Datasheet for ‘TTC Transit Delay Patterns’*

Open Data Toronto

Jerry Xia

December 3, 2024

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 provide a detailed record of transit delays across the Toronto Transit Commission (TTC) network in 2024. It was intended to fill a gap in publicly available, structured datasets on transit performance. This data facilitates the analysis of delays by transit mode (bus, subway, streetcar) and contextual factors like geography, time, and direction. The goal is to provide actionable insights for transit planners and researchers to improve system reliability and enhance commuter experiences.
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 City of Toronto and made available through Open Data Toronto, an official platform for public datasets.
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 creation of the dataset was funded and supported by the City of Toronto Toronto (2024). Open Data Toronto operates as part of the city’s broader transparency and open government initiatives.
4. *Any other comments?*
    - This dataset exemplifies the importance of publicly available resources for urban mobility research and planning.

[*] Code and data are available at: https://open.toronto.ca/.

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 comprises individual transit delay records. Each instance represents a specific delay event on the TTC network, capturing attributes such as mode (bus, subway, or streetcar), location, direction (bound), date, day of the week, and delay duration.

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

   - The dataset contains over 100,000 records of delay events, encompassing data for all three transit modes across a full calendar year.

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 transit delays reported by the TTC for 2024. It is not a sample; rather, it aims to be a complete record of delays across the network. Geographic coverage includes all TTC-served locations, but it may exclude certain unreported or unrecorded delay events.

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 features extracted from raw data, including mode, date, day of the week, bound, delay duration (in minutes), and location. These fields enable structured analysis of delay patterns.

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

   - Each instance is labeled with a unique identifier and includes fields for the recorded delay duration and associated attributes.

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 may have missing values for location or bound due to incomplete reporting by transit operators. These omissions likely reflect operational limitations or errors in data entry.

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, such as recurring delays at specific locations, are implicit but can be inferred through analysis of shared attributes like mode, location, or time.

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.*

   - Recommended splits include stratifying the data by transit mode or by time periods (e.g., training on delays from January to October and testing on November and December). Such splits facilitate predictive modeling and trend analysis.

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

   - Potential sources of noise include recording errors in delay duration or inconsistent naming of locations. Redundancies may arise from repeated reporting of the same delay event under different descriptions.

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 and available through Open Data Toronto. It does not rely on external resources, and there are no restrictions on its use beyond adherence to Open Data Toronto's licensing terms.

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 only contains publicly accessible transit performance data and does not include 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 contains neutral and factual transit delay information.

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 identify any sub-populations. Its focus is on transit delays and related operational attributes.

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 dataset does not contain personal or identifiable information.

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 include sensitive information.

16. *Any other comments?*

    - NA.

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 acquired through automated systems within the TTC network, which track and report delays in real-time. It was not inferred or reported by individuals but directly recorded.

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 TTC's internal monitoring systems, including GPS and operational logs, were used to record delay data. Validation occurs through automated consistency checks and manual reviews within TTC operations.

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 is a comprehensive record of TTC delays for 2024.

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 process was managed by the Toronto Transit Commission (TTC) as part of their routine operational monitoring. No external individuals, such as crowdworkers or contractors, were involved in the data collection.

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 throughout the calendar year of 2024 (Jan 1 - Oct 31). The collection timeframe aligns with the creation timeframe of the delay instances, as the dataset reflects real-time transit performance for that year.

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 processes were conducted, as the dataset does not involve human subjects or sensitive personal data. It comprises publicly accessible transit delay 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 dataset was obtained directly from TTC operational systems and published via Open Data Toronto. It does not involve data collected from individuals.

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.*

    - Notification was not applicable, as the dataset does not involve personal data or interactions with individuals.

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 was not applicable, as the dataset does not involve personal data or information from individuals.

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).*

    - Since no personal data is involved, consent and mechanisms for revoking consent are not applicable.

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 data protection impact analysis was conducted, as the dataset does not involve data subjects or personal information.

12. *Any other comments?*

    - NA.

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, preprocessing was performed to clean and standardize the dataset. This included harmonizing date formats, filtering out missing or erroneous entries, and categorizing locations into predefined groups (e.g., Dundas, Yorkdale, Other). Cardinal directions were standardized for the "bound" field, and redundant or inconsistent records were removed.

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 available through Open Data Toronto, ensuring that users can access the original unprocessed dataset for alternative analyses.

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 conducted using R, leveraging packages like `tidyverse` and `dplyr`. The code for cleaning and analysis is available on GitHub: https://github.com/Jerryx2020/TTC_Delay_Analysis.

4. *Any other comments?*

- Preprocessing steps were thoroughly documented to ensure reproducibility and transparency.

Uses

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

   - Yes, the dataset was used in this paper to analyze patterns in transit delays across TTC modes and locations, employing statistical methods and machine learning models (e.g., random forests) to identify key predictors of delays.

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 dataset and analysis code are linked to this project and can be accessed on GitHub: https://github.com/Jerryx2020/TTC_Delay_Analysis.

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

   - The dataset could be used for predictive modeling of transit delays, simulation of improved scheduling strategies, spatial analysis of delay hotspots, and temporal studies of seasonal trends in transit reliability.

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 focus on reported delays may exclude operational disruptions that were not officially logged, potentially underrepresenting certain types of delays. Users should interpret findings with this limitation in mind.

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 requiring personal or demographic data, as it does not contain such information.

6. *Any other comments?*

   - NA.

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 publicly available via Open Data Toronto.

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: https://open.toronto.ca/.

3. *When will the dataset be distributed?*

    - The dataset has already been distributed and is publicly accessible.

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 Data Toronto license, which permits free access and reuse with appropriate 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, there are no third-party-imposed restrictions.

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?*

    - NA.

Maintenance

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

    - The dataset is hosted and maintained by Open Data Toronto.

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

    - Open Data Toronto can be contacted via their website: https://open.toronto.ca/contact/.

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

- NA.

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 managed by Open Data Toronto and are communicated through their platform.

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.*

    - NA, as the dataset does not involve personal 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.*

    - Open Data Toronto does not explicitly maintain older versions of datasets. However, any updates or changes to the dataset are documented on the platform, and consumers are encouraged to check for the latest version when accessing 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.*

    - Contributions to the dataset are not directly accepted, as the dataset is curated and managed by Open Data Toronto. However, users can suggest improvements or raise concerns through the platform's contact page. Any such feedback is reviewed by the dataset administrators.

8. *Any other comments?*

    - NA.

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.

Toronto, City of. 2024. "Open Data Toronto." https://open.toronto.ca/.