Datasheet for 'PollinateTO Primary Project Garden Locations' Dataset*

An Overview of Garden Locations, Features, and Grant Information for PollinateTO Recipients

Aliza Abbas Mithwani

14 December 2024

The PollinateTO Primary Project Garden Locations dataset documents the locations and characteristics of pollinator gardens funded through Toronto's PollinateTO initiative. Analysis of this dataset reveals patterns in how garden funding is distributed across neighborhoods, including variations by ward, neighborhood type, and project characteristics. These findings highlight the potential influence of municipal funding priorities on urban biodiversity and community engagement. Understanding these patterns helps inform equitable urban greening efforts and supports the broader goal of fostering pollinator-friendly environments in cities.

Extract of the questions from Gebru et al. (2021). Questions answered using PollinateTO Primary Project Garden Locations dataset found on Open Data Toronto portal (Environment & Climate 2024).

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 examine the allocation patterns of grants provided by the PollinateTO initiative in Toronto. The aim is to understand how various factors—such as pollinator types, neighborhood characteristics, and whether gardens are Indigenous-led—influence the size and distribution of community gardens. This analysis can inform future decision-making and equity considerations in grant allocations.

2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*

*Code and data are available at: https://github.com/alizamithwani/PollinateTO.git.

- The dataset is published by the City of Toronto's Environment & Climate department under the Open Government License - Toronto. The contact for the dataset is pollinateto@toronto.ca.

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 funding for data creation likely stems from Toronto's municipal budget for environmental and sustainability initiatives. Specific grant details are not provided but may relate to urban biodiversity programs.

4. *Any other comments?*

   - The dataset provides a valuable foundation for analyzing urban greening efforts and community-based sustainability practices.

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 single PollinateTO-funded project, characterized by attributes such as location, grant size, garden size, urban density, and specific project details.

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

   - The dataset contains 513 instances.

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 represent the full set of PollinateTO-funded projects as of the date of collection. While it is unclear whether all relevant years and projects are included, the data spans multiple years and covers various wards and neighborhoods in Toronto. While the garden locations are publicly visible and accessible, some are on private property. This could affect analysis, particularly if any data points are not fully accessible for public engagement or observation.

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 structured tabular data, including features such as the year, ward, pollinator type, grant size, garden size, urban density, and whether the garden is Indigenous-led.

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

   - There is no explicit "target" variable. However, fields such as "Grant size" or "Garden size" could serve as targets in predictive analyses.

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.*

   - There may be missing values in some columns (e.g., urban density or Indigenous-led status). Reasons for missing data are unclear but could include incomplete reporting or lack of data collection for certain fields.

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 are defined. However, some relationships may exist implicitly, such as multiple projects within the same ward or neighborhood.

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 data splits are provided. Data consumers could create splits based on temporal factors (e.g., by year) or geographic boundaries (e.g., wards).

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

   - The dataset appears clean at first glance. However, potential inconsistencies in categorical variables (e.g., "Pollinator type" or "Ward") could introduce noise.

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 appears self-contained.

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 data pertains to public grants and community gardens and does not include sensitive or 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 is unlikely to contain any offensive or anxiety-inducing 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 identifies Indigenous-led gardens, which could be considered a subpopulation. Distributions by ward and neighborhood are also available.

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 data focuses on community gardens and grants rather than individual-level 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 sensitive data is included. The only potentially sensitive field is whether a garden is Indigenous-led, but this is framed as a programmatic focus rather than an individual attribute.

16. *Any other comments?*

    - The dataset is straightforward and seems designed for public use and analysis.

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 appears to have been collected through administrative records from the PollinateTO initiative. These records likely document grant approvals and associated project details.

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 collection likely involved manual curation from grant applications and project reports submitted to the PollinateTO initiative, with potential use of spreadsheets or database software for recording project information.

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 does not appear to be a sample but rather includes all available PollinateTO projects up to the date of collection.

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)?*

    - Municipal staff, researchers, or grant administrators likely conducted data collection as part of their roles. Specific information about personnel and compensation is not provided.

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 dataset spans multiple years, likely corresponding to the start of the PollinateTO initiative. The exact collection timeframe is unclear but seems aligned with the duration of the program.

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.*

    - Ethical reviews were likely unnecessary, as the dataset pertains to public grants and does not include sensitive personal 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)?*

    - The dataset does not directly relate to people.

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 as the dataset does not directly relate to people.

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.*

    - N/A as the dataset does not directly relate to people.

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 as the dataset does not directly relate to people.

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.*

    - N/A as the dataset does not directly relate to people.

12. *Any other comments?*

    - No other 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.*

    - Minor preprocessing may have been conducted, such as standardizing categorical variables (e.g., ward names or pollinator types). However, the dataset primarily appears to be raw or minimally processed administrative data.

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.*

    - This is unclear. The dataset provided does not include any accompanying raw data.

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

- No preprocessing software is specified. If preprocessing was conducted, it may have involved common data tools such as Excel or R.

4. *Any other comments?*

    - No other comments.

Uses

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

    - The dataset has not yet been used for published analyses but is intended for examining factors influencing grant allocations and the distribution of PollinateTO-funded gardens.

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 known repository exists at this time.

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

    - Since the dataset is openly available under a public license, it can be freely used for analysis, but any use of private property data might need to be considered in the analysis to ensure compliance with privacy expectations. The dataset could be used to: Analyze spatial patterns of urban greening efforts; Evaluate equity in grant distribution across neighborhoods; Identify trends in funding for Indigenous-led projects; Explore the impact of urban density on garden sizes and types.

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?*

    - Consumers should be cautious when interpreting variables like "Indigenous-led" gardens to avoid unintended implications about communities or funding fairness. Additionally, variations in reporting quality across wards could bias results.

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

    - The dataset should not be used to draw conclusions about individual-level behavior or to assess the personal characteristics of those involved in projects.

6. *Any other comments?*

- No other 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.*
   - Yes, the dataset is intended for public analysis and research purposes.
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 could be shared via public repositories (e.g., GitHub) or municipal open data portals. No DOI is currently available.
3. *When will the dataset be distributed?*
   - The distribution timeline is unspecified, but it is likely available immediately or soon after compilation.
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 published under the Open Government License - Toronto. The Open Government License - Toronto allows users to freely use, modify, and share datasets published by the City of Toronto, with the condition that the source is properly credited. The license permits both commercial and non-commercial uses, but disclaims any warranties regarding the accuracy or completeness of the data, and prohibits implying endorsement by the City of Toronto. There are no fees associated with the license, and users are free to create derivative works, as long as they meet the attribution requirement. Full details of the license can be found on the City of Toronto's Open Data Portal.
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 IP-based restrictions are indicated.
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 or regulatory restrictions apply.

7. *Any other comments?*
    - No other comments.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - Likely the municipal government of Toronto or affiliated researchers.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - Contact information is not provided but could potentially be found via municipal or research channels.
3. *Is there an erratum? If so, please provide a link or other access point.*
    - No errata are available at this time.
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)?*
    - The dataset could be updated annually to include new PollinateTO projects. Updates might be communicated via public portals or repositories. The dataset was last refreshed on October 15, 2024, and is updated annually.
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.*
    - The dataset does not relate to individuals, so retention limits are not applicable.
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.*
    - It is unclear if older versions will be maintained, though retaining historical records could aid longitudinal studies.
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.*
    - Collaboration mechanisms are not mentioned but could involve municipal partnerships or public data contribution tools.

8. *Any other comments?*

    - The dataset offers valuable opportunities for research into urban sustainability practices and grant distribution equity.

References

Environment & Climate. 2024. *PollinateTO Primary Project Garden Locations*. https://open.toronto.ca/dataset/pollinateto-primary-project-garden-locations/.

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.