Datasheet for Community Crime Statistics*

Haobo Ren

2024-11-29

The raw dataset obtained from Open Data Calgary(Calgary, n.d.) was recorded and updated monthly by the Calgary Police Service. The data is considered cumulative as late-reported incidents are often received well after an offence has occurred. An incident is either reported just after the crime happened, or reported on the Calgary Police Service(Service, n.d.).

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 transparent and accessible crime statistics for the city of Calgary. It serves to inform the public, researchers, and policymakers about trends and patterns in various types of crimes, enabling evidence-based decision-making and fostering accountability.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - The dataset is maintained by Open Calgary, the City of Calgary's open data platform. The data itself is collected and reported monthly by the Calgary Police Service (CPS).
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 maintaining Open Calgary and its datasets is provided by the City of Calgary as part of its commitment to open data initiatives.
4. *Any other comments?*
    - The dataset reflects the City of Calgary's effort to increase public trust and foster collaboration through open data sharing.

Composition

*Code and data are available at: https://github.com/HaoboRrrr/calgary_vehicle_crime_forcast

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 monthly record of reported crimes in Calgary, categorized by crime type (e.g., "Theft FROM Vehicle," "Theft OF Vehicle") and broken down by specific counts.

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

    - The dataset contains thousands of records spanning multiple years. Each record corresponds to a unique combination of a month, year, and crime type.

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 aims to include all reported crimes as recorded by the Calgary Police Service. However, it may not account for unreported crimes, which could introduce a reporting bias.

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 the following raw data fields: crime category, crime count, year, month, and a timestamp. These are sufficient for basic trend analysis and forecasting.

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

    - The primary variable of interest is the crime count, which serves as the response variable in most 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.*

    - The dataset does not include demographic information or contextual details such as socioeconomic factors, making it challenging to understand broader systemic influences on crime rates.

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 instances, such as trends across time or spatial groupings, are not directly represented but can be derived through analysis.

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.*

    - While no official splits are provided, a common approach is to use early years for training predictive models and recent years for validation or testing.

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

    - The dataset may contain reporting errors, such as inconsistent categorization of crimes. Additionally, some crimes might be double-counted due to human or system errors during data entry.

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 and does not rely on external resources for its core data. However, external datasets (e.g., socioeconomic or geospatial data) may be required for advanced analyses.

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 dataset does not contain personally identifiable or confidential information. It aggregates crime counts at the city level.

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.*

    - Some users may find crime-related data distressing, especially if they are directly affected by the crimes reflected in the dataset.

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.*

    - No explicit sub-populations are identified. The data is aggregated and does not include individual-level details.

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 dataset is fully anonymized and aggregated to prevent the identification of individuals.

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.*

    - While crime data can be sensitive, this dataset avoids including personal or private information, focusing only on aggregate statistics.

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.*

    - Each instance in the dataset represents reported crimes recorded by the Calgary Police Service (CPS). These reports are based on raw observational data, as police officers and administrative staff log crimes through official reporting systems. The data reflects the direct reporting of incidents, verified through CPS's internal processes to ensure accuracy before being released to the public.

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?*

    - The data was collected and processed using digital reporting systems employed by the Calgary Police Service. Officers log incidents into the CPS database, which categorizes them based on the type of crime and aggregates counts monthly. These systems undergo regular validation by CPS to ensure consistency in reporting and categorization. Open Calgary then publishes this data after further processing to remove any sensitive or identifiable 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)?*

    - This dataset is not a sample; it aims to include all reported crimes in Calgary. However, it does not account for unreported crimes, which means the dataset represents

a subset of total crime incidents. As such, it is comprehensive for reported incidents but not necessarily representative of the true prevalence of crime in the city.

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)?*

    - The dataset relies on the efforts of Calgary Police Service officers and administrative staff. Their compensation comes through standard city employment contracts. Open Calgary staff curate the data for public use, ensuring it meets quality and accessibility standards.

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 includes monthly crime reports from 2018 to the present. Each instance is recorded close to the time of the incident, ensuring timeliness. Open Calgary updates the dataset on a regular basis, typically within a month of receiving the raw data from CPS.

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.*

    - No formal institutional review board (IRB) review was conducted, as the dataset only includes aggregated, anonymized data. However, Open Calgary follows standard practices for protecting privacy and ensuring compliance with ethical data-sharing guidelines.

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 data was collected by the Calgary Police Service directly from reported incidents. Open Calgary acts as a third-party intermediary, curating and publishing the dataset on its open data platform.

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.*

    - Since the data reflects incidents reported to the police, individuals filing reports are inherently aware that their information is being collected. However, the dataset itself is fully anonymized and does not retain identifiable personal details.

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.*

    - By reporting crimes to the Calgary Police Service, individuals provide implied consent for their information to be used in law enforcement and administrative processes. No additional consent is required for inclusion in this anonymized and aggregated dataset.

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).*

    - As the data is anonymized, individuals cannot revoke consent for its use once submitted. However, specific information associated with personal cases can be updated or removed if inaccuracies are reported to the Calgary Police Service.

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 formal data protection impact analysis has been conducted. However, the dataset is designed to minimize risks to individuals by only including aggregate, anonymized statistics. This reduces the likelihood of any negative impact on the data subjects.

12. *Any other comments?*

    - The collection process for this dataset adheres to standard practices for public safety and transparency while prioritizing the privacy and confidentiality of individuals. The collaboration between the Calgary Police Service and Open Calgary ensures that the data is accurate, accessible, and responsibly shared.

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.*

    - Yes, several preprocessing and cleaning steps were performed before the data was made publicly available on Open Calgary. These steps include:

Aggregation: Crime counts are aggregated monthly by crime type to ensure consistency and comparability across time. Anonymization: Any personally identifiable information (e.g., addresses, names) is removed to protect individual privacy. Standardization: Crime categories

are standardized to ensure uniform classification across different data collection points. Error Handling: Duplicates and obvious data entry errors (e.g., negative crime counts) are flagged and removed. 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 collected by the Calgary Police Service is maintained internally but is not publicly available. The processed dataset hosted by Open Calgary is a cleaned and aggregated version intended for public use. Access to the raw data may require special permissions and compliance with CPS's privacy policies. 3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.* - The preprocessing of the data is managed internally by the Calgary Police Service and Open Calgary. While the exact tools used are not disclosed, similar tasks are typically handled using database management systems (e.g., SQL) and data cleaning tools. The publicly available dataset does not include software tools or scripts, but Open Calgary provides documentation for accessing and interpreting the data. 4. *Any other comments?* - The preprocessing steps focus on maintaining the usability and integrity of the dataset while prioritizing privacy and transparency. However, these steps may introduce limitations, such as the loss of granular details or contextual information, which could be valuable for more nuanced analyses.

Uses

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

    - Yes, the dataset has been used for various purposes, including:

Crime Trend Analysis: Researchers and policymakers have used the data to study historical crime trends in Calgary and assess the effectiveness of crime prevention measures. Resource Allocation: Law enforcement agencies have utilized the data to allocate resources efficiently based on observed crime patterns. Public Transparency: The dataset has been a key resource for increasing public awareness and understanding of crime dynamics in Calgary. 2. *What (other) tasks could the dataset be used for?* - The dataset has potential applications in a variety of tasks, including:

Predictive Modeling: Building machine learning models to forecast future crime trends. Urban Planning: Informing urban development strategies by identifying high-crime areas. Sociological Studies: Examining correlations between crime rates and socioeconomic or demographic variables. Community Safety Programs: Designing targeted safety initiatives based on seasonal or geographic crime patterns. 3. *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?* - Yes, the following aspects could impact future uses:

Aggregated Nature of Data: Since the data is aggregated by month and crime type, detailed analyses (e.g., hourly or neighborhood-level trends) are not possible. Underreporting: The dataset includes only reported crimes, which may lead to underestimation of actual crime rates. This could bias analyses or decision-making if not accounted for. Lack of Contextual Variables: The absence of socioeconomic or demographic data limits the ability to explore systemic factors influencing crime. 4. *Are there tasks for which the dataset should not be used? If so, please provide a description.* - Yes, the dataset should not be used for:

Identifying Individuals: The data is anonymized and aggregated, making it inappropriate for any task attempting to identify individuals or specific incidents. Definitive Cause-Effect Analysis: Without additional variables, the dataset cannot definitively establish causal relationships between crime rates and external factors. 5. *Any other comments?* - The dataset is a valuable resource for understanding crime trends, but users should approach its analysis thoughtfully, considering its limitations and potential biases. Open Calgary's ongoing updates and transparency efforts enhance its utility, but future improvements, such as geospatial details or additional contextual variables, could significantly expand its application scope.

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 freely available to the public through Open Calgary, which serves as the City of Calgary's open data platform. It is intended for use by researchers, policymakers, journalists, and the general public.

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 is distributed through the Open Calgary website, where users can download it in multiple formats, such as CSV or Excel. Additionally, an API is provided for programmatic access to the data. As of now, the dataset does not have a digital object identifier (DOI).

3. *When will the dataset be distributed?*

    - The dataset has been publicly available on the Open Calgary platform since its creation. It is updated monthly as new crime data is collected and processed by the Calgary Police Service.

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.*

- Yes, the dataset is distributed under the Open Government License - Canada. This license allows users to copy, modify, and use the data for any purpose, provided that proper attribution is given to the City of Calgary and the Calgary Police Service. No fees or restrictions are associated with the use of this dataset under this 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 third parties have imposed restrictions on this dataset. It is fully owned and managed by the City of Calgary and is made freely available to the public under the Open Government License.

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, there are no export controls or regulatory restrictions that apply to this dataset. It is freely accessible worldwide.

7. *Any other comments?*

    - The distribution model for this dataset ensures transparency and accessibility. The provision of both downloadable files and API access allows for flexible usage across different user types, from individual researchers to large-scale data applications.

Maintenance

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

    - The dataset is hosted and maintained by Open Calgary, with data updates provided by the Calgary Police Service.

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

    - For inquiries about the dataset, users can contact Open Calgary through their contact page or reach out to the Calgary Police Service for data-specific queries.

3. *Is there an erratum? If so, please provide a link or other access point.*

    - Currently, there is no centralized erratum. Errors, if identified, are typically corrected in subsequent updates to the dataset.

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)?*

- Yes, the dataset is updated monthly with new crime reports. Updates are managed by Open Calgary in coordination with the Calgary Police Service, and changes are communicated through the Open Calgary platform.

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 include personally identifiable information and is anonymized at the aggregation level. There are no specific retention limits, as the data represents aggregate statistics rather than individual records.

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.*

    - Older versions of the dataset are not explicitly maintained but remain accessible through the Open Calgary platform. Users can retrieve historical records directly from the dataset's archive.

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.*

    - There is no official mechanism for public contributions to the dataset. However, users are encouraged to share derivative works or insights publicly, provided they adhere to the Open Government License's attribution requirements. Any feedback or corrections can be submitted to Open Calgary for review.

8. *Any other comments?*

    - The maintenance and distribution strategy for this dataset reflects a commitment to open data principles. Continued collaboration between the Calgary Police Service and Open Calgary ensures its relevance and reliability for a wide range of applications.

References

Calgary, Open. n.d. *Open Calgary*. https://data.calgary.ca/.

Service, Calgary Police. n.d. *The City of Calgary - Home Page*. https://www.calgary.ca/cps.html.