Datasheet for 'Bicycle Theft'*

Zitong Guo

Invalid Date

This datasheet provides an overview of the Bicycle Theft dataset, detailing its contents, usage considerations, and limitations. It is designed to inform researchers and practitioners working on crime data and prevention strategies. The dataset encompasses data from multiple years and highlights patterns related to temporal and spatial factors influencing bicycle theft incidents.

Extract of the questions from Gebru et al. (2021).

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 a comprehensive analysis of bicycle theft incidents in urban areas. The primary aim is to understand temporal and spatial patterns of thefts, recovery rates, and the types of bicycles most commonly stolen. There was a gap in publicly available structured data that contained detailed information on bicycle characteristics, theft locations, and outcomes, which this dataset addresses.

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 Toronto Police Service, who compiled the crime reports related to bicycle theft and released the data through Open Data Toronto, a public open data initiative run by the City of Toronto.

3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*

- The dataset was funded by the City of Toronto as part of their ongoing public safety and transparency efforts through the Open Data Toronto initiative. No specific research grant is associated with the dataset.

*Code and data are available at: https://github.com/jennygzt/Bicycle-Theft-.git.

4. *Any other comments?* -The dataset is an essential resource for understanding urban crime trends and for planning crime prevention strategies. It also supports academic research and policy-making by providing insights into theft patterns, recovery rates, and areas most affected by bicycle theft.

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 instances in the dataset represent individual reports of bicycle theft incidents. Each instance contains information related to the date, time, and location of the theft, the characteristics of the stolen or recovered bicycle (such as make, model, type, color, and cost), and the status of the theft (whether the bicycle was recovered or not). There is a single type of instance: theft report data.

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

- The total number of instances corresponds to the number of theft incidents recorded in the dataset. You can determine this by checking the dataset (e.g., nrow(data) in R). Each instance represents a unique theft report.

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 includes all reported bicycle theft incidents recorded by the Toronto Police Service over a given time period. It is not a sample but represents the full dataset of recorded bicycle thefts within the specified time range.

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.* -Certain fields, such as BIKE_COST, OCC_HOUR, STATUS, or BIKE_MODEL, may contain missing values due to incomplete or inaccurate reporting. For example, the BIKE_COST may not be provided if the owner didn't know the exact value of the bicycle.

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

    - The STATUS field can be considered a label, indicating whether the bicycle was "STOLEN" or "RECOVERED". This label is useful for analysis of recovery rates and the success of law enforcement efforts in retrieving stolen bicycles.

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.*

    - Certain fields, such as BIKE_COST, OCC_HOUR, STATUS, or BIKE_MODEL, may contain missing values due to incomplete or inaccurate reporting. For example, the BIKE_COST may not be provided if the owner didn't know the exact value of the bicycle.

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 incidents are not explicitly stated. Each instance is independent, but potential relationships could exist, such as multiple thefts in the same area or the same bicycle being stolen multiple times, which would be inferred based on location or bike details.

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.*

    - If the dataset is used for machine learning tasks, a natural split could be by time (e.g., year or month) to observe trends over time. Earlier years could be used for training models, while more recent data could be used for validation and testing.

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

    - Possible errors include manual data entry mistakes or inconsistent geolocation data. There might also be redundancies, such as duplicated entries or variations in how bike models or types are reported.

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. However, it can be enriched by combining it with external datasets such as geospatial data or crime data for broader analysis. The dataset does not rely on any external resources for basic use, and there are no licensing restrictions on its usage as it is publicly available.

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. Personal data, such as names and exact addresses, has been anonymized or removed to ensure privacy. Only general location data (e.g., coordinates or neighborhood-level information) is provided.

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 data itself is not offensive, but it could raise concerns for some individuals, particularly if theft rates are high in certain neighborhoods. The presence of theft data might cause anxiety for those living in areas where the incidents are concentrated.

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 include demographic information about victims or subpopulations. It focuses on the thefts and the bicycles involved rather than the individuals reporting the incidents.

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.*

    - Since the dataset has been anonymized, individuals cannot be directly or indirectly identified. No personally identifiable information (PII) is included, and the dataset only provides general details about the incidents and locations.

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 does not include sensitive information like race, religion, or financial data. However, location data might be considered sensitive if used to analyze crime hotspots. This is mitigated by the anonymization of the data.

16. *Any other comments?*

- The dataset serves as a critical resource for policymakers, researchers, and urban planners to better understand crime trends and create strategies to reduce bicycle thefts. It is a valuable tool for data-driven analysis of urban safety.

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 reported directly by individuals to the Toronto Police Service through incident reports. These reports were compiled and structured for public release on Open Data Toronto. Since the data is reported to law enforcement, it is considered validated and verified through their internal processes before being published.

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 was collected via police reports submitted by citizens or officers at the scene of bicycle thefts. The reports were manually curated, processed, and then structured into a dataset by the Toronto Police Service before being uploaded to Open Data Toronto. Police departments follow standard procedures to validate crime reports before entering them into the system.

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 includes all reported bicycle theft incidents in Toronto and is not a sample. It represents the complete set of reported bicycle thefts made to the Toronto Police Service for a given time period.

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 was collected by the Toronto Police Service. Officers, investigators, and data management personnel were involved in gathering and processing the data as part of their official duties. No external contractors or crowdworkers were involved in the data 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 dataset covers bicycle theft incidents over a span of multiple years (typically from 2014 to the present). The collection timeframe corresponds directly to the incident dates, meaning that the data is contemporaneously collected as thefts occur and are reported.

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.*

    - Since this dataset is anonymized and published for public use by Open Data Toronto, an institutional review board (IRB) may not have been required. However, the Toronto Police Service ensures that any sensitive personal information is removed before the data is made publicly available.

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 indirectly from the Toronto Police Service and made available through Open Data Toronto. The data is not collected directly from individuals for this analysis; it is a secondary dataset made available to the public.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

    - Individuals who report bicycle thefts to the Toronto Police Service are typically aware that their reports will be recorded and processed. However, specific notifications regarding the inclusion of these reports in the Open Data Toronto portal may not be explicitly provided. The data is anonymized before publication, ensuring no personal identifying information is shared.

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.*

    - When individuals report a theft to the police, they implicitly consent to the collection and use of their data for investigative and reporting purposes. However, since the dataset is anonymized, no explicit additional consent is likely required for making the anonymized data publicly available through Open Data Toronto.

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).*

- As this dataset is anonymized and contains no personally identifiable information, individuals would not typically need to revoke consent. The Toronto Police Service anonymizes the data before publishing it, ensuring that individuals cannot be directly identified.

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.*

    - The data is anonymized, and any sensitive information is removed before public release. While a formal data protection impact analysis (DPIA) is likely not required for this type of dataset, the Toronto Police Service follows best practices in removing any personal identifiers before the data is made public.

12. *Any other comments?*

    - The Open Data Toronto bicycle theft dataset is a valuable public resource for understanding crime patterns in the city. It helps inform both policy decisions and public awareness about crime prevention and law enforcement efforts related to bicycle thefts.

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.*

    - The dataset likely underwent basic preprocessing before being made publicly available, including the anonymization of personal information (such as removing names or specific addresses). Additionally, missing values for certain fields (e.g., BIKE_COST or BIKE_MODEL) may not have been imputed, and the data could have been cleaned to standardize the format of certain fields like dates and bike descriptions.

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, unprocessed data is not publicly available due to privacy concerns, as it may contain personally identifiable information (PII). Only the anonymized, cleaned version is available through Open Data Toronto.

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 preprocessing was likely done by the Toronto Police Service or the City of Toronto before publication. There is no public access to the specific software or scripts used for this task. The processed data is available through Open Data Toronto at Open Data Toronto Bicycle Theft Dataset.

4. *Any other comments?*

   - The preprocessing steps taken likely focused on ensuring data consistency (such as standardizing date formats) and removing sensitive information to protect the privacy of individuals. These steps were essential for creating a publicly usable dataset without compromising personal privacy

Uses

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

   - TBD: The dataset is publicly available via Open Data Toronto and may have been used by researchers, policymakers, and law enforcement for urban planning, crime analysis, and public safety studies. Specific papers or projects may not be directly cited, but the data likely contributes to various analyses in crime prevention and urban security.

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.*

   - TBD: At this time, there may not be a centralized repository linking all papers or projects that use the dataset. However, you can check research papers or reports that cite Open Data Toronto as a source. The dataset itself is accessible via Open Data Toronto.

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

   - The dataset could be used for a variety of tasks, including: Crime trend analysis: Identifying hotspots and periods of higher bike theft activity. Predictive modeling: Building models to predict future thefts based on historical patterns. Urban safety planning: Assisting policymakers in designing interventions (e.g., bike locks, surveillance) based on theft locations and patterns. Bicycle theft prevention studies: Investigating factors that increase theft risk, such as bike type, location, or time of day. Correlation with other crime datasets: Exploring relationships between bicycle thefts and other types of crimes or socioeconomic factors.

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?*

- TBD: One consideration is that the dataset contains location information, which could lead to unfair treatment of certain neighborhoods if used for predictive policing. Users should be cautious of reinforcing biases or stigmatizing areas based on theft rates alone. Mitigating these risks could involve contextualizing the data with broader socioeconomic and environmental factors rather than treating crime rates as isolated indicators.

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 unfairly target specific neighborhoods or demographic groups based on bike theft incidents alone. It should also not be used for individual profiling, as it has been anonymized and is designed for aggregate analysis. Additionally, users should avoid using the dataset for commercial purposes that could harm individuals or communities (e.g., denial of services based on crime data).

6. *Any other comments?*

   - The dataset provides valuable insights for urban safety and crime prevention but should be used responsibly and in conjunction with other datasets (such as environmental, socioeconomic, or transportation data) to ensure a balanced understanding of crime trends.

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 Open Data Toronto portal, which distributes it to any third party, including researchers, policymakers, and the general public. It is intended for broad access and use.

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 users can download the data in various formats, such as CSV, directly from the portal. It does not have a DOI but is accessible through the Open Data Toronto Bicycle Theft Dataset website.

3. *When will the dataset be distributed?*

   - The dataset is already available and updated periodically by the City of Toronto through the Open Data Toronto portal. It has been available for several years and is refreshed with new data as incidents are reported.

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 distributed under the Open Government License – Toronto, which allows for free use, modification, and distribution of the data, provided that appropriate credit is given. More information can be found in the Open Government 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.*

    - There are no third-party IP-based restrictions on the dataset. The data is provided by the City of Toronto and is subject to the Open Government License.

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 freely available to users globally, provided they comply with the licensing terms specified by the City of Toronto.

7. *Any other comments?*

    - The Open Data Toronto portal serves as a valuable resource for public access to various datasets, including bicycle theft data, with no associated fees or complex licensing restrictions. The data is intended for research, policy-making, and public awareness purposes.

Maintenance

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

    - The dataset is maintained and hosted by the City of Toronto through the Open Data Toronto initiative. The Toronto Police Service contributes updates to the dataset as new data becomes available.

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

    - The Open Data Toronto team can be contacted through the Open Data Toronto contact page or via email at open@toronto.ca for questions or issues related to the dataset.

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

- There is no formal erratum, but any updates or corrections to the dataset are likely handled through periodic updates by the City of Toronto. Issues or discrepancies can be reported through the Open Data Toronto contact page.

4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

   - Yes, the dataset is periodically updated by the City of Toronto, reflecting newly reported incidents of bicycle theft. Updates are made available through the Open Data Toronto portal, and users can download the latest version directly from the site. No specific update notifications (such as mailing lists) are currently provided.

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 is anonymized, and no personally identifiable information (PII) is included. Therefore, there are no retention limits tied to individuals in the dataset.

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 may not be maintained, as Open Data Toronto generally provides the most up-to-date dataset. If necessary, users can retain copies of older versions for their own use, but there is no guarantee that previous versions will remain hosted or available on the platform.

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 is open for use and modification under the Open Government License – Toronto. However, there is no formal mechanism for third parties to contribute directly to the dataset itself. Any extensions or augmentations by third parties would need to be validated independently and are not directly supported by the City of Toronto.

8. *Any other comments?*

   - The dataset provides an important resource for studying crime patterns in Toronto, but it is essential to use it responsibly and ensure that its usage does not inadvertently harm or stigmatize specific neighborhoods or populations.

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.