Datasheet for 'Perceived mental health, by gender and other selected sociodemographic characteristics'*

Zijun Meng

Invalid Date

This is a datasheet for 'Perceived mental health, by gender and other selected sociodemographic characteristics'. By providing information about its motivation, composition, collection process, preprocessing/cleaning/labeling, uses, distribution, and maintenance, the datasheet can help users have a better understanding of this dataset.

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 dataset (Canada 2024b) was created to fill a gap in understanding the mental health problems in Canada.They are structured to provide insights that inform policy-making, academic research, and public understanding.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - Statistics Canada created the datasets.
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.*
    - They were funded by the Canadian government.
4. *Any other comments?*
    - No.

*Code and data are available at: https://github.com/FrankMengZJ/canada_mental_health

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 primarily represent individual responses from Canadian residents aged 15 and over, living in private households across the 10 provinces. These responses capture a variety of social and economic aspects.
2. *How many instances are there in total (of each type, if appropriate)?*
    - The exact number of instances varies from cycle to cycle, but each instance represents a survey response from an individual participant. The number typically extends into the tens of thousands to ensure a representative sample of the population.
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).*
    - It's a sample of the larger Canadian population, specifically designed to be representative geographically and demographically. Statistics Canada uses stratified sampling to ensure that all provinces and key demographic groups are proportionately represented. The representativeness is validated through statistical techniques and comparison 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 survey response (instance) in the dataset consists of both raw and processed data, encompassing answers to survey questions on topics ranging from health to social participation. This data is processed to ensure accuracy and reliability before being made available for analysis.
5. *Is there a label or target associated with each instance? If so, please provide a description.*
    - Each instance includes labels related to the survey questions, such as health status or social activity levels, which can be used for analytical purposes to assess trends and correlations within the population.
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.*

- Yes. Statistics Canada employs various methods to handle missing data, including statistical imputation where appropriate (Canada 2024a).

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 includes metadata that allows researchers to examine relationships, such as familial or household connections, when this information is relevant to the survey's focus.

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.

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

   - Despite Statistics Canada used several ways to minimize the errors, considering it is a large scale survey,there might be some errors.

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 and does not rely on external resources that could change over time, ensuring its stability and reliability for longitudinal studies.

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.*

    - Data confidentiality is strictly maintained, with all personal identifiers removed to ensure that individual respondents cannot be directly or indirectly identified.

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 does not contain offensive content as it is purely factual and focused on socio-economic factors.

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 demographic characteristics such as age and gender, which are crucial for analyzing trends across different segments of the population.

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 from the dataset due to the anonymization and aggregation of data prior to public release.

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.*

    - While the dataset may contain sensitive information, such as health status or personal economic conditions, all data is anonymized and presented in aggregate form to protect individual privacy.

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 primarily acquired through direct survey responses reported by subjects, covering various socio-economic aspects. Validation of this data includes consistency checks and comparison against other national data sources to ensure reliability.

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?*

- Data was collected using a combination of computer-assisted telephone interviewing (CATI) and online self-administered questionnaires. These methods are validated through pilot tests and methodological research to ensure effectiveness 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 dataset uses a probabilistic sampling design, stratifying by geographical and demographic characteristics to ensure a representative sample of the Canadian 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)?*

   - Data collection was conducted by trained interviewers employed by Statistics Canada, compensated as per government employment standards.

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 quarterly to ensure timely updates on social trends. This timeframe matches the creation timeframe of the data, as each data release corresponds directly to its collection period.

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.* -As a government survey, the CSS follows strict ethical guidelines reviewed by Statistics Canada's ethical review board, ensuring all practices comply with national standards for research ethics.
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, without involvement from 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.*

   - Individuals were notified about the data collection at the beginning of each survey through a standard script which explains the purpose of the survey and how the data will be used.

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.*

    - Consent to collect data was obtained verbally at the beginning of each survey session, with interviewers explaining the use of the data and participants agreeing to proceed.

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).*

    - Participants can request the removal of their data or decline to participate at any stage of the survey, with mechanisms in place to ensure their data is not used in subsequent analyses if consent is withdrawn.

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.*

    - An impact analysis is typically conducted to assess the potential effects of the dataset on subjects, ensuring no adverse consequences arise from its use. This analysis is part of Statistics Canada's standard practice for all its surveys.

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 data collected through the CSS used several preprocessing and cleaning steps to ensure accuracy and usability. This typically includes handling missing values, correcting data entry errors, and standardizing responses to ensure consistency across the dataset. Additionally, responses might be categorized or bucketed to facilitate analysis, particularly when dealing with ordinal or categorical data.

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.*

- Statistics Canada typically retains both the "raw" survey data and the processed versions. The raw data, which includes all original responses with minimal modification, is often used for methodological testing or to support further research that may require re-analysis with different parameters.

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 specific software used for data preprocessing, cleaning, and labeling by Statistics Canada is generally not publicly available.

4. *Any other comments?*

   - No.

Uses

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

   - CSS has been used for various research tasks, primarily focusing on understanding social trends and impacts of events like the COVID-19 pandemic on Canadian society.

2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*

   - No.

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

   - The CSS could be employed in predictive modeling for social policy outcomes, sociological research to track changes in cultural norms, and more extensive public health research.

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 way the CSS is collected and processed ensures a high level of data integrity and representativeness, which is crucial for equitable analysis. However, users must be cautious about interpreting the data without considering the context of the questions and the manner in which responses were gathered to avoid biases that could lead to unfair treatment or stereotyping of individuals or groups.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
    - The dataset should not be used for purposes that require medical or highly sensitive personal data, as it does not contain detailed personal health information or other sensitive personal details that are protected under privacy laws.
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.*
    - No.
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
    - On Statistics Canada. Yes, it's https://doi.org/10.25318/4510008001-eng.
3. *When will the dataset be distributed?*
    - February 13, 2024
4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*
    - No.
5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*
    - No.
6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
    - No.
7. *Any other comments?*

- No.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
   - Statistics Canada.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
   - https://www.statcan.gc.ca/en/reference/refcentre/index provides many ways to contact
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 is updated regularly, in line with the survey's quarterly cycle. Statistics Canada is responsible for these updates and communicates them through its official website and via release announcements that can be subscribed to via email or RSS feeds.
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.*
   - For datasets involving personal data, Statistics Canada adheres to strict privacy laws which dictate data retention policies. Participants are generally informed about the retention period during the survey consent process. This period aligns with the purpose of the data collection and is enforced through internal data governance policies.
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.*
   - No. On their official website.
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.

8. *Any other comments?*
    - No.

References

Canada, Statistics. 2024a. "Canadian Social Survey." https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=5354.

———. 2024b. "Perceived Mental Health, by Gender and Other Selected Sociodemographic Characteristics." https://doi.org/10.25318/4510008001-eng.

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.