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

Gauravpreet Thind

December 3, 2024

For this tutorial, the hypothetical datasheet is constructed based on the textbook by Wickham et al. (2019), with questions drawn from Gebru et al. (2021), using the open-source statistical 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 changes and examining the growing complexities 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 is a crucial national resource, open to the public and regarded as one of the most thoroughly analyzed data collections in the social sciences. NORC demonstrates its dedication to expanding access to GSS data through efforts such as the GSS Data Explorer, which encourages its use by legislators, policymakers, researchers, educators, and others.

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

    - The GSS is launched by the 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.*

    - The GSS is primarily supported by funding from the National Science Foundation (NSF).

[*] Code and data are available at: https://github.com/GauravT-crypto/Mental-Health-to-Wellness

4. *Any other comments?*

    - 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 U.S. adults. Its goal is to collect data that reflects contemporary American society, tracking and analyzing trends in opinions, attitudes, behaviors, social status, health, and other areas. The GSS dataset is designed to capture a wide range of information from individuals to better understand societal patterns and trends.

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

    - Generally, there are three types of response formats in the GSS: Likert scale, continuous, and categorical. Due to the vast size of the GSS dataset, which includes a large number of variables, it is not possible to list them all here.

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 in this research paper includes only a subset of variables from the original GSS data. The full GSS dataset is designed to be representative of the population, as it gathers responses from a large portion of the population. While there is some level of non-response, the data collected in recent decades remains broadly representative, as a significant number of responses have been gathered 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 instance in the dataset contains raw data, with non-responses converted into numerical values (such as -100) to facilitate filtering during analysis.

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

    - The GSS Data Explorer includes labels that specify the actual questions asked and the response options provided to participants, making it easier to analyze the data.

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 may be missing from individual instances, typically due to respondents either not answering certain questions or withdrawing from 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 days spent feeling mentally and physically unwell in the past 30 days is clearly specified, as linear regression analysis was used to assess the relationship between these figures.

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.*

   - Since the survey relies on self-reported data, the dataset may contain errors such as biases, noise, or redundancies. However, these issues are not addressed or speculated upon 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 sourced from the GSS Data Explorer website, with the following considerations:

a) There is assurance that the dataset will remain available and consistent over time, as it is created and maintained by NORC at the University of Chicago with funding.
b) Official archival versions of the complete dataset are available.
c) There are restrictions for external users, including the fact that GSS data can only be accessed through the GSS Data Explorer, and the process of obtaining the dataset can be replicated using 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 it does not include any personal information or identifiable content from individuals.

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.*

    - The dataset contains some health-related information that may be considered sensitive, such as health status and mental illness diagnoses. Additionally, demographic data, including gender and age, are also included.

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 to 2018, the General Social Survey (GSS) primarily used face-to-face interviews to collect data. However, in response to the COVID-19 pandemic, the GSS shifted mainly to web-based surveys for data collection between 2020 and 2021.

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 in-person interview method provided a strong data collection environment, enabling interviewers to explore responses more deeply and clarify any uncertainties, making it an ideal way to gather comprehensive data. In contrast, the online survey method allowed the GSS to collect a larger volume of data with less effort involved.

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, participated in the data collection process, and the participants were not compensated for completing 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 was collected from 1972 to 2022, and this timeframe aligns with the period during which the data associated with the instances was created.

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 the GSS Open Data Explorer website by selecting and extracting the relevant variables for study.

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.*

    - The paper includes an analysis of non-response rates, which could affect the dataset and its application to data subjects. The non-response rates for each question were notably high, potentially impacting the representativeness of the dataset and influencing the results of 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.*

   - The data was cleaned and relabeled for two main purposes: 1) to remove and process non-responses, and 2) to simplify variable names, making them easier to understand and replicate.

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 to address non-responses in advance, preparing it for future use.

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/GauravT-crypto/Mental-Health-to-Wellness/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 has been utilized for a variety of purposes by legislators, policymakers, researchers, educators, and others.

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.*

    - Currently, there is no central repository that links all papers or systems utilizing this dataset.

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

    - Legislation, policymaking, research, and education.

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 content on NORC websites is protected by NORC's copyright. According to their policy: "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 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 are protected by United States and international trademark and other intellectual property laws." For more details, visit their Terms and Conditions. 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 permitted to include the GSS raw data in the repository. To replicate the analysis using the raw dataset, users should follow our instructions to download the data directly from the Open Data 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 as new research and insights into labeling and instances are introduced.

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 is avaliable 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 are still hosted and maintained, with year labels assigned to each variable. These versions include descriptions of previous variables that have 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 have the authority to extend and conduct further surveys to expand 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 Grolemund, et al. 2019. “Welcome to the tidyverse.” *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.