Datasheet for 'Toronto Motorbike Collision and Theft Risk'*

Supplementary document to "Motor Vehicle Risks in Toronto: Uncovering Collision and Theft Patterns"

Yingke He

2024-12-19

This supplementary datasheet accompanies the motorbike risk analysis paper. Adopting the standardized format proposed by Gebru et al. (2021), it documents the motivation, composition, collection process, recommended applications, maintenance, and distribution of the cleaned dataset, available in the analysis data folder of the GitHub Repository. The goal is to enhance communication between dataset creators and users while promoting transparency and accountability within the statistical and risk analysis communities.

Table of contents

*Code and data are available at: https://github.com/ohyykk/Toronto_Motor_Viehicle/tree/main.

1 Acknowledgement

All the following questions are extracted from Gebru et al. (2021).

2 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 document and analyze incidents of motor vehicle collisions (KSI dataset) and thefts from motor vehicles within Toronto. Its primary purpose is to assist law enforcement, policymakers, and researchers in identifying trends, risk factors, and areas of concern. Specifically, it addresses gaps in understanding the temporal, spatial, and contextual factors contributing to serious road safety incidents and vehicle thefts. This data also provides a foundation for neighborhood-level analysis and supports the development of targeted interventions for public safety.

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

    - The datasets were created and published by the Toronto Police Service as part of their commitment to transparency and public accountability. The data is made available through Open Data Toronto, an initiative aimed at providing open access to public datasets for community use and research.

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 funding for the dataset's creation likely stems from the operating budget of the Toronto Police Service, supported by the City of Toronto. There is no specific mention of external grants or funding in the documentation, suggesting that it is a part of their routine operational and public outreach activities.

4. *Any other comments?*

    - The datasets serve as important tools for urban planning, public safety strategies, and academic research. However, users should note the limitations of the dataset, including potential discrepancies due to privacy adjustments and the lack of guarantees for completeness or timeliness. The initiative reflects broader efforts to utilize open data to enhance community outcomes and promote collaborative problem-solving.

3 Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

    - The dataset includes two types of instances: motor vehicle collisions and thefts from motor vehicles. Instances in the Motor Vehicle Collisions (KSI dataset) represent individual events where a person was killed or seriously injured, recorded along with details such as severity, type of collision, and location. Instances in the Theft from Motor Vehicles dataset represent theft events categorized by the value of stolen items and recorded with attributes like the date, time, and geographic location. These instances collectively enable an analysis of both theft and collision risks.

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

    - There are in total 32425 incidents for the theft data, and 18957 incidents for the collision data

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 reported incidents and does not include all possible instances. It represents events documented by the Toronto Police Service, meaning unreported or undocumented collisions and thefts are excluded. While it may not capture every incident, the dataset broadly reflects patterns across Toronto and includes diverse geographic coverage.

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 structured features rather than raw data. For collisions, features include variables such as collision type, severity, and location. For thefts, features include attributes like the value category of stolen items, date, time, and location. The data has undergone cleaning and formatting to ensure usability and consistency.

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

    - The raw dataset does not include explicit labels or targets. However, in this study, the composite risk score is used as an outcome variable, calculated by integrating probabilities of theft and collisions derived from the dataset's predictor variables, such as time of day, road conditions, and neighborhood characteristics.

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 instances have missing or incomplete information, such as geographic coordinates recorded as "0, 0" or locations labeled as "Not Specified Area (NSA)." These gaps occur due to unvalidated data or unavailable details at the time of reporting, and while they do not render the dataset unusable, they may affect the precision of certain analyses.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

   - Relationships between instances are not explicitly defined in the dataset. However, implicit relationships can be inferred, such as temporal or spatial clustering of thefts and collisions within specific neighborhoods or during certain time periods, which offer opportunities for broader contextual analysis.

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 provide predefined splits for training, validation, or testing. Any splits for analysis would need to be created by the user, potentially based on temporal factors (e.g., separating data by year) or geographic boundaries to maintain logical consistency and ensure representativeness.

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

   - The dataset contains potential sources of noise and errors, such as location offsets introduced for privacy and placeholder values like "None" or "NSA." Additionally, duplicates and inconsistencies in field entries were addressed during the cleaning process, but some minor errors may persist due to the limitations of the original data collection.

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 largely self-contained but relies on external frameworks such as Toronto neighborhood structures and definitions of theft categories. While these external resources are publicly available, their long-term stability is not guaranteed. The dataset itself is accessible through Open Data Toronto, ensuring its immediate usability without additional fees or licensing restrictions.

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 explicitly confidential information, as it is publicly available through Open Data Toronto. However, some details, such as locations of incidents, have been adjusted for privacy (e.g., offset to the nearest intersection) to prevent the identification of exact addresses or individuals. While this reduces the risk of revealing sensitive data, users must still handle the dataset responsibly to avoid inadvertent breaches of privacy.

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 might cause anxiety or discomfort for some individuals, as it involves serious incidents like fatalities, injuries, and thefts. While the data is presented in an aggregated, structured format, its subject matter could evoke emotional responses, particularly for those affected by similar events.

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 includes information that can indirectly identify sub-populations, such as geographic regions (e.g., neighborhoods) and temporal patterns (e.g., time of day, day of the week). However, it does not explicitly identify sub-populations based on demographic factors such as age or gender, as these variables are either not included or anonymized.

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.*

    - It is not possible to directly identify individuals from the dataset, as personal identifiers are excluded. However, indirect identification might be possible in rare cases by combining detailed temporal, geographic, or contextual information with external data sources. The privacy adjustments (e.g., location offsets) mitigate this risk significantly.

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 includes potentially sensitive information, such as geographic locations of incidents, which could reveal patterns related to crime or traffic safety in specific areas. While no direct identifiers are present, the nature of the data—especially involving fatalities or serious injuries—could be considered sensitive due to its implications for public safety and community well-being.

16. *Any other comments?*

    - The dataset serves as an important resource for public safety analysis and urban planning, but users must approach it with care, especially regarding privacy and ethical considerations. While efforts have been made to anonymize and adjust sensitive information, the potential for misuse underscores the importance of responsible data handling and interpretation.

4 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 collected through reports made to the Toronto Police Service. For motor vehicle collisions, the data reflects incidents documented by police officers based on on-site investigations or reports submitted by the public. For thefts from vehicles, the data was gathered from incident reports submitted by vehicle owners or other witnesses. The data is considered directly observable in terms of the reported events but may include some subjectivity or errors in reporting. Validation processes are not explicitly described, but police documentation likely involved standard procedures to ensure accuracy.

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 collection relied on manual reporting by law enforcement officers, public submissions (e.g., online forms or emergency calls), and administrative data entry systems used by the Toronto Police Service. These mechanisms are part of standard law enforcement operations, which are typically validated through internal quality assurance processes, though specific details about validation are not provided.

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 represents an extensive collection of reported incidents rather than a probabilistic or random sample. It includes all documented collisions and thefts meeting specific criteria (e.g., collisions resulting in fatalities or serious injuries). However, unreported incidents are excluded, meaning the dataset represents only the subset of events known to law enforcement.

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)?*

   - Data collection was conducted by Toronto Police Service personnel, including officers and administrative staff. Their compensation would be part of their standard salaries as law enforcement and administrative employees. No external crowdworkers or contractors were involved in the collection process.

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 for motor vehicle collisions spans from 2006 to the present, while the theft data covers a defined reporting period, typically several years. The creation timeframe aligns with the reporting and documentation of these events, as the data was recorded contemporaneously with the incidents.

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.*

   - There is no mention of specific ethical review processes for the dataset, as it is operational data collected as part of law enforcement duties. Ethical considerations likely adhere to standard privacy and data protection protocols, such as anonymizing sensitive information and offsetting geographic locations.

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 the Toronto Police Service from individuals reporting incidents, either in person, through emergency calls, or via online submissions. It was subsequently aggregated into structured datasets for public use through Open Data Toronto.

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.*

    - Individuals reporting incidents were likely aware of the data collection as part of the reporting process. While explicit notification language is not described, it is standard for law enforcement to inform individuals that their reports may be used for documentation and analysis.

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 for data collection is implicit in the act of reporting an incident to law enforcement. Explicit consent for public release of anonymized data was likely not sought, but the dataset complies with privacy regulations, including removing or offsetting personally identifiable information.

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).*

    - There is no mechanism for individuals to revoke consent for the use of anonymized, aggregated data in this context. Since the data is de-identified and part of public records, it is not subject to withdrawal by individuals under typical privacy frameworks.

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 formal data protection impact analysis (DPIA) or similar review for this dataset. However, the Toronto Police Service has implemented privacy measures, such as anonymizing data and offsetting locations, to minimize risks to individuals. The dataset's availability through Open Data Toronto implies adherence to public transparency and data protection policies, though the lack of a detailed impact assessment leaves room for further ethical analysis.

12. *Any other comments?*

    - While the dataset provides important ideas for public safety and urban planning, a more detailed impact analysis could address potential unintended consequences, such as misrepresentation of neighborhoods or misuse of sensitive information. Documentation of such an analysis would enhance confidence in the dataset's ethical use.

5 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, extensive preprocessing was performed to clean and standardize the data. This included renaming columns for consistency, removing duplicates, validating geographic coordinates, handling missing values, and converting data types (e.g., numeric or categorical). Specific examples include replacing "None" with NA, validating latitude and longitude ranges, and standardizing textual fields like division codes.

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 raw data is not explicitly described as being saved alongside the cleaned version. However, given the dataset's origins with Open Data Toronto, the raw data may still be accessible through their repository, albeit in a less processed form.

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

   - Yes, the cleaning process relied on open-source software, particularly R (R Core Team 2023) and associated libraries, such as tidyverse (Wickham, Averick, et al. 2023), Dplyr (Wickham, François, et al. 2023), and here (Müller 2022). These tools are freely available, and detailed preprocessing scripts can likely be reproduced or adapted using their documentation.

4. *Any other comments?*

   - The structured cleaning ensures that the dataset is ready for analysis, but consumers should verify cleaning steps for their specific use cases. Maintaining access to the raw data would also support greater flexibility and reproducibility in future research.

6 Uses

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

   - Yes, the dataset has been used to develop a composite risk score model in this study, integrating probabilities of theft and collisions to assess motorbike-related risks in Toronto. It has also likely been used for public safety analyses, urban planning, and policy development.

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.*

    - There is no specific repository linking papers or systems that use this dataset. However, Open Data Toronto serves as the primary source for accessing the dataset, and related studies may cite it in their publications.

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

    - The dataset could be applied to tasks such as traffic safety evaluations, urban mobility studies, insurance risk assessments, crime prevention strategies, and machine learning applications for predictive modeling of safety incidents.

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's reliance on reported incidents means it may underrepresent unreported events, particularly in neighborhoods with lower reporting rates. Privacy adjustments, such as location offsets, could affect the precision of geographic analyses. Users should carefully interpret results to avoid bias or misrepresentation of communities and ensure ethical use.

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

    - The dataset should not be used for tasks that could harm individuals or communities, such as creating unfair stereotypes about specific neighborhoods or using the data for discriminatory purposes. It is also unsuitable for highly granular analyses requiring exact locations due to privacy adjustments.

6. *Any other comments?*

    - This dataset is a important resource for research and policy but requires careful handling to respect privacy and avoid misuse. Clear documentation of limitations and preprocessing steps would further support ethical and effective use.

7 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 distributed to the public through Open Data Toronto, a platform designed to share datasets collected by the City of Toronto and its agencies. It is accessible to third parties, including researchers, policymakers, and the general public.

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 Open Data Toronto website, where it can be downloaded in formats such as CSV or accessed through APIs for integration into applications. It does not appear to have a specific Digital Object Identifier (DOI), but a permanent link is provided for access.

3. *When will the dataset be distributed?*

   - The dataset is already available for public access on the Open Data Toronto platform. Updates to the dataset are likely released periodically, depending on the reporting frequency and data processing timelines of the Toronto Police Service.

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.*

   - Yes, the dataset is distributed under the Open Government License - Toronto, which allows users to copy, modify, publish, and distribute the data, provided proper attribution is given. There are no fees associated with accessing or using the dataset under this license.

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-party IP restrictions are imposed on the dataset. However, the terms of the Open Government License require users to acknowledge the source of the data when using it in derivative works or publications.

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 export controls or regulatory restrictions apply to the dataset. It is openly available for access and use by anyone, provided they adhere to the terms of the Open Government License.

7. *Any other comments?*

- The dataset's open distribution model ensures broad accessibility, promoting transparency and enabling its use in research, policy-making, and public initiatives. However, users should remain mindful of privacy and ethical considerations when analyzing and sharing ideas derived from the dataset. Providing regular updates and improving metadata (e.g., DOI assignment) could enhance its usability and traceability.

#Maintenance

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

    - The dataset is hosted, supported, and maintained by Open Data Toronto, a platform managed by the City of Toronto. The data originates from the Toronto Police Service, which is responsible for collecting and curating the raw data.

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

    - Open Data Toronto provides contact information on its website for inquiries. General questions can often be directed to the City of Toronto's Open Data team via email at opendata@toronto.ca.

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

    - There is no specific erratum provided for this dataset. However, any issues or corrections might be addressed in updates to the dataset, and users can report discrepancies through Open Data Toronto's feedback mechanisms.

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 to include new data, correct errors, or modify existing entries. Updates are managed by the Toronto Police Service and communicated through the Open Data Toronto platform, typically indicated in the dataset's metadata or version history.

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.*

    - There are no explicit retention limits mentioned. However, since the dataset is anonymized and aggregated, it is unlikely to be subject to strict retention rules. Individual privacy is protected through anonymization and privacy offsets, making long-term retention less of a concern.

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.*

    - Open Data Toronto does not explicitly provide access to historical versions of the dataset. Updates replace older versions, and users are encouraged to download and archive versions if historical consistency is required. Obsolescence is not formally communicated beyond the release of updated data.

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.*

    - There is no formal mechanism for third-party contributions to the dataset itself. Users can create derivative works or analyses based on the data, but these are independent of the official dataset. Open Data Toronto does not currently validate or integrate external contributions.

8. *Any other comments?*

    - The dataset is maintained to ensure relevance and utility, but users must account for updates and potential changes in structure or format over time. Establishing clearer update policies, version control, and a mechanism for community contributions could enhance its usability and foster collaboration.

References

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.

Müller, Kirill. 2022. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

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

Wickham, Hadley, Mara Averick, Jennifer Bryan, et al. 2023. *Tidyverse: Easily Install and Load the Tidyverse*. https://CRAN.R-project.org/package=tidyverse.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2023. *dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.