Datasheet for 'Highrise Residential Fire Inspection Results'*

Motivation, Composition, Collection Process, Preprocessing, Uses, Distribution, and Maintenence of the Dataset

Yunkai Gu

November 30, 2024

This datasheet extracts questions from Gebru et al. (2021).

The dataset discussed in this datasheet and used in the study is 'Highrise Residential Fire Inspection Results' (City of Toronto 2024b), obtained from Open Data Toronto.

Relevant information could be found on the 'Fire Prevention – Inspection & Enforcement' page by City of Toronto (City of Toronto 2024a).

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 provide transparent data to present the results of fire safety inspections in all types of occupancies within Toronto conducted by the Fire Prevention Division, addressing violations of the Ontario Fire Code and other fire safety hazards within the authority of the Fire Protection and Prevention Act.
    - Its purpose is to increase public safety and awareness of fire safety, and to offer open, accessible and structured information to the public and researchers. This dataset filled the lack of resources related to detailed fire inspection outcomes and violation status of fire safety regulations for different types of properties 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 The Fire Prevention Division, and was provided access to the public by the City of Toronto's Open Data Portal.

*Code and data are available at: https://github.com/Kylie309/Toronto-Fire-Inspection.

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 creation and maintenance of the dataset are funded by the City of Toronto through its Open Data Portal.

4. *Any other comments?*

   - None.

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 in the dataset represents a fire inspection event for a residential property in Toronto and its result and outcome.
   - Each instance include details in three main aspects: property identifier (columns of 'PropertyAddress', 'ADDRESS_NUMBER', 'ADDRESS_NAME', 'PropertyType', 'propertyWard'), inspection information (columns of 'Enforcement_Proceedings', 'InspectionThread', 'INSPECTIONS_OPENDATE', 'INSPECTIONS_CLOSEDDATE'), and inspection outcomes (columns of 'VIOLATION_FIRE_CODE', 'VIOLATIONS_ITEM_NUMBER', 'VIOLATION_DESCRIPTION').

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

   - There are 96821 instances in total in the dataset. This number changes as the dataset is updated with new inspections.

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 raw dataset downloaded from the Open Data Toronto Portal is not a subset or sample from another larger data set. It contains the results of all reported fire inspections in Toronto, and the work completed in 2020 has allowed for the expansion of the portal to not only show residential high-rise buildings but inspection matters for multi-unit residential occupancy types. The types are recorded in the 'PropertyType' column.

- The cleaned dataset for this study, on the other hand, chose the cases whose inspection processes ended with all violations fixed within 2024. The columns selected to conduct study are the 'PropertyType', 'INSPECTIONS_OPENDATE', 'INSPECTIONS_CLOSEDDATE' and 'VIOLATIONS_ITEM_NUMBER' columns.

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 consist of details in three main aspects: property identifier (columns of 'PropertyAddress', 'ADDRESS_NUMBER', 'ADDRESS_NAME', 'PropertyType', 'propertyWard'), inspection information (columns of 'Enforcement_Proceedings', 'InspectionThread', 'INSPECTIONS_OPENDATE', 'INSPECTIONS_CLOSEDDATE'), and inspection outcomes (columns of 'VIOLATION_FIRE_CODE', 'VIOLATIONS_ITEM_NUMBER', 'VIOLATION_DESCRIPTION').

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

   - The target is the violation status of the fire safety checks for each residential property inspected.
   - Three columns act as the label: 'VIOLATION_FIRE_CODE' presents fire code under which violation was noted; 'VIOLATIONS_ITEM_NUMBER' presents the order number of violations by code, and 0 if no violations observed; 'VIOLATION_DESCRIPTION' presents description of fire code under which violation was noted.

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.*

   - Some individual instances do not have the information of inspection closed date due to incomplete records. Also, if there is no violation inspected for one case, 'VIOLATION_FIRE_CODE' and 'VIOLATION_DESCRIPTION' columns will include NA values.

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 defined explicitly.

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 does not include predefined splits. People downloaded the dataset could create splits by their own based on different aspects of the features recorded, such as property types or geographic locations.

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

    - The dataset may include errors arising from manual data entry, such as inconsistent formatting, especially for column 'VIOLATION_DESCRIPTION' which describes the details of fire code violated by the property inspected.

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 the portal does not include any external resources linked to it.

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 does not contain confidential information. It focuses on property-level data and excludes personal, sensitive details.

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 does not contain directly offensive, insulting or threatening information. However, inspection outcomes related to violations of fire safety checks might cause anxiety for some readers.

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 does not identify data by sub-populations. It is focused on property-level information.
    - For the property-level information, populations could be segmented by the property types. There are ten residential occupancy types included in the dataset: high-rise, low-rise, rooming houses, group homes, hotels and motels, detention centres, hospitals, nursing homes, residential cares and group homes designated a VO. The latter 4 are applicable vulnerable occupancy buildings.

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, either directly or indirectly, from the datase. The dataset does not include personal data that could be used to 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 sensitive data of this nature. It is limited to fire safety and inspection information.

16. *Any other comments?*

    - While the dataset only contains property-level information and excludes personal and sensitive data, users should interpret the findings within appropriate contextual frameworks to avoid misinterpretation.

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. It presents the results and outcomes of fire inspections in Toronto collected by Toronto Fire Services, recording the observations of fire safety checks of the buildings and violations of regulations during inspections.

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 is collected by manual human curation. It shows properties where inspections have been conducted by a TFS Fire Inspector, including properties where violations have been found which are required to be fixed for compliance at the time with the Ontario Fire Code, FPPA and Municipal Code.

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 raw dataset downloaded from the Open Data Toronto Portal is not a subset or sample from another larger data set. It contains the results of all reported fire inspections in Toronto, and the work completed in 2020 has allowed for the expansion of the portal to not only show residential high-rise buildings but inspection matters for multi-unit residential occupancy types. The types are recorded in the 'PropertyType' column.
- The cleaned dataset for this study, on the other hand, chose the cases whose inspection processes ended with all violations fixed within 2024. The columns selected to conduct study are the 'PropertyType', 'INSPECTIONS_OPENDATE', 'INSPECTIONS_CLOSEDDATE' and 'VIOLATIONS_ITEM_NUMBER' columns.

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)?*

   - Licensed fire inspectors employed by Toronto Fire Services were responsible for the data collection as part of their official duties.

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 includes the results of inspections over an ongoing timeframe, with updates reflecting new inspection activities. The earliest open date of the inspection case is 2017-01-02, and the newest inspection case has been closed at 2024-11-28.

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.*

   - Ethical review processes were not explicitly mentioned. However, data collection follows guidelines and privacy laws (e.g. City Information Management Policies and Legislation) to ensure ethical practices.

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 collected directly by fire inspectors manualy during fire safety inspections of residential buildings.

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.*

- The website did not provide explicit information about how TFS notified the residents in the properties for inspections. It could be inferred that the Fire Prevention Division would make notifications about fire inspections through official channels.

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.*

    - The dataset includes to property-level information rather than personal data. Therefore, explicit individual consent was not applicable. The inspections are conducted under relevant regulatory authority and privacy regulations.

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).*

    - Not applicable. The dataset does not involve personal data requiring consent.

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 explicit mention of a data protection impact analysis. However, the dataset excludes personal data and the inspections have been conducted under privacy laws and supervisions by official government departments, minimizing potential risks.

12. *Any other comments?*

    - None.

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.*

    - Specific preprocessing/cleaning/labeling steps for the data are not explicitly documented.

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.*

    - The truly "raw" data is not publicly available. The dataset provided on the Open Data Toronto Portal represents a cleaned table for public use.

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 the "raw" data is currently not available. For the cleaning process included in this study, the dataset downloaded from the Open Data Toronto Portal is cleaned by R, and the script is included in `scripts/03-clean_data.R`.

4. *Any other comments?*

    - None.

Uses

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

    - Currently, it is only posted to show open, accessible, transparent information of fire inspection results in Toronto by Toronto Fire Services to implement fire prevention and enforcement strategies and to increase public safety and awareness.

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.*

    - Haven't found one yet.

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

    - Other tasks the dataset could be used for includes: Analyzing trends in violations of fire safety regulations in properties across different neighborhoods and regions; Implementing targeted policy decisions in certain areas and for certain property types to improve fire safety measures; Visualizing regional trends of fire inspection results by mapping.

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 focuses on property-level data and does not include personal information. This means that its risk of violating privacy regulations and resulting in unfair treatment of individuals is minimized.
    - Careful contextual and regional analysis are recommended to avoid misinterpretation.

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

- For instance, the dataset should not be used for predicting the fire safety check outcomes for tagreted individuals only, as it contains no personal data.

6. *Any other comments?*

    - The dataset is a starting point for fire safety analysis and public awareness, but its utility can be enhanced by combining it with other datasets, such as demographic data or information related to properties, to provide a more complete understanding of Toronto's fire safety dynamics.

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 distributed to anyone since it is publicly available. Anyone could download it through the City of Toronto's Open Data Portal. The portal provides open, transparent and accessible information, aimed to increase public safety and awareness.

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 through the Open Data Toronto Portal. It can be downloaded directly from the portal in three formats: CSV, JSON and XML. The portal also provides code snippets for accessing a dataset via the API. No DOI is available for this dataset.

3. *When will the dataset be distributed?*

    - The dataset is already available and continuously updated on the Open Data Toronto Portal.

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.*

    - The data is distributed under the terms of the Open Government License - Toronto. The license writes that everyone is free to "copy, modify, publish, translate, adapt, distribute or otherwise use the Information in any medium, mode or format for any lawful purpose."
    - There's no fee associated with accessing the dataset.

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 parties have imposed IP-based or other restrictions on the dataset. It is freely available under the terms of the Open Government License - Toronto.

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 export controls or other regulatory restrictions applied to the dataset or to individual instances.

7. *Any other comments?*

   - None.

Maintenance

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

   - The creation and maintenance of the dataset are supported and funded by the City of Toronto through its Open Data Portal.

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

   - The contact information of the Fire Prevention Command offices is listed below:
   - North Command Wards: 6-York Centre, 8-Eglinton-Lawrence, 15-Don Valley West, 16 Don Valley East, 17-Don Valley North, 18-Willowdale, 5100 Yonge Street, Toronto, ON, M2N 5V7; Phone: 416-338-9150
   - East Command Wards: 14-Toronto-Danforth, 19-Beaches-East York, 20-Scarborough Southwest, 21-Scarborough Centre, 22-Scarborough-Agincourt, 23-Scarborough North, 24-Scarborough-Guildwood, 25-Scarborough-Rouge Park, 150 Borough Drive, Toronto, ON, M1P 4N7; Phone: 416-338-9250
   - South Command Wards: 9-Davenport, 10-Spadina-Fort York, 11-University-Rosedale, 12-Toronto-St.Paul's, 13-Toronto Centre, City Hall Gr Fl, W., 100 Queen St.West, Toronto, ON, M5H 2N2; Phone: 416-338-9350
   - West Command Wards: 1-Etobicoke North, 2-Etobicoke Centre, 3-Etobicoke-Lakeshore, 4-Parkdale-High Park, 5-York South-Weston, 7-Humber River-Black Creek, 399 The West Mall, Toronto, ON, M9C 2Y2; Phone: 416-338-9450

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

   - None. Any corrections or updates to the dataset are directly posted and published on the Open Data Toronto's page of the dataset.

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)?*

- Once there are new ongoing fire inspections and their results to be updated, the dataset would be changed by Toronto Fire Services. These changes will be directly communicated to the public through the Open Data Portal. Consumers can monitor the dataset page for version updates by viewing the "Data last refreshed" section on the page.

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.*

    - The dataset primarily contains property-level information rather than personal data. Any associated personal information, such as contact details from inspections, is handled with reference to Toronto's privacy and data retention policies, as outlined in their privacy guidelines.

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 typically maintained.

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.*

    - Since the dataset is complete created and contributed by Toronto Fire Services in City of Toronto, extensions and contributions cannot be directly made through the Open Data Portal. But potential questions, comments, and suggestions for improvement could be sent to the Open Data Team on Twitter, GitHub, Medium, or via e-mail opendata@toronto.ca.

8. *Any other comments?*

    - None.

References

City of Toronto. 2024a. "Fire Inspection Results." https://www.toronto.ca/city-government/accountability-operations-customer-service/access-city-information-or-records/fire-inspection-results/#listing/eyJxdWVyeVN0cmluZyI6IiIsInRvcCI6MTB9/1.

———. 2024b. "Highrise Residential Fire Inspection Results." https://open.toronto.ca/dataset/highrise-residential-fire-inspection-results/.

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.