Datasheet for Building Evaluation Scores in Downtown Toronto Study*

Yiyi Yao

December 14, 2024

This datasheet is the extract of the questions from Gebru et al. (2021). And it was put together with the help of Alexander (2023)

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 assess compliance with property standards in Toronto's multi-unit residential buildings under the RentSafeTO program. The purpose was to measure how building characteristics, such as year built, number of units, property type, and geographic location, influence evaluation scores. The goal was to support policy decisions, target maintenance efforts, and ensure better enforcement of housing standards across the city. For this study, the dataset meets the requirements for analyzing these factors, and there was no specific data gap that needed to be addressed.

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 the City of Toronto's RentSafeTO program. This initiative was established by Toronto Municipal Licensing & Standards to monitor and enforce compliance with property maintenance standards in multi-unit residential buildings.

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's creation was funded by the City of Toronto as part of its municipal operations under the RentSafeTO program. No specific external grants were used, as the program is city-funded.

*Code and data are available at: https://github.com/Yaoee111/condo_evaluation

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.*
    - Each instance represents a building in Toronto inspected under the RentSafeTO program. Instances include information about building characteristics, evaluation scores, and geographic location.
2. *How many instances are there in total (of each type, if appropriate)?*
    - There are 979 building instances in the dataset.
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 of multi-unit residential buildings in Toronto that were evaluated under the RentSafeTO program. It includes buildings meeting specific criteria for evaluation (e.g., number of units). While it does not cover all residential buildings in Toronto, it is representative of the population of buildings subject to RentSafeTO inspections.
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 consists of processed features, including year built, year registered, property type, number of units, number of storeys, geographic ward, and the evaluation score.
5. *Is there a label or target associated with each instance? If so, please provide a description.*
    - Yes, the target variable is the CURRENT.BUILDING.EVAL.SCORE, which represents the building's evaluation score based on compliance with property standards.
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.*
    - No, there is no missing information relevant to this study.
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 include explicit relationships between instances. Each instance represents a single building.

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 recommended data splits are provided in the dataset.

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

   - There are no known errors, sources of noise, or redundancies in the dataset.

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.

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

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

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 buildings by geographic sub-populations through the WARD variable, which indicates the location of each building within Toronto's ward boundaries.

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, it is not possible to identify individuals directly or indirectly from the dataset. It focuses on buildings rather than personal 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.*

    - No, the dataset does not contain sensitive data related to individuals. It focuses on building characteristics and compliance with property standards.

16. *Any other comments?*

    - No further comments.

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 directly observable during building inspections conducted under the RentSafeTO program. Evaluation scores were assigned based on inspectors' observations of compliance with property standards. Information on building characteristics, such as year built and number of units, was acquired from municipal records, which are validated as part of the city's administrative processes.

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 through in-person building inspections conducted by RentSafeTO inspectors, who used standardized checklists and procedures to assess compliance. Supplemental data, such as building age and size, was obtained from municipal property databases. The validation occurred through internal audits of inspection processes and cross-referencing with official records.

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 of multi-unit residential buildings in Toronto selected for RentSafeTO inspections. Buildings included in the program meet specific criteria, such as having three or more units. The sampling is not probabilistic but is determined based on eligibility for the RentSafeTO program.

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 was conducted by licensed municipal inspectors employed by the City of Toronto. Their compensation followed standard municipal pay scales for government employees.

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 from the launch of the RentSafeTO program in 2017 through subsequent inspections conducted up to 2024. The timeframe matches the creation timeframe, as data for each instance corresponds directly to the inspection date or the year it was recorded.

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.*

    - No formal ethical review process was conducted, as the dataset consists of administrative records and does not involve sensitive personal information.

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 dataset was obtained from the City of Toronto's open data portal. It was accessed directly through municipal records made available to the public.

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.*

    - Property owners were notified about inspections under the RentSafeTO program as part of the enforcement process. Notifications were issued through official letters and notices in accordance with municipal regulations.

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 was not explicitly required as the data collection was conducted under the authority of municipal bylaws, which mandate compliance with property standards. Property owners were informed about the program and its enforcement procedures when it was launched.

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).*

    - No

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.*

    - No

12. *Any other comments?*

    - No further 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.*

    - Yes, the data was preprocessed to select only the columns needed for the study and to remove rows with missing observations. The selected columns include YEAR.BUILT, PROPERTY.TYPE.CODE, WARD, CONFIRMED.UNITS, CONFIRMED.STOREYS, and CURRENT.BUILDING.EVAL.SCORE. Rows with missing data in any of these columns were excluded from the analysis to ensure consistency and completeness.

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.*

    - Yes, the "raw" data was saved and is publicly accessible. It can be accessed through the City of Toronto's open data portal.

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 for preprocessing the data was the R Programming Language (R Core Team 2023). It is freely available at https://cran.r-project.org.

4. *Any other comments?*

    - No further comments.

Uses

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

    - Yes, this dataset is publicly available and has been used for various analyses related to Toronto's RentSafeTO program. Researchers and policymakers have likely used it to study building compliance with property standards and identify patterns in inspection outcomes.

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 dataset could be used to explore additional relationships, such as the impact of socioeconomic factors on building maintenance or the effectiveness of specific RentSafeTO enforcement practices. It could also be used for geographic studies to compare compliance patterns across neighborhoods.

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?*

    - As the dataset reflects inspections conducted between 2017 and 2024, it may not represent current conditions in buildings, especially if maintenance or renovations have occurred since then. Consumers should use caution when making claims about current compliance or future predictions based on this data. Ensuring proper temporal context in analysis can help mitigate these risks.

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

    - The dataset should not be used to evaluate individual property owners or tenants, as it is intended for studying overall compliance patterns and not for identifying specific personal behaviors or circumstances.

6. *Any other comments?*

    - No further comments.

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.*

- Yes, the dataset is publicly available through the City of Toronto's open data portal. Any individual, company, institution, or organization can access it online.

2. *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 distributed via the City of Toronto's official open data portal in CSV format. It does not have a DOI.

3. *When will the dataset be distributed?*

   - The dataset has been publicly available since the launch of the RentSafeTO program and is updated periodically as new inspections are conducted.

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 further comments.

Maintenance

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

   - The City of Toronto is responsible for supporting, hosting, and maintaining the dataset through its official open data portal.

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

   - The City of Toronto's Municipal Licensing & Standards department can be contacted through the official city website for inquiries related to RentSafeTO data.

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)?*

    - Yes, the dataset is periodically updated by the City of Toronto as new inspections are conducted or new data becomes available. Updates are reflected directly on the open data portal without specific announcements.

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.*

    - No

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.*

    - Older versions of the dataset are not explicitly maintained, but the most current version is always available on the City of Toronto's open data portal.

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, this dataset is managed by the City of Toronto and cannot be directly extended or augmented by external contributors. However, researchers can build their own analyses using the dataset, but such contributions would be separate and not integrated back into the original dataset.

8. *Any other comments?*

    - No further comments.

References

Alexander, Rohan. 2023. *Telling Stories with Data: With Applications in R*. Chapman; Hall/CRC.

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.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.