Modeling Voter Preference Dynamics: Education, Income, and Gun Ownership in the 2020 US Presidential Election*

Yang Zhou

April 5, 2024

This paper explores the relationship between voting preferences and the predictors of education, income, and personal gun ownership in the United States through analyzing 2022 US election database. By examining data from Cooperative Election Study, we identify clear patterns indicating that higher income are more likely to vote for Biden. Besides, the higher personal gun ownership will also significantly increase the likelihood of voting for Biden. By contrast, the people with higher education level show less probability to vote for Biden. Our findings underscore the critical impact of socio-economic status and personal security concerns on electoral outcomes, shedding light on the underlying dynamics that shape voter behavior in contemporary American politics. Ultimately, this research contributes to our understanding of the complex factors that drive electoral decisions, offering valuable insights for policymakers, political strategists, and citizens aiming to foster more informed and equitable democratic processes.

Table of contents

*Code and data are available at: <https://github.com/yangzhoucoco/Political-support-in-the-United-State.git >

1 Introduction

2020 United States Presidential Election stood as an important political event, reflecting societal divisions and highlighting the influence of various demographic factors on voting behavior. This election, which culminated in the victory of Joe Biden over incumbent Donald Trump, has spurred a renewed interest in understanding the dynamics of voter preferences and the underlying factors that drive electoral decisions. Our study aims to dissect the relationship between voters' educational background, income levels, personal gun ownership, and their voting preferences, employing logistic regression analysis to reveal the potential associations between different factors.

The Cooperative Election Study (Schaffner, Ansolabehere, and Luks 2021)conduct a throughout survey encompassing a wide range of voter demographics and attitudes.CES collaborate with 60 researches teams to assembles sample comprising 60,000 cases. The cases are based on the survey of adult pariticipants in 2022 fall. YouGov was commissioned to help the teams to conduct national sample surbey. They conducted the survey before and after elections. The first round survey were completed from September 29 to November 8, 2022 and the second period is gathered from November 10 to December 15, 2022. By offering an throughout views

to different constituencies, voter demographics and voter behaviors, CES reveals the complexity of electoral behavior and voter demographics. By understanding voter behaviors,, we gain valuable insights that pave the way for refining and advancing democratic practices.

In addition, by leveraging data from the 2022 CES, this paper delves into how variables such as education, income, and gun ownership serve as predictors for electoral choices, specifically the likelihood of voting for Biden or Trump. I used a logistic regression model to identify the likelihood. This analysis not only reveal the individual impact of these factors but also contributes to a broader understanding of the socio-economic and cultural impact of American electoral behavior.

Based on our model, I found that individuals with middle incomes demonstrate a greater propensity to vote for Biden. Additionally, elevated levels of personal gun ownership significantly correlate with an increased likelihood of supporting Biden. Conversely, individuals possessing higher levels of education exhibit a reduced probability of voting for Biden.

To be more specific, this paper is structured as follow: in {#sec-data}, I will introduce the data variable and how the data is used for the overall analysis.Visualization would also be included. In {#sec-model}, I will set up the model to predict the relationship between the predictors and electoral outcomes. Model justification will provide a rationale for selecting logitics model. {#sec-results} is focused on explain the result of the model. Lastly, {#sec-Discussion} provides a discussion on what we found in the model. I will also talk about the weakness of this paper along with the further study on this topic.

2 Data

2.1 Source

The 2022 Cooperative Election Study (CES) (Schaffner, Ansolabehere, and Luks 2021) was conducted by YouGov. In order to ensure the representativesness of the data, the survey utilized sampling methodology, matching process, weighting procedure.

I will uses the programming language `R`(R Core Team 2023) to analyze the data. `dplyr`(Wickham et al. 2023), `tidyverse`(Wickham et al. 2019), `ggplot2`(Wickham 2016), `knitr` (**knitr?**), `here` (Müller 2020) , and `kableExtra` (Zhu 2024),`modelsummary`(Arel-Bundock 2022), and `rstanarm`(Brilleman et al. 2018) will help me in the model and visualization part.

2.2 Sampling

The survey interviewed 60,000 adults during two periods: pre-election from September 29 to November 8, 2022, and post-election from November 10 to December 15, 2022. YouGov

employed matched random sample methodology for the sampling process. Sample matching is used to create representative samples from non-random pools of respondents and it is suitable for online survey. First,it will draw a random target population and then, it is going to select the matching respondents from the pool which matched with the target sample's characteristics. In order to replicate the target sample's attributes, YouGov uses proximity matching to calculates the likelihood or closeness between characteristics of the target samples; they will adjust the weight of variables when they need.

2.3 Weightening

To correct the imbalance between the matched samples and the overarching target demographic, a two-phase weighting procedure is employed, ensuring the CES samples accurately mirror the population's diversity. The initial phase of this adjustment process involves entropy balancing, alongside iterative proportional fitting (often termed 'raking'), to align the sample with the population's demographic and political characteristics. This includes a comprehensive evaluation of variables and their interrelations. The approach guarantees that the common content of the survey represents each state accurately, incorporating adjustments for statewide electoral contests.

2.4 Vote Validation

Vote validation have been considered in the CES in order to verify the accuracy of the voting behaviors. The sample were matched to the TargetSmart database that include the registered voter responds. Only the records which have a high level of confidence could be matched. This process is helpful for identifying voters and the way they vote, such as absentee, early, and mail.Therefore CES can ensure the data is focused on voting behavior and accurately represent genuine voter engagement.

2.5 Variables

After installing `dataverse`(Kuriwaki, Beasley, and Leeper 2023), I use `get_dataframe_by_name` to analyze CES. I select four variables that most relevant to our topic, which are `votereg`, `presvote20post`, `educ`, `faminc_new`, `gunown`. `votereg` represents the voter registration status and only the voter who registered to vote would be considered. Besides, `presvote20post` represents the president that respondents vote in 2020 president election. `educ` means the education level of the respondents. In addition, `faminc_new` accounts for the family's annual income in 2021 year. `gunown` records the personal gun ownership by asking the respondent to answer whether they or anyone in their household own a gun. However, for the raw data, there are six different answers, including choosing ,Joe Biden, Donald Trump, Jo Jorgensen, Howie Hawkins, other, and did not vote for President. I only consider the respond of Joe Biden and Donald Trump. Therefore, when I clean the data, I select the group of Joe Biden

Table 1: Brief summary of variables types

votereg	presvote20post	educ	faminc_new	gunown
1	1	6	11	3
1	1	3	8	3
1	1	5	6	3
1	1	6	11	3
1	6	6	9	3
1	2	5	7	3
1	1	2	1	3
1	1	6	11	3
1	1	5	3	3
1	1	6	4	3

Figure 1: Brief summary of variables types

and Donald Trump. In addition, there are 6 different education level, ranging from no high school, high school graduate, some college, 2-year, 4-year, and post-grad. I care about all of the education level, so I take all of those types into consideration. Moreover, there are 16 ranges of family income and the differences between each range is 10,000. To simiplied the data, I classifed them into 4 different types. For the income group which less than $10,000, $10,000 - $19,999, $20,000 - $29,999, and $30,000 - $39,999 are defined as "Low" income group. For those who range from $40,000 to $79,999 can seem as "Middle" income group. $80,000 to $199,999 can be considered as "High" income group. For the family earn $200,000 to $500,000 or more is very high income group.

Figure 2 appears to be a bar chart displaying the number of respondents categorized by their income level, and which of the two candidates—Biden or Trump—they voted for. There are four income categories presented: Low, Middle, High, and Very High, with an additional category labeled "NA" that likely represents respondents whose income level was not available or not specified.

In addition, Figure 3 shows the number of respondents classified by their personal gun ownership status and which presidential candidate they voted for, either Biden or Trump. The categories of gun ownership are: "No one in the household owns a gun," "Personally own a gun," "Someone in the household owns a gun," and "NA" (which likely stands for data not available).

Figure 4 combines the predictors of both income and education and displays the voting preferences (for Biden in red and for Trump in blue) of respondents broken down by their level of education (some college, post-grad, no high school (HS), high school graduate, 4-year degree,

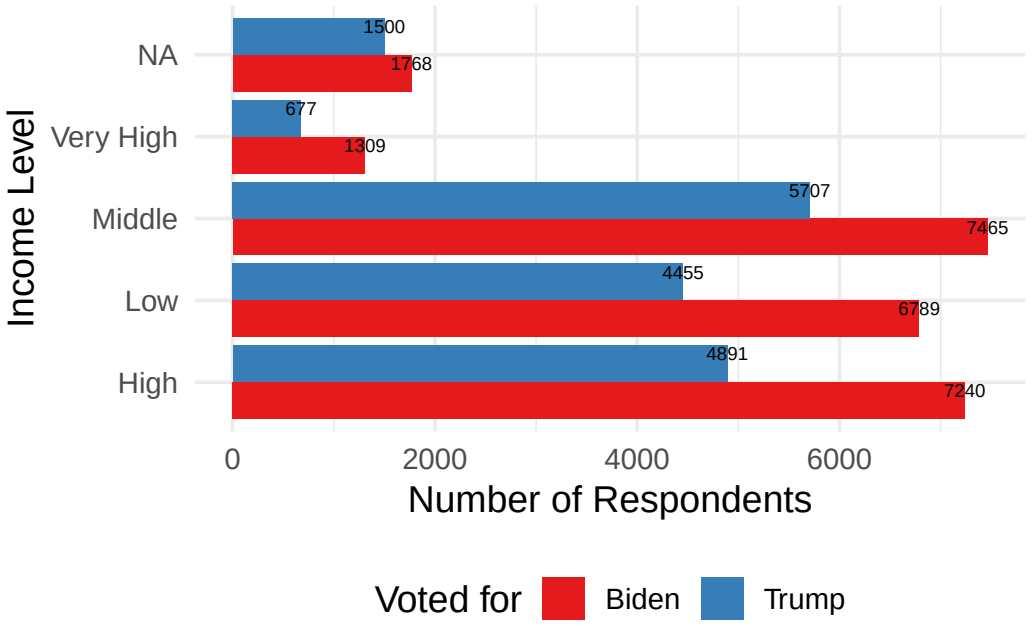

Figure 2: The distribution of presidential preferences, by income level

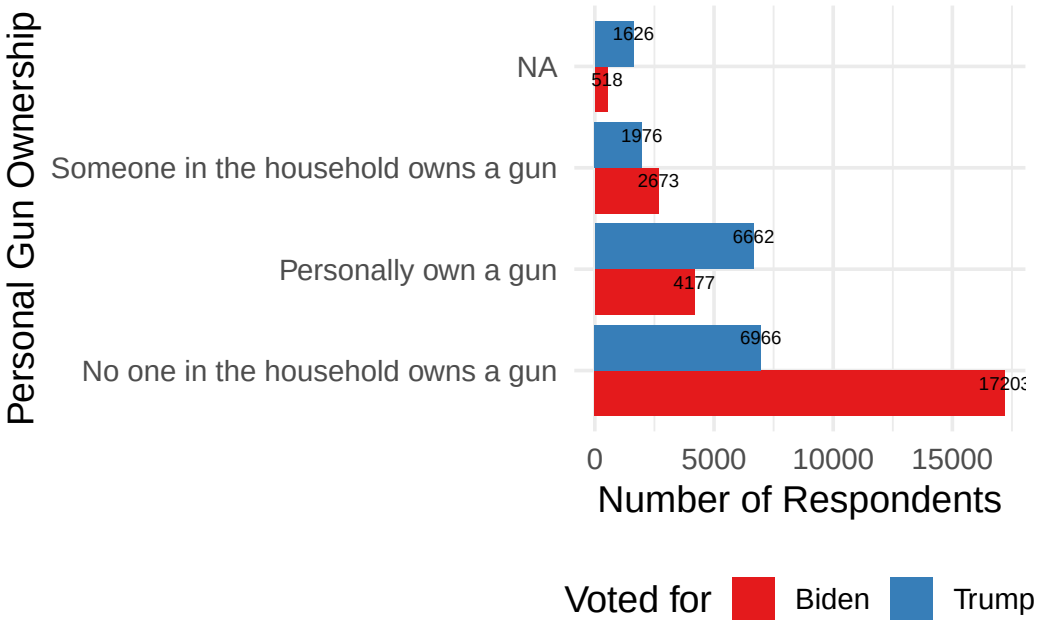

Figure 3: The distribution of presidential preferences, by gun ownership

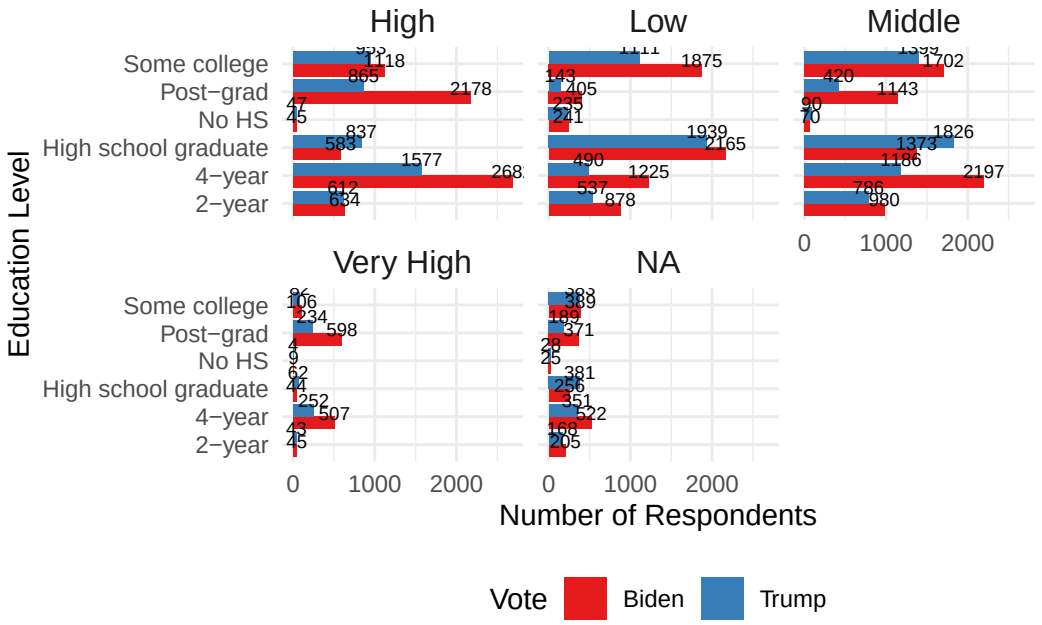

Figure 4: The distribution of presidential preferences, by education, and income level

2-year degree) and income level categories (high, very high, low, middle, and NA for data not available).

2.6 Measurements

I will interpret how `educ`, `faminc_new`, and `gunown` be measured. Respondents were asked a series of questions regarding their voting preferences, demographic information, and other relevant factors such as education, income, and gun ownership. The survey conduct by online questionnaires and each variable is correspond to a question. To be more specific, in terms of `educ`, the respondents are asked to answer "What is the highest level of education you have completed?" and the respondents can choose their education level. The answers are categorized into "no high school," "high school graduate," "some college," "2-year degree," "4-year degree," and "post-grad." However, the dataset fails to provide option for the respondents who graduate from secondary school or no education at all. `faminc_new` refer to the annual income for a family. People will be asked "Thinking back over the last year, what was your family's annual income?" The income ranges from less than $10,000 to more than $500,000. Some respondents might unwilling to disclose their actual income range. `gunown` indicates the status of gun ownership. In the course of gathering data on firearm ownership, the study must account for potential response biases. Specifically, individuals who may possess firearms

unlawfully could exhibit a tendency to provide untruthful responses, driven by a fear of legal consideration.

3 Model

3.1 Model set-up

Define y_i is the political preference of the respondent, where y_i equal to 1 indicates a preference for Biden and y_i equal to 0 represent a preference for Trump. The predictors include education_i, which indicates the education level of the respondent and income_i is the income level of the respondent. gunownership_i is the state of whether individual or the people he/she knows own a gun.

$$y_i|\pi_i \sim \text{Bern}(\pi_i) \tag{1}$$
$$\text{logit}(\pi_i) = \alpha + \beta \times \text{education}_i + \gamma \times \text{income}_i + \delta \times \text{gun}_i \tag{2}$$
$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\gamma \sim \text{Normal}(0, 2.5) \tag{5}$$
$$\delta \sim \text{Normal}(0, 2.5) \tag{6}$$

I set up a logistic regression model within a Bayesian framework to estimate model parameters, which particularly focusing on binary outcome variables such as political preference. To be more specific, I utilize `stan_glm()` function from `rstanrm` (**rstanrm?**) package, which can handles categorical variables efficiently by converting them into binary indicators. In order to optimize computational efficiency, I use a random sample of 500 observations to fit the model.

3.2 Model justification

Logistic regression is designed for binary outcome variables,especially when the dependent variables indicates two categories, in this case, the dependent variables are voting for Biden(1) or Trump(0). This model can provide a straightforward method for estimating the probability for voting for a specific candidate based on its predictors, including education level, income level, and gun ownership. In addition, Logistic regression is an ideal choice for explaining how unit changes in predictors variables affect the dependent variables.

Table 2: Explanatory models of political preferences based on education, income and gun ownership (n = 500)

	Support Biden
(Intercept)	−1.269
	(0.384)
educ4-year	−0.251
	(0.391)
educHigh school graduate	1.337
	(0.380)
educNo HS	1.131
	(0.936)
educPost-grad	−0.108
	(0.459)
educSome college	0.482
	(0.385)
income_groupLow	−0.207
	(0.285)
income_groupMiddle	0.316
	(0.266)
income_groupVery High	−0.954
	(0.680)
gunownPersonally own a gun	1.058
	(0.249)
gunownSomeone in the household owns a gun	0.280
	(0.346)
Num.Obs.	439
R2	0.157
Log.Lik.	−259.274
ELPD	−271.0
ELPD s.e.	9.3
LOOIC	542.1
LOOIC s.e.	18.6
WAIC	541.9
RMSE	0.45

4 Results

After conducting a logistic regression analysis on 500 samples, I summary the result in Table 2. According to Mustafa (2024), the coefficients of the model are referred to posterior means which provides a central tendency measure of parameter's posterior distribution. The coefficient represents the log odds of supporting Biden for the reference group while all other variables are held at their baseline values. I care about three predictors which are education level, income level, and gun ownership and focus on the relationship of political preference of different groups and their reference group. To be more specific, in terms of education level, the model interpret the relationship between the voters who are 4-year education, high school graduate, no high school, post graduate, some college and the reference group with a 2-year education. The reference group for income level is high income group range from $80,000 to $199,999 and its compared to other income group, including low income, middles income, and very high income. For the individual gun ownership, "no one in the household owns a gun" is the reference group. For example, the coefficient for voters with a 4-year education level is -0.251. This indicates that holding all other variables constant, the voters having a 4-year education compared to the reference group(2-year education) is associated with a slight decreases.

4.1 Education

According to Table 2, the respondents who have high-school degree(1.337), no high school degree(1.131), and voter have some college degree (0.482) are more likely to vote for Biden, whereas, the people with 4-year education(-0.251), or post-graduate degree (-0.108) are less likely to vote for Biden.

4.2 Income

The respondent with low income, ranging from less than $10,000 to $39,999, and very high income group, ranging from $200,000 to $500,000 or more, are less likely to vote for Biden. Their coefficients are -0.207 and -0.954 respectively. The middle income group is more likely to vote for Biden, ranging from $40,000 to $79,999.

4.3 Gun ownership

People own a gun(1.058) or know someone in the household owns a gun(0.280) have a larger likelihood to vote for Biden.

5 Discussion

5.1 Findings

After I used a Bayesian logistic regression model to find the relationship between Biden support and the predictors, including education level, income level, and individual gun ownership, I conclude three main findings shown below: 1. People have higher education level is less likely to vote for Biden than those of lower education level 2. Middle income group tend to vote for Biden 3. The voters who own gun or know someone who own gun exhibit a higher propensity to vote for Biden

5.2 People have higher education level is less likely to vote for Biden than those of lower education level

The summary data presented in Table 2 reveals a notable electoral dynamic: individuals possessing a four-year undergraduate degree or higher education credentials exhibit a lower propensity to support President Biden, in contrast to those whose highest educational attainment is a high school diploma or below, who show a marked tendency towards endorsing him. This pattern suggests that educational attainment may play a significant role in shaping political preferences, though it's important to acknowledge that a multitude of factors can influence an individual's voting decisions.

A critical consideration in unpacking these trends is the tax policy proposed by President Biden's administration, as outlined by (2024). The policy seeks to increase the marginal income tax rate to 39.6% for individuals whose annual earnings exceed $400,000. Notably, this income bracket is more likely to include individuals with advanced degrees. The people often command higher salaries because of their education level. Thus, the observed correlation between higher educational levels and diminished support for Biden might lead to considerations within this high earner regarding the potential impact of the proposed tax adjustments on their financial well-being.

Such concerns are not merely about the taxation but also about the perceived fairness and implications of these tax changes. High-earning individuals may worry about bearing a disproportionate tax burden. Additionally, this group might conduct the survey on the long-term economic impacts of higher taxes on investment, job creation, and overall economic growth.

Furthermore, this discussion highlights the complexity between education, income, and political ideology. While higher education often correlates with higher income, it also fosters exposure to diverse ideas and critical thinking skills that can influence political beliefs and policy preferences. Therefore, the relationship between education, income, and political preference might be complex.

5.3 Middle income group tend to vote for Biden

The analysis presented in Figure 2 compellingly demonstrates that Biden's support base comprises significantly of middle and high-income groups, with the middle-income segment being particularly dominate. This trend is substantiated by data in Table 2, which indicates positive correlation between support for Biden and belonging to the middle-income group. While there is no a specific analyses explicating the motivations behind this pattern, it is plausible to attribute this trend to Biden's policy that directly benefit middle-income families, such as the enhancement of tax credits. According to Parys (n.d.), Biden has supported expanding tax credits for middle-income families, such as Children Tax Credit, which a measure designed to alleviate the financial burden of child-rearing for families. Middle income group could benefit from this policy. Under Biden's administration, according to (2022a), there was an increased credit amount in 2021; the credit amount raised from $2,000 each child to $3,600 for each child who aged from 6 to 17. Unlike previous version, Biden support a fully refundable credit policy which aim to provide a direct financial assistance to families. This policy is really about helping middle-income families by giving them tax breaks and other benefits. The Biden administration is focusing on these families to help them with their money problems, which is why many of them support Biden.

5.4 The voters who own gun or know someone who own gun exhibit a higher propensity to vote for Biden

The data presented in Table 2 reveals an interesting trend: individuals who own guns or are acquainted with gun owners exhibit a higher likelihood of supporting President Biden, despite he advocates more stringent gun control measures. Biden seeks to implement universal background checks, which called Bipartisan Background Checks Act. (2022b) aims to expand background check to cover all sales and transfers of weapons, including private transaction, online sales, and gun show sales.

This apparent paradox, where gun owners show support for a candidate known for promoting restrictive gun policies. It is plausible that gun owners, while cherishing their rights, are equally concerned about the pervasive issue of gun violence concerns in the nation. Their support for Biden may because they recognize the measures such as universal background checks represent a balanced approach to regulate gun violence without infringing on the legitimate rights of the citizens to own firearms for personal security and recreational purposes.

5.5 Weaknesses and Next Steps

There are a few potential weaknesses for this paper. Firstly, I only conduct a logistic regression analysis on 500 samples for computational efficiency purpose. It is limited to reflect a representative sample, which might lead to a inaccurate conclusion on the population's attitudes and behaviors.

In addition, the model can identify the relationship or correlation between different variables but it does not necessarily means causality. To build up a causal relationship require more experimental design.

Moreover, select logistics model merely might fail to capture the overall interaction between predictors. In order to get a more detailed interpretation, some complex model might be helpful.

For solving this limitation, the next steps can be firstly, expand the sample size. To design a research in longitudinal approach might also helpful for identify causality. Additionally, introduce other models to further ensure the accuracy and interpretability will also be useful.

##Appendix **Motivation**

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description* -The Cooperative Election Study (CES) is dedicated to exploring American citizens' perceptions and accountability of their elected officials during elections. It delves into the intricacies of voting habits and the influence of political geography and societal factors on these behaviors. By gathering extensive data from a majority of legislative districts, the CES seeks to bridge the knowledge gap in the American electoral system through detailed analysis of congressional races and national polls.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?* -The CES initiative encompasses a collaborative effort of 62 research teams and organizations.
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
    - The project receives backing from the National Science Foundation (NSF), under grant number 2148907. This grant supports the collaborative efforts of numerous teams and organizations involved in the study.
4. *Any other comments?*
    - None

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

- The CES 2022 dataset includes responses from individual participants who took part in the survey. This survey was structured to collect a wide range of information about voters' behaviors, political views, and demographic details. The data encompasses details on participants' voting experiences, the context of their electoral participation, and various demographic factors.

2. *How many instances are there in total (of each type, if appropriate)?*

    - The CES study captured data from 60,000 respondents during the fall of 2022. These respondents were adult participants surveyed on a range of topics related to their electoral participation.

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

    - Instead of capturing every possible instances, the CES 2022 dataset prioritizes a selection drawn from a broader pool of adult citizens in the U.S. The approach utilized by YouGov for CES involves matched random sampling. This technique starts with identifying a target population count, typically encompassing all adults, often determined through extensive surveys like the American Community Survey. From this population, CES selects a random sample, referred to as the target sample. The accuracy of this sample's representativeness is enhanced through a two-stage weighting process. This process aims to correct any potential imbalances within the matched sample by adjusting the sample's distribution to align with the broader population across various demographic and political dimensions, utilizing entropy balancing and iterative proportional fitting for adjustments.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

    - The dataset encompasses a range of information, including demographic details, political beliefs and behaviors, electoral experiences, verification of voters, and data on device usage and participation. When it comes to political opinions, the dataset includes data on the political party affiliations of respondents, their choices for president, and their views on different policy matters and political figures.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

    - No.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

- The CES methodology includes matching respondents with external voter files, which can occasionally result in discrepancies. Errors in matching may arise from inconsistencies in names, addresses, or other personal information between the respondent and voter records, potentially leading to inaccuracies in correlating voting status with the individual respondent.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

    - The CES 2022 dataset is structured in such a way that it does not establish a direct model of relationships among individual responses. Instead, each entry in the dataset corresponds to the survey responses of an individual, encompassing a wide array of information including their demographic characteristics, political beliefs, voting actions, and other pertinent data.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

    - No.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

    - Random measurement error is identified as a potential source of inaccuracies within the dataset. This may occur when respondents provide incorrect answers, inadvertently leading to some individuals being misclassified.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

    - The CES 2022 dataset primarily consists of survey responses and does not directly link to external sources. While voter file matching is a part of its methodology, specific details on external databases like TargetSmart used for this purpose are not extensively covered. However, comprehensive data from the CES 2022, including vote verification for participants, is stored and accessible through the Harvard University Dataverse. This indicates a lack of detailed disclosure concerning the use of external resources for voter file matching and validation.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

    - The CES 2022 collection is made up of methodically organized survey responses from individuals, shedding light on their political perspectives, activities, and demographic background. This type of data collection is typically free from the constraints of legal confidentiality agreements, such as those found between doctors and patients or in legal communications.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

    - Survey questions relating to racial and sexual categories may be perceived as insensitive or biased by some respondents, potentially raising concerns about the manner in which these topics are approached.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - Several sub-population in the dataset. 1)Gender: The dataset categorizes respondents into "Man", "Woman", "Non-binary", and "Other" 2)Education:ranging from no high school diploma ("No HS") to postgraduate degrees 3)Age: age range from 18 to 65 and over 65 4)Race:"White", "Black", "Hispanic", "Asian", "Native American", "Middle Eastern", "Two or more races", and "Other" 5)Political Affiliation: including "Democrat", "Republican", and "Independent/Other".

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - No

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

    - The dataset indeed covers a broad spectrum of personal attributes including race, sexual orientation, religious beliefs, political views, transgender status, and health information. For instance, questions about sexual orientation offer choices like heterosexual/straight, lesbian/gay, bisexual, among others.

16. *Any other comments?*

    - No

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

    - Information for each entry in the CES 2022 dataset was mainly derived from direct survey inputs from participants. These surveys, managed by YouGov online, utilized a technique known as sample matching to select a subset of respondents that representatively reflects the broader target population. Additionally, parts of the dataset focusing on confirming voter registration details involved cross-referencing responses with the TargetSmart database, holding information on registered voters in the U.S., to verify the accuracy of participants' reported voter registration and activity in the elections.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

    - The CES 2022 utilized an online survey mechanism for data collection, implemented by YouGov. The "sample matching" methodology was applied, aligning with online panel access. This process involves selecting respondents who share similar attributes with the target demographic, using a formula that considers the aggregate differences across various characteristics. Adjustments are made to ensure the sample accurately mirrors the target population's traits.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

    - YouGov's methodology for creating the CES 2022 dataset involved a matched random sample approach, aiming to accurately represent the target population by matching survey participants to a predetermined demographic profile.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

    - The CES project was a collaborative effort involving 60 research teams and organizations. Specific details about participant compensation or involvement terms were not disclosed within the dataset's documentation.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old*

news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

- The data collection for CES 2022 spanned two distinct periods: the pre-election phase from September 29 to November 8, 2022, and the post-election phase from November 10 to December 15, 2022. This timeline was specifically chosen to capture respondents' perspectives and intentions related to the 2022 elections accurately.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

   - No specific information regarding the conduct of ethical review processes for the CES 2022 dataset has been provided.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

   - Data collection for the 2022 Cooperative Election Study (CES) was achieved through a survey administered by YouGov, which using the internet to reach respondents directly.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.* Yes, the CES data was indeed gathered directly from respondents via an online survey facilitated by YouGov, confirming the method of data collection.
9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

   - Participants in the 2022 CES explicitly gave their consent to be part of the study. They were presented with a clear inquiry about their willingness to participate, framed as "Consent to participate: Do you agree to participate in the study?" to ensure informed consent was obtained.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

    - There is no detailed information within the dataset regarding whether participants in the 2022 CES had the option to withdraw their consent at any stage after initial agreement or to restrict certain uses of their data.

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a*

description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

- No, it focuses on outlining the study's methodology, including sample matching and the procedures for collecting data.

12. *Any other comments?*

    - No

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*

    - Yes, the CES undertakes several data preprocessing steps to ensure high data quality, including cleaning and labeling. These steps are detailed within the methodology section, focusing on sampling, sample matching, weighting, and voter validation. Specifically, to address any discrepancies between the sampled group and the broader target population, the CES employs a two-step weighting process. Initially, entropy balancing is used to align the sample's distribution with key demographic factors such as age, gender, education, and race. Following this, iterative proportional fitting, or "raking," is applied to other variables like voter registration status and 2020 Presidential vote preferences, further enhancing the sample's representativeness.

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

    - No.

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

    - No.

4. *Any other comments?*

    - No.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*

    - The CES Guide 2022 does not provide specific examples of tasks for which the dataset has already been used.

2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*

    - The CES Guide 2022 does not explicitly mention a repository that links to paper.

3. *What (other) tasks could the dataset be used for?*

    - There are multiple task that the dataset can be used for. For example, to analyze voting behavior and study the impact of socioeconomic factor on political outcome.

4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*

    - Yes.The methodology section details efforts to create a representative sample and address potential biases through sample matching, weighting, and vote validation in order to avoid representation bias.The customer can ensure all research complies with applicable laws, ethical standards, and best practices for data privacy and protection.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

    - No.

6. *Any other comments?*

    - No.

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

    - Yes, it will. It indicated that the data for the Cooperative Election Study (CES) 2022. It is accessable in the Harvard University Dataverse. it is available to external parties beyond those directly involved in its survey.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

    - The CES 2022 dataset is shared via the Harvard University Dataverse, in order to ensure academic and research can access the data. The dataset's DOI is https://doi.org/10.7910/DVN/PR4L8P.

3. *When will the dataset be distributed?*

    - The distribution of the dataset occurs in two phases. Firstly, on March 20, 2023, the CES 2022 Common Content was released, encompassing data uniformly collected by all the contributing research teams. Subsequently, on September 8, 2023, vote validation data for all participants was added, significantly augmenting the dataset's value for election studies by confirming the accuracy of the voting behavior reported by survey respondents.electoral research by verifying the voting behavior reported by the survey participants.

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

    - No.

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

    - No.

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

    - No.

7. *Any other comments?*

    - No.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?* -The Harvard University Dataverse support the dataset.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

    - People can contact the dataset via the Harvard University Dataverse platform.

3. *Is there an erratum? If so, please provide a link or other access point.*

    - No

4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

- Do not have specific information about that.

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

   - Do not have specific information about that.

6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

   - Do not have specific information about that.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

   - Do not have specific information about that.

8. *Any other comments?*

   - No

References

2022a. *U.S. Department of the Treasury*. https://home.treasury.gov/policy-issues/coronavirus/assistance-for-american-families-and-workers/child-tax-credit#:~:text=The%20credit%20amount%20was%20increased,credit%20was%20made%20fully%20refundable.

———. 2022b. *Wikipedia*. Wikimedia Foundation. https://en.wikipedia.org/wiki/Bipartisan_Background_Checks_Act.

———. 2024. *The White House*. The United States Government. https://www.whitehouse.gov/briefing-room/statements-releases/2024/03/11/fact-sheet-the-presidents-budget-cuts-taxes-for-working-families-and-makes-big-corporations-and-the-wealthy-pay-their-fair-share/#:~:text=The%20President's%20Budget%20restores%20the,Loopholes%20for%20the%20Very%20Wealthy.

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

Brilleman, SL, MJ Crowther, M Moreno-Betancur, J Buros Novik, and R Wolfe. 2018. "Joint Longitudinal and Time-to-Event Models via Stan." https://github.com/stan-dev/stancon_talks/.

Kuriwaki, Shiro, Will Beasley, and Thomas J. Leeper. 2023. *Dataverse: R Client for Dataverse 4+ Repositories*.

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

Mustafa, Akif. 2024. "An Example of Bayesian Linear Regression." *Medium*. Medium. https://medium.com/@akif.iips/an-example-of-bayesian-linear-regression-c3bc8f8e2fa6.

Parys, Sabrina. n.d. "Earned Income Tax Credit 2023-2024: How to Qualify." *NerdWallet*. https://www.nerdwallet.com/article/taxes/can-you-take-earned-income-tax-credit.

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

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

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

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. http://haozhu233.github.io/kableExtra/.