Datasheet for 'Income and Gender: Forecasting the American Voters' Choices in 2018' Dataset*

Siqi Fei

March 13, 2024

The Cooperative Congressional Election Study (CCES) 2018(Schaffner, Ansolabehere, and Luks 2019) is a comprehensive dataset created to understand American electoral behavior, including views on Congress, voting practices, and demographic variations. Developed through a collaboration involving 60 research teams and funded by the National Science Foundation, this dataset encompasses responses from 60,000 individuals, collected via YouGov's online platform using a matched random sample methodology. It covers a broad range of topics from political opinions to detailed demographic information, aiming to provide insights into the electoral dynamics and representation in the U.S. The dataset's distribution, ethical considerations, and maintenance follow academic standards, ensuring its integrity and availability for research. It serves as a vital resource for political science research and offers potential for various analytical tasks within academic and educational settings.

Extract of the questions from Gebru et al. (2021).

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 2018 Cooperative Congressional Election Study (CCES) (Schaffner, Ansolabehere, and Luks 2019) was created with the purpose of studying how Americans view Congress, hold their representatives accountable during elections, their voting behavior, and their electoral experiences. It also aims to understand how these aspects vary with political geography and social context. The dataset was designed to construct a large sample capable of capturing variation across

*Code and data are available at: https://github.com/FXXFERMI/Political-support-in-the-United-States.git

a wide variety of legislative constituencies, enabling the measurement of voters' preferences within most states with a reasonable degree of precision.

2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*

   - The CCES was a collaborative effort involving 60 research teams and organizations. The project was led by Stephen Ansolabehere (Harvard University), Brian Schaffner (Tufts University), and Sam Luks (YouGov) serving as Principal Investigators. The project also involved a project administrator, Armelle Bernard, at Harvard.

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 2018 CCES was supported by the National Science Foundation under Award #1756447. This grant facilitated the comprehensive study and methodology applied in conducting the survey and compiling the dataset.

4. *Any other comments?*

   - The dataset serves as a rich resource for political science research, offering insights into electoral behavior, public opinion, and demographic trends within the context of Congressional elections. It continues a series of studies that began in 2006, building off of the MIT Public Opinion Research and Training Lab (PORTL) study from 2005, with support from the National Science Foundation for all even-year surveys since 2010. The dataset not only aids academic research but also provides a valuable tool for understanding the nuances of American electoral dynamics and representation.

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 instances in the CCES dataset represent individual survey responses from participants across the United States regarding their views on Congress, electoral behavior, and various demographic, social, and political questions. There are not multiple types of instances; each instance is a survey response from a single participant.

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

   - The 2018 CCES involved 60 teams, yielding a Common Content sample of 60,000 cases. Each case represents an individual survey response.

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 dataset is a sample of the larger set of eligible voters or the general adult population in the United States. The sample was constructed to be representative of the larger set through a methodology involving sample matching and weighting based on demographic characteristics derived from the American Community Survey and voter registration data. The representativeness was validated through comparison of the sample demographics and political preferences with known population benchmarks.

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 survey responses to a variety of questions, including demographic information, political opinions, voting behavior, and more. The data are structured and processed, with responses coded into specific categories or scales

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

    - Each survey question can be considered a "label" for that specific question. For example, questions about voting behavior have categorical responses indicating the participant's choices or actions

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 document does not explicitly detail missing information for individual instances, but in surveys like CCES, it's common to have some respondents skip questions or choose "Prefer not to say" options. The reasons for missing data can vary, including personal preference or the sensitive nature of questions.

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.*

    - While the dataset primarily consists of individual and independent responses, there are implicit relationships between instances through the aggregation of responses to explore broader social and political trends. However, these relationships are not explicitly made within the dataset but can be inferred through analysis.

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.*

    - The document does not mention specific recommended data splits for training, development/validation, testing purposes as it is primarily designed for research and analysis rather than predictive modeling or machine learning applications

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

    - The guide does not specifically address errors, noise, or redundancies, but it's common in large-scale surveys to encounter some level of response bias, measurement error, or inconsistencies due to the subjective nature of self-reported data

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 dataset is self-contained with survey responses collected and processed by the CCES team. There's no indication that it relies on external resources that might change over time

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 CCES dataset is designed to be anonymized, with no direct personal identifiers included. However, the nature of the data is such that it involves personal opinions and behaviors.

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.*

    - The dataset itself is unlikely to contain directly offensive material since it consists of responses to structured survey questions on political and social topics. However, topics discussed could be sensitive or controversial to some 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 identifies sub-populations by various demographic markers such as age, gender, race, education, and more. Distributions of these subpopulations are detailed in the breakdown of responses to specific questions, providing insights into the composition of the dataset

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.*

    - Direct identification of individuals is not possible from the dataset as it is anonymized and does not include personal identifiers. Indirect identification might theoretically be possible in very rare cases with unique combinations of attributes, but this risk is generally minimized through the design of the survey and data processing protocols

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 contains data that might be considered sensitive, such as political opinions, voting behavior, religious beliefs, racial or ethnic origins, and potentially other markers. This information is handled with care to ensure respondent anonymity and data security

16. *Any other comments?*

    - The CCES dataset is a valuable resource for understanding American political and social dynamics, designed with rigorous methodology to ensure its utility and integrity for academic and research purposes.

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 associated with each instance in the CCES dataset was directly reported by subjects through survey responses. The data collection involved structured questionnaires filled out by the participants, covering a wide range of topics from political opinions to demographic information. The methodology section of the document mentions using matched samples and weighting to ensure that the sample is representative of the larger population, suggesting some form of validation of the data

through comparison with known benchmarks or demographic data sources like the American Community Survey.

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 data was collected using online survey tools, specifically through YouGov's internet-based platform. The survey's methodology, including sample matching and weighting, was validated against known demographic benchmarks and voter registration data to ensure representativeness and accuracy.

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 used a matched random sample methodology. This involves selecting a random sample (target sample) from the target population and then selecting matching respondents from a pool of opt-in respondents to create a matched sample that mimics the characteristics of the target sample. This approach aims to ensure the representativeness of the survey sample to the larger population.

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 CCES was a collaborative effort involving 60 research teams, with specific individuals or groups responsible for different modules or aspects of the survey. The document does not specify how these individuals were compensated or their specific roles in the data collection process.

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 during two waves in the fall of 2018, specifically before and after the November elections. This timeframe matches the creation timeframe of the data, as the survey aimed to capture attitudes and behaviors relevant to the 2018 Congressional elections.

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.*

    - The document does not provide specific details about any ethical review processes conducted for the CCES. However, given the academic and institutional involvement in the study, it is likely that some form of ethical review was conducted, in line with standard practices for research involving human subjects.

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

    - The data was collected directly from individuals participating in the survey through an online platform. There's no indication that the data was obtained via third parties or other sources

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.*

    - The document does not detail the process of notifying participants or obtaining their consent. However, standard practices for survey research typically involve providing participants with information about the study and obtaining their informed consent before participation.

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.*

    - Details on how consent was obtained and whether participants could revoke their consent are not provided. Standard online survey practices usually include consent forms or agreements that participants must accept before proceeding with the survey.

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).*

    - The document does not mention whether an analysis of the potential impact of the dataset on data subjects was conducted. Such analyses are important for understanding and mitigating any risks associated with the use of the 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.*

    - The document does not specifically mention whether a data protection impact analysis or a similar assessment of the potential impact on data subjects has been conducted for the CCES dataset. In general, such analyses are crucial for large-scale datasets involving personal and political opinions to understand and mitigate any risks to privacy or potential harms that could arise from the use or misuse of the data. These analyses typically evaluate the sensitivity of the data, the context of its use, the potential for re-identification of individuals, and the measures in place

to protect data subjects' privacy and security. Without explicit mention in the document, it's unclear if such an analysis has been performed for the CCES dataset or what the outcomes were. However, given the academic and institutional context of the CCES, it's likely that considerations around data privacy and ethical use were part of the research design and execution proces.

12. *Any other comments?*

    - The collection and use of large-scale survey data like the CCES inherently involve ethical considerations, particularly regarding the privacy and anonymity of respondents. While the document does not detail specific measures beyond the methodology of data collection, users of the dataset should be aware of these ethical considerations and ensure their analyses respect the privacy and confidentiality of individuals' responses. Researchers typically employ techniques such as anonymization, data aggregation, and restricted access to sensitive variables to mitigate potential risks. Further, academic institutions and journals often require ethical review approvals for research involving human subjects, which likely applies to the CCES. Researchers using the CCES dataset or similar datasets are encouraged to follow best practices for ethical research and data use, including consideration of potential impacts on data subjects and the broader social implications of their work

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.*

    - The document outlines the methodology behind the survey, including sample matching and weighting, which can be considered forms of preprocessing to ensure data quality and representativeness. Specific details on cleaning or labeling of data, such as handling missing values or categorizing open-ended responses, are not provided.

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.*

    - There is no mention of whether "raw" data was saved alongside the preprocessed data or how the data was handled before analysis.

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 document does not specify whether the software or tools used for preprocessing, cleaning, or labeling the data are available for public use.

4. *Any other comments?*

    - None.

Uses

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

    - The CCES dataset has been widely used in political science research, particularly for studies on voter behavior, political attitudes, electoral dynamics, and representation in the United States. It provides rich insights into how Americans view Congress, their representatives, and the electoral process, as well as variations in political behavior across different geographies and social contexts.

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.*

    - While the document does not mention a specific repository, the Cooperative Congressional Election Study (CCES) is frequently cited in academic research, and related works can often be found through academic databases and Google Scholar by searching for the "Cooperative Congressional Election Study" or specific years of the study.

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

    - Beyond political science research, the dataset could be used for tasks in sociology, data science (e.g., modeling political preferences, predicting electoral outcomes), education (as a teaching resource on survey design and analysis), and even in machine learning for natural language processing tasks related to analyzing political discourse based on open-ended survey responses.

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?*

    - The dataset's composition and collection methodology are critical for ensuring fair treatment of individuals and groups. Users should be aware of the demographic weighting and sampling strategies to understand the representativeness of the data. Potential biases in self-reported data should be considered to avoid stereotyping or drawing unwarranted conclusions. Ensuring privacy and handling sensitive information respectfully can mitigate risks of harm. Awareness of these aspects and transparent reporting of methodologies used in analysis can help mitigate risks.

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

    - Given the dataset's focus on political behavior and attitudes, it should not be used for purposes that could compromise individual privacy or for commercial targeting based on political preferences. Additionally, care should be taken to avoid misuse that could contribute to political polarization or misinformation.

6. *Any other comments?*

    - None.

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 typically made available to researchers and the public through academic and research institutions involved in the CCES. Distribution to third parties would follow the terms set by these entities.

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

    - The CCES data is often distributed via official academic and research websites, with access usually provided through data repositories associated with the participating institutions or organizations. The dataset may have a DOI for citation and tracking purposes, as is common with academic datasets.

3. *When will the dataset be distributed?*

    - Academic datasets like the CCES are typically made available after the data collection and initial analysis phase are completed. Given the CCES is a biennial study, datasets are usually released in the year following the election year they cover. Specific distribution dates can vary based on the completion of data cleaning and analysis1111.

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.*

    - Academic datasets are generally distributed under licenses that allow for their use in research and education while protecting the intellectual property of the creators. The terms might restrict commercial use or require attribution when using the data in publications. The CCES data is likely accompanied by a similar license, emphasizing citation requirements and possibly restricting commercial uses. However,

specific terms would be detailed on the official CCES or associated institutional websites.

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.*

    - If third-party data or tools were used in creating the dataset, there might be additional restrictions based on those entities' terms of use. However, the CCES's methodology primarily involves direct data collection via surveys, likely minimizing third-party data restrictions. Any such restrictions would be specified in the dataset documentation or terms of use provided upon access.

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.*

    - Users interested in the CCES dataset should review the most current documentation available from the official source for details on access, usage rights, and any restrictions. This ensures compliance with legal and ethical standards and supports the integrity of research conducted with the dataset. Prospective users may need to agree to specific terms of use, which could include provisions for privacy protection, data security, and limitations on the use of the data for specific purposes.

7. *Any other comments?*

    - None.

Maintenance

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

    - The dataset is maintained by the institutions and researchers involved in the CCES, with Harvard University often playing a central role given its involvement in the project. Updates and maintenance are typically handled by the principal investigators and their teams.

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

    - Contact information for inquiries related to the CCES dataset is usually provided through the official CCES website or the websites of the principal investigators and their institutions.

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

- The CCES is conducted biennially, with each new iteration of the study potentially adding to or updating the dataset. Information about updates, corrections, or new releases is typically communicated through academic channels, the CCES website, and related mailing lists or forums.

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)?*

    - While the document does not specify a mechanism for external contributions to the dataset, academic collaborations and contributions are common in large-scale research projects like the CCES. Such contributions would likely undergo review by the principal investigators or a designated committee to ensure validity and consistency with the dataset's standards.

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.*

    - Data involving human subjects, especially data collected in the United States, must comply with privacy laws and ethical standards, such as those outlined by the Institutional Review Board (IRB). While not explicitly regulatory restrictions, these standards dictate how data can be collected, used, and shared to protect participants' privacy. Specific export controls or other regulatory restrictions would typically be mentioned in the dataset's access conditions if applicable.

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.*

    - Users interested in the CCES dataset should review the most current documentation available from the official source for details on access, usage rights, and any restrictions. This ensures compliance with legal and ethical standards and supports the integrity of research conducted with the dataset. Prospective users may need to agree to specific terms of use, which could include provisions for privacy protection, data security, and limitations on the use of the data for specific purposes.

7. *Any other comments?*

    - None.

References

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.

Schaffner, Brian, Stephen Ansolabehere, and Sam Luks. 2019. "CCES Common Content, 2018." Harvard Dataverse. https://doi.org/10.7910/DVN/ZSBZ7K.