Datasheet For Dataset: Daily Shelter & Overnight Service Occupancy & Capacity[*]

Aakash Vaithyanathan

December 14, 2024

This document provides a datasheet for the Toronto Daily Shelter & Overnight Service Dataset provided and maintained by OpenData Toronto. This document provides additional information about our dataset and answeres the preset questions provided by Timnit Gebru et al.

The following document provides a datasheet for the Daily Shelter & Overnight Service Occupancy & Capacity data set (Toronto Shelter & Support Services 2024). The template for the questions in the datasheet are provided by the Datasheet for Dataset paper (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 data set was created to track the daily active overnight shelter services across Toronto. The data set includes data about bed based and room based shelters with details about demographics, shelter types and services the shelter offers daily. It highlights the need for increased funding needed to improve the homelessness situation in Toronto. To address the gap in missing shelter occupancy data, the dataset was improved to include information about overnight shelter service type, capacity type (bed or room based) and metrics regarding occupancy rate and daily occupancy count of beds and their total available capacity.

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

[*] Code and data is available at: https://github.com/aakash2002/quantitative-analysis-canadian-shelters.

- The dataset was created by Toronto Shelter & Support Services. They can be contacted using this email.

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.**

   - City of Toronto funded this dataset.

4. **Any other comments?**

   - This datasheet is for the 2024 version of the dataset which is updated daily.

Composition

5. **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 contains instances that represent the daily overnight service count at shelters across various locations. The different types of instances offered are demographics, shelter types, shelter overnight service type, capacity type which are categorical that represent the various possible types for these. In our cleaned dataset, we use an interaction between the demographic and shelter type.

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

   - There are a total of 44300 instances in this dataset. After removing the room based shelter capacity types and missing values, we have 31484 instances remaining in the dataset for bed based capacity types.

7. **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 contains all possible shelter instances in the dataset. However, after filtering based on bed based shelter capacity type, we do not have any instances for ‘Family’ demographics. Since the dataset used in this study is for the year 2024, no instances for ‘COVID-19’ shelter service type was found in the data.

8. **What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.**

    - In our raw dataset, each instance consists of some key features like daily shelter service count, shelter type, shelter service type and demographic. Additional features included in the raw data are shelter location, daily bed occupancy rate and more.

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

    - For each instance, features like shelter service type, shelter type and demographic are non-binary categorical with countable distinct values these features can take.

10. **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.**

    - Since the dataset includes bed based and room based capacity type for shelters, instances that contain bed based data have missing values in these instances for room based shelter capacity like the room shelter occupancy rate.

11. **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 relationships between individual instances is mentioned in the dataset. In this study, we look at the interaction between the demographic and shelter type instances when constructing a negative binomial model for our analysis.

12. **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 such data splits are mentioned by the authors of this dataset.

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

    - The dataset is un-audited and compiled directly from an administrative database. Additionally, the dataset is only updated by 4 AM everyday except for the weekends. As such, at any point, it may not be a complete representation of the shelter data in Toronto.

14. **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 references was built on top of the archived version of the Daily Shelter Occupancy data provided by the Toronto Shelter & Support Services. This archived data was last updated in October 15, 2024 and is no longer refreshed. Other than this, the data does not rely on any external data source.

15. **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.**

    - Not applicable.

16. **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.**

    - Not applicable.

17. **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 contains feature for the demographic of the people seeking shelter service. The types of demographics available in our dataset are Men, Women, Youth, Family and Mixed Adults. Mixed Adults represent adults with fluid gender identity.

18. **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.**

    - Not applicable.

19. **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.**

    - Not applicable.

20. **Any other comments?**

    - None.

Collection process

21. **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 through the mandatory data recorded by the Toronto Shelter & Support Services under the Shelter Management Information System (SMIS) database. All shelters are required to record the daily instance details into the SMIS database which is refelcted by 4 AM everyday except for weekends.

22. **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 was collected using the Shelter Management Information System (SMIS) database and and verified through the Toronto Shelter & Support Services.

23. **If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?**

    - Not applicable.

24. **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 was done by Toronto Shelter & Support Services and the workers at various shelters reported in the dataset. The compensation for these workers is not specified.

25. **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 was collected for every day since 2021 and is continuing to be updated daily. At the time of writing this datasheet, the instances for December 2024 are being collected on a daily basis.

26. **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.**

    - Not applicable.

27. **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 retrieved the Open Data Toronto by the from Toronto Shelter & Support Services. This portal is maintained by the City of Toronto.

28. **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.**

    - Information regarding notifying individuals in the data collection process isn't mentioned. The dataset doesn't contain information about any specific individuals but rather contains for a demographic type on a daily basis.

29. **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.**

    - As discussed above, individual consent was not required.

30. **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).**

    - As discussed above, individual consent was not required.

31. **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.**

    - Not applicable.

32. **Any other comments?**

    - None.

Preprocessing/cleaning/labeling

33. **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 has undergone cleaning process. Firstly, we filter out to only include bed based shelter capacity type. We then proceed to removing any missing values from the rows of data. Lastly, we only select the requried feature for our study and drop the unnecessary column features.

34. **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 is available at the Open Data Toronto Portal. The raw dataset can be found here.

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

    - The preprocessing was done using R (R Core Team 2024) programming language and its packages like Janitor (Firke 2024), Arrow (Richardson et al. 2024) and Caret (Kuhn and Max 2008). The repository can be found here. The script performing the cleaning steps can be found at: scripts/03-clean_data.R

36. **Any other comments?**

    - None.

Uses

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

    - The dataset is used in a study to understand the Toronto demographic seeking shelter services for the year 2024. The study aims to understand the factors that influence the daily count of the demographics that seek shelters.

38. **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.**

    - The paper and all code used for this study can be found here.

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

    - This dataset can be used to study what locations of shelters are most popular or how does the occupancy rate vary by the month of the year. These types of analysis can help understand the regions or months that are at their highest capacity and how additional resources can be allocated to help alleviate the stress on shelters to better support the homeless population.

40. **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 includes unaudited data from the Shelter Management Information System (SMIS) database. It is important to note that the dataset at any given point of time may not be a valid representation of the current shelter state. The dataset is updated daily by 4 AM and isn't updated on the weekends. Since the data contains the daily shelter service count for various demographics, there are no risks of privacy, unfair treatment or legal concerns.

41. **Are there tasks for which the dataset should not be used? If so, please provide a description.**

- The dataset should not be used to draw conclusions regarding the 24 hour respite site shelter services or warming center based shelter service types as these account for only about 2 % and 1.2 % of observations. Due to the lack of sufficient observations, we must be cautious of the potential skew in data results that we may observe.

42. **Any other comments?**

    - None.

Distribution

43. **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 dataset is maintained by Toronto Shelter & Support Services and published on the Open Data Toronto portal. The dataset is published under the Open Government License - Toronto and is allowed by anyone to be used with appropriate citation.

44. **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 at the Open Data Toronto portal. The available formats for the data is CVS, XML and JSON.

45. **When will the dataset be distributed?**

    - The dataset is available at the Open Data Toronto portal and updated daily.

46. **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.**

    - The dataset is published under the Open Government License - Toronto, and can be used by anyone with appropriate citation.

47. **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.**

    - Not applicable.

48. **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.**

    - Not applicable.

49. **Any other comments?**

    - None.

Maintenance

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

    - The dataset is maintained by Toronto Shelter & Support Service. It is updated daily.

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

    - The team at Toronto Shelter & Support Services can be contacted here.

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

    - Not applicable.

53. **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 daily on the Open Data Toronto portal.

54. **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.**

    - Not applicable.

55. **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 version of the dataset is archived and no longer used. The archived version of the dataset was last updated on October 15, 2024.

56. **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.**

    - Since the dataset is under the Open Government License - Toronto, any individual can use the dataset and contribute. In order to make modifications, the individual would need to work with the Toronto Shelter & Support Services. For the detailed steps for how to begin contributing, the individual can contact the staff at Toronto Shelter Services here.

57. **Any other comments?**

    - None.

References

Firke, Sam. 2024. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

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.

Kuhn, and Max. 2008. "Building Predictive Models in r Using the Caret Package." *Journal of Statistical Software* 28 (5): 1–26. https://doi.org/10.18637/jss.v028.i05.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Toronto Shelter & Support Services. 2024. *Daily Shelter & Overnight Service Occupancy & Capacity*. https://open.toronto.ca/dataset/daily-shelter-overnight-service-occupancy-capacity/.