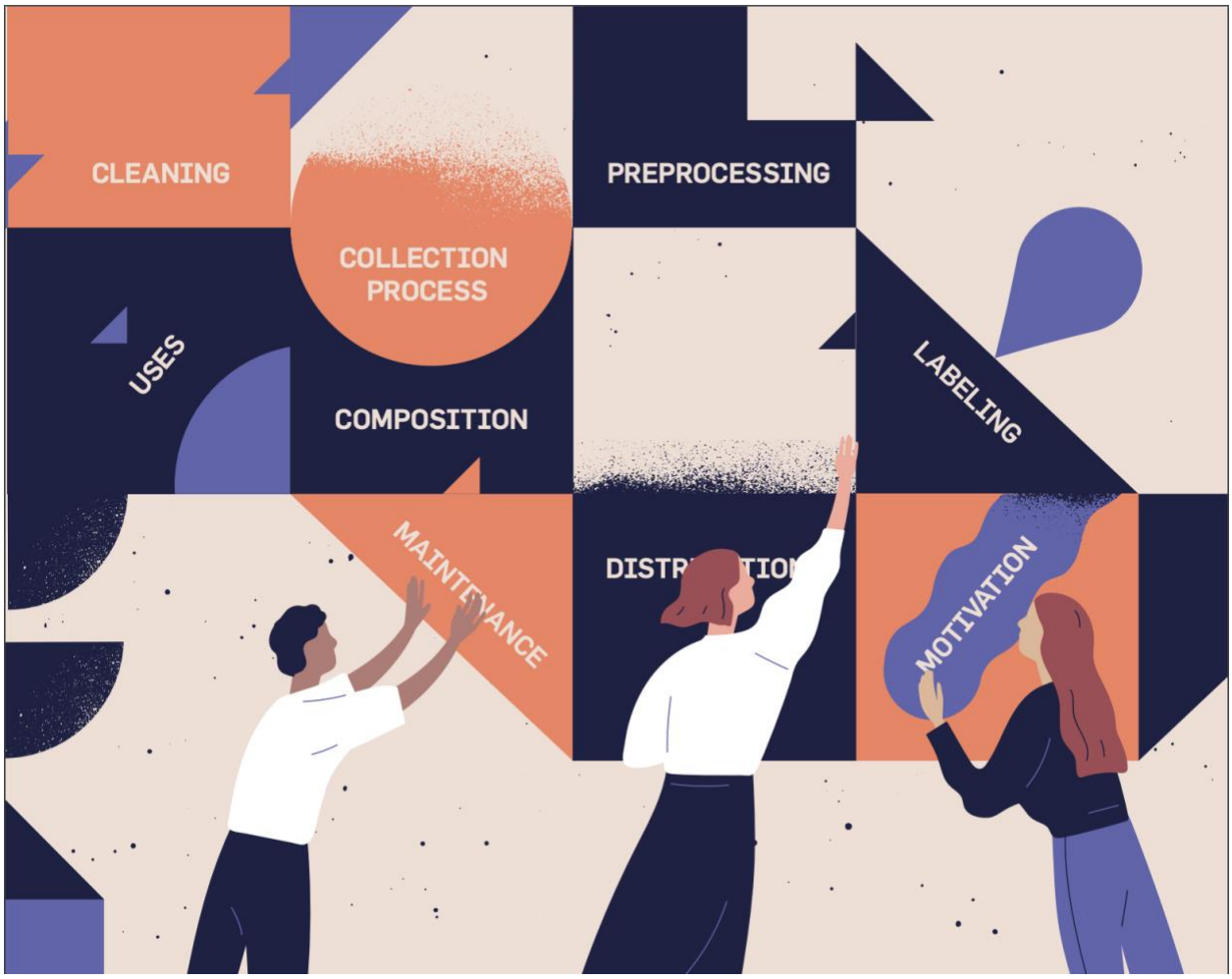
Datasheet for Datasets

Original Dataset from the Official Website of Norman Oklahoma Police Department: Norman Police Department

GitHub Repository for the Augmented Dataset: Augmented Dataset



In creating my dataset, I began with the primary incident data provided by the Norman Police Department. This data consisted of several key attributes for each incident: the time and date it occurred, a unique incident number, the specific location, the nature of the incident, and an originating identifier (incident ORI).

To enrich this data and make it more insightful for analysis, I embarked on a series of data augmentation and preprocessing steps. Here's a rundown of what I accomplished:

1. Day and Time Extraction: I used Python's 'datetime' library to parse the 'incident_time' field, extracting the day of the week and the hour of the day when each incident took place. This would allow for temporal analysis of incidents.

2. Location Ranking: I employed a frequency analysis on the 'incident_location' column to rank locations based on the number of incidents. The rankings were assigned such that locations with the same number of incidents received the same rank, and the ranking sequence was maintained for locations with unique incident frequencies.

3. Nature Ranking: Similarly, I ranked the nature of the incidents using a frequency count, allowing me to identify the most common types of incidents reported.

4. Determining Side of Town: Utilizing the Google Maps API, I geocoded each incident location to ascertain the side of town it occurred on. This was determined relative to a fixed central coordinate in Norman, Oklahoma, and the resulting cardinal and intercardinal directions (e.g., NE for northeast) were added to my dataset.

5. Weather Code Retrieval: To understand the weather conditions at the time of each incident, I called upon the Open-Meteo API. By providing it with the precise time and geocoded locations, I obtained historical weather codes that indicated the weather status during each incident.

6. EMSSTAT Flagging: I carefully scanned through the data to flag incidents involving an EMSSTAT response. This was determined by the incident ORI or by assessing adjacent records for any mentions of EMSSTAT at the same time and location.

The final dataframe I constructed is a rich tapestry of information, interweaving the original incident data with time-based insights, location intelligence, weather conditions, and emergency service involvement. To achieve this, I leaned heavily on the power of Python, especially pandas for dataframe manipulation, and the datetime library for parsing time information.

Moreover, I ensured compliance with the terms of service of the APIs used, recognizing that the enriched data remains subject to their respective licensing agreements. My meticulous documentation and transparent augmentation process are designed to enable reproducibility and to foster ethical utilization of the data. This enhanced dataset stands as a robust foundation for various analytical tasks, including identifying patterns and trends, determining hotspots, and potentially predicting future incidents.

Questions

Motivation

- **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.
  ⇨ To provide insights into public safety and police activities in Norman, OK.
- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**
  ⇨ Developed by analysts or researchers associated with the Norman Police Department, and the augmented dataset was created by me by using a bunch of python based libraries and APIs.
- **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.
  ⇨ Likely funded through city or police department budgets.

Composition

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
  ⇨ The instances in the dataset represent individual reports of incidents recorded by the Norman Police Department. Each instance includes details of a specific event, such as the time and location of the incident, the nature of the event, weather conditions of the location, and an incident number. There is a single type of instance the reported incidents though within these, there are multiple attributes detailing the specifics of each incident, which could include interactions between people or actions taken by the police department.
- **How many instances are there in total (of each type, if appropriate)?**
  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 (e.g., 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 (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

  ⇨ In the augmented dataset, a total of 8 instances are present, which include: day_of_week, time_of_day, weather_code, location_rank, side_of_town, incident_rank, nature, and EMSSTAT. This dataset incorporates additional information beyond what was originally provided by the police department.

- **What data does each instance consist of?** "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

  ⇨ Each instance in the dataset consists of structured data that likely originated from raw data sources such as incident reports or call logs. The data for each instance includes: - 'Date / Time': The timestamp of the incident. - 'Incident Number': A unique identifier for the incident. - 'Location': The specific or approximate location where the incident occurred. - 'Nature': A description of the incident, indicating the type or category of the event. - 'Incident ORI': An originating agency identifier, which could be used to reference the reporting agency or department. This structured data represents processed features ready for analysis, rather than raw unstructured data.

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

  ⇨ There isn't a traditional label or target in the sense used for supervised machine learning. However, the 'Nature' field in each instance could serve as a categorical label describing the type of incident reported, such as 'Traffic Stop' or 'MVA Non Injury'. This categorization could potentially be used as a target for analyses or predictive modeling focused on incident types. Additionally, if any form of outcome or resolution is recorded for each incident, that could also serve as a label or target for analysis.

- **Is any information missing from individual instances?** If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.
  ⇨ The dataset provided may have missing information in individual instances if certain details were not reported, not observed, or deemed not applicable at the time of the incident. For instance, specific location details might be generalized or missing if the incident occurred in a non-specific area. Also, certain fields could be incomplete due to reporting errors or privacy concerns that necessitate redaction. Any missing information would likely be a result of these factors or similar operational constraints during data collection.

- **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links**)? If so, please describe how these relationships are made explicit.
  ⇨ In the dataset, relationships between individual instances are not explicitly defined as in datasets designed for network analysis or collaborative filtering. However, implicit relationships might be inferred from the 'Location' or 'Time' fields if multiple incidents occur at the same place or time. These relationships could be indicative of patterns or correlations in the data but would require further analysis to make explicit.

- **Are there recommended data splits (e.g., training, development/validation, testing)?** If so, please provide a description of these splits, explaining the rationale behind them.

⇨ No, this dataset was not used for training any model, and it shouldn't directly be used for as it can lead to producing some biased results.

- **Are there any errors, sources of noise, or redundancies in the dataset?** If so, please provide a description.
  ⇨ No
- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., 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 (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? 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 depends on the supplementary analysis or enrichment, such as geolocation for the 'side_of_town' feature, may rely on external resources like Google Maps API. For such external resources: a) There are no guarantees that they will exist or remain constant over time; API services may change or become deprecated. b) There might not be official archival versions of the dataset including external resources unless specifically maintained by the data curator. c) Use of external resources like the Google Maps API comes with licensing terms and potential costs that future users must adhere to. Users of the dataset should be prepared to update their analysis if external resources change and must comply with any licensing requirements those resources impose.

- **Does the dataset contain data that might be considered confidential (e.g., 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 dataset likely does not contain data that would be considered confidential or protected by legal privilege, as it is a collection of police incident reports which are generally a matter of public record.

- **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 comprises police incident reports, which just fepic the nature of the crime and other columns in the dataset are repalted to time, location,weather code, etc, which are non offensive.

- **Does the dataset relate to people?** If not, you may skip the remaining questions in this section.
  ⇨ The dataset relates to nature of crime of the people and location of the crime.

- **Does the dataset identify any subpopulations (e.g., 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 explicitly identify subpopulations by characteristics such as age or gender.
- **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?** If so, please describe how.
  ⇨ No

Collection Process

- **How was the data associated with each instance acquired?** Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
  ⇨ The data was obtained through a process of indirect inference and derivation, facilitated by Python libraries such as requests, pypdf, and APIs including Google Maps and OpenMeteo. Specifically, location details were sourced from the Google Maps API, while historical weather conditions were accessed via the OpenMeteo API. Rigorous validation was conducted utilizing the built-in mechanisms of these APIs, guaranteeing the accuracy and reliability of the information gathered for each data instance.

- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?** How were these mechanisms or procedures validated?
  ⇨ The dataset likely stems from the meticulous documentation efforts of law enforcement personnel, systematically logging details during and post-incidents. Each entry was likely inputted into a police database via a standardized form, capturing crucial information such as time, location, nature, and incident number. My generated dataframe enhanced the original dataset by harnessing APIs to imbue the data with enriched context. Leveraging the Google Cloud API, we accessed geocoding services to ascertain the 'side_of_town' for each incident, translating street addresses into geographical coordinates to delineate their relative location within the town. Furthermore, the Open-Meteo API played a pivotal role in gathering historical weather conditions, appending a 'weather_code' to each instance based on the incident's date, time, and location, thereby shedding light on the environmental backdrop of each event.

- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**
  - ⇨ The sampling strategy for the initial dataset appears to be deterministic, based on the systematic collection of official reports from the Norman Police Department. The subsequent data augmentation, involving weather conditions and town side determinations via API calls, was applied to the entire sampled dataset, not influencing the initial sampling method but enriching the collected data deterministically.

- **Were any ethical review processes conducted (e.g., 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.
  - ⇨ As the dataset comprises police incident reports, which are generally public records, an ethical review process was not required for its creation.

- **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**
  - ⇨ The data was obtained from the Norman Police Department's publicly available records, which are likely to have been collected directly by law enforcement officers during their duties. Additionally, some data was enriched using third-party services like Google Maps API for geographical information and the Open-Meteo API for historical weather data.

- **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 dataset consists of police incident reports, which are typically considered public records and so it did not require the explicit consent of the individuals involved for their release.

Preprocessing/cleaning/labeling

- **Was any preprocessing/cleaning/labeling of the data done (e.g., 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 remainder of the questions in this section.
  - ⇨ The dataset underwent several preprocessing steps to enhance its utility for analysis. This included extracting the day of the week and hour of the day from the 'incident_time' data, ranking incident locations and nature types based on frequency, and determining the side of town using Google Maps API for geocoding. Additionally, weather conditions were assigned using Open-Meteo API, and a boolean 'EMSSTAT' was derived from the presence of EMSSTAT in the 'incident_ori' or nearby records. This preprocessing served to structure the raw data into a more informative and accessible format.
- **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** If so, please provide a link or other access point to the "raw" data.
  - ⇨ The raw data was preserved alongside the processed dataset to facilitate verification and future analyses. It provides a foundational reference for the applied preprocessing steps and ensures data integrity for subsequent use. Access to this raw data typically requires direct contact with the dataset custodians or an official request through proper channels provided by the original data source.
- **Is the software used to preprocess/clean/label the instances available?** If so, please provide a link or other access point.
  - ⇨ **No, I haven't used any software here, instead I have used Python based libraries and functions for data labelling and cleaning.**

Uses

- **Has the dataset been used for any tasks already?** If so, please provide a description.
  ⇨ The dataset has been utilized to conduct a range of analyses, such as identifying patterns in incident timings and locations, assessing nature frequencies, and correlating incidents with specific days of the week. It has also supported the extraction of weather conditions and the determination of emergency statuses related to incidents, providing a comprehensive overview of public safety dynamics within the region.

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

⇨ Beyond its current applications, the dataset could be leveraged for a variety of other tasks such as:

1. Urban Planning: By analyzing the frequency and nature of incidents, urban planners could identify areas needing improved infrastructure or increased safety measures.
2. Machine Learning: Data scientists could use the dataset to train models to predict future incidents based on patterns observed, aiding in proactive policing and community safety measures.
3. Sociological Research: Researchers could study the correlation between incident types and various socio-economic indicators to understand underlying causes of crime.
4. Resource Allocation: Emergency services could optimize their resource distribution based on the analysis of incident times and locations to ensure quicker response times.
5. Public Awareness Programs: The dataset could inform the creation of public awareness campaigns by highlighting prevalent incident types and areas.

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 future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

⇨ The composition and collection process of the dataset could indeed affect its future uses. If the data skews towards particular locations or nature of incidents, it could potentially bias analyses and lead to unfair resource allocation or stigmatization of certain areas. Future users should ensure they are aware of such biases and take steps to mitigate them, such as:

1. Applying stratified sampling to ensure a more representative distribution of data.
2. Augmenting the dataset with additional data to balance out any over- or under-represented categories.
3. Conducting thorough exploratory data analyses to identify and understand any inherent biases before using the data for predictive modeling or decision-making processes.
4. Being transparent about the dataset's limitations and potential biases when communicating findings to stakeholders or the public.

- **Are there tasks for which the dataset should not be used?** If so, please provide a description.
  ⇨ The dataset should not be used for tasks that require personally identifiable information (PII) as it may have been removed or anonymized to protect individuals' privacy. It also should not be used for predictive policing, which could lead to biased enforcement actions or reinforce systemic inequalities. Additionally, the dataset should not be employed for any form of automated decision-making without thorough scrutiny for biases that could impact certain groups unfairly. Users must also ensure they comply with legal and ethical guidelines, especially when dealing with sensitive data related to criminal incidents.

Distribution

- **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** If so, please provide a description.
  ⇨ There is no agreement as such so as of now this dataset will not be distributed

- **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?** Does the dataset have a digital object identifier (DOI)?
  ⇨ The dataset will be available under a private repository on GitHub and it can be accessed after requesting the owner.

- **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 distribution of the dataset will be subject to a specific copyright or intellectual property license, which outlines how the data can be used, shared, and modified. The terms of use may also include privacy considerations, especially for data involving individuals. These terms will be provided alongside the dataset, often in a file named LICENSE or TERMS OF USE. It's crucial to review these terms to understand any restrictions, such as the prohibition of commercial use or requirements to credit the source. Some datasets may also require a fee for access, especially if they contain proprietary or commercially valuable information. To fully understand the scope of permissions and any associated costs, future users should consult the official licensing terms provided with the dataset or via the issuing authority's website.

Maintenance

- **Who will be supporting/hosting/maintaining the dataset?**
  ⇨ The dataset is managed and curated by the Norman Police Department, ensuring regular publication of diverse reports and updates. The augmented data I generate will be hosted on my GitHub repository, with ongoing support from me. Availability will be contingent on the continued functionality and adherence to policy of the openmaeto and Google Maps APIs.

- **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**
  ⇨ The owner, curator, or manager of the dataset can be reached through the official communication channels provided by the Norman Police Department. For inquiries related to the augmented data, please contact the designated data owner via email at harshakparmar12@gmail.com.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

⇨ The dataset should be regularly updated since it is sourced from the Norman Police Department daily. As incident summaries are constantly evolving, the content of the dataset may vary with each update.

- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were 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.
  - ⇨ If the dataset relates to individuals and includes retention limits on their data, it is typically in compliance with privacy policies and legal frameworks. For the dataset in question, individuals may have been informed that their data would be retained for a specific period, after which it would be deleted or anonymized. Enforcement of these limits would be the responsibility of the data curating entity, often through automated systems that flag data for review or deletion after the retention period lapses. Users must refer to the dataset's privacy policy or contact the data curator for precise details on retention practices.

- **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 users.
  - ⇨ Yes, older versions of the dataset will continue to be supported, hosted, and maintained to ensure that researchers and users who have developed applications or conducted studies based on these versions can continue their work without disruption. They will be archived with clear versioning indicated and can be accessed through a repository where each version is timestamped and provided with a changelog. Users will be notified of new versions and the status of older ones through regular communication channels such as update logs on the dataset's website, mailing lists, or official announcements on platforms where the dataset is hosted. The goal is to maintain a transparent record of the dataset’s evolution over time.

- **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 other users? If so, please provide a description.
  - ⇨ If others wish to contribute to my dataset, they can initiate a collaboration request through GitHub (GitHub Username: hpamdeoxys).