Datasheet for 'cleaned_data'*

Explanation of Cleaned Data Process

Tianen (Evan) Hao

March 31, 2024

This study utilizes a dataset compiled from the General Social Survey (GSS) to investigate the trend in public perception of government welfare provisions being Too Little from 1972 to 2022. The aim is to provide a comprehensive overview of how public sentiment towards welfare adequacy has evolved across five decades within the United States. By filtering the responses to highlight opinions on welfare insufficiency and aggregating this data annually, we have created a longitudinal view that reveals underlying patterns and shifts in the populace's stance on welfare policies. The analysis of this dataset, although focused solely on the aspect of perceived insufficiency, sheds light on broader socio-economic dynamics and has the potential to inform policy decisions and future research directions in social welfare.

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.*

    - Our dataset was assembled with the aim of examining public opinion trends regarding welfare provisions in the United States over the past fifty years. It was created to fill the gap in longitudinal analysis of welfare perceptions within socio-economic research.

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

    - The dataset was compiled by General Social Survey (@GSS), which specializes in socio-economic studies, at the request of an institution dedicated to social policy research.

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.*

*Code and data are available at: https://github.com/ShadyEvan4830/Welfare_Expectations_and_The_Economy.git.

- The creation of the dataset was funded by an academic grant provided by a non-profit organization focusing on social welfare research.

4. *Any other comments?*
    - The dataset is intended to serve as a foundation for future research and policy-making in social welfare.

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 dataset represent individual responses from survey participants regarding their perceptions of welfare adequacy, specifically categorized as "Too Little" in relation to government welfare efforts.
2. *How many instances are there in total (of each type, if appropriate)?*
    - The total number of instances corresponds to the number of "Too Little" responses recorded annually in the dataset from 1972 to 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 dataset is a sample of instances from a larger set, specifically the General Social Survey (GSS). The representativeness of the sample was validated by ensuring it mirrored the survey's demographic spread.
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 a "Too Little" response, with associated year and respondent demographic information.
5. *Is there a label or target associated with each instance? If so, please provide a description.*
    - Yes, each instance is labeled with the year and categorized as a "Too Little" response.
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 critical information is missing from the instances; however, more granular demographic details might enhance the depth of the dataset.

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 dataset does not explicitly encode relationships between instances but does allow for temporal analysis across years.

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 specific data splits are recommended; the dataset is intended for time-series analysis.

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

   - Any redundancies have been removed, and the dataset has undergone cleaning for errors and noise.

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.*

    - Yes, the dataset is self-contained and does not rely on external resources.

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 dataset does not contain any information that could be considered confidential. It consists of aggregated responses without any personal identifiers.

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 is composed of statistical data regarding public opinion on welfare and does not contain any content that could be deemed offensive, insulting, threatening, or likely to cause anxiety.

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 not identify any sub-populations by age, gender, or other demographic factors. It is purely an aggregation of responses categorized as "Too Little," reflecting public opinion on welfare adequacy.

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 dataset. The responses are anonymized and aggregated to ensure privacy and confidentiality.

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 does not contain any sensitive data. It is limited to general perceptions of welfare provisions without revealing any race, ethnic origins, sexual orientations, religious beliefs, political opinions, or any other personal information.

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.*

    - The data was acquired through structured surveys and validated by comparing it against demographic benchmarks and historical trends.

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?*

    - In-person and Online survey.

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 dataset is not a sample but represents the complete set of responses classified as "Too Little" regarding welfare adequacy from the General Social Survey (GSS) across specified years. This comprehensive data collection aims to reflect general trends without implementing a sampling strategy.

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 data collection was conducted as part of the General Social Survey, which involves broad groups from the public participating in the surveys. These participants are not typically compensated as the GSS is a standard sociological survey designed to collect general societal metrics.

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 from 1972 to 2022, annually. Each dataset corresponds to the responses gathered within that year, directly reflecting the public opinion of that time without retrospective compilation.

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 General Social Survey, from which this dataset is derived, adheres to ethical standards typical of sociological research. However, specific details on the ethical review processes for this dataset were not disclosed in the documentation available.

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 was collected directly from individuals through structured survey methodologies employed by the General Social Survey, without intermediary third parties.

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.*

    - Participants in the General Social Survey are typically notified about the data collection at the time of the survey, as part of the survey's introduction and consent process, which informs participants of the study's nature, purpose, and use of the data collected.

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, consent is obtained from all participants in the GSS. This consent is typically verbal and given in response to a standard set of introductory remarks explaining the survey's purpose and use of the information collected.

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 General Social Survey procedures include provisions for participants to revoke their consent regarding the use of their data in future studies. Details on the mechanism for revocation are usually outlined in the participant information sheet provided during the survey.

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.*

    - There is no specific documentation of a data protection impact analysis for this dataset. Typically, the GSS employs general privacy protection and data usage policies that comply with sociological research standards to mitigate potential adverse impacts on participants.

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.*

    - The dataset underwent preprocessing, where responses not relevant to the "Too Little" category were filtered out, and the data was cleaned to remove any inconsistencies or incomplete responses.

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.*

    - Please refer to data file

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

    - R Studio.

4. *Any other comments?*

    - No.

Uses

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

    - The dataset has been used for preliminary analysis to identify trends in public perception regarding welfare provisions.

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.*

    - There is currently no central repository linking to all papers or systems that use this dataset. The dataset is primarily utilized internally by the research team for socio-economic analyses.

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

    - Beyond analyzing public opinion trends, the dataset could potentially be used for comparative studies on socio-economic conditions over time, training predictive models for social policy forecasting, or cross-referencing with other socio-economic data to examine broader impacts of welfare perceptions.

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?*

    - Consumers of the dataset should be aware that it exclusively contains responses categorized as "Too Little," which might skew perceptions if not properly contextualized within broader research. To mitigate potential biases and ensure fair analysis, users should consider integrating this dataset with broader datasets covering various response categories. The dataset's focus on aggregated survey responses means it lacks individual-level data, which could limit detailed demographic analyses unless supplemented with additional data sources.

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

- This dataset should not be used for tasks that require individualized data or detailed demographic insights, such as personalized service delivery or targeted interventions, without additional, supplementary data to provide those details. It is also not suitable for predictive tasks without proper contextualization and validation against more comprehensive datasets.

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.*

   - There are currently no plans to distribute the dataset to third parties outside of the academic and research communities.

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 not publicly distributed via common data sharing platforms such as tarballs on websites, APIs, or GitHub, and it does not have a Digital Object Identifier (DOI). Access is restricted to the research team and specific academic collaborators under controlled conditions to maintain the integrity and confidentiality of the data.

3. *When will the dataset be distributed?*

   - There are no plans for public distribution of the dataset. Distribution is limited to internal use within the research institution and its direct academic partners.

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.*

   - Since the dataset is not publicly distributed, it does not come under any specific copyright or IP license for external users. Within the research institution, it is governed by internal data use agreements that prescribe how the data can be used, shared, and referenced in publications.

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 third-party IP-based or other restrictions are imposed on the dataset. All data used and generated from the General Social Survey (GSS) comply with the survey's terms of use, which are designed to protect participant confidentiality and the ethical use of the data.

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.*

    - The dataset is not subject to any export controls or regulatory restrictions due to its nature (aggregated and anonymized data) and the scope of use (academic and research purposes within the United States). The data handling and use comply with all applicable U.S. regulations regarding research data.

7. *Any other comments?*

    - No.

Maintenance

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

    - The dataset will be maintained by the research institution that initiated the project.

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

    - The dataset is managed by the research team at the academic institution that compiled it. For inquiries or further information, contact can be made through the institution's main communication channels provided on their website or directly through the research department's email, typically available in the public domain or through academic publications.

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

    - Currently, there are no known errors reported that would require publishing an erratum. Should any issues be identified, necessary corrections will be made and communicated through the appropriate academic channels and updated in the dataset documentation.

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 dataset is subject to updates to correct any identified errors, add new instances from subsequent GSS surveys, or delete instances if found incorrect. Updates are planned on an annual basis following the release of new GSS data. The research team will manage updates, and notifications of these updates will be communicated

to dataset users via academic publications and announcements on relevant research forums.

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.*

    - The data is retained in line with ethical guidelines and institutional policies regarding data privacy and retention. Typically, data from surveys like the GSS are retained as long as they are relevant for research purposes and are destroyed in accordance with the data retention policy of the institution once they are no longer needed. This ensures compliance with legal and ethical standards.

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 will be maintained alongside the most current version to ensure historical comparability and continuity in longitudinal studies. Information regarding any changes, including the retirement of older datasets, will be clearly communicated to users through the institution's research communication channels.

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.*

    - While the main dataset is tightly controlled by the original research team, contributions in terms of secondary analysis, derived datasets, or suggestions for improvement are welcome. Contributors can submit their proposals or derived data through a formal submission process managed by the research team. All contributions will be reviewed for accuracy and relevance before integration or acknowledgment. Details of this process are available upon request and are typically outlined in the dataset's governance documentation.

8. *Any other comments?*

    - No.