The Generational and Demographic Determinants of Voting Behavior: Evidence from the 2022 CES*

Urban Residents Are 2.3 Times More Likely, College Graduates 1.9 Times More Likely to Support Democrats

Jinyan Wei

December 3, 2024

Over the past decades, American voting behavior has revealed distinct generational and demographic patterns. This study explores how age cohorts and demographic characteristics influence partisan preferences in the 2022 election cycle. Using data from the 2022 Cooperative Election Study (CES), we analyze the interaction of generational differences, socioeconomic status, and regional variation in shaping voting behavior. Employing binary logistic regression and decision tree models, we identify significant differences across age cohorts, with older voters demonstrating distinct partisan preferences compared to younger generations. These patterns vary further by gender, education level, and geographic region. Our findings highlight the critical roles of generational experiences and demographic factors in contemporary American political alignment, emphasizing the evolving nature of partisan identification. The results underscore the necessity of considering generational and demographic factors in understanding electoral behavior.

Table of contents

*Code and data are available at: [https://github.com/jeno0403/Voter-Behavior-2022-CES).

1 Introduction

The determinants of voting behavior in the United States have been a focus of political science research, particularly as the electorate becomes more diverse and polarized. Among these factors, educational attainment and racial identity play a significant role in shaping political preferences. Understanding how these and other demographic characteristics influence voting behavior is essential for analyzing current trends in political alignment and participation.

This study examines how educational attainment, racial identity, and other demographic variables influence partisan voting patterns. Using data from the 2022 Cooperative Election Study Schaffner, Ansolabehere, and Shih (2023), which includes responses from 60,000 registered voters across all U.S. states, this analysis explores the relationships between these factors and the likelihood of supporting Democratic or Republican candidates. The CES dataset integrates validated voter registration data with detailed demographic and socioeconomic information, forming the basis for this research.

The analysis uses logistic regression and decision tree models to assess the effects of education, race, income, and urbanicity on voting preferences. Logistic regression quantifies the linear relationships between predictors and voting behavior, while decision trees capture non-linear interactions. Together, these methods allow for a detailed examination of voter preferences and highlight demographic and socioeconomic patterns within the electorate.

The findings indicate that the relationship between educational attainment and voting preferences varies across demographic groups. For example, higher education levels are associated with greater support for Democratic candidates among White and Black voters, but the pattern differs for Hispanic and Asian voters, suggesting the influence of cultural and contextual factors. These results emphasize the need to consider multiple factors when analyzing voter behavior in the United States.

The remainder of this paper is organized as follows: Section 2 describes the data sources and variable measurements, Section 3 details the logistic regression methodology and model specifications, Section 4 presents the empirical findings, and Section 5 concludes with implications and future research directions. Additional details and validation steps are provided in Appendix- A, Appendix- B, and Appendix- C.

2 Data

2.1 Overview

This analysis examines voting behavior in the 2022 U.S. election using R programming (R Core Team 2023). The dataset, drawn from the 2022 Cooperative Election Study (Schaffner, Ansolabehere, and Shih 2023), provides a structured basis for examining how demographic, socioeconomic, and geographic variables—such as race, age, education, income, and urbanicity—shape political behavior. The CES dataset combines validated voter registration data with extensive survey responses, offering a detailed view of partisan preferences. Guided by Alexander (2023) with Data by Alexander (2023), this study employs statistical modeling to understand the influence of these factors on partisan alignment.

Several R packages were employed for data preparation, modeling, and visualization. Data cleaning and manipulation were facilitated by `tidyverse` (Wickham et al. 2019) and `dplyr` (Wickham et al. 2023), while `arrow` (Richardson et al. 2024) managed parquet files for efficient storage and compatibility.The `here` (Müller 2020) package aided in managing file paths, improving the reproducible workflow of our study. Logistic regression models were implemented and validated using `caret` (Kuhn and Max 2008), with performance assessed through Receiver Operating Characteristic (ROC) curves using `pROC` (Robin et al. 2023). Geographic trends were analyzed with the `maps` (Brownrigg et al. 2023) package, and `ggplot2` (Wickham 2016) was used for detailed visualizations. Report generation relied on `knitr` (Xie 2014) and `kableExtra` (Zhu 2024) for clear and reproducible documentation.

By applying statistical techniques and a reproducible workflow, this analysis examines voting behavior in the 2022 U.S. election. The structured use of R tools ensures transparency and accuracy, enabling a thorough evaluation of the effects of demographic, socioeconomic, and geographic factors on voter preferences. This approach identifies significant patterns within the electorate, contributing to a better understanding of the forces shaping partisan alignment.

2.2 Measurement

The process of transforming real-world voting behavior into a structured dataset requires a systematic and rigorous approach. This study draws on data from the 2022 Cooperative Election Study (CES), a large-scale survey designed to capture voter preferences and demographic characteristics. The CES employs matched random sampling techniques stratified by demographics and geography to ensure a representative sample of the U.S. electorate. Respondents are recruited using diverse methods, including online panels and targeted outreach, reflecting the population's diversity.

Survey items are carefully designed to capture a wide range of voter information, including political preferences, demographic details, and socioeconomic characteristics. These questions aim to provide insights into factors shaping voter behavior, such as education, race, gender, and urbanicity.

After responses are collected, the dataset undergoes thorough cleaning and validation procedures. This involves addressing inconsistencies, rectifying missing values, and recategorizing variables like education and race into standardized groups to facilitate meaningful comparisons. While the CES dataset includes weighting to adjust for sampling biases—accounting for disparities in age, gender, and state representation—this study does not apply these weights. Instead, the analysis is conducted using unweighted data, focusing on the raw relationships between demographic factors and voting behavior.

The resulting dataset provides a rich source of information for analyzing the determinants of voter preferences. Each entry reflects an individual's demographic characteristics and voting behavior, offers a detailed understanding of how key factors influence partisan alignment in the United States. This structured methodology ensures that the dataset accurately captures the nuances of the electorate, providing a foundation for robust statistical analysis.

2.3 Variables

The dataset incorporates a range of variables to capture demographic, socioeconomic, and geographic characteristics of registered voters. These variables are categorized as follows:

2.3.1 Outcome Variable

The primary outcome variable, `vote_choice`, is binary, indicating whether a respondent supports a specific candidate. This allows us to analyze the factors influencing voter choice in the 2022 election.

2.3.2 Predictor Variables

Key predictors include:

- **Age_cohort**: Categorical variable with four groups (18-29, 30-49, 50-64, 65-90)
- **Education**: Highest educational attainment, categorized as "no high school," "some college," "high school graduate," "2-year College Degree","4-year College Degree" or "Postgraduate Degree".
- **Gender**: A categorical variable capturing self-identified gender (Male or Female).
- **Race**: Self-identified racial or ethnic group, categorized as "White," "Black," "Hispanic," "Asian," "Native American," "Middle Eastern," and "Other."
- **Urbanicity**: A categorical variable classifying respondents as residing in urban, suburban, rural or town areas.
- **Religion**: Religious affiliation, measured alongside attendance frequency.

- **Region**: A categorical variable indicating geographic location (e.g., Northeast, Midwest, South, or West).
- **Income_tier**: Household income level.

These variables were selected to reflect key factors identified in the literature as significant predictors of voting behavior. By including a diverse set of predictors, the analysis captures nuanced dynamics in voter preferences and behavior across different demographic and geographic groups.

2.4 Relationships between varaibles

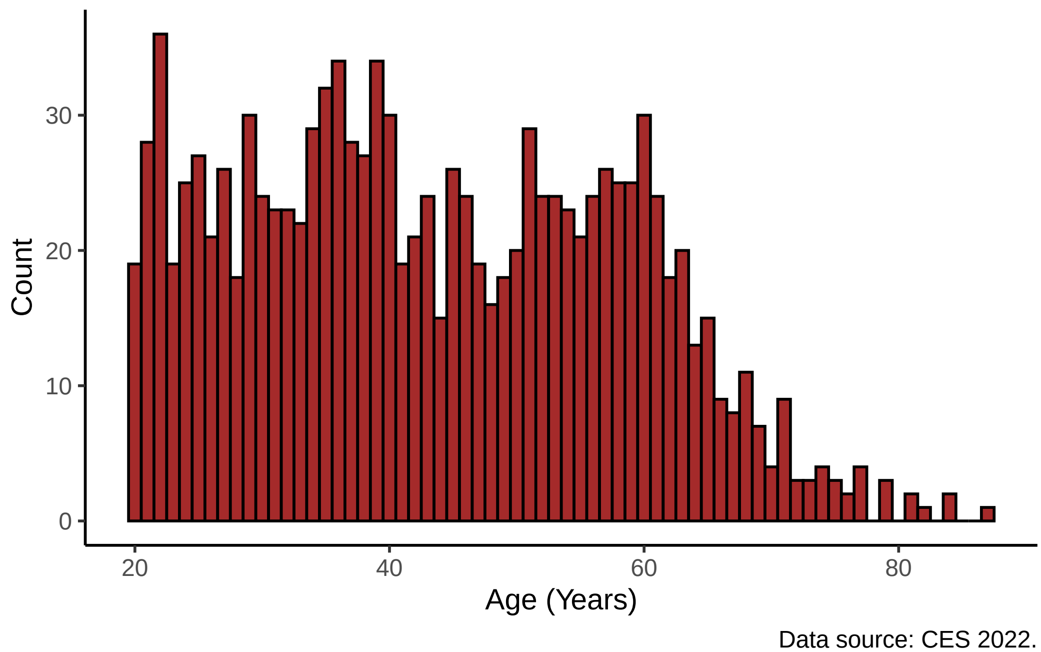


Figure 1: Distribution of Respondent Age

Figure 1 displays the age distribution of respondents. The count of respondents peaks in the 40 to 60-year range, with a noticeable drop-off in the higher and lower age brackets. The distribution appears somewhat uniform across the middle age ranges, with several spikes indicating larger groups of respondents at specific ages. The age range spans from 20 to approximately 85 years, with fewer respondents in the younger and older age groups.

Figure 2 illustrates the distribution of party identification across different age groups (18-29, 30-49, 50-64, and 65+) segmented by gender. depicts the intersection of gender, age, and party identification. It uncovers patterns of political affiliation across different age brackets for male and female respondents. Notable trends include the strong Democratic identification in younger age groups and the rise of Republican identification as age increases. Independents

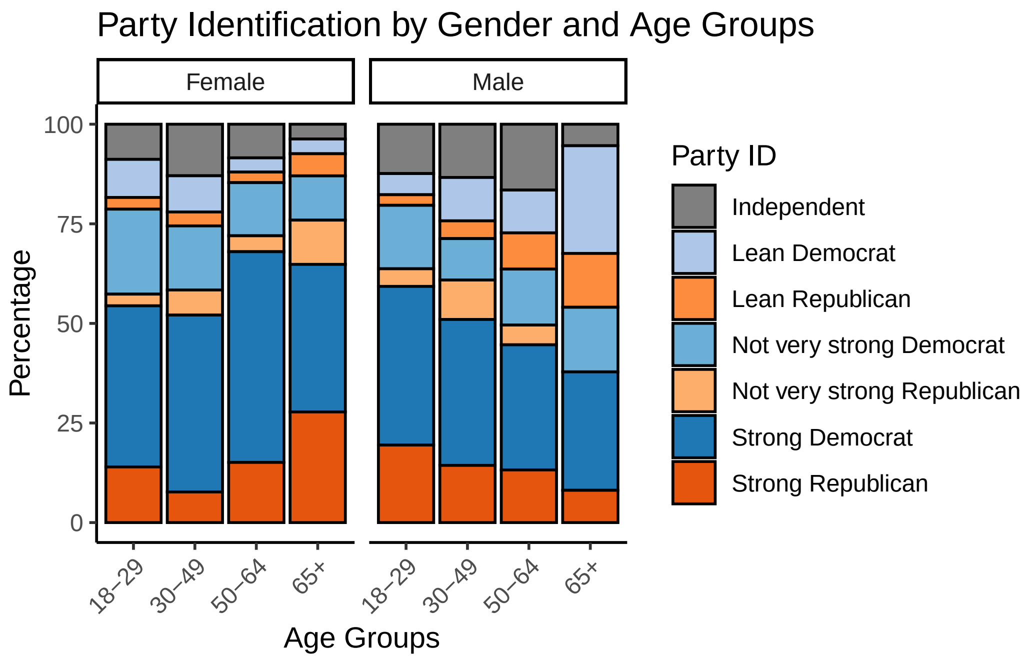


Figure 2: Party Identification by Gender and Age Groups.

are more evenly distributed, emphasizing the variability of non-affiliated voters across age and gender demographics. This visualization highlights how age and gender contribute to partisan tendencies within the electorate.

Figure 3 visualizes the proportion of Democrat voters across the United States based on polling data. The color gradient on the map indicates the proportion of Democrat voters, with darker shades representing a higher percentage of Democrat support. The states with the darkest blue colors indicate a strong preference for the Democratic party, while lighter colors represent weaker support. States in gray either have missing data or are not included in the polling sample. The visual highlights regional variations in voting preferences, with some states consistently supporting the Democratic party, while others, particularly in the South and parts of the Midwest, show lower levels of support.

3 Model

OOur modeling approach seeks to examine how demographic, socioeconomic, and geographic factors collectively influence partisan voting preferences during the 2022 U.S. election cycle. For this analysis, we employ a logistic regression model to predict the likelihood of voting for the Democratic party (`vote_choice = 1`) compared to voting Republican or for other parties (`vote_choice = 0`). This model is implemented using the `glm()` function in R, applying a

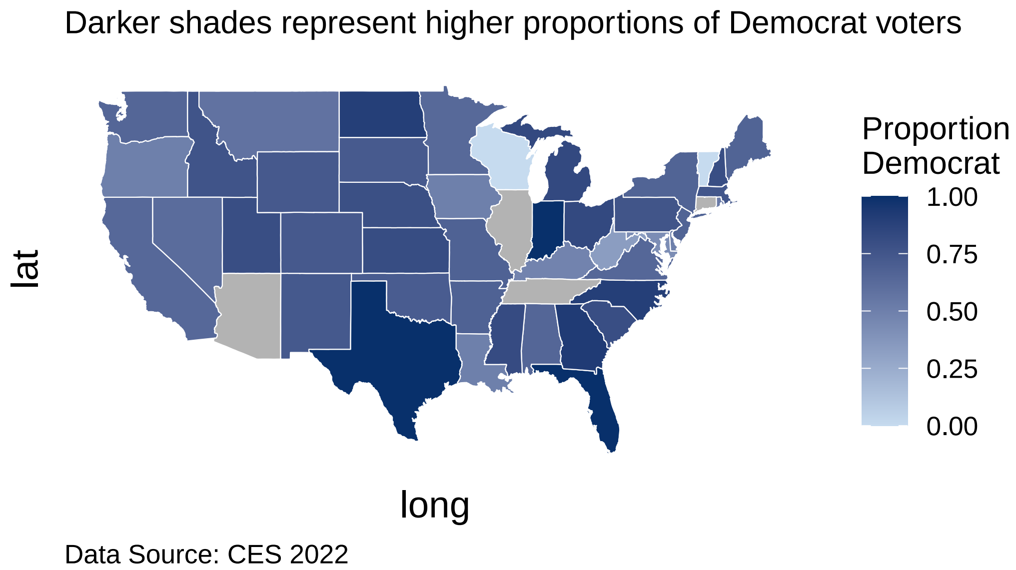


Figure 3: Proportion of Democrat Voters by State shows the distribution of Democrat voters across the United States, with darker shades representing a higher proportion of Democrat voters in each state.

binomial distribution with a logit link function to capture the binary nature of the voting outcome.

The predictors used in the model include a combination of demographic, socioeconomic, and regional variables. `age_cohort` divides respondents into generational groups (18–29, 30–49, 50–64, and 65–90), reflecting differences in life stage and political priorities. `gender` accounts for self-identified gender categories, while `education` captures the highest level of educational attainment, ranging from high school or less to postgraduate degrees. `income_tier` is used to approximate socioeconomic status, while `religion` represents self-identified religious affiliation and its potential influence on political behavior.

Race and geographic context are central to this analysis. `race` captures the self-reported racial or ethnic identity of respondents, while `urbanicity` differentiates urban, suburban, and rural areas. `region`, categorized as Northeast, Midwest, South, and West, provides a broader geographic context. These variables collectively allow the model to capture the intersectional dynamics of race, geography, and demographic factors in shaping partisan preferences.

The logistic regression model assumes that the probability of voting for the Democratic party, given these predictors, follows a logistic distribution. This framework enables the estimation of the effects of individual variables and their interactions, particularly between race and geography, on the log-odds of voting Democrat. By utilizing this approach, the model provides insights into the nuanced ways in which race, urbanicity, and demographic factors influence voting behavior, highlighting specific regional and intersectional trends.

3.1 Model Set-Up

The model predicts the likelihood of voting for the Democrat party by constructing a logistic regression model using the following predictor variables:

- `age_cohort`: Categorical variable categorizes individuals into distinct age groups based on their age. This approach replaces the continuous `age` variable with four categorical cohorts:
  - **18-29 years**: Represents younger voters, often including students or those early in their careers.
  - **30-49 years**: Middle-aged individuals, typically established in their careers or managing families.
  - **50-64 years**: Pre-retirement voters who may prioritize healthcare, pensions, or economic stability.
  - **65+ years**: Older voters who are often retired and focus on social security and Medicare.
- `income_tier`: Categorical variable indicating the respondent's income level.
- `education`: The highest level of education attained by the respondent.
- `gender`: Categorical variable capturing the respondent's gender.
- `religion`: Categorical variable indicating the respondent's religious affiliation.
- `region`: Categorical variable indicating geographic location (e.g., Northeast, Midwest, South, or West).
- `race`: Categorical variable representing the respondent's racial or ethnic background.
- `urbanicity`: Variable indicating whether the respondent resides in an urban, suburban, or rural area.

Grouping ages into cohorts simplifies interpretation by analyzing broad age-based voting trends rather than year-to-year changes. We assume that Age and voting behavior does not follow a linear relationship. This transformation allows the model to identify distinct patterns among groups.

3.1.1 Logistic Regression Model

The logistic regression model predicts the probability of voting for the Democrat party based on the predictors listed above. The model is specified as:

$$
\begin{aligned}
y_i|\eta_i &\sim \text{Bernoulli}(P(y_i = 1)) \\
P(y_i = 1) &= \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} \\
\eta_i &= \beta_0 + \beta_1 \cdot \text{Age Cohort}_i + \beta_2 \cdot \text{Income Tiers}_i \\
&\quad + \beta_3 \cdot \text{Education}_i + \beta_4 \cdot \text{Gender}_i \\
&\quad + \beta_5 \cdot \text{Religion}_i + \beta_6 \cdot \text{Race}_i + \beta_7 \cdot \text{urbanicity}_i + \beta_8 \cdot \text{Region}_i
\end{aligned}
$$

Where:

- y_i is the binary outcome variable, where $y_i = 1$ indicates voting Democrat and $y_i = 0$ indicates voting Republican.
- β_0 is the intercept term for baseline log-odds.
- β_1 represents the effect of age cohort (18-29, 30-49, 50-64, 65-90)
- $\beta_2, \beta_3, \dots, \beta_8$ are the coefficients for each predictor, indicating their impact on the log-odds of voting Democrat.

The model is implemented in R using the `glm()` function with a binomial family and logit link function to estimate the probability of Democratic party support as a function of demographic and generational characteristics.

3.1.2 Decision Tree Model

The decision tree model employs the CART (Classification and Regression Tree) algorithm, which recursively partitions the dataset into subsets to maximize homogeneity in the voting outcome. Similar to the logistic regression model described in Section 3.1.1, the decision tree uses the same set of predictor variables, including demographic, socioeconomic, and geographic factors. Each split in the tree is determined by a single predictor variable, and the splitting criterion minimizes impurity, measured using the Gini index:

$$G(N) = 1 - \sum_{k=1}^{K} p_k^2$$

where $G(N)$ is the impurity of node N, K is the number of classes, and p_k is the proportion of observations in class K at node N.

At each node, the data is split based on a predictor variable X_j and a threshold t to create two subsets:

$$N_{\text{left}} = \{X \mid X_j \leq t\}, \quad N_{\text{right}} = \{X \mid X_j > t\}$$

The terminal nodes represent the final predictions. The predicted class at each terminal node T_m is determined by majority voting:

$$\hat{y} = \arg\max_k \{p_k\}$$

where $\hat{y}$ is the predicted class, and p_k is the proportion of class k in T_m.

The model is implemented in R using the `rpart()` function. The resulting tree is interpretable, with each branch representing a decision rule, and the terminal nodes providing the predicted class, the proportion of Democrat voters, and the percentage of observations in the node.

3.1.3 Model Justification

Scholars in political science have long recognized that demographic factors such as age, gender, education, race, and urbanicity, alongside socioeconomic and geographic contexts, significantly shape voting behavior in the United States. This analysis examines these dynamics by including key predictors to understand how demographic, socioeconomic, and geographic variables influence partisan voting behavior.

This study employs a logistic regression model to predict the likelihood of voting Democrat (coded as 1) versus voting Republican or for other parties (coded as 0). Logistic regression is particularly suited for modeling binary outcomes, allowing us to estimate the odds of supporting the Democratic party based on a range of demographic, socioeconomic, and geographic predictors. Key predictors include age cohort, gender, education, income tier, religion, race, urbanicity, and region.

The model is estimated using maximum likelihood estimation (MLE), which identifies parameter values that maximize the likelihood of observing the data. This approach ensures robust and interpretable estimates of the relationships between predictors and voting preferences. To evaluate the model's performance, diagnostics such as the Akaike Information Criterion (AIC) and Receiver Operating Characteristic (ROC) curves are employed, providing measures of model fit and predictive accuracy. These diagnostics ensure the reliability of the results and support a nuanced interpretation of how demographic and geographic variables shape partisan alignment. Further methodological details are provided in Section Section B.

4 Results

4.1 Model Results

The logistic regression results presented in Table 1 and Table 2 highlight the significant demographic, socioeconomic, and contextual factors influencing voting preferences in the 2022

Table 1: Summary of Logistic Regression Model Predicting Voting Choices Based on Demographic and Contextual Factors: An Analysis of CES 2022 Data

	(1)
(Intercept)	1.654
	(0.719)
age_cohort(29,49]	0.412
	(0.206)
age_cohort(49,64]	0.341
	(0.220)
age_cohort(64,90]	0.289
	(0.311)
genderMale	−0.538
	(0.156)
income_tierUpper income	−0.493
	(0.344)
urbanicityUrban	0.675
	(0.162)
regionNortheast	−0.268
	(0.292)
regionSouth	−0.466
	(0.222)
regionWest	−0.323
	(0.190)
Num.Obs.	1174
AIC	1187.9
BIC	1350.1
Log.Lik.	−561.963
F	6.820
RMSE	0.39

Note: The table omits coefficients for race, education,and religion for simplicity. The logistic regression model predicts voting preferences using demographic, socioeconomic, and contextual variables. Analysis was conducted using the CES 2022 dataset.

Table 2: Summary of Logistic Regression Model Predicting Voting Choices Based on Demographic and Contextual Factors: An Analysis of CES 2022 Data

	(1)
(Intercept)	1.654
	(0.719)
education4-year College Degree	0.895
	(0.332)
educationPostgraduate Degree	1.133
	(0.452)
income_tierUpper income	−0.493
	(0.344)
religionAtheist	1.110
	(0.659)
religionMuslim	−1.586
	(0.715)
religionNothing in particular	−0.884
	(0.411)
religionProtestant	−1.583
	(0.414)
religionRoman Catholic	−1.108
	(0.430)
religionSomething else	−1.380
	(0.449)
raceBlack	1.265
	(0.544)
raceHispanic	0.413
	(0.572)
raceWhite	−0.865
	(0.532)
Num.Obs.	1174
AIC	1187.9
BIC	1350.1
Log.Lik.	−561.963
F	6.820
RMSE	0.39

Note: The table retains only significant variables and combines some categories for clarity. Analysis uses CES 2022 data.

Table 3: Summary of Tree Model Based on Demographic and Contextual Factors: An Analysis of CES 2022 Data

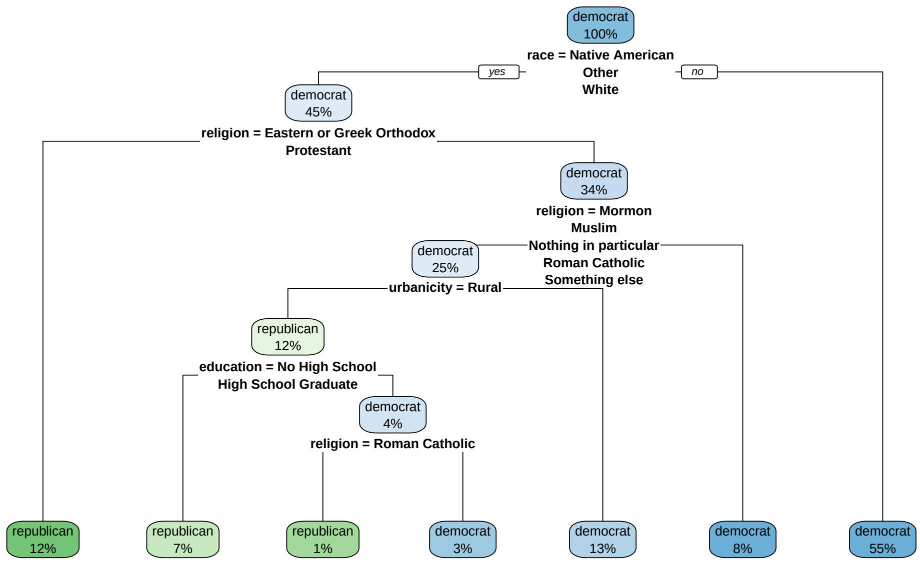


Table 4: Key Metrics for Logistic Regression Predictions

	Metric	Value
Accuracy	Overall Accuracy	0.7734
	Confidence Interval (95%)	(0.7484, 0.7971)
Kappa	Kappa Statistic	0.3998
AccuracyPValue	P-Value for Accuracy Improvement	1.21e-06

Table 5: Key Metrics for Decision Tree Predictions

	Metric	Value
Accuracy	Overall Accuracy	0.7905
	Confidence Interval (95%)	(0.766, 0.8134)
Kappa	Kappa Statistic	0.4407
AccuracyPValue	P-Value for Accuracy Improvement	5.99e-10

U.S. election. Table 1 focuses on the broad demographic and geographic variables. Urbanicity emerges as a key predictor, with urban residents significantly more likely to vote Democrat (coefficient: 0.675, ($p < 0.001$)). Gender also plays a role, as male voters are less likely to support Democratic candidates compared to females (coefficient: -0.538, ($p < 0.01$)). Among age cohorts, voters aged 30–49 exhibit a modest positive association with Democratic support (coefficient: 0.412, ($p < 0.05$)), while older cohorts show weaker, non-significant trends. Regional effects are limited, with only the South demonstrating a statistically significant negative relationship with Democratic support (-0.466, ($p < 0.05$)).

Table 2 delves deeper into the roles of education, religion, and race. Higher education levels strongly correlate with Democratic support, particularly among voters with a 4-year college degree (coefficient: 0.895, ($p < 0.01$)) or postgraduate education (coefficient: 1.133, ($p < 0.05$)). Racial identity also significantly affects preferences, with Black voters showing a strong positive association with Democratic voting (coefficient: 1.265, ($p < 0.05$)), while White voters demonstrate a negative association (-0.865). Religion proves to be another critical factor, as Protestant and Catholic voters exhibit reduced likelihoods of Democratic support (coefficients: -1.583 and -1.108, respectively, ($p < 0.01$)). Individuals with no religious affiliation are also less likely to vote Democrat (-0.884, ($p < 0.05$)). Together, these tables reveal the nuanced ways in which structural and identity-based factors shape partisan alignment.

The results presented in Table 4 and Table 5 summarize the key metrics for both the logistic regression and decision tree models, providing a quantitative comparison of their predictive performance in the context of the 2022 U.S. election. For the logistic regression model, the overall accuracy is 77.34%, with a confidence interval of (74.84%, 79.71%), and a moderate Kappa statistic of 0.3998, indicating fair agreement between predicted and actual classes. The decision tree model, on the other hand, demonstrates slightly higher accuracy at 79.05%, with a narrower confidence interval of (76.60%, 81.34%), and an improved Kappa statistic of 0.4407, reflecting better agreement. Both models show statistically significant p-values for accuracy improvement over random guessing, underscoring their reliability.

Table 3, the decision tree plot, visually illustrates the hierarchical structure of how key predictors—such as race, religion, urbanicity, and education—interact to shape voting preferences. The tree begins with race as the primary split, followed by religion, urbanicity, and education, capturing the layered and complex relationships among these variables. Terminal nodes provide probabilities for Democratic or Republican support, offering interpretable insights into patterns of partisan alignment.

5 Discussion

5.1 Key Findings and Implications

This analysis highlights the significant role of demographic and contextual factors in shaping voting preferences. `Urbanicity` emerges as one of the strongest predictors, with urban residents significantly more likely to vote Democrat (coefficient: 0.675, $p < 0.001$). Gender differences are also notable, with male respondents showing a lower likelihood of Democratic support compared to female respondents (coefficient: -0.538, $p < 0.001$). These findings reinforce the importance of both demographic characteristics and geographic context in shaping partisan alignment.

`Education` continues to play a pivotal role, with individuals holding a four-year college degree (coefficient: 0.895, $p = 0.007$) or a postgraduate degree (coefficient: 1.133, $p = 0.012$) demonstrating significantly higher Democratic support. Even individuals with some college experience show a positive association (coefficient: 0.518, $p = 0.047$), emphasizing the influence of educational attainment on voter preferences. Conversely, religious affiliation shows strong negative effects for some groups. For example, Protestant (coefficient: -1.583, $p < 0.001$) and Roman Catholic respondents (coefficient: -1.108, $p = 0.010$) are less likely to vote Democrat, reflecting entrenched trends in faith-based voting behavior.

Interestingly, regional indicators largely lack statistical significance except for the South (coefficient: -0.466, $p = 0.036$), suggesting that geographic variation in voting preferences may be better explained by urban-rural divides or other demographic factors. `Race` also influences partisan alignment, with Black respondents showing a significantly higher likelihood of voting Democratic (coefficient: 1.265, $p = 0.020$). These findings underscore the complexity of voter behavior and the need for nuanced analyses of demographic and socioeconomic influences.

The decision tree model provides complementary insights by visually representing the hierarchy of influential predictors. Race emerges as the primary determinant, with initial splits differentiating voters based on racial identity. White, Native American, and Other racial groups are more likely to vote Democrat (node probability: 71%), while Black and Hispanic groups exhibit even stronger Democratic alignment in subsequent branches (node probability: 86%). Religion plays a critical role in further stratifying voters, with Protestant and Roman Catholic respondents favoring Republican candidates, while Jewish and unaffiliated voters lean Democratic.

Urbanicity and education also stand out in the decision tree results. Urban residents exhibit a high probability of Democratic support (node probability: 84%), contrasting with rural voters who show stronger Republican alignment (node probability: 55%). Educational attainment reinforces Democratic preferences, with postgraduate-educated individuals displaying the highest likelihood of Democratic voting (node probability: 86%). Conversely, high school graduates and those with lower education levels align more with Republican candidates (node probability: 35%).

Together, these findings from both models reinforce the layered and complex interplay of demographic, socioeconomic, and geographic factors in determining voting preferences. The logistic regression model quantifies the strength of individual predictors, while the decision tree model provides an intuitive framework for understanding their hierarchical interactions.

5.2 Implications for Policy and Political Strategy

The results emphasize the need for tailored political strategies that address the distinct preferences of urban and rural voters. Campaigns targeting urban voters should focus on policies that resonate with younger, more diverse, and more highly educated populations, as these groups are more likely to align with Democratic platforms. Conversely, addressing economic and cultural concerns could help campaigns engage rural voters and individuals in higher income tiers, who lean more Republican.

Religious affiliations also present opportunities for targeted outreach. Campaigns should adopt culturally sensitive messaging when engaging with Protestant and Catholic communities, which may help reduce the Democratic Party's deficits in these voter groups. Similarly, addressing educational disparities and improving access to higher education could reshape long-term political alignments, particularly in regions or demographic groups where educational attainment is relatively low.

Finally, urban-rural divides continue to shape voting behavior, with urban voters exhibiting significantly different political preferences compared to their rural counterparts. For policymakers, recognizing these divides and crafting inclusive policies that address the needs of both urban and rural communities is critical. At the same time, disparities in education and race remain pressing issues. Expanding access to higher education, particularly for marginalized communities, could reduce inequalities and reshape voting patterns in the long term.

In conclusion, this study demonstrates the importance of integrating demographic, socioeconomic, and geographic variables into analyses of voter behavior. It calls for future research to further explore the intersections of these factors, particularly the role of education and race in influencing political preferences. Such insights can guide more effective policy development and political strategy, ensuring that diverse voter needs are addressed.

5.3 Data and Temporal Limitations

A key limitation of this study lies in the temporal and demographic scope of the dataset. The analysis focuses on the 2022 Cooperative Election Study (CES), capturing data from U.S. residents regarding voting behavior during the 2022 election cycle. However, this timeframe does not encompass more recent political shifts, such as the 2024 election, or longer-term trends that might affect voting outcomes in the future. The absence of data from the subsequent election period or extended timeframes may lead to an underrepresentation of emerging patterns or the longer-term impacts of policy changes, which could influence voter behavior in future elections.

Additionally, the dataset utilized for this analysis is based on aggregated data at the state level, which could mask localized effects or intra-state disparities. Differences in political dynamics within individual states or between urban and rural areas, for example, could alter voting patterns and are not fully captured in the analysis. Future studies could benefit from more granular data, such as county-level information, to better understand how localized factors influence voting behavior. This would offer a more comprehensive understanding of the complex relationships between demographic factors, political policies, and voting preferences.

Moreover, relying solely on self-reported demographic information presents another limitation. While race, education, and other demographic factors are critical to understanding voting behavior, self-reports can be subject to bias or misinterpretation. The potential for respondents to misidentify their race or education level could skew the results, particularly for minority groups or those with non-traditional education pathways. Future research should consider supplementary qualitative approaches, such as interviews or focus groups, to address these challenges and ensure more accurate data collection methods.

5.4 Weaknesses and Future Directions

While this study provides meaningful findings into how demographic factors influence voting behavior, there are several limitations that should be addressed in future research. One key limitation is the use of data from a single point in time (2022), which does not account for changes in voter preferences over time. This analysis assumes that the patterns observed during this period are static, but voting behavior can change due to shifts in political climate, policy changes, or societal events. Future research could use longitudinal data from multiple elections to examine how demographic variables affect voting behavior over time, offering a deeper understanding of long-term trends.

Additionally, the study relies on self-reported demographic data, which can be subject to biases such as misreporting or respondents' unwillingness to disclose certain information. For example, individuals may not accurately report their race, gender, or educational background, leading to potential inaccuracies in the analysis. To improve data accuracy, future research could consider using administrative data or official government records, which may provide more reliable information on demographic characteristics. Combining quantitative surveys with qualitative research methods, such as interviews or focus groups, could also provide richer insights into how people perceive and act on the demographic factors influencing their vote.

These efforts will support the development of more nuanced, evidence-based policies aimed at increasing voter engagement and addressing disparities in political representation across different demographic groups. By building on these limitations, future studies can enhance our understanding of the complex dynamics of voter behavior and contribute to more inclusive democratic processes.

Appendix

A Additional data details

A.1 Dataset and Graph Sketches

Sketches depicting both the desired dataset and the graphs generated in this analysis is available in the GitHub Repository `other/sketches`.

A.2 Data Cleaning

The CES 2022 dataset was carefully cleaned and processed to prepare it for analysis. Key variables such as `vote_choice`, `age_cohort`, `income_tier`, `education`, `gender`, `religion`, `race`, `urbanicity`, and `region` were retained, reflecting the primary demographic, socioeconomic, and geographic factors relevant to voting behavior. Observations with missing or invalid values in these variables were excluded to ensure consistency and accuracy in the analysis. Additionally, only registered voters (`votereg == 1`) were included to align the dataset with the study's focus on electoral participation.

Categorical variables, such as `age_cohort`, `gender`, and `region`, were transformed into factors to enable proper handling in the logistic regression model. Continuous variables like `age` were grouped into cohorts to facilitate meaningful comparisons across age groups. The cleaned dataset was saved in both CSV and Parquet formats, ensuring compatibility with statistical tools and efficient data management. These preprocessing steps provided a robust foundation for the logistic regression analysis, enabling a comprehensive research of how demographic and contextual variables influence voting behavior.

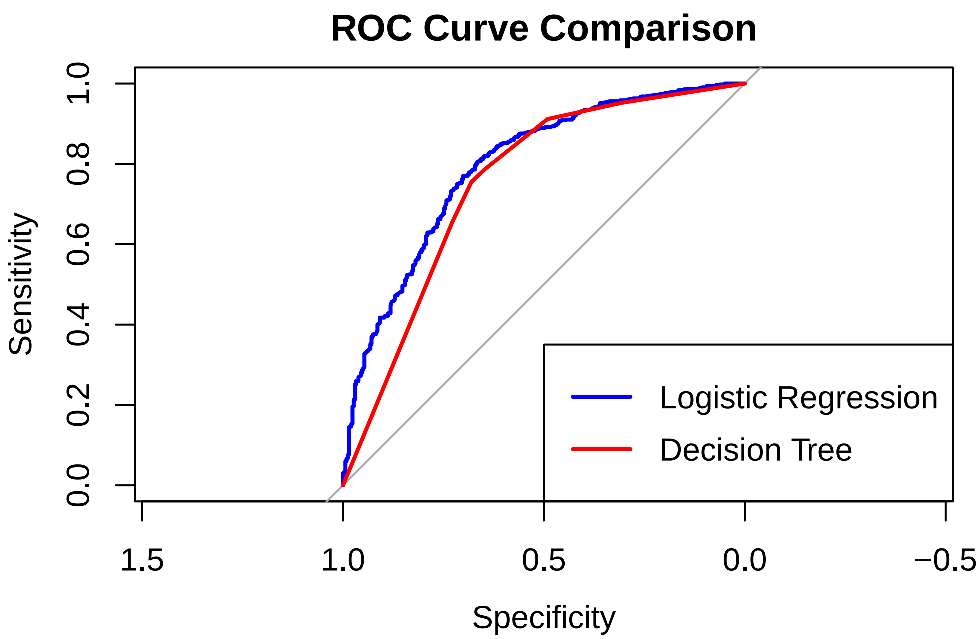
A.3 Data Source Acknowledgment

The data utilized in this study was sourced from the Harvard Dataverse. Access to the data and its use comply with the terms outlined in the Harvard Dataverse data use agreement. Specifically, the data was used for academic research purposes, with acknowledgment of the original data contributors as the source. The data contributors and Harvard Dataverse, however, do not assume responsibility for any analyses, interpretations, or conclusions drawn from the data by the authors of this study.

B Model details

B.1 Model Validation: K-Fold Cross-Validation & ROC Curve & Log Loss and Pruning

```
  parameter  Accuracy     Kappa AccuracySD    KappaSD
1      none 0.7674634 0.3847542 0.03083402 0.08155464
```




```
Log-Loss: 0.4769
```

The logistic regression model was evaluated using 10-fold cross-validation to assess its predictive performance. The model achieved an accuracy of 0.762, indicating that it correctly classified approximately 76.2% of the observations. The kappa statistic was 0.372, reflecting fair agreement between the model's predictions and the actual outcomes. The log-loss, which measures the model's calibration, was calculated as 0.477, suggesting the predicted probabilities align reasonably well with observed outcomes. The ROC curve comparison showed that the logistic regression model demonstrated stronger discriminatory ability compared to the decision tree, with an AUC that highlights its effectiveness in distinguishing between outcomes. Despite these strengths, there remains room for improvement in enhancing classification accuracy and reducing residual error, potentially by incorporating additional predictors or exploring more complex modeling techniques.

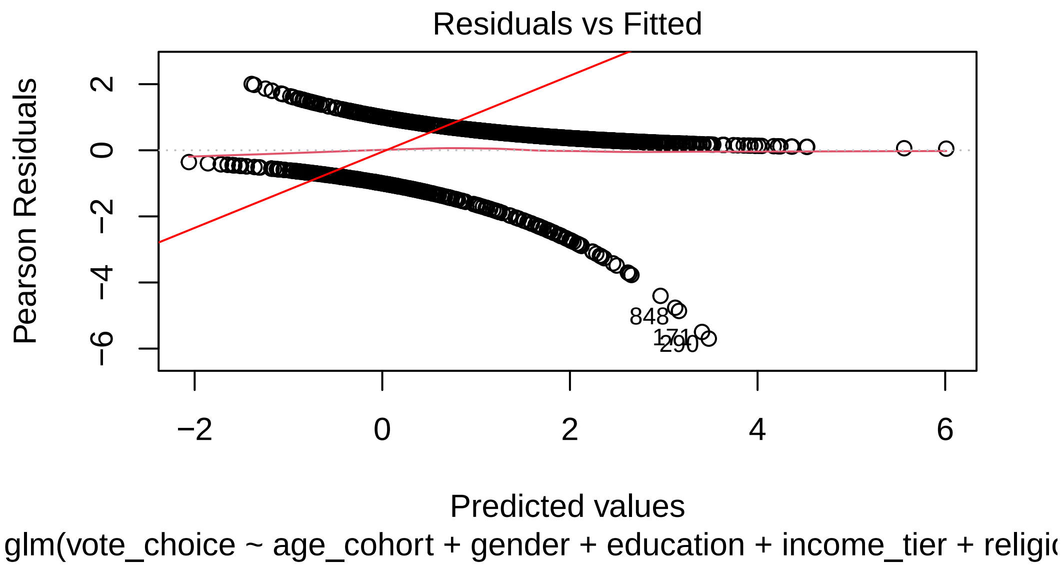


Figure 4: Diagnostics of Logistic Rregerssion model for vote support using residual vs fitted plot

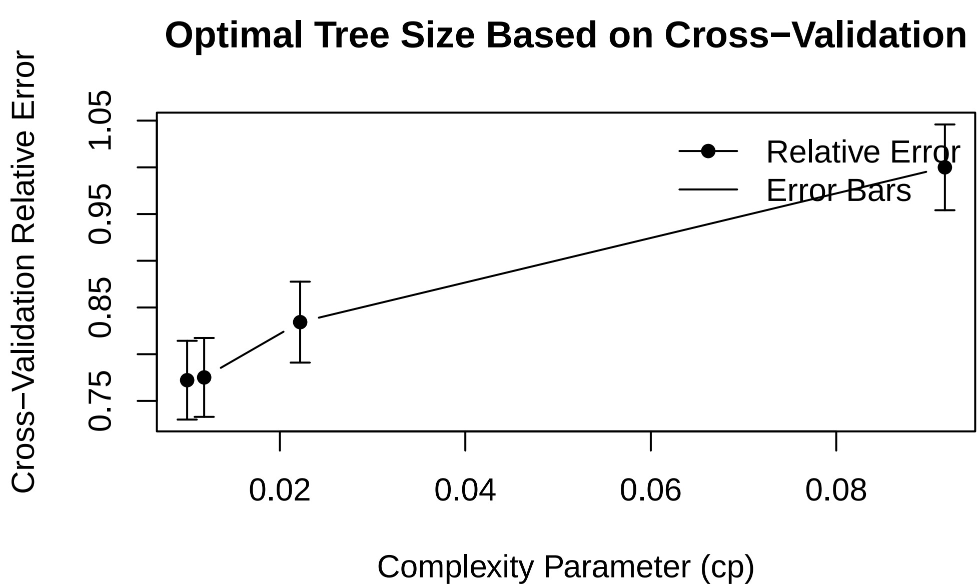


Figure 5: Cross-Validation Results for Decision Tree Pruning

B.2 Diagnostics

Table 6: Complexity Parameter Table for Decision Tree Model

CP	Number Splits	Relative Error	X Error	X Std
0.0917160	0	1.0000000	1.0000000	0.0458998
0.0221893	2	0.8165680	0.8343195	0.0433068
0.0118343	5	0.7396450	0.7751479	0.0422082
0.0100000	6	0.7278107	0.7721893	0.0421507

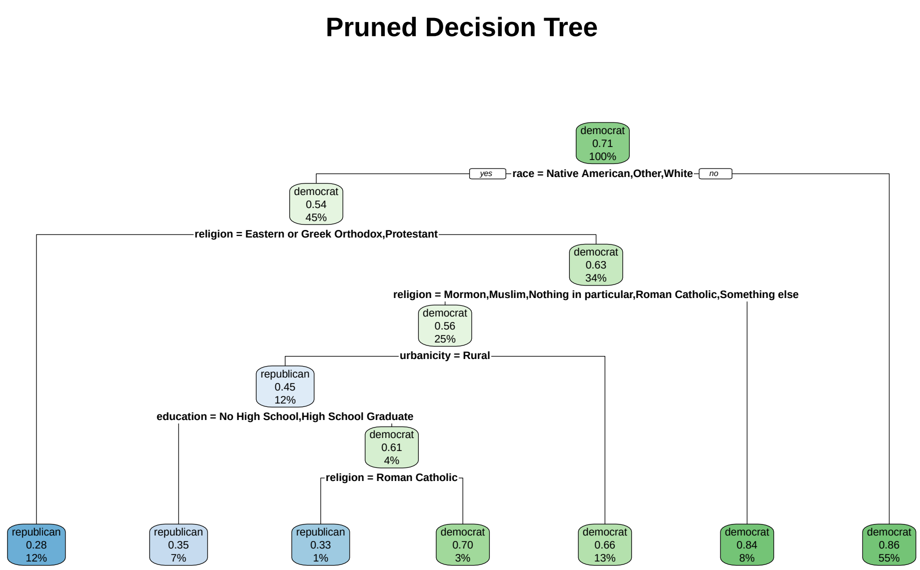


Figure 6

The Residual vs. Fitted plot (Figure 4) for the logistic model in Section 3.1.1 shows residuals plotted against the fitted values. Residuals, representing the differences between observed and predicted outcomes, offer insight into model fit. Ideally, they should be randomly scattered around the zero line, indicating no systematic errors or poor fit.

The Complexity Parameter (CP) Table (Table 6) and the Cross-Validation Error Plot (Figure 5) for the decision tree model in Section 3.1.2 highlight how the decision tree balances complexity and performance. The optimal CP value of 0.010000 minimizes cross-validation error (X Error = 0.7721893) while maintaining low variability, as shown in the error bars. The Pruned Decision Tree (Figure 6) , based on this CP, provides a simplified structure that identifies key predictors such as race, religion, and education, avoiding overfitting and improving interpretability. These diagnostics confirm that the decision tree achieves a good balance between accuracy and simplicity.

C Idealized Methodology for A Survey-Based Qualitative Studies

Our study explores the relationship between political attitudes and voter behavior in the U.S., utilizing data from the Cooperative Election Study by Schaffner, Ansolabehere, and Shih (2023). By combining observational survey analysis with targeted qualitative insights, we aim to capture both quantitative trends and nuanced voter experiences. While the CES dataset provides a comprehensive framework to assess demographic, geographic, and ideological patterns, qualitative surveys supplement these findings by examining factors such as voter motivations, information sources, and decision-making processes. This combined approach enhances our understanding of systemic influences and personal perspectives, enabling more robust interpretations of observed trends and informing evidence-based political strategies and policies.

C.1 Introduction

This appendix outlines the methodology used to analyze the 2022 Cooperative Election Study (CES) dataset. The CES provides a rich source of survey data encompassing political attitudes, behaviors, and demographic characteristics of U.S. residents. This methodology ensures a systematic approach to data preparation, analysis, and interpretation, facilitating robust and replicable results.

Our survey focuses on individuals across different socioeconomic and demographic groups to understand how economic conditions, educational attainment, and racial identity shape voting behavior and political preferences. By integrating quantitative data from the 2022 Cooperative Election Study (CES) with qualitative insights, this study aims to uncover the nuanced interplay of economic realities and demographic characteristics. The findings will inform evidence-based interventions, addressing systemic inequities in political participation and advancing our understanding of how diverse lived experiences influence electoral outcomes.

C.2 Objective

The objective of this study is to investigate the socioeconomic, demographic, and systemic factors shaping voter behavior and political preferences in the United States. By focusing on the interplay between economic conditions, educational attainment, and racial identity, the study aims to uncover how these factors influence voting decisions and participation. Understanding how economic realities and demographic traits interact to affect political behavior is important for addressing systemic barriers to equitable representation and democratic engagement. The findings will guide the development of evidence-based strategies to enhance voter participation, promote inclusive political representation, and inform policies that address socioeconomic disparities in the electorate.

C.3 Sampling Approach

In this analysis, we use matched random sampling, a methodology employed by Schaffner, Ansolabehere, and Shih (2023) through YouGov and Team (2022). This method ensures a representative sample of the U.S. population by matching respondents from an opt-in panel to a target sample drawn from demographic benchmarks such as the Bureau (2019). The matched random sampling approach is particularly effective for large-scale studies, leveraging statistical adjustments to mitigate biases and improve sample representativeness.

Matched random sampling is ideal for our study as it allows us to capture a diverse range of socioeconomic and demographic characteristics across the electorate. This approach ensures that key variables, such as education, income, race, and urbanicity, are well-represented in the dataset, enabling robust analysis of the interplay between these factors and voting behavior. By combining rigorous sampling techniques with advanced weighting methods, we enhance the reliability and validity of our findings, ensuring that they reflect the broader U.S. electorate. These strengths make matched random sampling an ethical and effective method for examining the systemic, economic, and demographic factors shaping political preferences and participation.

C.4 Target Population

Our target population comprises U.S. voters across diverse socioeconomic, educational, and racial demographics, as captured in the 2022 Cooperative Election Study by Schaffner, Ansolabehere, and Shih (2023). Specifically, we focus on individuals whose voting behavior is shaped by their economic conditions, educational attainment, and racial identity. This includes voters from various states, regions, and urbanicity levels to ensure a comprehensive analysis of how these factors interact to influence political preferences and participation.

C.5 Sample frame

The sample frame for this study is derived from the 2022 Cooperative Election Study dataset by Schaffner, Ansolabehere, and Shih (2023), which includes responses from 60,000 individuals recruited through YouGov's matched random sampling methodology. The CES sample frame utilizes a politically representative modeled frame based on the 2019 American Community Survey by Bureau (2019), voter registration files, and demographic data such as age, gender, race, education, and region.

This sample frame ensures coverage of diverse demographic and socioeconomic groups across the United States, enabling the study to examine how variables such as economic conditions, educational attainment, and racial identity interact to influence voting behavior. The sample frame is specifically structured to include a wide range of political preferences and behaviors, providing the representativeness necessary to draw meaningful insights about the interplay between systemic factors and voter participation.

C.6 Sample

We aim to survey 1,000 respondents who meet our defined sample criteria: eligible U.S. voters representing diverse racial, economic, and educational backgrounds. The sample will specifically target individuals from varying income levels, educational attainment, and racial groups to explore the interplay of these factors in shaping voting behavior. Participation will be voluntary, with respondents required to answer survey questions truthfully and comprehensively to ensure data quality and depth.

C.7 Recruitment of Respondents

To recruit participants for this study, we will use a stratified sampling approach in collaboration with an online survey platform, such as YouGov, which specializes in matched random sampling to ensure representativeness. The initial recruitment will focus on demographic and geographic diversity, ensuring proportional representation of racial groups, income brackets, and urbanicity levels.

Outreach materials will emphasize the study's purpose of understanding socioeconomic influences on voting behavior while assuring respondents of confidentiality and anonymity. Eligible participants will be invited to complete a screening survey to ensure alignment with the study's inclusion criteria. The survey will be distributed online for accessibility and convenience, with respondents able to participate from any location while maintaining their privacy. To encourage participation, respondents will receive modest compensation, further ensuring meaningful engagement with the study.

This approach enables the collection of a representative dataset, capturing the nuances of economic and demographic factors influencing voter behavior.

C.8 Handling Non-response bias

Non-response bias is a critical concern in survey research, particularly when studying voter behavior and socioeconomic factors. Participants who do not respond or drop out may differ significantly from those who complete the survey, leading to skewed conclusions. To address this, we will emphasize the importance of the study, ensure anonymity, and provide a straightforward, user-friendly survey experience that takes approximately 5–10 minutes to complete. Outreach efforts will also include reminders and incentives, such as modest compensation, to encourage higher participation rates, particularly among underrepresented groups.

C.9 Respondent Validation

To ensure the reliability and credibility of the collected data, we will implement a rigorous respondent validation process. Eligibility screening questions will verify participants' age, voting eligibility, and demographic characteristics such as income, education, and racial identity. Responses will be reviewed for completeness, logical consistency, and alignment with inclusion criteria. Additionally, weights will be applied to adjust for imbalances in demographic representation, ensuring the data accurately reflects the broader U.S. electorate. By leveraging a reputable survey platform like YouGov and integrating advanced quality checks, we aim to maintain data integrity and draw meaningful, representative insights.

C.10 Ethical Concerns

This study involves exploring sensitive topics such as voting behavior, socioeconomic disparities, and demographic influences, necessitating an ethical framework to safeguard participants' privacy and ensure fairness. Recognizing the potential discomfort participants might experience when sharing personal information about their economic conditions, political preferences, or demographic characteristics, the survey will provide clear explanations of its purpose and allow participants to skip questions or withdraw at any time without consequences.

Strict confidentiality measures will be in place to protect participants' identities, with responses securely stored and anonymized to prevent re-identification. Recruitment through reputable platforms and organizations will foster trust, and respondents will be informed of their rights throughout the process. Additionally, transparency in reporting and ethical data usage will ensure the findings are used responsibly to advance understanding without perpetuating harm or bias. This ethical framework underscores our dedication to conducting inclusive, respectful, and socially responsible research.

C.11 Proposed Survey Design

Investigating the relationship between education, economic conditions, and voter behavior necessitates a meticulously designed survey to ensure accurate and unbiased data collection. Given the politically polarized environment, this study acknowledges the potential influence of social desirability bias and the perceived sensitivity of questions addressing socioeconomic and political preferences. Drawing on best practices outlined by Stantcheva (2023) and Fowler (1995), this survey integrates strategies to minimize bias, enhance respondent comfort, and optimize data accuracy.

This survey examines the intersection of economic conditions, education, and racial identity with voting preferences. It employs neutral, inclusive phrasing and an anonymous, online format to reduce respondent apprehension. Inspired by proven methodologies, the design incorporates randomized response options, opt-out choices (e.g., "Prefer not to say"), and a combination of multiple-choice and open-ended questions to collect both structured data and nuanced insights. To encourage honest and thoughtful participation, the survey employs a "contribution" framework in its introductions, emphasizing how participants' input helps improve understanding of voting behavior and inform public policy. This approach fosters trust, engagement, and accurate responses, ensuring high-quality data that reflects the diverse realities of voter experiences.

C.12 Solutions to Response Bias in Our Survey

Drawing on recommendations from Stantcheva (2023) and Fowler (1995), this survey addresses common response biases, including moderacy bias, extreme response bias, response order bias, acquiescence bias, experimenter demand effects (EDE), and social desirability bias (SDB). This study focuses particularly on moderacy bias, extreme response bias, response order bias, and SDB. Detailed definitions of these biases are provided in Appendix- C.14.

To mitigate these biases, the survey implements the following strategies:

- **Addressing Extreme/Moderacy Bias**: A minimum of five response options is included for scale-based questions, offering participants a range of detailed choices to discourage defaulting to extreme or neutral answers.
- **Mitigating Response Order Bias**: Randomizing response options in nominal questions eliminates potential biases caused by the order of presented choices.
- **Minimizing Social Desirability Bias (SDB)**: Emphasizing anonymity and confidentiality throughout the survey reassures participants and reduces SDB. The introduction clearly outlines the survey's purpose as an academic study on voter behavior and ensures that responses will remain confidential and solely used for research. This anonymous online format creates a safe space for participants to share their perspectives without fear of judgment or stigma.
- **Encouraging Honest Feedback**: A feedback section at the survey's conclusion invites participants to express concerns or provide additional insights. This fosters trust and strengthens the data's reliability and depth.

By employing these strategies, this survey ensures a robust and reliable approach to studying voting behavior, collecting comprehensive data that contributes meaningfully to understanding the sociopolitical dynamics shaping voter preferences.

C.12.1 Survey Link

The survey has been implemented using Google Forms. You can access it here: Survey Link.

C.13 Copy of Survey on Generational and Demographic Determinants of Voting Behavior

Welcome Section

Introduction: Welcome to our study on voting behavior across different generations,regions, and demographic backgrounds. Your participation will help us understand how age, geographic location, and demographic characteristics influence political affiliation and voter preferences.Rest assured that your responses are anonymous and will only be used for academic research purposes.

This survey is conducted by researchers at the University of Toronto and is part of a broader study on voting behavior and generational change. It consists of 17 questions and should take approximately 10–15 minutes to complete.

Please answer the questions honestly. If you experience any discomfort while completing the survey, you may stop at any time.

Contact Information: Jinyan Wei Email: jinyan.wei@mail.utoronto.ca

Section 1: Demographics and Background Information

1. What is your age?
   - Under 18
   - 18–24
   - 25–34
   - 35–44
   - 45–54
   - 55-64
   - 65+
   - Prefer not to say
2. What is your gender?
   - Man
   - Woman
   - Non-binary
   - Prefer not to say
3. What is your highest level of education?
   - No HS
   - High school graduate or equivalent
   - Some college
   - Bachelor's degree
   - Graduate or professional degree
   - Prefer not to say

4. What is your race or ethnicity?
    - White
    - Black or African American
    - Hispanic or Latino
    - Asian
    - Native American or Alaska Native
    - Middle Eastern or North African
    - Other (please specify): ______________
    - Prefer not to say
5. What is your household income level(in dollar)?
    - Below 10,000
    - 10,000 - 24,999
    - 25,000 - 49,999
    - 50,000 - 74,999
    - 75,000 - 99,999
    - 100,000 - 149,999
    - 150,000 - 199,999
    - 200,000 or above
    - Prefer not to say
6. What is your primary region of residence?
    - Northeast
    - Midwest
    - South
    - West
    - Prefer not to say
7. Do you live in an urban, suburban, or rural area?
    - Urban
    - Suburban
    - Rural

Section 2: Understanding Political Affiliations

1. What is your current political affiliation?
    - Democrat
    - Republican
    - Independent

- Prefer not to say

2. In the last election, did you vote?
    - Yes
    - No
    - Prefer not to say

3. How important is your political affiliation to your identity?
    - Not important
    - Slightly important
    - Moderately important
    - Very important
    - Extremely important

4. How likely are you to participate in the next election?
    - Very likely
    - Somewhat likely
    - Not likely
    - Prefer not to say

5. How important is a candidate's stance on social issues (e.g., healthcare, education)?
    - Very important
    - Somewhat important
    - Not important

6. How would you rate the influence of regional factors (e.g., state policies) on your voting behavior?
    - High influence
    - Moderate influence
    - Low influence

Section 3: Generational and Regional Perspectives

1. How much do you agree with the following statement: "Generational differences significantly influence voting behavior."

- Strongly agree
- Agree
- Neutral
- Disagree
- Strongly disagree

2. Do you believe urban-rural divides play a major role in political alignment?
   - Yes
   - No
   - Not sure
3. Do you feel adequately informed about the policies of political candidates in your region?
   - Yes
   - No
   - Not sure
4. Do you think your age cohort has different priorities compared to older or younger generations when it comes to voting?
   - Yes
   - No
   - Unsure

Section 4: Resources and Support

If you feel distressed or need support after completing this survey, the following resources are available:**

1. Mental Health America (MHA)
   - Website: www.mhanational.org
   - Services: Online screening tools, support networks, and educational materials.
2. National Helpline for Mental Health
   - Hotline: 1-800-662-HELP (4357)

- Services: Free, confidential referrals for mental health support.

Section 5 : Feedback

1. Do you have any concerns or feedback regarding the survey, surveyor, or entity?
    - Your feedback is important to us and will help ensure transparency and trust in the research process.

Section 6: Thank You

Thank you for taking the time to complete this survey. Your honest feedback is invaluable and will help us better understand and address the experiences of women who have faced similar circumstances. We deeply appreciate your participation and the courage it takes to share your experiences.

C.14 Response Bias Definition

In survey research, response bias refers to the systematic tendency of survey respondents to answer questions inaccurately or falsely, leading to distorted data and potentially invalid conclusions. Response bias can arise from various factors, including question phrasing, respondent motivations, and survey administration methods. Definitions below are defined by Fowler (1995) and Converse and Presser (1986).

Some common types of response bias include:

- **Social Desirability Bias**: The tendency of respondents to answer questions in a manner that will be viewed favorably by others, often leading to overreporting of socially desirable behaviors and underreporting of undesirable ones.
- **Acquiescence Bias**: Also known as "yea-saying," this is the inclination of respondents to agree with statements regardless of their content, resulting in a disproportionate number of affirmative answers.
- **Extreme Response Bias**: The propensity to use the extreme ends of a response scale, such as "strongly agree" or "strongly disagree," more frequently than the middle options, which can skew the data toward more polarized responses.
- **Moderacy Bias**: The tendency to avoid extreme responses and consistently select middle or neutral options on a scale, potentially masking true variations in opinions or behaviors.
- **Question Order Bias**: Occurs when the sequence of questions influences responses, as earlier questions can provide context that affects answers to subsequent ones.

- **Nonresponse Bias**: Arises when certain groups are underrepresented because they are less likely to respond to surveys, leading to results that do not accurately reflect the target population.

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

Brownrigg, Raymond E., Thomas P. Minka, Alex Deckmyn, and Robert Hijmans. 2023. *Maps: Draw Geographical Maps*. https://CRAN.R-project.org/package=maps.

Bureau, U. S. Census. 2019. "American Community Survey." https://www.census.gov/programs-surveys/acs/.

Converse, Jean M., and Stanley Presser. 1986. *Survey Questions: Handcrafting the Standardized Questionnaire*. Beverly Hills, CA: SAGE Publications.

Fowler, Jr., Floyd J. 1995. *Improving Survey Questions: Design and Evaluation*. Applied Social Research Methods Series Volume 38. Thousand Oaks, CA: SAGE Publications.

Kuhn, and Max. 2008. "Building Predictive Models in r Using the Caret Package." *Journal of Statistical Software* 28 (5): 1–26. https://doi.org/10.18637/jss.v028.i05.

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/.

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://CRAN.R-project.org/package=arrow.

Robin, Xavier, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller. 2023. *pROC: Display and Analyze ROC Curves*. https://CRAN.R-project.org/package=pROC.

Schaffner, Brian, Stephen Ansolabehere, and Marissa Shih. 2023. "Cooperative Election Study Common Content, 2022." Harvard Dataverse. https://doi.org/10.7910/DVN/PR4L8P.

Stantcheva, Stefanie. 2023. "How to Run Surveys: A Guide to Creating Your Own Identifying Variation and Revealing the Invisible." *Annual Review of Economics* 15: 205–34. https://doi.org/10.1146/annurev-economics-091622-010157.

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.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

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.

YouGov, and Cooperative Election Study Team. 2022. "2022 Cooperative Election Study: Methodology Report." https://cces.gov.harvard.edu/.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.