Datasheet for 'Bike Share Toronto Ridership Data'*

Haowei Fan

December 3, 2024

The dataset is named "Bike Share Toronto Ridership Data" (Toronto Parking Authority 2024). It was sourced from Open Data Toronto (Gelfand 2022), uploaded by the Toronto Parking Authority (Toronto Parking Authority, n.d.), and collected by Bike Share Toronto (Bike Share Toronto, n.d.). The questions in the datasheet originate from "Datasheets for Datasets"(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 provide insights into the usage patterns of the Bike Share Toronto program, with the aim of supporting urban planning, transportation research, and promoting sustainable transportation in Toronto.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - The dataset was created by Bike Share Toronto, in collaboration with the City of Toronto's Open Data Team.
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 its open data initiative to promote transparency and innovation in public services.
4. *Any other comments?*
    - None.

Composition

*Dataset are available at: https://open.toronto.ca/dataset/bike-share-toronto-ridership-data/.

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 bike trips made using the Bike Share Toronto service. Each instance includes data such as trip duration, start and end station, start and end times, and bike ID.

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

    - The dataset contains data on 28,017,329 bike trips.

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 contains all recorded bike trips during the specified time period, representing the complete set of Bike Share Toronto ridership data.

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 raw data related to bike trips, including fields such as trip duration, start and end station, start and end times, and bike ID.

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

    - No, there is no specific label or target associated with each instance.

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.*

    - There may be occasional missing data for specific fields, such as start or end station, due to technical issues during data collection.

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 trips are not explicitly represented in the dataset, but users can infer patterns by analyzing common start and end stations or trip times.

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.*

    - No specific data splits are recommended. However, users may split the data by time periods (e.g., month, season) for analysis purposes.

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

    - There may be some noise in the dataset, such as incorrect timestamps or missing values, due to technical issues during 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 self-contained, but additional information about Bike Share Toronto can be found on the City of Toronto's open data website.

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, the dataset does not contain any confidential information.

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.*

    - No, the dataset does not contain any offensive or threatening content.

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.*

    - No, the dataset does not contain any demographic information about riders.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - No, it is not possible to identify individuals from the dataset.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

    - No, the dataset does not contain any sensitive data.

16. *Any other comments?*

    - None.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

    - The data was directly acquired through the Bike Share Toronto system, which records trip details automatically when bikes are rented and returned.

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 using automated software systems and sensors installed at bike docking stations.

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 is not a sample; it contains all recorded bike trips within the specified timeframe.

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 data collection was automated, and no manual human involvement was required.

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 was collected continuously over a period of time, with the timeframe matching the actual rental times of the bikes.

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.*

    - No ethical review process was conducted, as the dataset does not contain personal or sensitive information.

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 Bike Share Toronto's automated systems.

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.*

    - Riders are informed that their trip data may be used for analysis through Bike Share Toronto's terms of service.

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 is provided when riders agree to the terms of service upon signing up for the Bike Share Toronto program.

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 specific mechanism for revoking consent, as the data collected does not contain personally identifiable 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.*

    - No such analysis has been conducted, as the dataset does not contain personal data.

12. *Any other comments?*

    - None.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*
    - Minimal preprocessing was done, such as removing incomplete records or correcting obvious errors.
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.*
    - Yes, the raw data is available on the City of Toronto's open data portal.
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 specific preprocessing software was used beyond standard data cleaning tools.
4. *Any other comments?*
    - None.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - The dataset has been used for analyzing bike-sharing patterns, understanding peak usage times, and informing city planning and transportation policies. For example, the University of Toronto's School of Cities (Liu and Allen 2023) has studied the general patterns of bike-sharing in the Toronto region.
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.*
    - The University of Toronto's School of Cities conducted the study, and the URL for their research is as follows:https://schoolofcities.github.io/bike-share-toronto/growth.
3. *What (other) tasks could the dataset be used for?*
    - he dataset could be used for tasks such as predicting future bike demand, identifying underserved areas, or analyzing the impact of weather on bike usage.
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?*

- There are no known risks associated with the dataset, as it does not contain personal or sensitive information.

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 any tasks attempting to infer personal information about riders, as it lacks demographic data.

6. *Any other comments?*

   - None.

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 on the City of Toronto's 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)?*

   - The dataset is available for download as a CSV file on the City of Toronto's open data portal. The dataset does not have a DOI.

3. *When will the dataset be distributed?*

   - The dataset is already 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 distributed under the City of Toronto's Open Data License, which allows free use with attribution.

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 restrictions apply.

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.

7. *Any other comments?*
    - None.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - The dataset is maintained by the City of Toronto's Open Data Team and Bike Share Toronto.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - Contact information is available on the City of Toronto's open data portal.
3. *Is there an erratum? If so, please provide a link or other access point.*
    - No erratum is available.
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)?*
    - The dataset is being updated monthly by the City of Toronto's Open Data Team, with updates announced on the open data portal.
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 contain personal data, so no retention limits apply.
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 are not specifically maintained, but historical data remains available as part of the dataset.
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 external contributions.

8. *Any other comments?*
    - None.

References

Bike Share Toronto. n.d. "How It Works." https://bikesharetoronto.com/how-it-works/.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. "Datasheets for Datasets." *arXiv Preprint arXiv:1803.09010*. https://arxiv.org/abs/1803.09010.

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

Liu, Michael, and Jeff Allen. 2023. "Exploring Bike Share Growth in Toronto." https://schoolofcities.github.io/bike-share-toronto/growth.

Toronto Parking Authority. 2024. "Bike Share Toronto Ridership Data." https://open.toronto.ca/dataset/bike-share-toronto-ridership-data/.

———. n.d. "About Us." https://parking.greenp.com/about/about-us/.