Datasheet for 'Understanding Bicycle Theft in High-Risk Neighborhoods: Key Predictors for Targeted Risk Assessment'*

Understanding Theft Patterns and Risks

Tommy Fu

December 3, 2024

This datasheet documents the 'Bike Thefts Data,' a dataset aimed at understanding patterns and risks of bike thefts in urban areas. It covers the motivation, composition, collection, preprocessing, uses, and limitations of the dataset, providing critical insights for researchers, policymakers, and stakeholders interested in bike security.

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 explore bike theft patterns, identify high-risk areas, and understand the impact of various factors such as premises type, time of day, and bike value. It aims to inform theft prevention strategies and urban planning. Futher, the Toronto Shapefile dataset was created by the University of Toronto in the Map and GIS Library (University of Toronto n.d.).
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 compiled by Toronto Police Open Data (Toronto Police Service 2024) focused on urban transportation safety and bike security.
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 direct funded creation can be found.

[*] Code and data are available at: https://github.com/YichengFu/bike_thefts_analysis.git

4. *Any other comments?*

   - This informative dataset enables Bayesian regression analysis and provides insights for Theft Patterns and Risk Factors in Toronto's High-Theft Areas

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 represents reported incidents of bike thefts in Toronto, including information about the location (e.g., neighborhoods, premises types), time (e.g., time of day, date), bike cost, and other contextual factors. Each instance corresponds to an individual theft report.

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

   - The dataset contains approximately 36,125 unique instances of reported bike thefts.

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 bike thefts from Toronto Police records. While it captures a broad range of thefts across neighborhoods, it may not represent unreported incidents. The geographic and temporal coverage ensures relevance for urban theft analysis.

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 data about a reported bike theft. This includes:

- Temporal data: Occurrence date, time, and report date.
- Geographic data: Latitude, longitude, and neighborhood identifiers.
- Incident specifics: Type of offense, location type, premises type.
- Bike details: Make, model, type, speed, color, and cost. Outcome: Theft status (e.g., stolen, recovered).

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

   - Yes, the STATUS column acts as a label, indicating the outcome of the theft (e.g., "STOLEN," "RECOVERED").

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.*

    - Yes, some columns have missing data:

- BIKE_MODEL (many missing values, possibly because this detail wasn't reported).
- BIKE_COST and BIKE_COLOUR (some missing 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.*

    - No explicit relationships between instances exist. Each row represents an independent theft report.

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.*

    - There are no predefined splits in the dataset. Users may define splits based on temporal or geographic data for training/testing in predictive models.

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

    - Possible noise includes missing or incomplete information. Some entries might have errors in reporting (e.g., incorrect dates or geographic coordinates).

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. It does not rely on external links or 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 personal or confidential information about individuals appears in the dataset. It contains anonymized, aggregate-level data.

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 itself is not offensive. However, data related to theft and crime might be sensitive for certain users.

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 identifies incidents by neighborhoods but does not classify by subpopulations like age or gender.

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, the data is anonymized and does not contain personally identifiable information (PII).

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 sensitive personal data is included. It focuses on theft incidents and their characteristics.

16. *Any other comments?*

    - The dataset is well-structured and suitable for geographic and temporal analysis of bike thefts.

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.*

    - Data was reported by individuals to the Toronto Police and compiled into the police records. It is directly observable and validated through police procedures.

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 police reports, likely supplemented by digital systems for data entry.

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 all reported bike thefts but excludes unreported incidents, limiting it to recorded crimes.

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 Toronto Police Service staff.

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 spans multiple years. The OCC_YEAR and REPORT_YEAR columns provide temporal coverage.

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.*

    - Likely not applicable, as this is an operational dataset derived from police records.

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 is derived directly from Toronto Police Open Data records.

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.*

    - As this dataset is based on police reports, individuals were inherently aware of the data collection during the reporting process.

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.*

    - By reporting thefts, individuals implicitly consented to the collection and recording of this data.

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, as the dataset anonymizes the information and focuses on incidents rather than individuals.

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 indication of a formal impact analysis specific to the dataset's release.

12. *Any other comments?*

    - The dataset appears to align with standard practices for public release of anonymized crime data.

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.*

    - No explicit preprocessing or cleaning appears to have been performed on this dataset before release. Columns like BIKE_MODEL, BIKE_COST, and BIKE_COLOUR contain missing values. Data such as dates and times appear in a structured format but may require transformation for specific analytical purposes.

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 dataset provided seems to represent the raw or minimally processed data as collected by the Toronto Police Service. No additional raw version is referenced.

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

    - No information about specific preprocessing or software tools is included. Users may need to conduct their own cleaning and transformation.

4. *Any other comments?*

    - The dataset is a suitable starting point for analysis but requires user-driven preprocessing to handle missing values and potentially derive new features for advanced tasks.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - No explicit mention of prior usage is included. The dataset is likely used for urban planning, crime prevention, and academic research
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 reference to a repository of publications or systems using this dataset.
3. *What (other) tasks could the dataset be used for?*
    - Potential uses:

- Predictive modeling: Identifying high-risk areas for bike theft.
- Spatial analysis: Correlation of thefts with neighborhood characteristics.
- Temporal analysis: Seasonal or hourly trends in bike theft.
- Policy evaluation: Effectiveness of theft-prevention measures.

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?*
    - Missing data, especially for critical fields like BIKE_COST, may affect the quality of certain analyses. Users should address these gaps through imputation or exclusion strategies.
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 draw conclusions about unreported thefts or directly infer causation due to the absence of comprehensive contextual data.
6. *Any other comments?*
    - The dataset provides a good foundation for analyzing reported bike theft trends but requires careful consideration of its limitations.

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 dataset is publicly available through the Toronto Police Open Data portal.

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

    - Distribution occurs through an online portal, potentially as CSV files. No DOI or permanent identifier is referenced.

3. *When will the dataset be distributed?*

    - The dataset is already available to the public.

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 likely subject to an open data license provided by the Toronto Police Service, which permits public use.

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.*

    - There are no indications of external restrictions on its use.

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, the dataset is free from such restrictions.

7. *Any other comments?*

    - The dataset's accessibility supports transparency and public engagement with municipal crime data.

Maintenance

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

    - The Toronto Police Service is responsible for maintaining the dataset.

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

    - Contact details for the Toronto Police Service's Open Data available on their official website.

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

- No errata have been referenced or linked.

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 are based on periodic data releases by the Toronto Police Service. The frequency and notification mechanism are unclear.

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 explicit retention policy is indicated, but anonymization ensures compliance with privacy standards.

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.*

    - No information is provided, though users are encouraged to access the latest version from the Toronto Police 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.*

    - The dataset does not support direct contributions but allows users to conduct independent analyses.

8. *Any other comments?*

    - Clear documentation and periodic updates would enhance the dataset's usability.

References

Toronto Police Service. 2024. "Bicycle Thefts Open Data." Toronto, Canada. https://data.torontopolice.on.ca/datasets/TorontoPS::bicycle-thefts-open-data/about.

University of Toronto. n.d. "Introduction to GIS Using r." https://mdl.library.utoronto.ca/technology/tutorials/introduction-gis-using-r.