

Datasheet for U.S. General Social Survey (GSS) dataset*

Mingjia Chen

April 21, 2024

For the purpose of this tutorial, the hypothetical datasheet is constructed based on the textbook by Wickham et al. (2019) and questions are extracted from Gebru et al. (2021), using open source statistically programming language R (R Core Team 2023). The General Social Survey (GSS) dataset (“General Social Survey” 2024), initiated by NORC at the University of Chicago and primarily funded by the National Science Foundation, has been diligently tracking societal shifts and studying the increasing intricacies of American culture since 1972.

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 GSS stands as a vital national asset, publicly accessible and esteemed as one of the most extensively scrutinized repositories of data in the social sciences. NORC’s commitment to broadening access to GSS data is exemplified through initiatives like the GSS Data Explorer, fostering utilization by legislators, policymakers, researchers, educators, and beyond.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
 - GSS is initiated by National Opinion Research Center (NORC) at the University of Chicago.
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.*
 - GSS is primarily funded by the National Science Foundation (NSF).
4. *Any other comments?*

*Code and data are available at: https://github.com/MjChen120/Mental_to_Physical_Health.git.

- N/A

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.*
 - Since 1972, the General Social Survey (GSS) has been conducting nationally representative surveys of adults in the United States. By gathering data on contemporary American society, the GSS aims to track and elucidate trends in opinions, attitudes, behaviors, social status, health status, and more. The GSS dataset focuses on collecting various data points from individuals to understand societal trends and patterns.
2. *How many instances are there in total (of each type, if appropriate)?*
 - There are in general three types of instances in total: Likert scale, continuous, and categorical responses. The number of variables could not be listed here as GSS is a large social dataset, consisting of enormous amount of variables.
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 used for the research paper only contains a few variables from the original GSS data (larger set). The original GSS dataset is representative of the population as it collects responses from a high percentage of population; although having some degree in non-responsive rate, the data in recent decades should be representative to the population as huge amount of responses were collected without biases.
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 instances consist of raw data, where non-responses were transformed into number (for example, -100) that could be filtered out when analyzing.
5. *Is there a label or target associated with each instance? If so, please provide a description.*
 - Labels indicating the actual question asked and options offered to the respondents associated with each instance are included in the GSS Open Explorer for analysis purposes.

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.*
 - Some information are missing from individual instances, this usually happened due to respondents' non-responding to certain questions or withdraw the survey.
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 number of counted days of felling mentally unwell and physically unwell in the past 30 days are made explicit, as we used linear regression analysis for evaluating the relationship between the numbers.
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.*
 - N/A
9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*
 - Due to the fact that the survey is in a self-report manner, errors such as biases, noises, or redundancies could take place in the dataset. They are not speculated in this particular research paper.
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 extracted from GSS Open Explorer website; a) there is guarantees that they will exist, and remain constant, over time since the dataset is created and maintained by NORC at University of Chicago with funding, b) official archival versions of the complete dataset, and c) there are restrictions associated with external resources that apply to a dataset consumer, such as GSS data can only be accessed from the Open Explorer and GSS dataset obtaining can be replicated with the same steps.

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 is protected by legal privilege, as personal contents of individuals are not included in the data.
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.*
 - No.
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.*
 - N/A
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 it is not possible to indentify individuals.
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.*
 - Some of the health data is contained that might be considered sensitive such as health status and mental illness diagnosis. In addition, demographic data such as gender and age are also included in the dataset.
16. *Any other comments?*
 - N/A

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.*
 - From 1992 until 2018, the General Social Survey (GSS) relied on face-to-face interviews to gather data. In response to the COVID-19 Pandemic, between 2020 and 2021, the GSS transitioned primarily to web surveys for data collection.

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 interview method offered a robust data collection environment, allowing interviewers to delve deeper into responses and clarify any uncertainties. It provided an optimal means of gathering comprehensive data. On the other hand, using online survey method allows GSS to collect larger amount of data with less effort required.
3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*
 - N/A
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)?*
 - Americans in general were involved in the data collection process. The participants were not paid for the survey.
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 is collected from the year 1972 to 2022. This timeframe does match the creation timeframe of the data associated with the instances.
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.*
 - N/A
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 is obtained from GSS Open Data Explorer website, by selecting and extracting desired variables for studying.
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.

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).*
 - N/A
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.*
 - In the paper, an analysis on non-response rates that could impact the dataset and its use on data subjects was conducted. The non-response rates for each question were high. The rates could be influencing the representiveness of the dataset and the analysis.
12. *Any other comments?*
 - N/A

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.*
 - Cleaning and relabeling of the data was done for 1) removing and processing the non-responses and 2) have variable names easier for understanding and replicating.
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 preprocessed with non-responses ahead of time for future uses.
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
 - https://github.com/MjChen120/Mental_to_Physical_Health/tree/main/scripts
4. *Any other comments?*

- N/A

Uses

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

- The data have been used for numerous purposes for legislators, policymakers, researchers, educators, and beyond.

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.*

- One of my group project paper have used this dataset: https://github.com/MjChen120/Labour_Mark

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

- Legislations, policy making, researching, and educating.

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?*

- N/A

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

- N/A

6. *Any other comments?*

- N/A

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.*

- TBD

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 on GSS Open Data Explorer.

3. *When will the dataset be distributed?*

- N/A
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.*
 - All contents of NORC websites is under NORC copyright protection: “No part of the contents of NORC websites may be reproduced, stored, or transmitted in any form or by any means, electronic or mechanical, in whole or in any part, without the express written consent of NORC. Requests for permissions should be directed to: commhelp@norc.org. “NORC” and associated graphic images are service marks of NORC and subject to United States and international trademark and other intellectual property protections.” Link: https://gssdataexplorer.norc.org/terms_and_conditions
 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.*
 - Yes. We are not supposed to include GSS raw data we used in the repository. For replicating with the raw dataset, users should follow our instructions to download the data from the Open Explorer website.
 7. *Any other comments?*
 - N/A

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
 - National Opinion Research Center (NORC) at the University of Chicago.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
 - commhelp@norc.org
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 dataset will be updated when research and new knowledge of labeling and instances are introduced at the time.
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. All the dataset are available from 1972 until now.
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 is still being hosted and maintained by having year labels on each variables and older variables used with descriptions ever appeared in the dataset.
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. Only NORC will be able to extend and conduct further survey to extend and build on the dataset.
8. *Any other comments?*
 - N/A

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.
- “General Social Survey.” 2024. *General Social Survey*. NORC. <https://gss.norc.org/get-the-data/stata>.
- R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Golemund, et al. 2019. “Welcome to the tidyverse.” *Journal of Open Source Software* 4 (43): 1686. <https://doi.org/10.21105/joss.01686>.