Datasheet for 'Examining the Influence of Premises Type and Time of Day on Violent Crime in Toronto'*

Tim Chen

December 2, 2024

Provides access to a dataset of violent and non-violent crime occurrences in Toronto, highlighting contextual and temporal factors. This datasheet supports reproducible research using the dataset.

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?*
   - The dataset was created to enable analysis of violent and non-violent crime patterns in Toronto, focusing on premises type and time of day.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity?*
   - The dataset was created and published by Open Data Toronto, sourced from the Toronto Police Service.
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.*
   - No direct funding information is provided.
4. *Any other comments?*
   - The dataset enables Bayesian analysis and spatial-temporal crime prevention strategies.

*Code and data are available at: https://github.com/timchen0326/crime-analysis-toronto

2 Composition

1. *What do the instances that comprise the dataset represent?*
    - Each row represents a reported crime event in Toronto.
2. *How many instances are there in total?*
    - There are 37,061 instances in the cleaned analysis dataset.
3. *Does the dataset contain all possible instances or is it a sample?*
    - The dataset is a subset of reported crimes in Toronto, specifically focusing on July, known for peak social activity.
4. *What data does each instance consist of?*
    - Each instance includes crime type (violent or non-violent), premises type, and time of day.
5. *Is there a label or target associated with each instance?*
    - Yes, the target variable is `VIOLENT_CRIME`, a binary classification of crime type.
6. *Is any information missing from individual instances?*
    - Some location coordinates may be missing for unverified crimes.
7. *Are relationships between individual instances made explicit?*
    - No explicit relationships exist beyond shared locations or time patterns.
8. *Are there recommended data splits?*
    - No specific splits are recommended; analysis may involve custom temporal or premises-based splits.
9. *Are there any errors, sources of noise, or redundancies in the dataset?*
    - Self-reported and verified data may introduce biases or inaccuracies.
10. *Is the dataset self-contained, or does it link to external resources?*
    - It is self-contained and available via Open Data Toronto.
11. *Does the dataset contain data that might be considered confidential?*
    - No, all data is anonymized.
12. *Does the dataset contain data that might be offensive, insulting, threatening, or cause anxiety?*
    - It involves sensitive topics like violent crimes, which could be distressing.

13. *Does the dataset identify any sub-populations?*
    - It categorizes data by premises type and time of day.
14. *Is it possible to identify individuals from the dataset?*
    - No, all data is aggregated and anonymized.
15. *Does the dataset contain data that might be considered sensitive?*
    - Yes, it contains crime-related data.
16. *Any other comments?*
    - The dataset supports urban safety planning.

3 Collection process

1. *How was the data associated with each instance acquired?*
    - Data was collected from police-reported incidents, verified by the Toronto Police Service.
2. *What mechanisms or procedures were used to collect the data?*
    - Data was extracted from police reports and structured into digital records.
3. *If the dataset is a sample, what was the sampling strategy?*
    - It includes crimes reported during July, focusing on peak seasonal activity.
4. *Who was involved in the data collection process?*
    - Data was collected and structured by the Toronto Police Service.
5. *Over what timeframe was the data collected?*
    - The dataset reflects crimes reported during July, year unspecified in metadata.
6. *Were any ethical review processes conducted?*
    - Not reported.
7. *Did you collect the data directly, or obtain it via third parties?*
    - Data was obtained from Open Data Toronto.
8. *Were the individuals in question notified about the data collection?*
    - As this is aggregated public data, individual consent was not applicable.
9. *Any other comments?*
    - The dataset supports public safety efforts.

4 Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done?*
    - Yes, data was cleaned to remove missing values and reformat variables such as premises type and time of day.
2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data?*
    - Yes, raw data remains accessible via Open Data Toronto.
3. *Any other comments?*
    - Data cleaning focused on consistency for Bayesian modeling.

5 Uses

1. *Has the dataset been used for any tasks already?*
    - Yes, it was analyzed to assess crime risk by premises type and time of day.
2. *What (other) tasks could the dataset be used for?*
    - Urban planning, crime prevention, and spatial analysis.
3. *Any other comments?*
    - No restrictions on creative use cases.

6 Distribution

1. *Will the dataset be distributed to third parties?*
    - It is publicly available via Open Data Toronto.
2. *How will the dataset be distributed?*
    - Through Open Data Toronto as downloadable files.
3. *Any other comments?*
    - The dataset encourages public analysis for safety planning.

7 Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - Open Data Toronto manages its availability.
2. *Will the dataset be updated?*
    - Updates depend on Toronto Police Service and Open Data Toronto processes.
3. *Any other comments?*
    - Regular updates are not guaranteed but are likely in future iterations.