Datasheet for Open Data Toronto Ward Budget Dataset*

Comprehensive Overview of Ward Budgets (2021-2024)

Aviral Bhardwaj

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 comprehensive overview of the Ward Budgets in Toronto from 2021-2024. It aims to offer detailed insights into budget allocation and spending patterns across different wards, helping identify potential disparities or inequities in resource allocation.
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 Aviral Bhardwaj, with the source being Open Data Toronto (City of Toronto 2024)
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 project is self-funded by the creator, with no external grant support.
4. *Any other comments?*
    - There are no additional comments.

Composition

*Code and data are available at: https://github.com/Aviral-03/InfrastructureCausalModel-TO.

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 represents budget allocations for Toronto wards. It includes multiple instance types such as wards, budget categories, budget amounts, and years (2021-2024).

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

    - The dataset contains 25 wards in Toronto, covering 4 years (2021-2024), for a total of 100 ward-year combinations.

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 appears to be a comprehensive collection of ward budgets, covering all 25 wards in Toronto for the specified time period. It represents official budget data from Open Data Toronto (City of Toronto 2024)

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.*

    - Raw data contains these headers: Program/Agency Name, Project Name, Sub-Project Name, Total 10 Year, Ward Number, Ward, Category

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

    - Instances are labeled by ward identifier, budget category, year, and budget amount.

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 information is missing from the instances.

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.*

    - Instances are related through consistent ward identifiers, comparable budget categories across years, and geographical relationships between wards.

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.*

    - The data can be naturally split by ward, year, and budget category.

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

    - Potential considerations include budget rounding, categorization inconsistencies, and currency fluctuations. It is recommended to cross-reference with official sources.

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 primarily based on Open Data Toronto (City of Toronto 2024) and recommends keeping the original source documentation. There are no known external resource restrictions.

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 contains public budget data with no personally identifiable information. Data is aggregated at the ward level.

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 dataset contains no offensive or anxiety-inducing content, as it is purely factual budget 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 is divided by ward, covering all 25 Toronto wards, but it does not provide demographic breakdowns within the wards.

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.*

- There is no direct individual identification within the dataset, as it is aggregated at the ward level.

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 contains public financial data and no sensitive personal information. Budget allocations are public records.

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 was directly sourced from Open Data Toronto (City of Toronto 2024), an official government data portal.

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 dataset was likely obtained through a software API or downloaded from Open Data Toronto (City of Toronto 2024), with manual compilation and verification for 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 provides a comprehensive collection, including all 25 Toronto wards and covering all years from 2021 to 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 was collected by Open Data Toronto (City of Toronto 2024) and was made available by Government of Toronto.

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 budget planning and allocation data was collected over the years 2021-2024, matching the creation timeframe of the dataset. Each year on February 1st, the City of Toronto releases the budget for the upcoming 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.*

    - The dataset follows ethical guidelines, as it consists of publicly available data with no personal privacy concerns, adhering to open government data principles.

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 obtained from Open Data Toronto, an official government 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.*

    - N.A

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.*

    - Public budget data is not subject to individual consent requirements, as it is publicly available information.

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).*

    - N.A

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.*

    - Public budget data does not require a data protection impact analysis, as it is publicly available information.

12. *Any other comments?*
    - N.A

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.*
    - Potential preprocessing steps include standardizing budget category names, ensuring consistent formatting, and converting data to appropriate types, and combining each year's data into a single dataset.
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 on Open Data Toronto, and the preprocessed data is available in the dataset. Link: 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.*
    - `opendatatoronto` package in R was used to download the data. The preprocessing steps were done in R and Python.
4. *Any other comments?*
    - N.A

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - The dataset can be used for urban policy analysis, budget allocation research, comparative ward-level budget studies, and local government funding analysis.
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.*
    - No
3. *What (other) tasks could the dataset be used for?*
    - TBD

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 is based on official government records and should be used responsibly, avoiding any unfair treatment or misrepresentation of the data. It is recommended to cross-reference with official sources and consider the context of budget allocation.

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

    - No.

6. *Any other comments?*

    - No.

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 likely to be shared for research or educational purposes, with no commercial distribution intended.

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 can be shared via platforms such as GitHub or a research repository. It is also recommended to assign a DOI for academic citation.

3. *When will the dataset be distributed?*

    - The dataset is available now and can be distributed upon request.

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 follows Open Data Toronto's licensing terms, with a recommendation to use a Creative Commons Attribution license and to cite the original source.

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 are known to be associated with the dataset.

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 are known to apply to the dataset.

7. *Any other comments?*

    - No

Maintenance

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

    - The dataset is maintained by Aviral Bhardwaj, and annual updates are recommended and raw data is available on Open Data Toronto (City of Toronto 2024).

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

    - Contact information has been provided in the author section.

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

    - No.

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)?*

    - It is suggested that the dataset be reviewed annually, with updates to include new budget years and verification of data accuracy. Updates will be communicated through the dataset repository.

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.*

    - No personal data is included in the dataset, so there are no retention limits.

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.*

- No

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 are welcome, with a suggested peer review process and validation against official sources.

8. *Any other comments?*

    - No.

References

City of Toronto. 2024. "Budget - Capital Budget Plan by Ward (10-Yr Approved)." Data Set. City of Toronto Open Data Portal; City of Toronto. open.toronto.ca/dataset/budget-capital-budget-plan-by-ward-10-yr-approved/.

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.