Datasheet for 'Motor Vehicle Collisions in Toronto'*

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

Yitong Wang

December 14, 2024

The Toronto Motor Vehicle Collisions dataset provides documentation of traffic incidents resulting in fatalities or serious injuries within the city from 2006 to present. Police department collects detailed information about collision locations, environmental conditions, and involved parties through standardized reporting procedures. Each record captures essential details including geographic coordinates, time information, road conditions, and injury severity classifications, while carefully balancing public safety interests with privacy protection by excluding personal identifiers. This dataset serves multiple stakeholders including urban planners, researchers, and community organizations, supporting evidence based decision making for road safety improvements.

Extract of the questions from (Gebru et al. 2021). The Motor Vehicle Collisions involving Killed or Seriously Injured Persons dataset (City of Toronto 2024) in this discussion and study is obtained from opendatatoronto (Gelfand 2022). Some background information comes from Killed or Seriously Injured Glossary (Services 2006) and Motor Vehicle Collision Report Guide (Transportation 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 enable analysis of serious and fatal motor vehicle collisions in Toronto. The dataset fills an gap in providing transparent, detailed information about severe traffic incidents, it combines both environmental factors and driver factors to understand the pattern of injury occurrence and severity to help

*Code and data are available at: https://github.com/demainwang/Toronto_vehicle_collision_analysis.

inform infrastructure improvements, and ultimately prevent future collisions. This data collection effort aligns with Toronto's broader commitment to improving road safety and public transparency.

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 transportation services division in collaboration with Toronto police service, who are responsible for collecting and documenting collision data at the scene and through subsequent investigations.

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.*

    - No specific grantor fund this dataset, this dataset based on the record of collision cases.

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 documents a motor vehicle collision in Toronto from 2006 to 2023 where someone was killed or seriously injured. The records include environmental conditions, location details, and circumstantial factors for each incident.

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

    - The dataset includes 14,976 records, each representing an individual involved in a collision.

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 aims to contain all reported serious injury and fatal collisions that occurred within Toronto city and were reported to Toronto Police Service. It can be considered as a subset of all motor traffic collision happen in Toronto, but this dataset focus on those case involve human injury.

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 record contains event specifics (date, time, location), environmental factors (weather, road conditions), collision details (type, impact), geographic coordinates such as latitude or longitude, involved driver related information (age group, driving conditions), injury severity, road type, and traffic control presence.

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

   - Records are labeled by outcome severity (level of injury).

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.*

   - All records are not include the exact speed of traffic when incident happened. Some records have incomplete fields, particularly in older cases. Weather conditions occasionally show missing values when officers could not definitively determine conditions at the time of reporting. Personal identifying information is omitted to protect privacy, including names, exact ages. Exact incident location have been deliberately offset to the nearest road intersection node to protect the privacy.

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 dataset does not explicitly link separate incidents even if they might share common characteristics. Location based relationships may link incidents occurring at the same intersection or road segment. Temporal relationships may connect collisions occurring during similar conditions or time periods.

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 data naturally divides into temporal splits, with recommendations to separate analysis by year to account for changes in reporting practices and road safety measures. The dataset can be split geographically by district or ward for local analysis. However, no official training and testing splits exist as this represents real world incident data rather than a machine learning dataset.

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

   - Initial reporting might include estimation errors in environmental conditions or exact collision timing, the recogniztion of driver conditions are highly rely on officer's subjective thinking, might be bias.

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 connects to several external systems while maintaining its core integrity. It links to the City of Toronto's geographic information system for location validation and mapping. The dataset references standardized coding systems for collision types and injury classifications used by the Toronto Police Service.

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 contains sensitive information about traffic collisions involving deaths and serious injuries. While direct personal identifiers are removed, the combination of location, time, and incident details could potentially allow identification of individuals involved in high-profile cases. Medical information appears only in broad categories of injury severity rather than specific 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 documents serious and fatal traffic collisions, which could cause distress to readers, particularly those who have experienced similar incidents or lost loved ones in traffic collisions. The descriptions remain factual and technical, focusing on circumstances.

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.*

    - It includes basic demographic information but avoids detailed personal characteristics. The data allows analysis of collision patterns affecting different road users while maintaining individual privacy. Geographic distribution of incidents might correlate with neighborhood demographics.

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.*

    - While the dataset removes direct personal identifiers, the combination of location, date, time, and incident characteristics might enable indirect identification in some cases, particularly for notable incidents that received media coverage.

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 several categories of sensitive information. Location data shows where serious incidents occur, which could affect neighborhood perceptions. The timing and circumstances of fatal collisions represent sensitive details for affected families and communities. The data includes broad age groups of those involved, though it excludes other demographic details to protect privacy.

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.*

    - Toronto Police Service officers collect data directly at collision scenes through standardized reporting procedures. Officers observe the scene, take measurements, and document environmental conditions.

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?*

    - Police officers use electronic reporting systems installed in their vehicles to document collision details on-site. They follow standardized protocols established by the Toronto Police Service for collision investigation and reporting.

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

    - This dataset represents the complete record of collisions involving fatalities or serious injuries in Toronto. Every qualifying incident undergoes documentation and inclusion in the dataset.

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)?*

    - Toronto Police Service officers serve as primary data collectors as part of their regular duties. All personnel involved receive compensation through their standard government employment arrangements.

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 covers collisions from 2006 to present, with continuous updates as new incidents occur. The collection timeframe matches actual collision occurrences.

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 data collection operates under the Toronto Police Service's established protocols for incident reporting, The public release process includes privacy impact assessments to protect individual rights, including hide exact location, names and age.

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 Toronto Police Service obtains information directly through officer investigations at collision scenes. In cases of serious injuries, medical authorities provide injury classification updates while maintaining patient privacy. The dataset represents official police records rather than third party collected information.

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.*

    - When police officers investigate collisions, they inform involved parties about official reporting requirements and data collection. The public nature of serious collision investigation means participants understand that basic incident information becomes part of public safety records, though personal details remain protected.

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 data collection occurs under the legal authority of the Toronto Police Service to investigate and document serious collisions. The public release version removes personal identifiers to protect privacy while maintaining the public safety benefits of collision data analysis.

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).*

- Given the public safety mandate for collision reporting, the dataset does not revocate related information.

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 was not official analysis of potential impact.

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.*

    - Not specific mention of details of data processing, but it made change for location data, which deliberately offset to the nearest road intersection node to protect the privacy.

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 original raw data not published due to the sensitive information about people's privacy.

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

    - Not specific mention about the software, but generally it may use the police internal system to handle the data.

Uses

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

    - Not specific mention of the previous using of data, but transportation services may uses this dataset to identify high risk locations requiring safety improvements. City planners incorporate the data into Vision Zero safety initiatives.

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.*

    - Have not found it.

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

- The data could support urban planning decisions, emergency response resource allocation, and public education about road safety.

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 collection of this dataset mostly based on fact, but as mentioned some recording like driver conditions can have officer's subjective bias.

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

    - The dataset should not serve for individual incident liability determinations or personal identification purposes.

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 City of Toronto makes the dataset publicly available through its open Data portal, enabling access by researchers, planners, and community organizations.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

    - The Open Data Portal provides CSV and GeoJSON format downloads. Users can access the data through web interfaces and API connections.

3. *When will the dataset be distributed?*

    - The dataset updates as new verified records become available.

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 dataset is permitting for broad application for public benefit.

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.*

- This government dataset is public and not restriction by other parties.

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.*

    - Not such restrictions.

Maintenance

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

    - No specific mention, but The police and transportation Services division may maintain the dataset with technical support from the city's information technology department.

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

    - Users can contact the open Data team through the portal's feedback system.

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

    - The open Data Portal maintains a change log documenting significant corrections and updates.

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)?*

    - Updates add when new incidents are avaible and revise existing records based on finalized investigations.

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 does not involve the exact personal information.

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.*

    - Nope, the new data will be added to the combinations of old data in one file.

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.*

- Nope, this is the government dataset based on the collision information provided by police.

References

City of Toronto. 2024. "Motor Vehicle Collisions Involving Killed or Seriously Injured Persons." https://open.toronto.ca/dataset/motor-vehicle-collisions-involving-killed-or-seriously-injured-persons/.

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.

Gelfand, Sharla. 2022. "Opendatatoronto: Access the City of Toronto Open Data Portal." https://CRAN.R-project.org/package=opendatatoronto.

Services, Toronto Police. 2006. "Killed or Seriously Injured (KSI) Glossary." https://www.tps.ca/files/download/1581967027/19551/.

Transportation, Ontario Ministry of. 2023. "Motor Vehicle Collision Report (MVCR) Guide." https://intra.stage.ecollision.mto.gov.on.ca/eCollision/pdf/MVCRGuide.pdf.