Datasheet for Toronto Cycling Infrastructure Analysis Dataset*

Steven Li

1 December 2024

This datasheet documents the creation and composition of a dataset analyzing the causal impact of cycling infrastructure improvements on Bike Share Toronto ridership between 2017-2023. Following the standardized format proposed by Gebru et al. (2021), this datasheet details the processing of over 7.4 million bike share rides across 1,191 bikeways, focusing specifically on infrastructure changes during 2019-2021. The dataset combines and cleans records from Toronto's Open Data Portal, including bike share trips, station locations, and cycling infrastructure details, to create a structured dataset for difference-in-differences analysis. The final dataset contains monthly ridership metrics, infrastructure classifications, and temporal indicators designed to evaluate how protected lanes, on-road lanes, and shared roadways influence cycling activity.

All of the following questions are extracted 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 analyze the causal effect of cycling infrastructure improvements on Bike Share Toronto ridership between 2017-2023, specifically examining how different types of bikeways influence system usage.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - Created by Steven Li at the University 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.*

*Code and data are available at: https://github.com/stevenli-uoft/Toronto_BikeShare_Causality.

- No external funding received.

4. *Any other comments?*

   - No additional comments.

Composition

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

   - Each row represents a bikeway-month observation tracking ridership metrics 24-months before and after infrastructure changes.

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

   - 71,460 observations tracking 1,191 bikeways over 60 months each.

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).*

   - Dataset includes 1,191 of the roughly 1,500 bikeways in Toronto with associated Bike Share stations within 100m, and Bike Share ridership from 2017-2023.

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 contains:

     - bikeway_id: Unique identifier for each bikeway segment
     - construct_year: Year when the bikeway was either upgraded or newly installed
     - period: One of three periods (pre/treatment/post) relative to infrastructure changes
     - treatment: Boolean indicating if bikeway received improvements during 2019-2021
     - year_month: Time variable tracking month and year
     - calendar_year: Year of observation
     - relative_month: Months relative to treatment (-24 to +24)
     - bikeway_type: Infrastructure classification (Protected Lanes: Cycle tracks, multi-use trails, bi-directional cycle tracks. On-Road Lanes: Painted bike lanes. Shared Roadways: Sharrows, signed routes, park roads)

  - sub_treatment_type: Classification of infrastructure change (Upgraded: Existing infrastructure improved. Newly-Installed: New infrastructure added. No Treatment: Control group)
  - monthly_rides: Raw count of bike share trips starting near bikeway
  - monthly_rides_adj: Seasonally adjusted ride count

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

   - Treatment status (TRUE/FALSE) indicates whether bikeway received infrastructure improvements during 2019-2021.

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.*

   - No missing values in final analysis dataset after cleaning.

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.

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 recommended splits for analysis dataset.

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

- Spatial precision:

  - Only captures rides starting within 100m of bikeways
  - May miss riders who use bikeways but start further away
  - Some riders within 100m may not use the adjacent bikeway

- Temporal uncertainty:

  - Treatment year only known to annual precision
  - Construction/upgrade timing within year unknown
  - Seasonal patterns affect ridership (addressed through adjustment)

- Treatment classification:

  - Control group randomly assigned to pseudo-treatment years
  - Some bikeways may have received minor unreported improvements

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.*

    - Self-contained parquet file.

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 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 sensitive 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.*

    - Bikeways categorized by:

        - Infrastructure type (Protected Lanes, On-Road Lanes, Shared Roadways)
        - Treatment type (Upgraded, Newly-Installed, No Treatment)

    - Sub-treatment categorized by:

        - Newly-Installed: New installation of a bikeway
        - Upgraded: Upgrades to an existing bikeway

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 individual identification possible.

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 data included.

16. *Any other comments?*

    - No Additional Comments

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 sourced from City of Toronto Open Data Portal (Gelfand 2022):

    - Bike Share ridership data
    - Bike Share station locations
    - Cycling Network (Bikeways) data

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?*

    1. Initial Data Download:

        - Used `opendatatoronto` package (Gelfand 2022) in R (R Core Team 2023) to access raw data

        - Downloaded separate datasets for:

            - Bike Share ridership (2017-2023)
            - Bike Share station locations
            - Cycling Network (Bikeways)

    2. Spatial Processing:

        - Using the sf (Pebesma and Bivand 2023) package for spatial processing
        - Converted station locations to spatial points using sf package
        - Transformed coordinates to UTM zone 17N for accurate distance calculation
        - Calculated distances between stations and bikeways
        - Filtered to stations within 100m of bikeways

    3. Temporal Processing:

        - Standardized timestamps across datasets
        - Created relative time variables for diff-in-diff analysis
        - Generated seasonal indices for ridership adjustment

4. Data Integration:

    - Joined ridership data with nearby bikeways
    - Created treatment/control groups based on 2019-2021 changes
    - Generated monthly aggregates of ridership

All processing was validated through unit tests and intermediate data checks, using R packages including `tidyverse` (Wickham et al. 2019), `dplyr` (Wickham et al. 2023), and `sf` (Pebesma and Bivand 2023) for spatial analysis.

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

    - Full population of relevant bikeways included.

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 automated through R scripts.

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.*

    - Original data spans 2017-2023
    - Data processing completed December 2024

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 required for public data.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

    - All data obtained through Toronto Open Data Portal.

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.*

    - Not applicable for public administrative data.

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.*

    - Not applicable for public administrative 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 for public administrative data.

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.*

    - Not applicable for public administrative data.

12. *Any other comments?*

    - No additional comments.

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.*

    1. Station Data Cleaning:

        - Standardized coordinate systems
        - Removed stations with invalid coordinates
        - Generated unique station-bikeway pairs within 100m

    2. Ridership Data Standardization:

        - Unified column names across years
        - Converted timestamps to consistent format
        - Filtered invalid trip records
        - Added missing station IDs for 2017 Q3-Q4

    3. Bikeway Classification:

        - Categorized infrastructure types from detailed classes
        - Created treatment indicators based on construction/upgrade years
        - Assigned control bikeways to balanced pseudo-treatment years

4. Temporal Adjustments:
    - Created seasonal indices based on monthly patterns
    - Adjusted ridership for seasonality
    - Generated relative time variables (-24 to +24 months)
5. Final Integration:
    - Joined cleaned datasets
    - Generated monthly aggregates
    - Created final analysis variables

All cleaning steps are documented in the `03-clean_data.R` script.

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.*
    - Raw data preserved in data/01-raw_data directory
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
    - All processing done in R
4. *Any other comments?*
    - None

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - Analysis of causal impact of cycling infrastructure on bikeshare usage
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.*
    - GitHub: https://github.com/stevenli-uoft/Toronto_BikeShare_Causality
3. *What (other) tasks could the dataset be used for?*

- Evaluating infrastructure investment impacts
- Planning future cycling network expansion
- Studying seasonal cycling patterns

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?*

    - Results should be interpreted considering 100m station proximity threshold

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

    - Should not be used to assess individual rider behavior

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.*

    - Publicly available via GitHub

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

    - Parquet file on GitHub

- No DOI assigned

3. *When will the dataset be distributed?*

    - Available December 2024

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.*

    - No, MIT 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 additional restrictions or controls

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 additional restrictions or controls

7. *Any other comments?*
    - None

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - Maintained by Steven Li

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - Email: stevency.li@mail.utoronto.ca

3. *Is there an erratum? If so, please provide a link or other access point.*
    - No known errors

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)?*
    - No planned updates

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.*
    - Not applicable (public data)

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.*
    - Single version maintained

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.*

- Contributions welcome via GitHub pull requests

8. *Any other comments?*

    - None

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.

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

Pebesma, Edzer, and Roger Bivand. 2023. *Spatial Data Science: With applications in R*. Chapman and Hall/CRC. https://doi.org/10.1201/9780429459016.

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, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, et al. 2019. *Welcome to the tidyverse*. *Journal of Open Source Software*. Vol. 4. https://doi.org/10.21105/joss.01686.

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