# Toronto Homicide Occurrences*

Shifts in crime patterns in Toronto and the different motivational factors to homicide

Ayoon Kim

27 April 2022

**Abstract**

This report presents an analysis of the City of Toronto's data on homicide occurrences. The results indicate that education, income, and living environment has an affect on homicide occurrences. In order to perform an analysis of the relationship between Toronto's homicide occurrences and such external factors, the statistical programming language R is used. The results contribute to our understanding of the issue, enabling to identify patterns to Toronto's homicide occurrences.

## 1 Introduction

Only heavy crimes such as first degree murder, second degree murder, infanticide or manslaughter are considered homicide. The definition of homicide is usually consistent across nations. Hence, compared to any other crimes, homicide has more international consensus on the meaning and nature of homicide. Therefore, it is important to examine homicide since it is considered as a "comparable and reliable barometer of violence in society".

As severe as it is, homicide is relatively rare in Canada compared to other countries, but in recent years, Canada's homicide rate has been higher. However, due to COVID-19, there has been stay-at-home measures and encouraged "physical distancing". This caused people to spend more time at home and caused businesses to close temporarily or adapt to different methods of services. These changes in society and economy has led to a shift in crime patterns across Canada. This report will help identify such shifts in crime patterns in Toronto and the different motivational factors that could lead to homicide.

The data originates from the Toronto Open Data Portal, which is the official source for Toronto Open Data from city divisions and agencies. The objective of Open Data is to help make the city more accountable, transparent, participatory and accessible. This paper includes data set on police annual statistical report on homicides. The original data includes event unique id, hood id, division, homicide type, object id, id, geometry, neighbourhood, occurrence date, and occurrence year. However, this report will focus on the division, hood ID, homicide type, neighbourhood, occurrence date, and occurrence year. The data set is converted into a usable data set with R using only the information that is necessary for analysis.

The paper begins by explaining the motivation behind the data and the analysis. Followed by the composition of the data obtained from the Toronto Open Data Portal. Afterwards, we go in depth about the methodology. Our data section contains plots that summarize the variables and an accompanying discussion. Followed by graphs generated from simulated data and its implications to understand how homicide occurrences changed over the years and to identify patterns to Toronto's homicide occurrences. The paper concludes with an explanation of possible weaknesses and limitations.

*Code and data are available at: https://github.com/19akim/Final_paper.git

## 2 Data

The data is obtained from the Toronto Open Data Portal, which is an open source delivery tool that allows people to access data. The Open Data Portal for Toronto first launched in 2009. It consists of various useful datasets, which is updated regularly for more accurate and practical information.

The data used for our analysis is the 'Police Annual Statistical Report - Homicides'. The dataset includes all homicide occurrences in Toronto from 2004 to 2020. As mentioned earlier, it contains event unique id, hood id, division, homicide type, object id, id, geometry, neighbourhood, occurrence date, and occurrence year. However, the dataset is cleaned to make analysis more clear and simple.

Also, this data has limitations. In order to protect the privacy of the involved people in the occurrences, the location of the occurrences have been offset to the nearest road intersection. This can possibly affect the count of occurrences reported since it may not reflect the geographies accurately.

## 3 Results and Discussion

The plots below summarize the variables in our dataset, which show the trends in homicide occurrences in Toronto. Figure 1 shows the distribution of homicide types in total. It shows that the number of shooting occurrences are the highest compared to stabbing and other types of homicide. Figure 2 shows the same pattern where each year, there is most number of shooting occurrences compared to the other homicide types. However, Figure 2 also displays an up and down pattern over the years. Starting from 2004 there is an increase until 2007. Since 2007 there is a decrease until 2012 where it starts increasing again until 2018. After 2018, the number of homicide occurrences starts to decrease once more. It is assumed that COVID-19 has an affect on such decrease in recent years from 2019. Figure 3 represents the distribution of homicide occurrences according to police divisions. The graphs show that 31 Division has the most number of homicide occurrences. It is noticeable that divisions in the edge of Toronto generally have more number of homicide occurrences compared to divisions in the central area of Toronto.

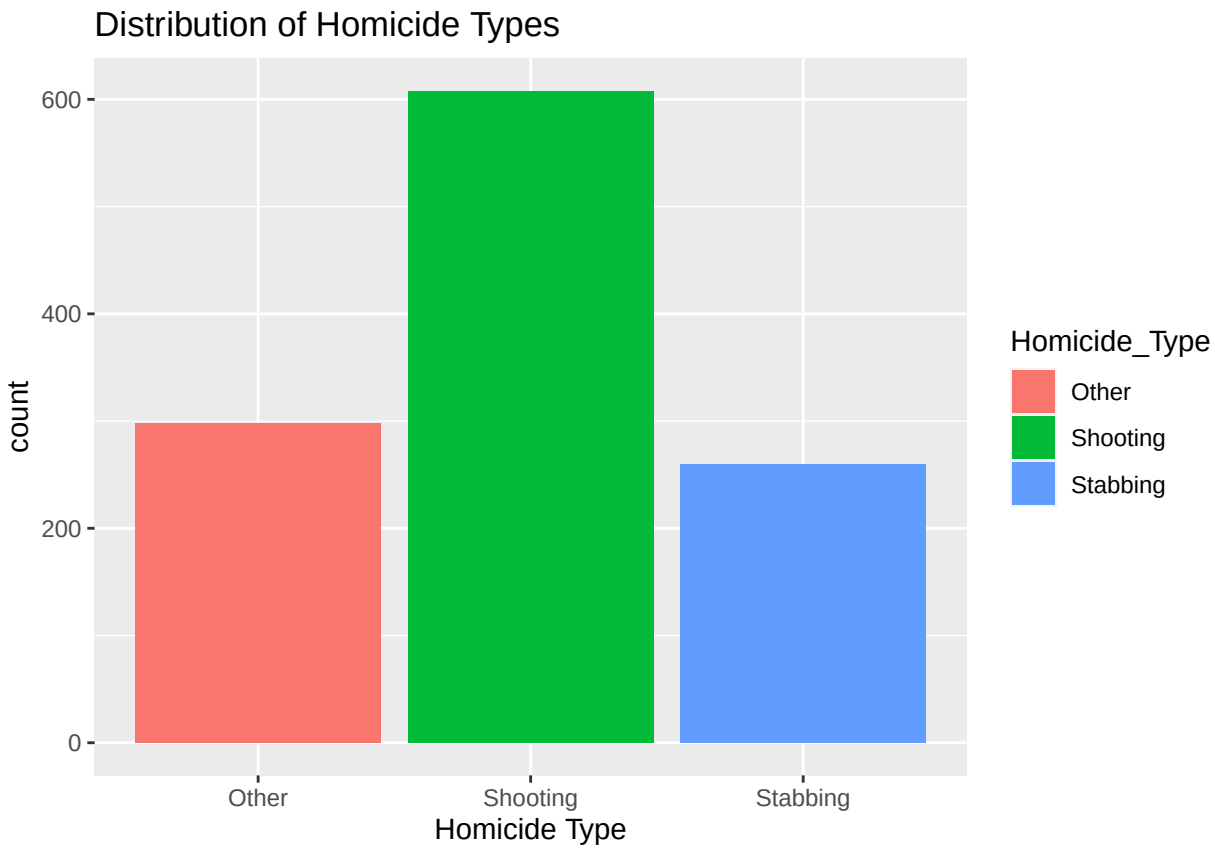



Figure 1: Distribution of Homicide Types (Shooting, Stabbing, Other)

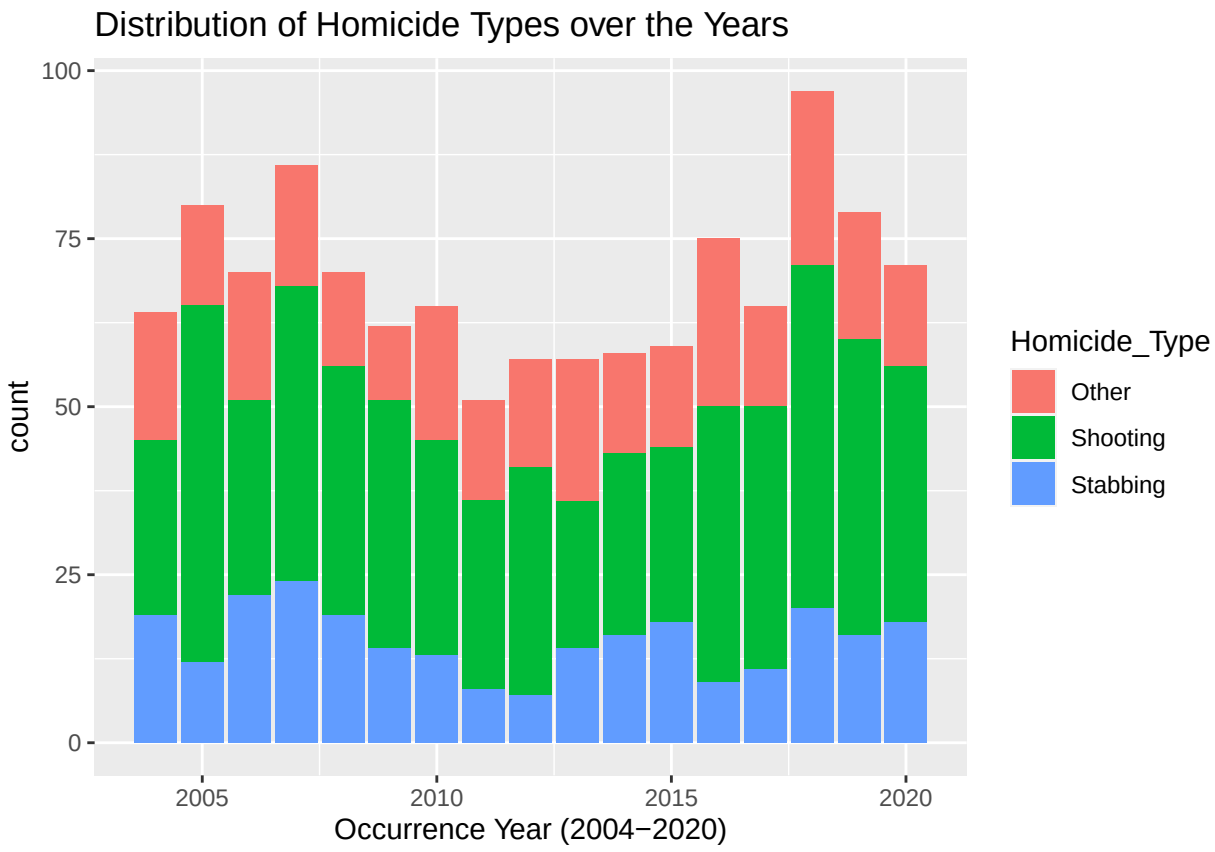



Figure 2: Distribution of Homicide Types over the Years (2004-2020)

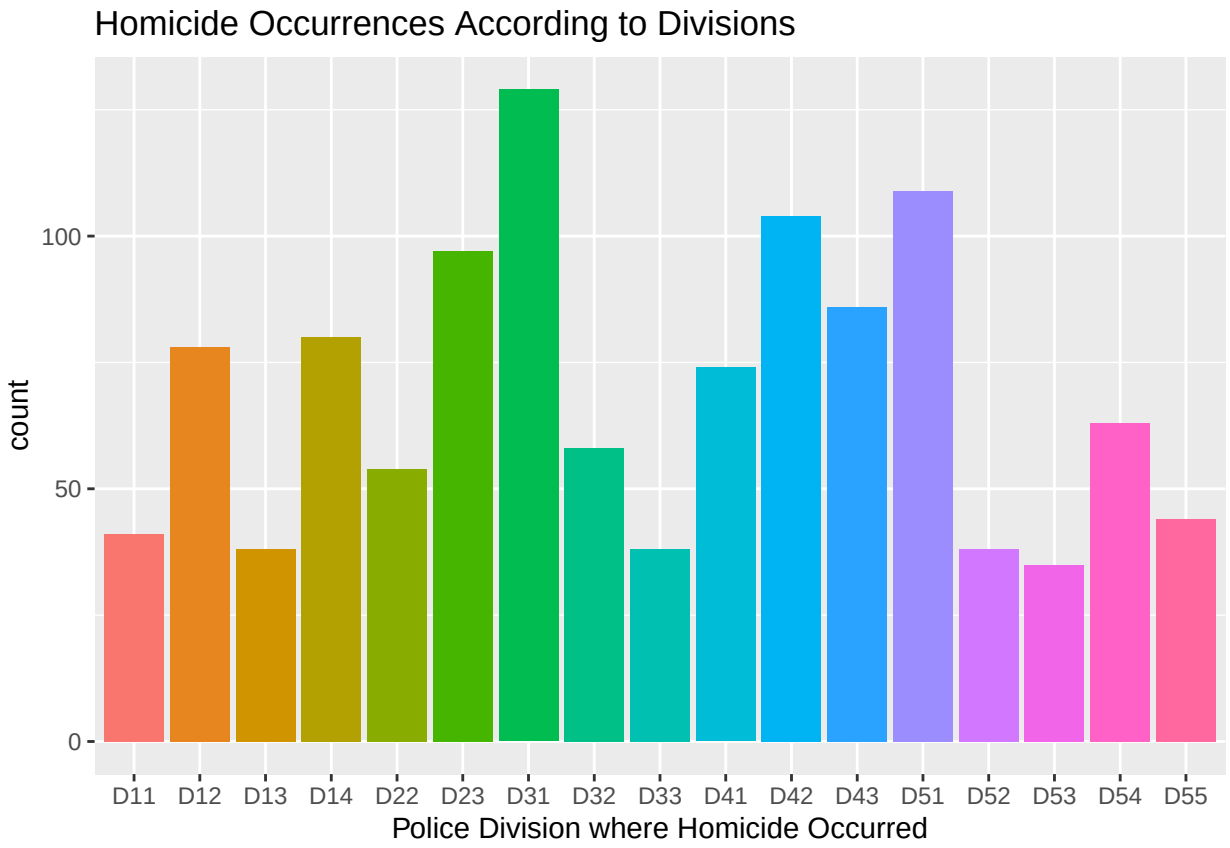
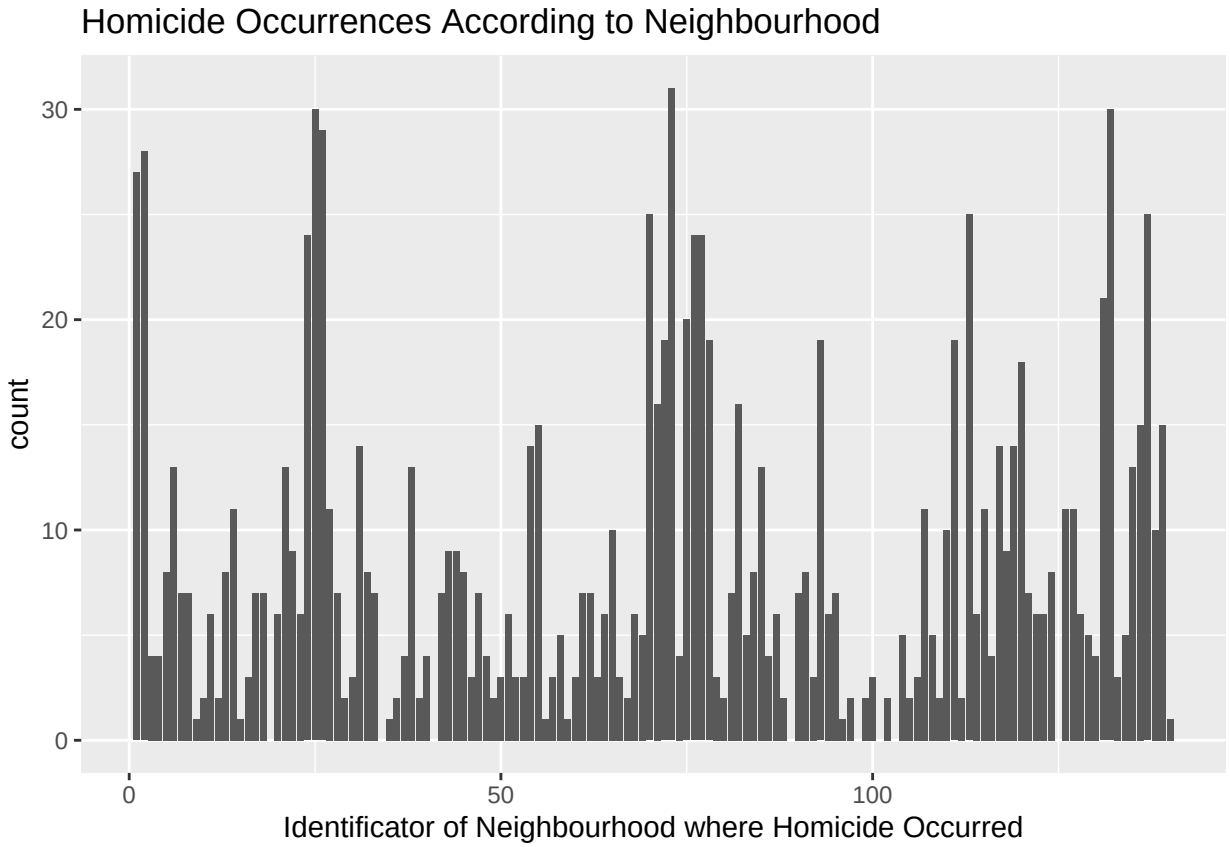



Figure 3: Distribution of Homicide Occurrences According to Police Division


Homicide Occurrences According to Neighbourhood
count
30
20
10
0
0
50
100
Identificator of Neighbourhood where Homicide Occurred

### 3.1 Weaknesses and next steps

There are some possible weaknesses in regards to this analysis. The data has limitations. In order to protect the privacy of the involved people in the occurrences, the location of the occurrences have been offset to the nearest road intersection. This can possibly affect the count of occurrences reported since it may not reflect the geographies accurately.

# Appendix

### 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 identify patterns to Toronto's homicide occurrences. Information about the meaning and nature of homicide was a gap that needed to be filled.

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 the Toronto Police Service's Annual Statistical Report. This is publicly available through the Toronto Police Service Public Safety Data Portal or the City of Toronto's Open Data Portal.

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.*

- None

### 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.*

    - The instances represent the homicide occurrences.

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

    - There are 1166 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).*

    - It is not representative of the larger se because it needs to cover a more diverse range of instances.

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.*

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

    - There is no label or targe associated with each instance.

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 is no missing information from individual 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.*

    - There are no relationships between individual instances.

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.*

    - There are no recommended data splits.

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

    - There are no errors, sources of noise, or redundancies in the dataset.

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 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.*

    - The data is aggregated, so it does not affect confidentiality of individual responses.

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 consist of data that might be offensive, insulting, threatening or axniety-inducing.

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 does not consist of any sub-populations.

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.*

    - It is not possible to identify individuals directly or indirectly as the raw data is provided in aggregated format.

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 consists of data that might be considered sensitive. It contains data that reveals locations/neighbourhoods.

16. *Any other comments?*

- None

**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 associated was acquired through the survey conducted by DHS.
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?*
   - Manual human curation was used to collect the data
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
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)?*
   - University students across all universities in Ankara, Turkey was primarily used to collect data. No information was reported about compensation.
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 data was collected from the first week of August 1998 to the end of November 1998.
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 review processes were not conducted
7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*
   - We obtained the data via the Demographic and Health Surveys website: dhsprogram.com
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.*
   - The interviews with data collectors were conducted on a voluntary basis. No notice is available.
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.*
   - The individuals consented to the collection and use of their data. No information of consent agreement language is provided.
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).*

- A mechanism to revoke consesnt was not provided

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.*
    - No analysis of the potential impact of the dataset and use on data subjects was conducted.
12. *Any other comments?*
    - None

**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.*
    - The data was obtained originally in the PDF format. The table was extracted manually to a usable dataset via R. This usable dataframe was applied in RStudio for conducting analysis.
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 obtained is saved in inputs/data/homicide.csv
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
    - R Software - https://www.R-project.org/
4. *Any other comments?*
    - None

**Uses**

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - No
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?*
    - The datset could be used to examine other factors regarding homicide occurrences.
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?*
    - None
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 examine factors other than homicide occurrences in Toronto.

6. *Any other comments?*
    - None

**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.*
    - No, the dataset is only avaiable for personal uses only.
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 will be distributed using Github.
3. *When will the dataset be distributed?*
    - The dataset will be distributed in April 2022.
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 will be released under the MIT license
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
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
7. *Any other comments?*
    - None

**Maintenance**

1. *Who will be supporting/hosting/maintaining the dataset?*
    - Ayoon Kim will be responsible for supporting, hosting and maintaining this dataset.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - Can be contacted via email and Github.
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)?*
    - Not at the moment.

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.*
    - Since we use aggregate data, the data does not relate to a specific individual.
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.*
    - No mechanisms are in place as of now
8. *Any other comments?*
    - None

# A Additional details

# B References

Open data dataset. City of Toronto Open Data Portal. (n.d.). Retrieved April 26, 2022, from https://open.toronto.ca/dataset/police-annual-statistical-report-homicide/

Armstrong, A., & Jaffray, B. (2021, November 25). This annual JURISTAT article presents 2020 Homicide Data. short and long-term trends in homicide are examined at the national, provincial/territorial and census metropolitan area levels. gang-related homicides, firearm-related homicides, intimate partner homicides, and homicides committed by youth are also explored. this juristat also presents data for which complete information regarding indigenous identity has been reported for both victims and accused persons, regardless of gender. Government of Canada, Statistics Canada. Retrieved April 26, 2022, from https://www150.statcan.gc.ca/n1/pub/85-002-x/2021001/article/00017-eng.htm

City of Toronto. (2022, April 12). Neighbourhood Profiles. City of Toronto. Retrieved April 26, 2022, from https://www.toronto.ca/city-government/data-research-maps/neighbourhoods-communities/neighbourhood-profiles/

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Wickham, Hadley. 2016. Ggplot2: Elegant Graphics for Data Analysis.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." Journal of Open Source Software 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2022. Knitr: A General-Purpose Package for Dynamic Report Generation in r. 18