Datasheet for 'A dataset'*

My subtitle if needed

First author Another author

2024-04-19

First sentence. Second sentence. Third sentence. Fourth sentence.

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 explore the impact of public transit infrastructure enhancements in Toronto, with a particular focus on the Route 508 Lake Shore streetcar. It hopes to address the gap in understanding how access to improved public transit could influence traffic congestion in urban environments.

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 Maria Mangru using data provided by the The City of Toronto's Transportation Services Division.

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.*

- Unknown. The dataset was sourced from the City of Toronto's Transportation Services Division, which is typically funded by municipal budgets rather than specific grants.

4. *Any other comments?*

- No additional comments.

Composition

*Code and data are available at: LINK.

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 instance within the dataset represents a specific time and location intersection within Toronto, capturing various modes of traffic volume, such as cars, buses, pedestrians, and cyclists.

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

- Over 200 instances in the cleaned dataset.

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 intended to represent all traffic volume counts at selected city intersections and is not a random sample; it's a comprehensive aggregation across multiple years.

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 counts of death including demographic data (race and sex) and year. Is there a label or target associated with each instance? If so, please provide a description.*

- Each instance includes aggregated traffic volume counts for different modes of transport at various intersections in Toronto in addition to the date this instance was observed.

5. *Is there a label or target associated with each instance?*

- Yes, instances could be labeled with unique identifiers such as the intersection ID and the date and time of the traffic counts.

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.

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.*

- The data includes both temporal and spatial data for each instance such as the time and location coordinates.

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.

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

- As with any real-world dataset, some level of noise, error, or redundancy is likely but none have been observed to date.

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 does not rely on external resources for its primary analyses.

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 consists of aggregated traffic counts that are non-confidential and derived from public domain sources.

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 is strictly numerical and traffic-related, with no content that could be considered offensive or anxiety-inducing.

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 identify any sub-populations, as it deals with vehicles and traffic, not individuals.

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, individuals cannot be identified from this dataset as it does not contain personal data.

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 information pertaining to individuals is included. The dataset only contains information related to traffic volumes and patterns.

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.* - The data for each instance were obtained from automatic traffic recorders and manual counts at intersections. These data collection methods are considered direct observations of traffic volumes.

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 were collected using a combination of hardware apparatuses, such as automatic traffic recorders (ATR), and manual turning movement counts (TMCs). The procedures and equipment were likely standardized and validated by the city's transportation services to ensure accuracy.

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 a comprehensive collection of traffic volumes at various intersections within Toronto, not a sample from a larger set.

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 performed by the City of Toronto's Transportation Services Division, which employs trained personnel for such tasks. Compensation details for city employees are typically determined by municipal guidelines and salary scales.

5. *Over what timeframe was the data collected?*

- The dataset encompasses traffic volume data from 2010 to 2024, which likely matches the creation timeframe for the instances.

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.*

- For data collection that involves public transportation metrics without personal identifiers, ethical review processes are generally not applicable.

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

- Data were collected directly from the traffic volumes and movements at intersections, not 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.*

- Not applicable as the dataset consists of traffic counts, not personal data.

9. *Did the individuals in question consent to the collection and use of their data?*

- Not applicable as no individual personal data was collected.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?*

- Consent and mechanisms to revoke consent are not applicable to this dataset.

11. *Has an analysis of the potential impact of the dataset and its use on data subjects been conducted*

- Given the dataset contains aggregated non-personal data regarding traffic volumes, an analysis of the impact on data subjects would not be applicable.

12. *Any other comments?*

- No additional comments. The dataset is designed for urban planning and transportation analysis and does not involve personal data collection.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done?*

- Yes, preprocessing and cleaning of the data were done. This included aggregating traffic volumes to daily totals, cleaning anomalies, filtering data for relevant time periods, and aligning the dataset for comparative analysis before and after the implementation of transit improvements.

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data?*

- Yes. Both the raw and cleaned data were saved to this project.

3. *Is the software that was used to preprocess/clean/label the data available?*

- Yes, R programming language and its various packages were used for data preprocessing, cleaning, and analysis. The specific scripts used for this can be found on this GitHub repository (https://github.com/MariaMangru/Traffic-Congestion-on-Toronto-s-Shoreline).

4. *Any other comments?*

- No additional comments.

Uses

1. *Has the dataset been used for any tasks already?*

- The dataset has been used for the current study, which investigates the impact of public transit improvements on urban car congestion, utilizing traffic volume data to examine changes in traffic patterns over time.

2. *Is there a repository that links to any or all papers or systems that use the dataset?*

- This repository (https://github.com/MariaMangru/Traffic-Congestion-on-Toronto-s-Shoreline) is the only one to day which uses this dataset.

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

- Beyond studying traffic congestion, the dataset could potentially be used for urban planning, environmental impact assessments, and public policy formulation regarding transportation and mobility in urban areas. It may also be relevant for studies on the socio-economic implications of traffic 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?*

- The dataset is specific to traffic volumes in Toronto and reflects the unique urban and transit infrastructure of the city. Users should be cautious about generalizing findings to other cities without considering local differences.

5. *Are there tasks for which the dataset should not be used?*

- The dataset should not be used for tasks that require personal, sensitive, or confidential data as it is composed of aggregated non-personal traffic volumes.

6. *Any other comments?*

- No additional comments.

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 available through the NYC Open Data Profile.

2. *When will the dataset be distributed?*

- The dataset is uploaded and distributed annually through the NYC Open Data Profile.

3. *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.*

- None that are known.

4. *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.*

- None that are known.

5. *Will the dataset be distributed to third parties outside of the entity on behalf of which the dataset was created?*

- The dataset, as obtained from the Toronto OpenData portal, is already publicly available. As such, it can be accessed by third parties for analysis and research purposes.

6. *How will the dataset be distributed? Does the dataset have a digital object identifier (DOI)?*

- The dataset is distributed through the Toronto OpenData portal, typically in CSV format. It does not have a DOI as it is continuously updated by the city's transportation services division and not a static dataset.

7. *When will the dataset be distributed?*

- The dataset is currently available and is periodically updated by the Toronto OpenData Portal.

8. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?*

- The dataset from the Toronto OpenData portal is under an open government license, which allows for free use, modification, and sharing of the data, as long as the City of Toronto is credited for the data.

9. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances?*

- There are no known IP-based or other restrictions imposed by third parties on the dataset.

10. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?*

- There are no known export controls or regulatory restrictions that apply to this dataset, as it consists of non-sensitive, aggregated traffic volume data.

11. *Any other comments?*

- Any researcher or user of this dataset should ensure proper attribution as per the licensing terms of the City of Toronto and should verify that they are using the most up-to-date data available.

Maintenance

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

- The dataset is maintained by the City of Toronto's Transportation Services Division.

2. *How can the owner/curator/manager of the dataset be contacted?*

- Inquiries regarding the dataset can be directed to the City of Toronto's Transportation Services team at (VolumeCollision_Requests@toronto.ca).

3. *Is there an erratum?*

- There is no official erratum; however, the dataset is subject to updates and corrections as part of its regular maintenance schedule.

4. *Will the dataset be updated?*

- The dataset is regularly updated by the Transportation Services Division. Updates, including changes in traffic volumes or infrastructure, are typically communicated through the OpenData portal.

5. *Are there applicable limits on the retention of the data associated with the instances?*

- The dataset does not include personal data, and therefore retention limits associated with personal data do not apply.

6. *Will older versions of the dataset continue to be supported/hosted/maintained?*

- The OpenData portal maintains a history of updates, allowing for access to previous versions of the dataset. However, users are encouraged to use the most current data.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?*

- Contributions to the city's open data, such as corrections or enhancements, can be done through communication with the Open Data team. However, contributions are subject to validation and incorporation by the city staff.

8. *Any other comments?*

- Users of the dataset should regularly check the OpenData portal for updates and note the date of the version they are using in their analysis to ensure reproducibility.

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.