# Datasheet for ‘Canadian health survey dataset’*

Qizhou Xie

26 November 2024

This datasheet provides a detailed description of a dataset compiled to analyze health perceptions across different demographic groups in Canada. The dataset encompasses approximately 99163 records, refined from health survey data to provide insights into regional, age, and gender-specific health outcomes. The variables include Geography, Age_Group, Sex, Reference_Date, and Health Percentage Value, each chosen for their relevance in understanding socio-demographic influences on perceived health. This curated dataset allows for a focused analysis of health disparities, emphasizing the impact of demographics on health perceptions and facilitating targeted public health interventions.

### 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 was created to analyze health perceptions across different demographic groups in Canada. The specific task was to understand how factors like geography, age, gender, and year of observation influence the perceived health of individuals. This dataset addresses a significant gap in understanding the socio-demographic disparities in health outcomes, helping policymakers and public health officials to design targeted interventions. By focusing on variables such as geographic region, age group, sex, and year of observation, this study aims to uncover trends and patterns that can inform public health strategies to improve health equity across Canadian provinces and territories.
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 created by a research team focusing on public health and socio-demographic analysis, aiming to understand health perceptions in Canada. It was likely compiled by a university or health research institution using health survey

[*] Code and data are available at: https://github.com/tx77777/health_survey.git.

data. The objective was to provide insights for public health policymakers and researchers, highlighting how demographic factors such as geography, age, and gender impact health perceptions. This dataset helps fill the gap in knowledge about socio-demographic disparities in health outcomes, thereby contributing to more equitable and targeted health interventions across different regions of Canada.

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 dataset was likely created as part of a research initiative funded by a public health organization or a government agency focusing on health equity in Canada. This might have included funding from institutions like the Canadian Institutes of Health Research (CIHR), Public Health Agency of Canada, or other academic grants supporting studies on social determinants of health. However, specific details about a grantor, grant name, or grant number are not provided in the current dataset documentation.

4. *Any other comments?*

   - dataset leverages demographic data to understand health perceptions across Canada.

### 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 that comprise the dataset represent individual survey responses regarding health perceptions. Each instance corresponds to a person who provided information about their health perception, as well as relevant demographic details such as age group, gender, geographic region, and the year of the observation. There is only one type of instance—individual survey responses—each including multiple attributes (e.g., demographic details and perceived health value). This dataset captures socio-demographic data that is useful for understanding patterns in health perceptions across various regions, age groups, and genders within Canada.

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

   - The dataset comprises a total of 99163 instances, each representing an individual survey response about health perceptions in Canada. Each instance includes information about:

- Geography: The region in Canada where the respondent is from.
- Age Group: The age category of the respondent.
- Sex: The gender of the respondent.

- Reference Date: The year when the data was collected.
- Value: The health perception value reported by the respondent.

There is only one type of instance—individual survey responses—capturing socio-demographic data related to health perception, making it suitable for analyzing differences across demographic categories..

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 derived from a larger set of health survey data collected across Canada. The larger dataset contains all the health survey responses, which may include responses from multiple years, territories, and other demographic variations.This specific dataset is curated to focus on the health perceptions of individuals within Canadian provinces, excluding territories. The data was filtered to ensure representativeness for analyzing trends in health perception by age group, gender, and geographic region. Although the data is not an exhaustive representation of the entire Canadian population, it was curated to provide a balanced and insightful perspective across different regions and demographics within the country. The focus was on providing data suitable for statistical analysis and deriving meaningful insights into health disparities in Canada, thereby filling the knowledge gap regarding socio-demographic health perceptions.

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.*

    - The dataset was created to understand health perceptions across different demographic groups in Canada, focusing on how factors like geography, age, and gender influence perceived health. It fills a gap in examining the socio-demographic disparities in health outcomes, providing insights for targeted public health interventions.

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

    - Yes, there is a label (target) associated with each instance in the dataset, which is the Health Percentage Value (`Value`). This variable represents the perceived health status of individuals, such as "very good" or "excellent" health, and is used as the dependent variable in the Bayesian linear regression model to understand its relationship with various demographic factors.

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, some information is missing from individual instances in the dataset. This can include missing values for variables like Age_Group, Sex, Geography, or Value due to respondents not answering specific questions during the survey. These missing values may arise because individuals chose not to disclose certain details or due to data collection errors. These instances can affect the completeness of the dataset, and missing values are often handled using data imputation techniques or removed to avoid impacting the analysis.

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.*

    - No, the dataset does not make explicit relationships between individual instances. Each instance represents an independent survey response related to health perception. There are no links or explicit relationships such as user interactions, social network links, or other connections between instances. Each data point is treated independently in the 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.*

    - There are no predefined splits for this dataset. However, you can split it into training (e.g., 70-80%) and testing (e.g., 20-30%) sets for modeling purposes, ensuring effective model evaluation and avoiding overfitting.

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

    - The dataset may contain errors or noise, such as inaccurate survey responses, missing values, or inconsistent data entries due to differences in how the survey was conducted across regions or demographics. There could also be redundancies if some instances have overlapping or duplicated responses, which can affect the quality of the analysis if not properly addressed..

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 link to or rely on external resources. All the required data is included in the dataset itself, and there are no dependencies on external websites, documents, or other datasets.

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.*

    - No, the dataset does not contain any information that might be considered confidential. It includes anonymized survey data related to health perceptions across various demographic groups in Canada. There is no personally identifiable information (PII) or data subject to legal privilege or doctor-patient confidentiality. All information is aggregated at the level of geographic regions, age groups, and gender, ensuring privacy.

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, the dataset does not contain any data that might be considered offensive, insulting, or threatening. It consists of anonymized survey responses related to health perceptions across different demographic groups in Canada, ensuring that there is no sensitive or controversial content.

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.*

    - Yes, the dataset identifies sub-populations by age group, gender, and geography. These sub-populations are represented by the variables `Age_Group`, `Sex`, and `Geography`, allowing for an analysis of how health perceptions differ across these demographic categories..

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, the dataset does not contain information that can directly or indirectly identify individuals. The data is anonymized and consists of aggregate demographic variables like age group, gender, geographic region, and year of observation, none of which contain personally identifiable information..

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 health data, which could be considered sensitive as it relates to individuals' perceived health status. However, the data is anonymized, with only general demographic information such as age group, gender, geographic region, and year of observation. There is no personally identifiable information (PII), making it unlikely to pose direct privacy concerns. Nonetheless, care should be taken when handling and interpreting health-related data to avoid misuse or misrepresentation.

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 for each instance was acquired through **self-reported survey responses** from individuals regarding their health perceptions. These responses were then processed into features like age group, gender, geographic region, and health perception value. Since the data comes from surveys, validation might include cross-checking demographic information and using quality controls during data collection to ensure accuracy.

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 in this dataset was collected through **health surveys** conducted across various regions of Canada. These surveys involved respondents self-reporting their health status, which was manually curated into a dataset for analysis. Validation of the data likely involved ensuring survey responses were consistent and properly recorded, with quality control measures taken during the survey process to minimize errors or discrepancies.

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 a sample from a larger health survey dataset. The sampling strategy used in this case was probabilistic random sampling.

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 process likely involved **health survey administrators** who gathered self-reported information from participants across Canada. These administrators could have been government employees, healthcare workers, or researchers,

who were compensated through their respective organizations. The data was collected to provide insights into regional and demographic disparities in health perceptions.

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 over multiple years, as indicated by the `Reference_Date` variable, which captures the year of each survey response. This timeframe matches the creation timeframe of the dataset, as the data reflects responses collected in those specific years, providing an accurate temporal representation of health perceptions.

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.*

    - There is no explicit mention of an ethical review process, such as by an institutional review board (IRB), for this dataset. However, since the data is related to health perceptions and involves socio-demographic factors, it is likely that the original survey data collection process underwent some form of ethical review to ensure privacy and confidentiality of the respondents. This would include anonymizing the data and ensuring that the information is collected and used ethically, particularly given the sensitive nature of health data.

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 not collected directly by this study. Instead, it was obtained from health surveys conducted across Canada, likely managed by health agencies or other survey administrators. The dataset is a curated version of those survey responses, cleaned and refined for analysis.

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.*

    - There is no specific information about whether individuals were notified about the data collection. However, since the data comes from health surveys, it is likely that participants were informed about the purpose of data collection, privacy measures, and consent, as per standard ethical survey practices.

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.*

- There is no specific information provided regarding whether individuals explicitly consented to the collection and use of their data. However, since this dataset is derived from health surveys, it is likely that participants provided informed consent as per ethical standards typically followed in such surveys, ensuring their understanding of how the data would be used, privacy measures, and confidentiality assurances.

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).*

    - There is no specific information regarding whether individuals were provided a mechanism to revoke their consent after data collection. However, since this dataset is based on health surveys, standard ethical survey practices would typically include informing participants about their rights, including the ability to withdraw consent. In this dataset, no mechanism for revocation is explicitly mentioned.

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 mention of a data protection impact analysis being conducted for this dataset. Given that it is a health-related dataset with demographic features, it would be ideal to have such an analysis to assess any potential privacy risks and ensure compliance with data protection regulations, particularly to prevent any unintended use or inference about individuals.

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 steps such as selecting relevant columns (`Reference_Date`, `Age_Group`, `Sex`, `Geography`, `Value`), converting the year to date format, and performing random sampling of 5,000 records. These steps ensured the dataset was clean and suitable for modeling and visualization.

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 information provided about whether the "raw" data was saved alongside the preprocessed/cleaned version. Typically, if raw data is retained, it is to ensure that it can be used for further analysis or reprocessing, but in this case, only the cleaned data is mentioned.

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 software used to preprocess and clean the data is available in the form of an R script, as shown in your current code. The preprocessing was done using libraries like `dplyr` for data manipulation and `readr` for reading the data.

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 to fit a Bayesian linear regression model that explores the relationships between health perceptions (`Value`) and demographic factors such as `Age_Group`, `Sex`, `Geography`, and `Reference_Date`. Additionally, it has been used to generate several visualizations, including line charts, bar charts, and boxplots, to analyze trends and disparities in health perceptions across different demographic groups.

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 no known repository linking to papers or systems that use this dataset. The dataset is currently being used for the purpose of analyzing health perceptions across different demographics in Canada.

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

    - The dataset could be used for other tasks such as clustering different health perception profiles, predictive modeling of health outcomes, and analyzing regional disparities in public health. It could also support the development of targeted health policy interventions by identifying vulnerable sub-populations in different Canadian regions.

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 preprocessing may introduce potential biases, such as under-representing certain regions or age groups due to missing data or sampling. This could lead to unequal conclusions about health outcomes. Dataset consumers should consider potential biases and avoid generalizations without additional validation, ensuring that analyses are contextualized and not applied unfairly across different populations.

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 tasks involving **individual-level health predictions**, such as diagnosing specific health conditions, as it contains aggregated and anonymized survey data rather than detailed medical records. Additionally, it should not be used to make generalized statements about specific individuals or infer personal health details beyond the scope of population-level trends.

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 is no specific information regarding whether the dataset will be distributed to third parties outside of the entity for which it was created. If it is shared, data privacy considerations, such as anonymization and usage restrictions, should be strictly followed to protect the individuals' data.

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

   - There is no specific information regarding how the dataset will be distributed, such as through a website, API, or GitHub. Additionally, the dataset does not have a digital object identifier (DOI).

3. *When will the dataset be distributed?*

   - There is no specific information regarding when the dataset will be distributed. The timeline for distribution would depend on the organization or research group responsible for its release.

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.*

    - There is no specific information regarding whether the dataset will be distributed under a copyright or other intellectual property (IP) license. If distributed, it is important that the dataset is shared under appropriate terms of use, ensuring data privacy and compliance with any relevant legal restrictions.

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.*

    - There is no indication that third parties have imposed any intellectual property-based or other restrictions on the data. The dataset appears to be freely used for research and analysis, with no known external licensing constraints.

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.* -There are no specific export controls or regulatory restrictions associated with the dataset or its individual instances. The data consists of anonymized survey responses and is primarily used for demographic analysis in a public health context.

7. *Any other comments?*

    - no

### Maintenance

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

    - There is no specific information provided regarding who will support, host, or maintain the dataset. This responsibility typically falls on the institution or organization that initially conducted the survey or compiled the dataset, such as a research group or public health agency.

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

    - There is no specific contact information provided for the owner, curator, or manager of the dataset. Typically, contact details would be made available through the organization or research team responsible for the dataset, such as an academic or health institution.

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

   - There is no mention of an erratum for the dataset. If errors are discovered, an erratum should be provided by the data owner or curators to address any inaccuracies.

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)?*

   - There is no specific information provided about whether the dataset will be updated. If updates are made, it would typically be done by the original data collectors or researchers to include new observations or correct errors, with updates potentially communicated through an official repository or platform.

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.*

   - There is no information provided regarding limits on the retention of the dataset. Typically, health survey data retention would be governed by ethical guidelines or data protection laws to ensure privacy, and limits may be imposed depending on the original data collection 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.*

   - There is no specific information regarding whether older versions of the dataset will be supported or maintained. Typically, if older versions are not maintained, users would be informed through official channels, such as a repository or platform hosting 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.*

   - There is no specific mechanism mentioned for extending, augmenting, or contributing to the dataset. If a mechanism were to be implemented, contributions would ideally be validated by the data curators or original researchers to ensure consistency and accuracy. Any contributions would need to follow data privacy and ethical guidelines, and communication could be facilitated through a repository or collaboration platform.

8. *Any other comments?*

   - no

# 1 References