Datasheet for ‘Toronto Bids Awarded Contracts Dataset’*

Mandy He

December 1, 2024

This dataset contains information on competitive contracts awarded by the City of Toronto. It includes contract types such as tenders, requests for proposals, and more, spanning a range of services and goods. The data aims to increase transparency in procurement processes and facilitate analysis of contract trends and small business participation. It provides insights into city spending, supporting research into public procurement and equity in contract awards.

Extract of the questions from Gebru et al. (2021).

Answers of the questions are extract and referenced from City of Toronto (2024b), City of Toronto (2024a), City of Toronto (2024c).

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.*
    - This dataset was created to provide insight into Toronto’s public procurement processes. It addresses gaps in publicly available data on awarded contracts, enabling analysis of spending trends and equity in awarding processes.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - This dataset is published by the City of Toronto’s Open Data program.
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 is funded and maintained by the City of Toronto as part of its transparency and public accountability initiatives.

[*] Code and data are available at: https://open.toronto.ca/dataset/tobids-awarded-contracts/

4. *Any other comments?*

    - The dataset is intended for public use and research into government procurement efficiency, small business participation, and spending patterns.

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 instance represents a contract awarded by the City of Toronto, including details such as award date, value, type, and whether it was awarded to a small business.

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

    - There are 707 rows of observation and 13 variables.

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 captures all competitive contracts awarded, though it may exclude certain non-competitive or emergency procurement.

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 includes award details (e.g., _id, unique_id, Document Number, RFx (Solicitation) Type, High Level Category, Successful Supplier, Awarded Amount, Award Date, Division, Buyer Name, Buyer Email, Buyer Phone Number, Solicitation Document Description) in structured format.

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

    - TBD

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 lack award amounts or vendor classifications due to incomplete reporting.

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 contracts and categories (e.g., goods/services) are represented explicitly.

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.*

    - None explicitly provided, though data can be split by fiscal year or award type for analysis.

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

- Occasional data entry errors or inconsistencies in vendor classifications. For instance, the same supplier's name might be entered differently (e.g., "KMPG" and "KPMG LLP"), leading to potential redundancy and challenges in analysis. This highlights a need for data cleaning and normalization to ensure accuracy and consistency.

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 relies solely on the City of Toronto's internal reporting systems.

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 does not contain sensitive or confidential information. Although buyer names and email addresses are included, these details are not confidential as they are publicly available in contract offerings and other procurement-related documentation.

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 does not contain offensive 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.*

    - The dataset itself does not include explicit sub-populations such as age or gender. However, vendors can be classified into categories such as small or large businesses by researching supplier names and applying criteria like market capitalization or employee count. This classification would need external validation and is not directly available within the dataset.

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 can be identified 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 sensitive data is included.

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 is collected directly from City of Toronto procurement records.

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 is extracted from the city's internal procurement systems and formatted for public release. All solicitations were publicly advertised in accordance with Chapter 195, Purchasing By-law. Contracts were awarded to the lowest bidder meeting specifications for Request for Quotations and Tenders, and the highest scoring proponent in the case of Request for Proposals. Competitive Contracts are posted on the Toronto Bids Portal and are available for 18 months after the date the contract is created.

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 comprehensive, including all competitive contracts awarded.

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 is curated by City of Toronto staff.

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.*

    - Covers contracts awarded from 2022 to the present.

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.*

    - Not applicable; the dataset contains public procurement 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)?*

    - Data is directly sourced from the City of Toronto's 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.*

    - Not applicable as data pertains to contracts, not individuals.

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 cleaning; award amounts are standardized, and vendor (small, large) classifications added.

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 is retained as a CSV that can be found in folder data –> 01-raw_data –> Awarded Contracts.csv. It is also accessible via the Open Data Toronto portal (https://open.toronto.ca/dataset/tobids-awarded-contracts/).

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

    - Publicly available ETL (Extract, Transform, Load) tools.

Uses

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

    - Analysis of procurement trends and small business participation.

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.*

    - N/A

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

    - Research into equitable contracting, fiscal transparency, and public sector efficiency.

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

    - The dataset is specific to Toronto's procurement processes and may not be applicable to other jurisdictions without careful adaptation.

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.*

    - Available publicly 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)?*

    - Accessible as CSV files from the Open Data Toronto portal.

3. *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 Government License – Canada, which allows free use, modification, and sharing of the data, provided attribution is given to the source. Additional details about the license and its terms can be found here: https://open.toronto.ca/open-data-license/.

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.*

    - The dataset is licensed under the Open Government License – Canada, allowing users to copy, modify, distribute, and use the data for any lawful purpose, including commercial use. There are no specific export controls or regulatory restrictions beyond the terms of this license. For more details on the terms and conditions of this license, please refer to the official documentation at the following link: https://open.toronto.ca/open-data-license/.

Maintenance

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

    - The City of Toronto maintains and updates the dataset.

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

    - The dataset is published by the Purchasing & Materials Management division. For inquiries, you can contact them via email at supplychain@toronto.ca. Additionally, queries can be directed to the Open Data Toronto team for further assistance.

3. *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 updated daily via the Open Data Toronto portal. Updates include the addition of new instances and corrections to existing data. The dataset is directly maintained by the City of Toronto's Open Data team. Consumers can access the latest version of the data on the portal, though updates are not necessarily communicated.

References

City of Toronto. 2024a. "Open Data License - Canada." https://open.toronto.ca/open-data-license/.

———. 2024b. "Toronto Bids - Awarded Contracts." https://open.toronto.ca/dataset/tobids-awarded-contracts/.

———. 2024c. "Toronto Bids Portal." https://www.toronto.ca/business-economy/doing-business-with-the-city/searching-bidding-on-city-contracts/toronto-bids-portal/#awarded.

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.