Datasheet for the 'Comprehensive Election Study Common Content 2022'*

The survey data set used for 'US Election Analysis'

Talia Fabregas

April 18, 2024

This data sheet was put together with the guidance of Gebru et al. (2021) and the book "Telling Stories with Data" (Alexander 2023).

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.*
    - This dataset was created to study the voting behavior of American adults leading up to the November 2022 midterm elections. The CES aims to study how American adults view their representatives during elections, how they have voted, and how political behavior is influenced by geographic and demographic factors (Schaffner, Ansolabehere, and Shih 2023).
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - This dataset was put together by Talia Fabregas for research purposes at the University of Toronto. The original dataset, the CES Common Content 2022, was created by 62 research teams and organizations. Each team had its own Principal Investigator, and individual research groups put together their own team surveys.
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 creation of the dataset was not funded.
4. *Any other comments?*
    - No

*Code and data are available at: https://github.com/taliafabs/us-election-analysis.git

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.*

    - Each instance in the dataset represents 1 voter in the United States who responded to the CES 2022 survey questions

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

    - There are 10,000 instances in total. Each instance in the dataset is an American adult who responded to a CES survey in late 2022.

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).*

    - The CES 2022 dataset is designed to be representative of all US adults; not every survey response is included in the data set to make it representative (Schaffner, Ansolabehere, and Shih 2023).

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.*

    - Each instance consists of responses to survey questions about demographics, income, education, and political preferences

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

    - While there are questions about political preferences, there is no explicit label or target associated with each instance in the original data set. However, in the context of my US Election Analysis, I've created a vote_biden target binary variable which indicates whether each respondent will vote for Joe Biden (D) or Donald Trump (R) in November. The purpose of this target is to fit a logistic regression model to predict vote choice of US adults based on state, whether Biden previously won that state, age, race, sex, highest level of education, and whether the respondent lives in an urban or rural area.

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.*

- No. As far as my research is concerned, I did not find any missing information in the data set.

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.*

   - Relationships between individual instances are not made explicit.

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.*

   - There are no recommended data splits provided.

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

   - There are no clear errors, sources of noise, or redundancies in the data set. There are, however, individuals whose political preference is not necessarily what the logistic regression model would predict using the state that they live in, their age, race, sex, and highest level of education.

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 data set is not self contained.

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 data set does not contain data that might be considered confidential.

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.*

    - Due to the current level of political polarization in America, data about individuals'

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.*

- The dataset does identify sub-populations. Information about individual instances including, but not limited to, age, race, sex, whether they identify as lgbtq+, marital status, income, zip code, region, and state of residence are included in this data set. The data set includes …

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.*

    - It is not possible to identify individuals directly or indirectly from the data set. While information about state, region, city, race, age, gender, marital status, religious beliefs, political preferences, etc. is present in this dataset, it is very hard to connect these traits back to a particular individual.

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.*

    - Yes. This data set reveals sensitive data, such as the race, ethnic origins, sexual orientation, religious beliefs, political preferences, marital status, and income data, of individual instances. However, it is not done so in an identifiable manner. These are simply variables in a large survey data set that simply describe the individual, without identifying them in any way, shape, or form.

16. *Any other comments?*

    - No further comments

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.*

    - The data is mainly recruited from CCES respondents who are YouGov panelists that sing up for notifications about new surveys, however some respondents are recruited from online advertisements or outside survey providers (Schaffner, Ansolabehere, and Shih 2023).

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?*

- CCES survey participants can respond to the approximately 20-minute questionnaire on a smartphone, laptop, mobile device, or over the phone.

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

   - The CCES 2022 data set is designed to be representative of all American adults, and some respondents are not included in the final dataset to ensure that it is representative (Schaffner, Ansolabehere, and Shih 2023). I used a random subset of 10,000 of the 60,000 respondents for my US Election Analysis report.

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)?*

   - CCES survey respondents are adults in the United States. Most respondents are YouGov participants who are registered to receive notifications about new surveys and they are not paid for their responses, but they receive YouGov points that can be exchanged for giftcards. A

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 was collected in 2022, in the months leading up to the US midterm elections. Subjects were recruited in the Fall of 2022. Responses to the pre-eleciton questionnaire were collected between September 29 and November 8, 2022. Responses to the post-election questionnaire were collected between November 10 and December 15, 2022 (Schaffner, Ansolabehere, and Shih 2023).

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

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

   - I obtained the data from Harvard Dataverse and downloaded it via a URL.

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 individuals willingly responded to the questionnaire

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.*

    - Yes

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).*

    - No

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

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, I removed rows with missing values in the race, pid3, pid7, gender4, educ, birthyr, and state columns. While there is no one-size-fits-all approach for dealing with missing values, rows with missing values represented a very small proportion of the data set

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.*

    - The raw data was saved in addition to the cleaned data. It can be accessed at this link:https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910/DVN/PR4L8P

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

    - The software that was used to clean the data is R programming languaage (R Core Team 2023) and the dplyr (Wickham et al. 2023), janitor (Firke 2023), and arrow (Richardson et al. 2023) packages.

4. *Any other comments?*
    - Data cleaning steps are outlined in the data section and appendix of my report.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - Yes, this dataset has
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.*
    - No
3. *What (other) tasks could the dataset be used for?*
    - This data set could be used to analyze vote preference
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?*
    - This data set was put together in 2022, so its intended use was not necessarily forecasting the 2024 US presidential election. Using it to forecast the 2024 election is beneficial due to its size, but it contains no questions about 2024 preferred presidential candidate.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
    - No
6. *Any other comments?*
    - Schaffner, Ansolabehere, and Shih (2023) provides a warning about analyzing subsamples in its Guide 2022. Caution is advised because even if only 0.5% of respondents provided a mistaken response to a survey question, it could result in hundreds of resondents being miscategorized (Schaffner, Ansolabehere, and Shih 2023).

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.*

- The dataset is available for download via URL on the Harvard Dataverse. Therefore, any company, institution, organization, or researcher outside of the Cooperative Election Study that would like to access this data set can do so.

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

   - The dataset is distributed via the Harvard Dataverse website. It can be downloaded via URL and its digital object identifier (DOI) is doi.org/10.7910/DVN/PR4L8P,

3. *When will the dataset be distributed?*

   - The final release of the dataset was published on September 8, 2023

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.*

   - There are no export controls or other regulatory restrictions that apply to the dataset or individual instances that I am aware of.

7. *Any other comments?*

   - No

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

   - The final release of this dataset was published on Harvard dataverse on September 8, 2023.

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

   - The manager of the dataset, Brian Schaffner, can be contacted via the "Contact Owner" button on the dataset's Harvard Dataverse page.

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)?*

    - The data set is a final release, so it will no longer be updated.

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.*

    - No

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.*

    - Older versions of the dataset are still available via Harvard dataverse.

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.*

    - No. The dataset was put together by 62 research teams, each with its own Principal Investigator. It is designed to be representative of U.S. adults in 2022, so it can no longer be modified.

8. *Any other comments?*

    - No

References

Alexander, Rohan. 2023. *Telling Stories with Data*. "University of Toronto". https://www.tellingstorieswithdata.com.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://github.com/sfirke/janitor.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.

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. 2023. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

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

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.