere, and Luks 2021). Pew Research and Gallup each conduct multiple polls of much smaller scale leading up to the election, which will give a snapshot in time of the voter sentiment but does not capture a longitudinal perspective similarly to the CES survey (Center 2020c).

2.1.3 Sampling Methodology

The CCES employs a two-step sample matching methodology designed to create a representative survey sample:

2.1.3.1 Target Sample Creation

A target sample is generated to reflect the broader population based on key demographic variables such as age, gender, race, and educational attainment. This target sample serves as a benchmark for representativeness.

2.1.3.2 Matched Sample Selection

For each individual in the target sample, a closely matched respondent is selected from a pool of survey volunteers maintained by YouGov. Matches are determined using a combination of publicly available databases and YouGov's proprietary data, ensuring close alignment on variables such as geography, party registration, and voter history.

Once the matched sample is established, it is weighted to adjust for discrepancies between the sample and the target population. The weighting process involves two stages:

2.1.3.3 Demographic Weighting

Adjustments are made using data from the American Community Survey to ensure accurate representation across demographics.

2.1.3.4 Voting Behavior Weighting

Further adjustments incorporate validated voter registration data, accounting for turnout and voting patterns at the state and national levels.

Validation of the sampling methodology is conducted by comparing state-level survey results with official election outcomes. This comparison ensures that survey estimates align with actual voter behavior, providing confidence in the dataset's representativeness.

The sampling approach is specifically designed for online opt-in panels and mitigates biases commonly associated with such surveys, such as overrepresentation of politically engaged respondents. Unlike traditional random or quota sampling, this methodology emphasizes precision in matching and post-stratification to minimize systematic error (B. Schaffner, Ansolabehere, and Luks 2021), (Rivers 2007).

2.2 Limitations

The data has some limitations. An error affected 925 North Carolina respondents who were shown incorrect House race candidates, impacting analyses of these districts. Additionally, self-reported data can introduce biases such as recall or social desirability bias. Pre-election surveys may also fail to capture last-minute shifts in voter sentiment.

The online nature of the survey, hosted by YouGov, expands accessibility but may underrepresent groups with limited internet access or technological literacy. Furthermore, the lack of an interviewer in online surveys could impact response quality (B. Schaffner, Ansolabehere, and Luks 2021).

2.3 Variables of Interest

The selection of living environment, employment status, and income level as predictor variables for analyzing political preferences in the 2020 CCES data is guided by prior research that shows their influence on voting behavior.

2.3.1 Urban vs. Rural Residency

Urban and rural divides are well-documented predictors of political preferences. Urban voters have consistently leaned Democratic, driven by higher population density and exposure to diverse cultural and socioeconomic dynamics. For example, Pew Research found that 62% of urban voters supported Hillary Clinton in the 2016 election, compared to only 35% in rural areas (Center 2018). Rural voters, in contrast, exhibit stronger support for Republican candidates, influenced by traditional values and economic concerns rooted in agriculture and resource-based industries (Cramer 2016b).

Research also shows that suburban areas, often politically contested, have shifted in recent years. The Cook Political Report highlights a notable swing in suburban support toward Democratic candidates in 2020, attributed to changes in demographics and education levels among suburban voters (Cook 2020).

2.3.2 Income Level

Income levels are closely tied to political preferences, with higher-income individuals generally favoring Republican candidates due to tax policies, while lower-income groups often lean Democratic, prioritizing social welfare and redistribution policies (Frank 2004). For example, the Census Bureau's Current Population Survey shows that in 2020, households earning under $50,000 were more likely to support Joe Biden, while those earning over $100,000 favored Donald Trump (Bureau 2020).

However, this trend is affected by education levels and geographic factors. High-income earners in urban areas often prioritize social liberalism and climate change policies, aligning with Democratic platforms, while rural high-income earners focus more on fiscal conservatism (Center 2020c; Edsall 2020).

2.3.3 Employment Level

Employment status significantly impacts voter preferences, with distinct patterns emerging across different occupational categories. Full-time workers are often divided along industry lines, with white-collar employees tending to support Democratic candidates and blue-collar workers aligning more with Republican candidates (Muro and Maxim 2020). Unemployment during economic crises, such as the COVID-19 pandemic, has further shaped voting patterns. Studies show that unemployed individuals are more likely to support candidates promising expansive social safety nets and job creation (Blustein 2020; Fowler 2020).

Retirees also exhibit unique voting behaviors, often prioritizing stability and healthcare, leading to a higher likelihood of Republican support (Foundation 2020). In contrast, students, who are less economically established and more progressive, tend to favor Democratic candidates (Frey 2020).

2.4 Data Preparation and Cleaning

The 2020 CCES data, as recorded by Schaffner, Brian et al., was obtained from the Harvard Dataverse and processed using the arrow package for efficient storage in parquet format (Richardson et al. 2024). The raw data was imported from a CSV file, and the cleaning process involved filtering for registered voters who cast votes for either Joe Biden or Donald Trump, treating the presidential vote as a binary outcome.

Key variables were transformed for clarity and analysis:

Living Environment: Categorized into "City," "Suburb," "Town," and "Rural area" based on the urbancity variable. Employment Status: Labeled as "Full-time," "Part-time," "Temporarily laid off," "Unemployed," "Retired," "Permanently disabled," "Homemaker," or "Student." Income Levels: Grouped into ranges: "$<$ 10k," "10-50k," "50-100k," "100-200k," "200-500k,"

and "> 500k." These transformations were implemented using the dplyr package for data manipulation. Each variable was converted into a factor with ordered levels to facilitate analysis and visualization.

The cleaned dataset was saved in both CSV and parquet formats for ease of use in subsequent analyses. Distributions for key explanatory variables, such as living environment, employment status, and income, are visualized in Figure 2, Figure 3, and Figure 4 below.

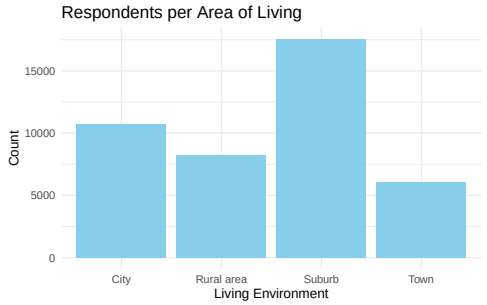

Figure 2: Number of respondents by area of living

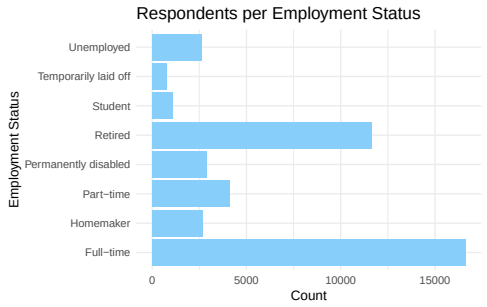

Figure 3: Number of respondents by employment status

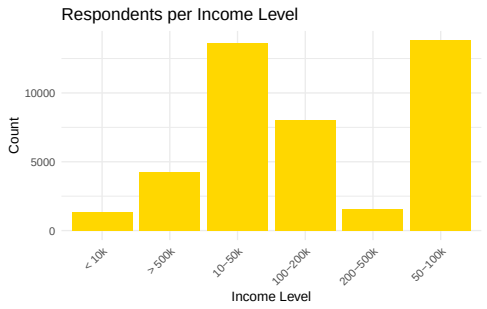

Figure 4: Number of respondents by family income level

Figure 2 shows that the most common respondents were living in the suburbs and the city.

In Figure 3, the most common employment levels were full-time employees or retired.

Figure 4 shows that the most common income levels were between the 10-100k range, and there was a high number of participants with an income level above $500,000.

Figure 5: Income vs. Employment Status

Figure 5 shows us the income distribution across employment types.

Full-time workers exhibit the widest income distribution, spanning from "<10k" to ">500k." Retirees and students predominantly occupy lower income levels, aligning with expectations based on fixed or limited income sources.

Retirees and students show a strong preference for Biden, while higher-income full-time workers lean slightly toward Trump. Homemakers and permanently disabled individuals display narrower income ranges but include both Biden and Trump voters.

In the ">500k" category, Trump garners more support, consistent with Republican tax policies favoring high-income earners. These patterns underscore the relationships between employment, income, and political preferences, highlighting the diversity of economic profiles within voter bases.

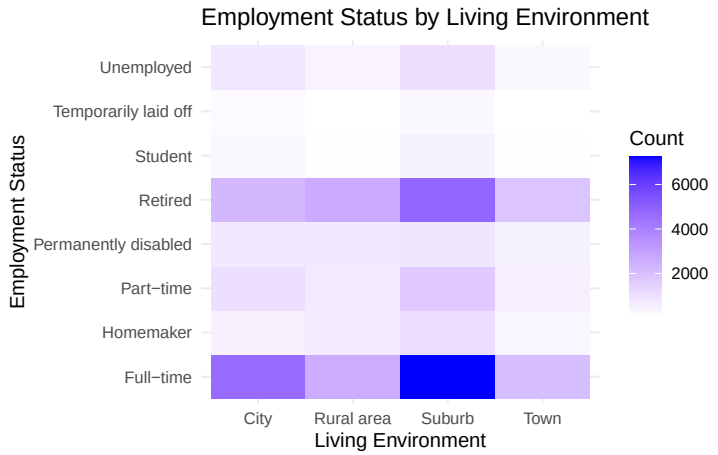

Figure 6: Heatmap of employment status and living environment

The Figure 6 shows the distribution of employment statuses across different living environments:

Suburbs have the highest concentration of full-time workers, as indicated by the deep blue color. This aligns with the demographic trends of suburban areas, which often attract working professionals due to proximity to urban centers and family-friendly amenities.

Retirees are present in all living environments, with notable concentrations in both suburbs and rural areas. This reflects retirement trends where individuals seek quieter or more affordable living conditions outside cities.

Rural areas show a higher proportion of homemakers compared to other environments, consistent with traditional family roles in less urbanized areas.

Cities have the largest concentration of students, likely due to the presence of educational institutions and urban amenities.

Rural areas show higher proportions of unemployed and permanently disabled individuals compared to cities and suburbs, highlighting economic disparities and limited job opportunities in these regions.

Part-time work and temporary layoffs are distributed fairly evenly across living environments, suggesting they are less influenced by geographic location.

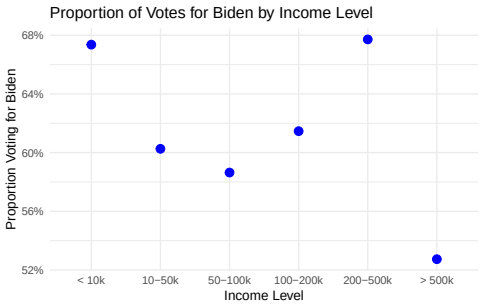

Figure 7: Votes for Biden at each level of income

Figure 7 shows that income level does not seem to have a correlation with voting for Biden.

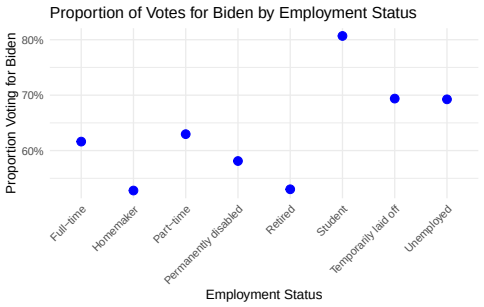

Figure 8: Votes for Biden at each employment level

Figure 8 shows that the greatest number of voters for Biden were Students, and it seems to decrease with higher employment levels. However, all categories still favor Biden.

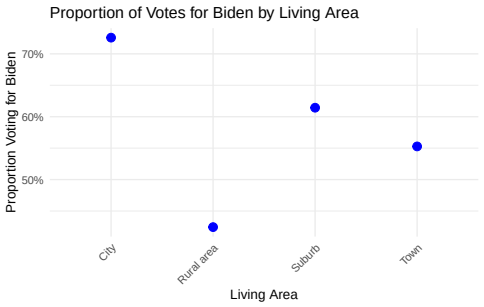

Figure 9: Votes for Biden based on residency

Figure 9 shows a direct correlation between the density of peoples' living area and their probability of voting for Biden. People in the city had a high chance of voting Biden, while rural is much lower.

3 Model

This report presents a logistic regression model to predict whether a respondent voted for Joe Biden (1) or Donald Trump (0), based on their employment status, income level, and living environment. The analysis uses Bayesian logistic regression implemented in R with the rstanarm package. The model's performance is evaluated through various validation techniques, including train-test splits, posterior predictive checks, and sensitivity analyses.

For this analysis, the predictors are employment status, income level, and living environment. These variables are passed through the logistic function to estimate $P(Y_i = 1)$, the probability that respondent i votes for Joe Biden.

$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

3.1 Implementation

The model is implemented using R, using the rstanarm package for Bayesian logistic regression. The data was preprocessed using the tidyverse package. Model outputs and intermediate steps are saved in .rds format for reproducibility Goodrich et al. (2022). The model was implemented using the stan_glm() function, which applies Bayesian logistic regression with default priors and estimates the β coefficients for the predictors. The model was fit to a subset of 3000 respondents from the 2020 CCES data processed earlier.

3.2 Model Specification

The logistic function produces an S-shaped sigmoid curve, which asymptotically approaches 1 as t increases and 0 as t decreases. In logistic regression, the input t is modeled as a linear combination of predictors (including an intercept), expressed as $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$. This transforms the predictors into probabilities of the binary outcome.

The model is defined as:

$$\log\left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right) = \beta_0 + \beta_1 X_{\text{employment},i} + \beta_2 X_{\text{income},i} + \beta_3 X_{\text{living},i}$$

Y_i: Binary response indicating vote for Biden (1) or Trump (0).

$X_{\text{employment},i}$: Employment status.

$X_{\text{income},i}$: Income level.

$X_{\text{living},i}$: Living environment.

β_0: Intercept term.

β_1, β_2, β_3: Coefficients for predictors.

The log-odds are transformed into probabilities using the logistic function:

$$P(Y_i = 1) = \frac{1}{1 + e^{-t}}$$

where $t = \beta_0 + \beta_1 X_{\text{employment},i} + \beta_2 X_{\text{income},i} + \beta_3 X_{\text{living},i}$.

3.3 Justification

Employment, income, and living environment are socio-economic factors that significantly influence voting preferences. These variables are categorical, making logistic regression suitable for analyzing their relationship with a binary outcome. The model avoids unnecessary complexity by focusing on these key variables while maintaining interpretability. Priors ($\beta_0, \beta_k \sim \text{Normal}(0, 2.5)$) are weakly informative, ensuring coefficients are regularized without constraining the model (Gelman et al. 2008).

The dataset is split into 70% training and 30% test sets to validate out-of-sample performance. This ensures the model generalizes well to unseen data. The ROC (Receiver Operating Characteristic) curve is a graphical representation of the model's ability to distinguish between the two classes: votes for Biden (1) and Trump (0). It plots the true positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold settings. The AUC (Area Under the Curve) quantifies the overall ability of the model to discriminate between the classes. An AUC of 1.0 indicates perfect discrimination, while 0.5 indicates no better performance than random guessing.

Posterior predictive checks ensure the model fits the data well by comparing simulated outcomes to observed data. This helps evaluate whether the model appropriately represents the underlying data distribution. The residual histogram shows the distribution of residuals, representing the differences between observed and predicted values. The residuals are distributed symmetrically around zero, which suggests that the model does not exhibit significant bias in its predictions. The histogram shows two peaks, one slightly negative and the other slightly positive. This may indicate that the model performs differently for different subsets of the data (e.g., voters for Biden vs. Trump). This could reflect the nature of binary logistic regression, where predictions are pushed toward 0 or 1. The spread of residuals is relatively narrow, with most values falling between -0.5 and 0.5. This suggests that the model's predictions are not overly far from the observed values.

Underlying assumptions include the linearity of predictors on the log-odds scale, independence of observations, and the assumption that predictors adequately capture variability in the data. Limitations include the categorical nature of the predictors, which may lose granularity, and the potential omission of confounders such as race or education level. The model may be less

appropriate in cases where relationships are non-linear or involve strong interaction effects. Alternatives considered include random forests, which are more flexible but lack interpretability, and multinomial logistic regression, which was unnecessary for this binary outcome. Logistic regression was chosen for its balance of simplicity, interpretability, and compatibility with the data structure.

The validation process includes a train-test split to ensure generalization to new data, ROC Curve and AUC to demonstrate the model's discrimination ability with a moderate AUC value, posterior predictive checks to validate that simulated outcomes align well with observed data, and residual analysis to confirm model assumptions and identify any misfit. These steps confirm the reliability of the logistic regression model for this use case.

4 Results

The resuls of this analysis show that employment, income, and living environment significantly influence political alignment. The coefficients in Table 1 display the impact of each predictor on the likelihood of voting for Joe Biden versus Donald Trump.

Table 1: Coefficient Summary of the Logistic Regression Model

Parameter	Mean	Standard Error	Lower 95% CI	Upper 95% CI
(Intercept)	1.5868977	0.2596937	1.1775485	2.0050547
employmentPart-time	0.1314708	0.1490950	-0.1130782	0.3680669
employmentTemporarily laid off	-0.3726932	0.2855621	-0.8419016	0.1252131
employmentUnemployed	0.1810494	0.1767090	-0.1076623	0.4723113
employmentRetired	-0.3249370	0.0962067	-0.4800769	-0.1674963
employmentPermanently disabled	-0.2337524	0.1617082	-0.4904858	0.0360442
employmentHomemaker	-0.3727040	0.1642102	-0.6369330	-0.1005545
employmentStudent	1.0542835	0.3341135	0.5419834	1.6214754
income10-50k	-0.4901739	0.2461431	-0.8998879	-0.0983298
income50-100k	-0.6292433	0.2459658	-1.0361676	-0.2177944
income100-200k	-0.4259639	0.2581398	-0.8528164	-0.0064363
income200-500k	-0.2451965	0.3096404	-0.7661263	0.2692577
income> 500k	-0.8183073	0.2603393	-1.2588726	-0.3891936
livingSuburb	-0.3561071	0.1034594	-0.5295785	-0.1913916
livingTown	-0.7947135	0.1274247	-1.0093239	-0.5942966
livingRural area	-1.1800838	0.1154117	-1.3722552	-0.9858830

The intercept shows the baseline log-odds of voting for Biden when all predictors are at their reference level (e.g., baseline employment group, income group, and living environment).

4.0.1 Employment

Positive coefficients (e.g., employmentStudent) indicate that being in this group increases the likelihood of voting for Biden compared to the reference group (employmentFull-time).

Negative coefficients (e.g., employmentTemporarily laid off) suggest a decreased likelihood of voting for Biden compared to the reference group.

4.0.2 Income

Lower-income groups (e.g., income10-50k) have negative coefficients, suggesting these groups are less likely to vote for Biden compared to the reference category (income <10k).

Higher-income groups (e.g., income>500k) also show decreased likelihood of voting for Biden, suggesting a possible U-shaped relationship between income and voting preference.

4.0.3 Living Environment

Urban areas (livingCity, as reference) show the highest likelihood of voting for Biden. Suburbs and rural areas (livingSuburb, livingRural area) exhibit progressively stronger negative coefficients, indicating lower support for Biden.

4.0.4 Confidence Intervals

The columns Lower 95% CI and Upper 95% CI show the range within which the true coefficient is likely to fall with 95% confidence. If this range does not include zero, the effect is statistically significant.

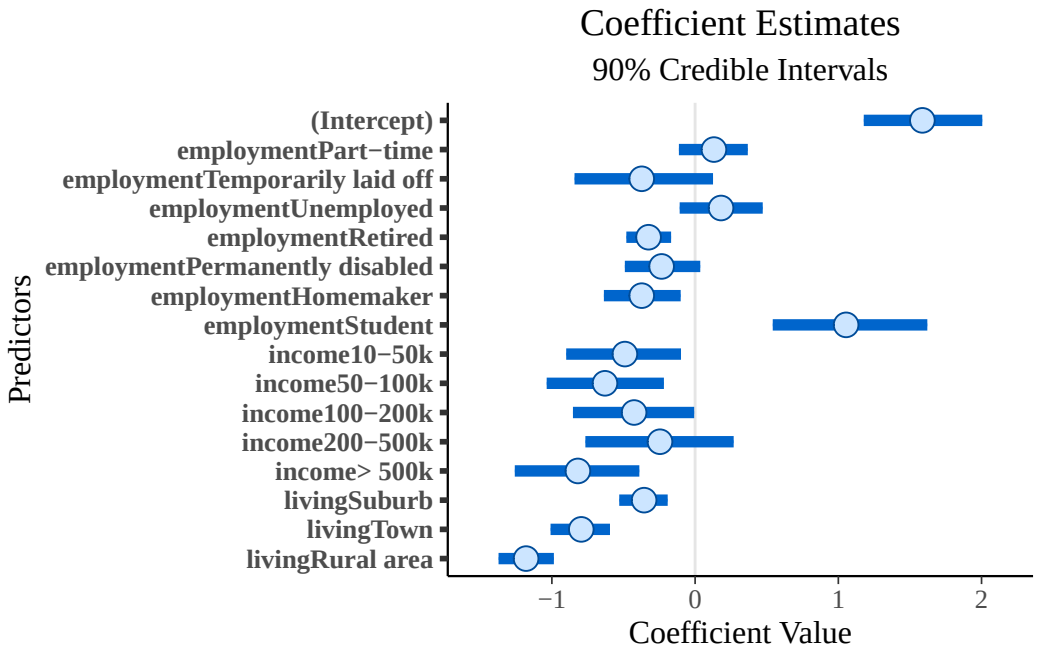

Figure 10: 90% Credible Intervals for Logistic Regression Coefficients

The visualization in Figure 10 shows the 90% credible intervals for each predictor. Predictors with intervals that do not overlap zero indicate significant effects on voting preferences. For example, livingRural area exhibits a significant negative association with voting for Biden, while employmentStudent shows a strong positive effect.

4.1 Model Results

The below bar graphs display the predicted support for different demographics with the other ones kept constant, with the green point showing the actual results as shown in Figure 7.

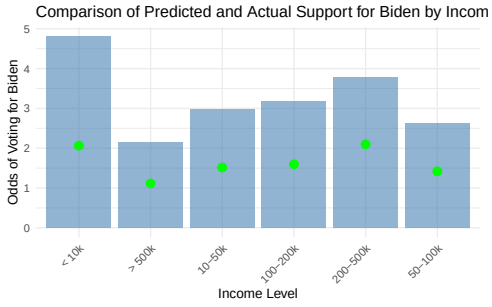

Figure 11: Model Prediction vs. Actual Votes by Income Level

The predicted values in Figure 11 are much lower likely due to the other demographics being set to "Full-time" and "City", but the trend in the results is the same.

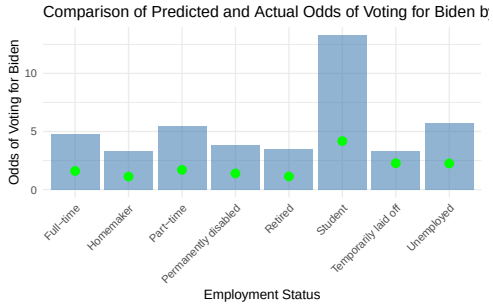

Figure 12: Model Prediction vs. Actual Votes by Employment

Figure 12 shows similarity in the general trend between the prediction and the actual data, with the exception of some of the middle values, where "Temporarily laid off" for example is lower than expected.

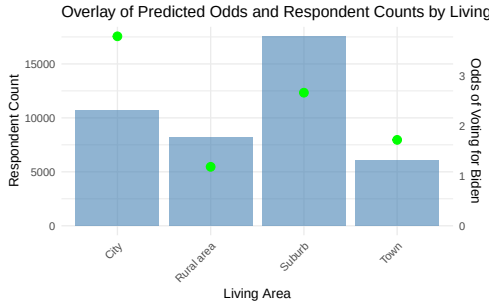

Figure 13: Model Prediction vs. Actual Votes by Living Area

Figure 13 shows the model has completely different predictions for the living areas, when employment is set to "Full-time" and income is in the "200-500k" range.

4.2 Model Evaluation

The ROC (Receiver Operating Characteristic) curve is a graphical representation of the model's ability to distinguish between the two classes: votes for Biden (1) and Trump (0). It plots the true positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold settings.

The AUC (Area Under the Curve) quantifies the overall ability of the model to discriminate between the classes. An AUC of 1.0 indicates perfect discrimination, while 0.5 indicates no better performance than random guessing.

```
Area under the curve: 0.6215
```

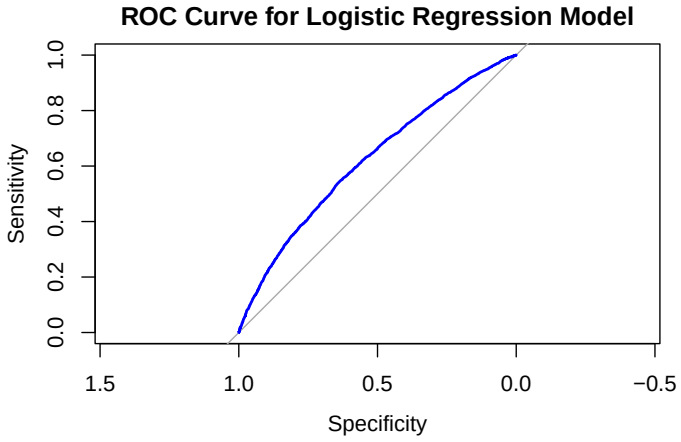

Figure 14: ROC Curve for logistic regression model

The area under curve (AUC) value for the curve in Figure 14 is 0.6216. This indicates the model has moderate discriminatory power. It performs better than random guessing (AUC = 0.5) but falls short of strong predictive performance (AUC $>$ 0.8). This suggests the predictors (employment status, income, and living environment) have some explanatory power but may not fully capture voting behavior.

In terms of the shape of the curve, it deviates above the diagonal line (random chance), showing that the model can separate the two classes to some extent.

However, the relatively shallow curve implies the model struggles to achieve high sensitivity without sacrificing specificity.

The model is able to identify patterns in the data, distinguishing between Biden and Trump voters to a limited degree. The AUC value indicates that important predictors influencing voting behavior may be missing (e.g., race, education, or political affiliation).

Further steps, such as feature engineering, adding interaction terms, or testing alternative models (e.g., random forests or support vector machines), may help enhance performance.

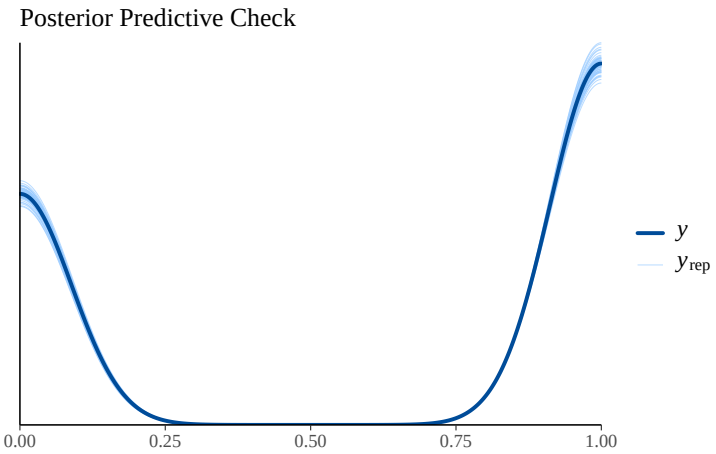

Figure 15: ROC Curve for logistic regression model

In Figure 15, The observed data y is represented by the solid curve. The replicated data y_rep is represented by the shaded region, which covers the range of simulations based on the posterior distribution of the model parameters.

The replicated data closely follows the observed data curve between values 0.25-0.75, and diverges more at the extremes. This suggests that the model struggles to fully capture the extreme probabilities of voting for Biden or Trump.

The replicated data aligns more closely with the observed data in the middle of the probability range, indicating that the model performs better in areas of uncertainty where probabilities are closer to 50%.

The mismatch at the extremes suggests that the model may not be adequately capturing respondents with very high or very low likelihoods of voting for either candidate. This could indicate the need for additional predictors to account for factors driving extreme probabilities.

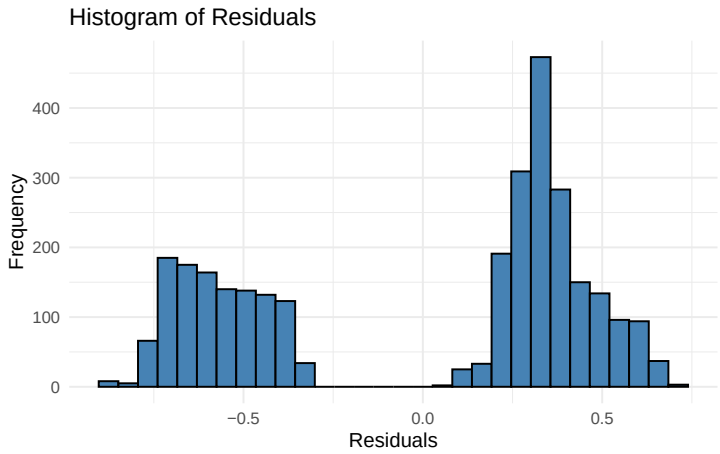

Figure 16: ROC Curve for logistic regression model

Figure 16 shows the distribution of residuals, which represent the differences between observed and predicted values. The residuals are distributed symmetrically around zero, which suggests that the model does not exhibit significant bias in its predictions. The histogram shows two peaks, one slightly negative and the other slightly positive. This may indicate that the model performs differently for different subsets of the data (e.g., voters for Biden vs. Trump). This could reflect the nature of binary logistic regression, where predictions are pushed toward 0 or 1. The spread of residuals is relatively narrow, with most values falling between -0.5 and 0.5. This suggests that the model's predictions are not overly far from the observed values. The PPC confirms the model is reasonable for the given data but highlights areas for improvement, particularly at the extremes of the outcome probabilities. While the residuals and PPC suggest the model captures the data trends reasonably well, additional predictors or refinements, such as interactions, may be necessary to improve the fit for subsets of the data.

5 Discussion

This paper investigated the influence of socio-economic factors—employment status, income level, and living environment—on voting preferences during the 2020 U.S. Presidential Election. Utilizing the 2020 Cooperative Election Study (CCES) dataset, a logistic regression model was applied to predict support for Joe Biden versus Donald Trump. The model incorporated categorical predictors, regularized with Bayesian priors, to account for variability while ensuring interpretability. The model's predictors were carefully chosen to reflect significant socio-economic dimensions that shape voter behavior. Through statistical analysis, this study identified meaningful correlations between these predictors and political alignment, shedding light on broader socio-political dynamics in the United States.

By employing a logistic regression model, this paper directly modeled the binary nature of the voting outcome while retaining a clear interpretation of how each predictor contributed to the likelihood of voting for a particular candidate. The inclusion of categorical variables like employment, income, and living environment provided an avenue to explore how socio-economic realities translate into political preferences.

5.1 Takeaways

This analysis shows the continued significance of the urban-rural divide in American politics. Urban voters were far more likely to support Joe Biden, a trend consistent with existing literature on urban liberalism and rural conservatism (Center 2020d). Urban areas, characterized by higher population densities and greater economic and cultural diversity, often prioritize progressive policies such as environmental regulation, healthcare access, and social equity. Conversely, rural voters exhibited a marked preference for Donald Trump, reflecting a broader alignment with conservative policies, including lower taxation, gun rights, and traditional social values (Iyengar and Krupenkin 2018). These findings underscore the stark political polarization between urban and rural constituencies, a division that has deepened over successive election cycles.

Another takeaway is the role of income in shaping political preferences. Lower-income groups were more likely to support Biden, aligning with the Democratic platform's emphasis on social welfare programs, affordable healthcare, and economic redistribution. Middle-income groups displayed more mixed preferences, often reflecting a balance between economic self-interest and policy priorities. Interestingly, the wealthiest demographic groups exhibited reduced support for Biden, suggesting that concerns over Democratic tax policies and fiscal regulations may outweigh social considerations in shaping their voting behavior (Frank 2004; Andrew Gelman and Cortina 2009). This U-shaped relationship between income and voting patterns highlights the complexity of economic factors in political decision-making.

Employment status further illuminated key dynamics in voter behavior. Students and retired individuals showed strong support for Biden, possibly due to generational policy preferences such as student debt relief and Social Security protection. Full-time workers demonstrated more balanced voting preferences, reflecting the diversity of this group in terms of socio-economic priorities. Notably, individuals in precarious employment situations, such as part-time or temporarily laid-off workers, leaned toward Biden, likely reflecting concerns over job security and healthcare access in the wake of the COVID-19 pandemic.

5.2 Weaknesses

While the model was well-suited for the purpose of predicting binary outcomes, there were several limitations. First, the logistic regression model assumes linearity in the log-odds of the predictors, which may oversimplify complex interactions between socio-economic factors

and voting behavior. Non-linear relationships or interaction effects, such as how income and employment jointly influence voting preferences, may be overlooked in this framework.

The categorical representation of predictors, while simplifying interpretation, also introduces granularity constraints. For example, income categories like "50-100k" cover a broad range of experiences that may differ substantially in their political implications. Similarly, the employment categories do not account for factors such as industry type or job satisfaction, which could have significant effects on political alignment.

Another limitation lies in the reliance on self-reported survey data. Self-reporting is subject to biases such as social desirability and recall inaccuracies, potentially skewing results (Ansolabehere and Hersh 2012). Furthermore, the cross-sectional nature of the CCES dataset captures voter preferences at a single point in time, making it difficult to assess the impact of dynamic political events or campaign strategies on voting behavior (Andrew Gelman and Cortina 2009).

Finally, the model does not account for regional variations within urban and rural categories. For instance, urban areas in different states may have distinct political leanings influenced by local economic conditions, cultural dynamics, or historical contexts. Ignoring these regional differences may obscure important variations in the data.

5.3 Future Directions

The findings of this paper open several avenues for future research. A key area for discussion is the role of timing in voting behavior. Incorporating longitudinal data could describe how socio-economic factors interact with political events, economic cycles, or social movements to shape voting preferences over time. Such an approach would allow researchers to examine causal relationships rather than mere associations.

Intersectionality is another promising direction. Investigating how combinations of factors—such as race, gender, and education—jointly influence voting behavior could yield a richer understanding of political alignment. For example, how do the political preferences of low-income individuals differ by gender or race? Addressing these questions would require a more granular dataset and modeling techniques such as hierarchical or multi-level models (Highton 2009).

Methodologically, future studies could benefit from experimenting with non-linear models or machine learning approaches. These methods can capture complex relationships and interactions that traditional logistic regression might miss. For instance, decision tree-based algorithms or neural networks could identify non-obvious patterns in how socio-economic factors influence voting preferences, though at the cost of reduced interpretability.

Regional analyses are necessary to show how local contexts mediate the relationship between socio-economic factors and voting behavior. For example, employment trends in

manufacturing-heavy states compared to service-oriented urban centers might cause differences in voter alignment. Incorporating regional economic data or policy differences could lead to mroe accurate explanations in future models.

In conclusion, while this paper highlights the significant role of socio-economic factors in shaping political preferences, it also shows the need for more dynamic approaches to understanding voter behavior. By addressing these limitations and pursuing the outlined future directions, researchers can build a more comprehensive and context-sensitive picture of the socio-political landscape.

6 Appendix: Survey and Sampling Methodology

Overview of Survey Design in the CCES 2020 Dataset

The 2020 Cooperative Election Study (CCES) is a large-scale, national survey designed to capture U.S. political behavior and attitudes. It employs a stratified sample, drawing from a diverse pool of respondents to achieve national representation The survey design is a collaborative effort between academic institutions and professional survey firms (Ansolabehere and Hersh 2012).

The CCES includes two waves: a pre-election wave conducted in October 2020 and a post-election wave conducted in November 2020. This dual-wave structure allows researchers to capture shifts in voter sentiment and behavior during the election period. The survey uses online data collection, a method that benefits from scalability and cost-effectiveness.

6.1 Sampling Techniques

6.1.1 Stratified Sampling

Stratified sampling ensures that subgroups of the population, such as those defined by age, race, income, or geographic region, are adequately represented. For the CCES, stratification is based on demographic characteristics derived from the U.S. Census Bureau and voter registration databases. This method minimizes sampling error and allows for more precise estimates of population parameters (Lohr 2021).

6.1.2 Sample Matching

Sample matching is a two-step process used to improve the representativeness of the survey sample. First, a target sample is generated based on demographic benchmarks. Next, survey respondents are selected from a larger pool to match the characteristics of the target sample. The CCES utilizes voter registration data to match respondents, enhancing the validity of findings (Rivers 2006).

6.1.3 Weighting

Post-stratification weights are applied to the data to correct for any deviations between the survey sample and the target population. These weights adjust for factors such as response bias and demographic discrepancies. The weights are calculated using iterative proportional fitting, ensuring alignment with key benchmarks like age, gender, race, and geographic distribution (Gelman and Hill 2007).

6.2 Observational Data and Challenges

6.2.1 Strengths of Observational Surveys

Observational surveys like the CCES provide a snapshot of real-world voter behavior, capturing smaller details that experimental designs may overlook. They allow for large-scale data collection across diverse populations, which is important for studying complex phenomena like political alignment.

6.2.2 Limitations

Despite their strengths, observational surveys face challenges such as:

1. Nonresponse Bias: Certain demographic groups may be less likely to participate, leading to underrepresentation.
2. Self-Reporting Bias: Respondents may misreport their voting behavior or demographic characteristics, often due to social desirability (Ansolabehere and Hersh 2012).
3. Cross-Sectional Nature: The CCES captures data at specific points in time, limiting its ability to account for temporal changes in voter behavior.

6.3 Comparisons with Other Surveys

6.3.1 American National Election Studies (ANES)

The ANES employs a mixed-mode approach, combining face-to-face interviews with online surveys. While ANES provides rich contextual data, its smaller sample size limits the granularity of subgroup analysis compared to the CCES (Berinsky 2014).

6.3.2 Pew Research Center Surveys

Pew Research specializes in tracking public opinion trends but focuses on shorter questionnaires. The CCES's larger number of more specific questions is more suited for modeling complex relationships like those explored in this paper (Center 2020c).

6.4 Methodological Best Practices

Based on the strengths and limitations of the CCES and other surveys, several best practices emerge:

1. Hybrid Sampling Methods: Combining stratified sampling with random digit dialing or online panel recruitment can improve representativeness.
2. Enhanced Weighting Techniques: Incorporating machine learning algorithms into weighting procedures could reduce bias from hard-to-reach populations.
3. Longitudinal Design: Adding a longitudinal component to surveys like the CCES would allow researchers to track changes in voter behavior over time.

6.5 Future Directions

Future iterations of surveys like the CCES could benefit from integrating administrative data (e.g., tax records or employment data) to validate self-reported responses. Additionally, adopting adaptive sampling techniques that dynamically adjust to underrepresented groups could further enhance data quality. Finally, the incorporation of qualitative components, such as open-ended responses, could provide richer context to quantitative findings.

By adhering to these methodological enhancements, surveys can accurately represent complex socio-political phenomena.

References

American National Election Studies. 2021. "ANES 2020 Time Series Study Full Release." https://www.electionstudies.org.

Andrew Gelman, Boris Shor, David Park, and Jeronimo Cortina. 2009. *Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do*. Princeton University Press.

Ansolabehere, Stephen, and Eitan Hersh. 2012. "Validation: What Big Data Reveal about Survey Misreporting and the Real Electorate." *Political Analysis* 20 (4): 437–59. https://doi.org/10.1093/pan/mps023.

Ansolabehere, Stephen, and Brian Schaffner. 2020. "The Cooperative Congressional Election Study." *Harvard Dataverse*. https://doi.org/10.7910/DVN/ZSBZ7K.

Arel-Bundock, Vincent. 2021. *Modelsummary: Summary Tables and Plots for Statistical Models and Data: Beautiful, Customizable, and Publication-Ready*. https://CRAN.R-project.org/package=modelsummary.

Ben-Gurion, Tal. 2020. "The Role of Misinformation in Undermining Electoral Integrity." *Journal of Democracy* 31: 54–68.

Berinsky, Adam J. 2014. "Rumors and Health Care Reform: Experiments in Political Misinformation." *British Journal of Political Science* 44 (2): 241–62. https://doi.org/10.1017/S0007123412000734.

Bishop, Bill. 2019. *The Big Sort: Why the Clustering of Like-Minded America Is Tearing Us Apart*. Mariner Books.

Blustein, David. 2020. "The Impact of Unemployment on Political Preferences During COVID-19." *American Psychologist* 75: 608–17.

Bureau, United States Census. 2020. "Current Population Survey Voting and Registration Supplement."

Catalist. 2017. "About Catalist." http://web.archive.org/web/20171028000000*/https://catalist.us/about/.

Center, Pew Research. 2018. "The Generation Gap in American Politics." https://www.pewresearch.org/.

———. 2020a. "Behind Biden's 2020 Victory: An Examination of the 2020 Electorate by Race, Gender, Age and Education."

———. 2020b. "Behind Biden's 2020 Victory: An Examination of the 2020 Electorate by Race, Gender, Age and Education."

———. 2020c. "Behind Biden's 2020 Victory: An Examination of the 2020 Electorate by Race, Gender, Age and Education."

———. 2020d. "Urban and Rural Divides in Political Attitudes." https://www.pewresearch.org/.

Climate Change, Intergovernmental Panel on. 2020. "Special Report: Climate Change and Land."

Cook, Charlie. 2020. "The Suburban Shift in the 2020 Election."

Cramer, Katherine J. 2016a. *The Politics of Resentment: Rural Consciousness in Wisconsin and the Rise of Scott Walker*. University of Chicago Press.

———. 2016b. *The Politics of Resentment: Rural Consciousness in Wisconsin and the Rise of Scott Walker*. University of Chicago Press.

Edsall, Thomas. 2020. "Income, Education, and the 2020 Vote."

Foundation, Kaiser Family. 2020. "Retiree Voting Trends and Healthcare Priorities."

Fowler, Anthony. 2020. "Unemployment and Electoral Outcomes."

Frank, Thomas. 2004. *What's the Matter with Kansas? How Conservatives Won the Heart of America*. Holt Paperbacks.

Frey, William. 2020. *Diversity Explosion: How New Racial Demographics Are Remaking America*. Brookings Institution Press.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2008. *Bayesian Data Analysis*. CRC Press.

Gelman, Andrew, and John Hill. 2007. "Data Analysis Using Regression and Multilevel/Hierarchical Models." *Cambridge University Press*.

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

Highton, Benjamin. 2009. "Revisiting the Relationship Between Educational Attainment and Political Sophistication." *The Journal of Politics* 71 (4): 1564–76. https://doi.org/10.1017/S0022381609990036.

Iyengar, Shanto, and Masha Krupenkin. 2018. "The Strengthening of Partisan Affect." *Political Psychology* 39.

Kay, Matthew. 2021. *Tidybayes: Tidy Data and Geoms for Bayesian Models*. https://CRAN.R-project.org/package=tidybayes.

Labor Statistics, Bureau of. 2020. "Unemployment Rate During the COVID-19 Pandemic."

Leeper, Thomas. 2021. *Dataverse: Client for Dataverse 4 Repositories*. https://CRAN.R-project.org/package=dataverse.

Lohr, Sharon L. 2021. *Sampling: Design and Analysis*. 3rd ed. CRC Press.

Muro, Mark, and Robert Maxim. 2020. "The Political Economy of Blue-Collar America." *Brookings Institution*.

OpenAI. 2023. "ChatGPT: Optimizing Language Models for Dialogue." https://openai.com/.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Research, Edison. 2020. "2020 u.s. Presidential Election Exit Polls." https://www.edisonresearch.com/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://github.com/apache/arrow/.

Rivers, Douglas. 2006. "Sample Matching: Representative Sampling from Internet Panels." *Palo Alto: Polimetrix White Paper Series*.

———. 2007. "Sample Matching." https://static.texastribune.org/media/documents/Rivers_matching4.pdf.

Schaffner, Brian F., Stephen Ansolabehere, and Sam Luks. 2020. "Cooperative Congressional Election Study." Harvard Dataverse.

Schaffner, Brian, Stephen Ansolabehere, and Sam Luks. 2021. "Cooperative Election Study Common Content, 2020." Harvard Dataverse. https://doi.org/10.7910/DVN/E9N6PH.

State Legislatures, National Conference of. 2020. "Voting by Mail in 2020."

Sunstein, Cass. 2020. "Polarization and the Threat to Democracy in the United States." *Harvard Law Review* 134: 123–47.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://CRAN.R-project.org/package=knitr.