Datasheet for 'Crime Rates in Toronto'*

An Analysis of Temporal and Spatial Crime Trends

Weiyang Li

2024-12-23

This datasheet documents the dataset used to analyze crime trends in Toronto from 2014 to 2023. It explains the motivation behind creating the dataset, details its composition and collection process, discusses preprocessing methods, and provides guidance on its intended uses and limitations.

1 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 analyze crime trends in Toronto, with a focus on temporal and spatial variations in crime rates. This fills the gap in understanding crime distributions at the neighborhood level over a decade.

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 Weiyang Li as part of an academic project at the University of Toronto.

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.*

    - This work was self-funded as part of a graduate-level research project.

4. *Any other comments?*

    - None.

*Code and data are available at: https://github.com/doravmony/static-dynamic-crime.

2 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 reported crime in a specific Toronto neighborhood, categorized by type and year.

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

    - The dataset includes over 13,000 rows, representing annual summaries of crime counts for various crime types across Toronto neighborhoods.

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?*

    - The dataset includes all available records for reported crimes in Toronto from 2014 to 2023.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features?*

    - Each instance consists of cleaned and processed features, including crime counts, crime rates per 1,000 population, and neighborhood population data.

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

    - The main targets are crime counts and crime rates per 1,000 population.

6. *Is any information missing from individual instances?*

    - Instances with missing data for population or crime counts were removed during preprocessing.

7. *Are relationships between individual instances made explicit?*

    - No direct relationships between instances are modeled, but data can be grouped by neighborhood or year for analysis.

8. *Are there recommended data splits?*

    - Yes, a 70:30 split into training and test datasets is recommended for modeling.

9. *Are there any errors, sources of noise, or redundancies in the dataset?*

- Some minor inconsistencies may exist due to underreporting or variations in data collection methods.

10. *Is the dataset self-contained?*

    - Yes, the dataset is self-contained and does not rely on external resources.

11. *Does the dataset contain data that might be considered confidential?*

    - No confidential data is included.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?*

    - The dataset includes crime data, which may cause anxiety for some users.

13. *Does the dataset identify any sub-populations?*

    - The dataset includes crime data grouped by neighborhood, but no demographic sub-populations are identified.

14. *Is it possible to identify individuals?*

    - No, the dataset is anonymized and does not include identifiable information.

15. *Does the dataset contain data that might be considered sensitive in any way?*

    - The dataset contains crime data, which could be considered sensitive.

16. *Any other comments?*

    - None.

3 Collection Process

1. *How was the data associated with each instance acquired?*

    - Data was sourced from publicly available government repositories, such as Toronto Police Service and Statistics Canada.

2. *What mechanisms or procedures were used to collect the data?*

    - Data was downloaded via APIs or directly from data portals and cleaned using R scripts.

3. *If the dataset is a sample from a larger set, what was the sampling strategy?*

- Not applicable; the dataset includes all available instances.

4. *Who was involved in the data collection process?*
   - The dataset was collected and processed by the author.
5. *Over what timeframe was the data collected?*
   - The dataset covers crimes reported between 2014 and 2023.
6. *Were any ethical review processes conducted?*
   - Not applicable; the dataset is based on publicly available data.
7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources?*
   - Data was obtained from third-party government sources.
8. *Were the individuals in question notified about the data collection?*
   - Not applicable; the data is anonymized.
9. *Did the individuals in question consent to the collection and use of their data?*
   - Not applicable; the data is anonymized.
10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent?*
    - Not applicable.
11. *Has an analysis of the potential impact of the dataset and its use on data subjects been conducted?*
    - Not applicable; the dataset is anonymized.
12. *Any other comments?*
    - None.

4 Preprocessing/Cleaning/Labeling

1. *Was any preprocessing/cleaning/labeling of the data done?*
   - Yes, missing values were removed, columns were normalized, and crime rates were calculated per 1,000 population.
2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data?*
   - Yes, raw data is saved in "data/01-raw_data/".
3. *Is the software that was used to preprocess/clean/label the data available?*
   - Yes, R scripts for preprocessing are available in the "scripts" directory.
4. *Any other comments?*
   - None.

5 Uses

1. *Has the dataset been used for any tasks already?*
   - Yes, for statistical modeling and analysis of crime trends in Toronto.
2. *Is there a repository that links to any or all papers or systems that use the dataset?*
   - Yes, the repository is available at https://github.com/doravmony/static-dynamic-crime.
3. *What (other) tasks could the dataset be used for?*
   - The dataset could be used for predictive modeling, policy analysis, and resource allocation studies.
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's reliance on reported crimes may underestimate actual crime rates.
5. *Are there tasks for which the dataset should not be used?*
   - The dataset should not be used to draw conclusions about individual behavior.
6. *Any other comments?*
   - None.

6 Distribution

1. *Will the dataset be distributed to third parties outside of the entity?*
    - Yes, it is publicly available.
2. *How will the dataset be distributed?*
    - Via GitHub and other open platforms.
3. *When will the dataset be distributed?*
    - Immediately upon publication.
4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license?*
    - Yes, under the MIT license.
5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances?*
    - No.
6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?*
    - No.
7. *Any other comments?*
    - None.

7 Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - The dataset is hosted and maintained by the author.
2. *How can the owner/curator/manager of the dataset be contacted?*
    - Via email at weiyang.li@mail.utoronto.ca.
3. *Is there an erratum?*
    - No.

4. *Will the dataset be updated?*
    - Updates will be provided if new data becomes available.
5. *If the dataset relates to people, are there applicable limits on the retention of the data?*
    - Not applicable; the data is anonymized.
6. *Will older versions of the dataset continue to be supported/hosted/maintained?*
    - Older versions will remain accessible.
7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?*
    - Contributions can be made via GitHub pull requests.
8. *Any other comments?*
    - None.