# Datasheet for Unveiling Vulnerabilities*

## An Analysis of Causes of Death Across Age Groups and Genders Within Toronto's Homelessness Population

Faiza Imam

April 23, 2024

This datasheet describes the motivations and composition of homesless death by cause data set obtained from OpenData Toronto. It also describes collection process, cleaning process, uses, distribution, and maintenance of the data sets.

**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 "Deaths of People Experiencing Homelessness" was created to track the deaths of people experiencing homelessness more accurately and understand their causes. The specific task was to improve data collection and analysis in order to gain insights into the mortality rates and causes of death within the homeless population in Toronto.
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 Toronto Public Health (TPH) as part of their initiative to track and analyze homeless deaths. TPH oversees the data collection, analysis, and reporting, with support from the Shelter, Support and Housing Administration (SSHA) and various health and social service agencies that assist the homeless population. The Office of the Chief Coroner of Ontario (OCCO) also verifies some of the data.
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 funding for the creation of this dataset is not specified in the provided information.
4. *Any other comments?* -There are no additional comments provided at this time.

**Composition**

*Code and data are available at: https://github.com/fazim07/deaths_gender

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

    - The instances in the dataset represent deaths of people experiencing homelessness in Toronto. Each instance likely includes information such as the cause of death, age group, year of death, gender.

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

    - The total number of instances in the dataset is not specified in the information provided. The file contains 253 entries in total.

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).*

    - Is the dataset a sample or representative of a larger set? The dataset represents reported deaths of people experiencing homelessness in Toronto and is not necessarily a random sample. It reflects data collected and reported by Toronto Public Health (TPH), the Shelter, Support and Housing Administration (SSHA), and other community partners, and is subject to limitations such as unknown causes of death, missing Indigenous status information, and unreported deaths.

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 likely consists of structured data fields such as cause of death, age group, year of death, gender, and possibly additional information related to the circumstances of the death.

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

    - The main target or focus of the dataset is the cause of death among people experiencing homelessness, with additional variables providing context such as age group, gender, and year of death.

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, there are limitations mentioned regarding missing information such as unknown causes of death, missing Indigenous status, missed or unreported age groups, races, ethnicities, and unreported deaths.

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 information from Opendata provided does not mention explicit relationships between individual instances.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

    - No specific data splits or recommendations are mentioned in the provided information. Are there any errors, noise, or redundancies in the dataset?

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

    - The dataset is subject to limitations such as unknown causes of death, missing Indigenous status information, and unreported deaths, which could be considered sources of noise or potential 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 appears to be self-contained, based on the provided information on Opendata Toronto data portal.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

    - The dataset likely contains sensitive information regarding individuals' deaths, but specific details about confidentiality are not provided but can assumed to be confidential.

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 contains information related to deaths, which may be sensitive but is not explicitly described as offensive or threatening.

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 includes sub-populations based on age groups and genders. The gender variable covers three entries; male, female and unknown. The *unknown* entries are not classfied nor described. The age group variable covers four demographs, <20, 20-39, 40-59, 60+.

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.*

    - The dataset likely does not contain personally identifiable information such as names or specific identifiers, there is an *id* number which is associated with every entry. Does the dataset contain sensitive data?

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 sensitive data related to deaths, causes of death, and demographic information, but specifics about sensitive categories like race, sexual orientation, etc., are not provided. The cause of death variable covers 9 common cuases of death within Toronto's homeless population.

16. *Any other comments?*

    - No comment

**Collection process**

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

    - The data was acquired by TPH collecting reports of deaths among the homeless population in Toronto. Therefore there maybe some gaps due to individual being miscounted or uncounted for.

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?*

    - No addtional information regarding data collection is provided in the description.

3. *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)?*

    - Toronto Public Health, no other informaiton is provided.

4. *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.*

    - Data collection was a binannual collection effort.

5. *Any other comments?*

    - No Comments

**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.*

    - No informaiton regarding data cleaning is provided

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.*

    - Raw data is saved here;https://github.com/fazim07/deaths_gender/tree/master/data/raw_data

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

    - Statistical programming language R was used to code and clean the data; https://github.com/fazim07/deaths_gender/blob/master/scripts/02-data_cleaning.R

4. *Any other comments?*

    - No comment

**Uses**

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - No tasks have been provided to indicate as such. Possible tasks may be municipal reports regarding this data.
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 repository is provided to suggest as such.
3. *What (other) tasks could the dataset be used for?*
    - Dataset could be used to analyze further trends across ethnicity, race and sociao-economic status. Additionally, the data can provide insight into the policy and vital information for the Government when making laws.
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?*
    - All the data from set is confidential and contains death counts and sensitive data on cause of death that may be trigger to future users.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
    - No, unless the intention is malicious and the intend to spread misinformation.
6. *Any other comments?*
    - No comment

**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.*
    - While being hosted on Opendata portal, its free access to the public hence a third party may have access to it.
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
    - The data set comes in a CSV, JSON or Excel file no DOIs.

3. *When will the dataset be distributed?*
    - The latest update or distribution was on Apr 17, 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.*
    - Dataset is licensed under Open Government Licence – Toronto - https://open.toronto.ca/open-data-license/
5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*
    - No third party IP based or other restrictions on the data association and instances
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.*
    - Cannot access who these individuals are that have passed.
7. *Any other comments?*
    - No comment

**Maintenance**

1. *Who will be supporting/hosting/maintaining the dataset?*
    - Opendata Toronto portal will host and support the dataset while Opendata Toronto updates the data set.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - Can be contacted via email from Opendata Toronto
3. *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)?*
    - Yes, this data updates in a end of year manner adding new collected satistics of the homelessness populations. The data collection will be collected by TPH and maintained and stored in Opendata Toronto.

4. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

    - The data refelects on the homeless peoples deaths, however, other than the cause, age and gender there is no other additional personal information provided hence no need for retention.

5. *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 data are supported and maintained as the study is continues to be collected and updated by the end of the year by the TPH.

6. *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.*

    - TBD

7. *Any other comments?*

    - No Comments