Datasheet for 'Toronto_Shelter'*

YiZhuo Li

27 September 2024

The 'Toronto_Shelter' dataset tracks shelter usage in Toronto, including occupancy rates, available beds, and shelter types, updated daily by the city's Housing and Homelessness Services department. It's self-contained, lacks sensitive information, and supports policy-making and resource allocation for homelessness services. Available through an API on the City of Toronto's official site, the dataset can be used to analyze trends and predict shelter demand but should not be used for individual-level analysis. Missing values marked as "NA" may require consideration during analysis.

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 census aims to monitor the capacity, utilization, and turnover in Toronto's shelter system to understand the changing population of people experiencing homelessness and to inform policy decisions.

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

    - The data is collected and managed by the City of Toronto, specifically by the Housing and Homelessness Services department.

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 initiative is typically funded through the city's budget allocations and relevant housing policy funds.

*A GitHub Repository containing all data, R code, and other files used in this investigation is located here: https://github.com/eason1218/Toronto_Shetler.git

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 dataset includes information on Toronto shelter system usage, such as daily occupancy rates, shelter types (e.g., for families, men, women), the number of residents, and available beds.

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

    - Dataset is updated daily at 4 a.m., with new additions each day.

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).*

    - No, the dataset is not a sample; it is the complete data directly uploaded from the shelters.

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 now includes: Date: A timestamp for when the data was recorded. Occupancy Values: Two different columns for occupancy rates; some cells are missing values (NA). Location Information: Specific details about the shelter's name and type. Actual Beds/Occupancy: Counts for the actual number of beds available or occupied. Sector: Categories such as "Youth," "Mixed Adult," "Men," or "Women" indicating the type of population served by the shelter.

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

    - No explicit target or label is identified, but columns such as "Actual Beds" and "Occupancy" could be considered potential targets for analysis.

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 cells contain "NA," indicating missing values in columns like occupancy rates or bed counts. The extent and pattern of missingness need to be evaluated further.

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.*

    - Relationships between instances are not explicitly marked. However, they can be inferred based on shared fields such as the location or sector.

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.*

    - The dataset structure is not inherently split into different subsets (e.g., for training and testing). The data could be split based on shelter types, occupancy rates, or other factors for specific analyses.

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

    - There is missing data in some columns, as shown by the "NA" values. This could be considered a source of noise or incompleteness. It's important to inspect the data further to check for other inconsistencies or redundancies.

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 any external resources.

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 confidential information.

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 content that might be offensive or cause anxiety.

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.*

- No, the dataset does not identify any specific sub-populations.

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 could be used to directly or indirectly identify 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.*

    - No, the dataset does not contain any sensitive information.

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.*

    - Data is typically reported daily by shelter staff and recorded in the city's data systems to ensure the accuracy of the shelter census data.

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 is typically reported daily by shelter staff and recorded in the city's data systems to ensure the accuracy of the shelter census data.

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 not a sample from a larger set; it is directly reported by shelter staff.

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)?*

    - Shelter staff are responsible for data entry, compensation for shelter staff comes from their regular wages.

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 dataset is updated daily.

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.*

    - The dataset is provided and processed by official agencies of the City of Toronto.

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 comes from the City of Toronto's official open data website.

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.*

    - Data is reported daily by shelter staff and entered into the city's data systems.

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 involvement of individual personal data.

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 may undergo cleaning and validation to ensure accuracy, which could include addressing missing data or correcting input errors.

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.*

    - It is not specified whether the raw data is saved.

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 availability of preprocessing software is not mentioned.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - The dataset is primarily used for policy-making, resource allocation, and monitoring the effectiveness of services for those experiencing homelessness. It may also be used to analyze trends and plan improvements to the shelter system.
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, there's no repository for papers or systems using the dataset.
3. *What (other) tasks could the dataset be used for?*
    - The dataset could be used for tasks such as analyzing shelter usage trends, predicting future shelter demand, assessing the impact of policy changes, and understanding the demographics of homelessness to improve targeted services.
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?*
    - No major concerns, but avoid misinterpreting macro trends for policy-making.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
    - Should not be used for individual-level analysis or assumptions about causes of homelessness.

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.*
    - The Shelter Census data is usually publicly available on the City of Toronto's official website for use by the public, researchers, and policymakers.
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
    - Distributed via API.

3. *When will the dataset be distributed?*
    - Already publicly available.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - The Housing and Homelessness Services department of the City of Toronto is responsible for maintaining the Shelter Census dataset.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - Email: 311@toronto.ca, Phone: 416-392-2489.
3. *Is there an erratum? If so, please provide a link or other access point.*
    - No additional information is provided.
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)?*
    - No additional information is provided.

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.