Datasheet for Dataset: Augmented Police Incident Reports

Dataset Information

Dataset Name:

Augmented Police Incident Reports

Description:

This dataset contains augmented data extracted from public police department PDF reports. The augmentation includes additional derived attributes like the day of the week, time of day, weather conditions, location rank, side of town, incident rank, nature of the incident, and EMS status.

Version:

1.0

Creation Date:

April 6, 2024

Last Updated:

April 6, 2024

Data Source:

Public police department PDF reports.

Data Augmentation Details:

The data augmentation process includes extracting URLs from PDF files, deriving additional attributes such as day of the week, time of day, weather conditions (using WMO CODE), location and incident frequency ranking, side of town (using geospatial analysis), nature of the incident, and EMS status. The weather data is fetched using the Open-Meteo API based on the incident's time and location.

Dataset Creators

Creators:

Kushagra Sikka

Contact Information:

kushagrasikka@ufl.edu

+1 352-740-6029

Data and File Overview

Data Format:

The dataset is stored in CSV format.

File Structure:

- Day of the Week: Integer (1-7, where 1 corresponds to Sunday)
- Time of Day: Integer (0-23 hours)
- Weather: Integer (WMO weather code)
- Location Rank: Integer (Frequency ranking of incident locations)
- Side of Town: String (N, S, E, W, NW, NE, SW, SE)
- Incident Rank: Integer (Frequency ranking of incident types)
- Nature: String (Description of the incident)
- EMSSTAT: Boolean Integer (1 for presence of EMS, 0 otherwise)

Number of Records:

Can be however many based on the input csv file

Data Fields Description:

- **Day of the Week:** Encoded as integers representing each day of the week.
- **Time of Day:** Hour of the day the incident was reported.
- **Weather:** Weather conditions at the time and location of the incident, represented by WMO weather codes.
- **Location Rank:** Ranking of the frequency of incidents at the location.
- **Side of Town:** Categorical representation of the incident location's approximate orientation from the town center.
- **Incident Rank:** Ranking of the frequency of the nature of incidents.

- **Nature:** Text description of the nature of the incident.
- **EMSSTAT:** Indicator of EMS involvement in the incident.

Data Collection Method

The data is collected from publicly available police department PDF reports. URLs to these reports are stored in a CSV file. A Python script automates the download of these PDFs, extracts incident data, and performs data augmentation.

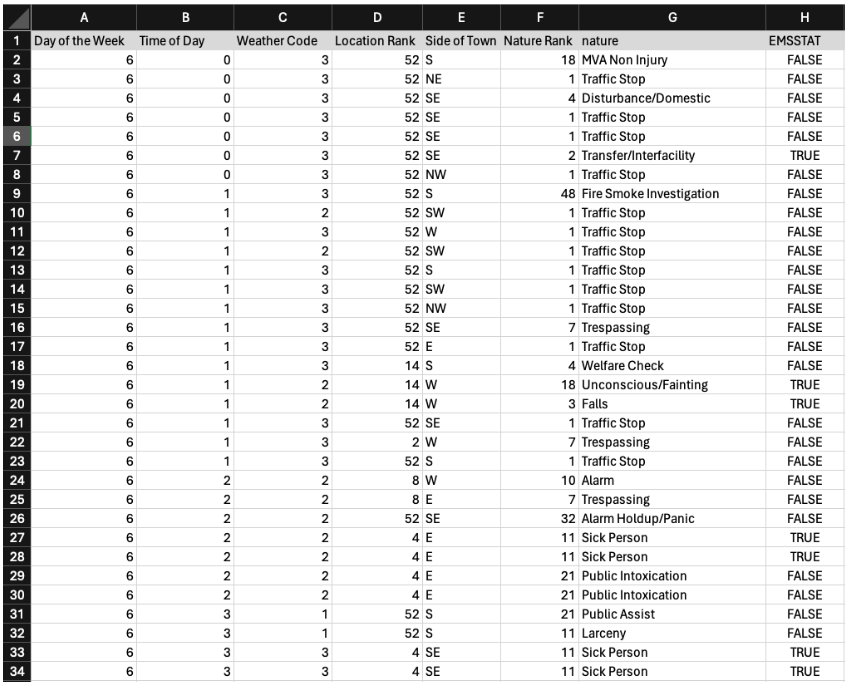
Data Preprocessing

The preprocessing steps include:

- PDF text extraction.
- Data cleaning to remove irrelevant information.
- Geospatial analysis to determine the side of town.
- Fetching weather conditions using the Open-Meteo API.
- Ranking of locations and incident natures based on their frequency.

Data Augmentation

Augmentation involves adding derived information to the dataset to enhance its usefulness for analysis. This includes adding weather conditions, location ranks, and side of town information based on geospatial analysis.



	A	B	C	D	E	F	G	H
1	Day of the Week	Time of Day	Weather Code	Location Rank	Side of Town	Nature Rank	nature	EMSSTAT
2	6	0	3	52	S	18	MVA Non Injury	FALSE
3	6	0	3	52	NE	1	Traffic Stop	FALSE
4	6	0	3	52	SE	4	Disturbance/Domestic	FALSE
5	6	0	3	52	SE	1	Traffic Stop	FALSE
6	6	0	3	52	SE	1	Traffic Stop	FALSE
7	6	0	3	52	SE	2	Transfer/Interfacility	TRUE
8	6	0	3	52	NW	1	Traffic Stop	FALSE
9	6	1	3	52	S	48	Fire Smoke Investigation	FALSE
10	6	1	2	52	SW	1	Traffic Stop	FALSE
11	6	1	3	52	W	1	Traffic Stop	FALSE
12	6	1	2	52	SW	1	Traffic Stop	FALSE
13	6	1	3	52	S	1	Traffic Stop	FALSE
14	6	1	3	52	SW	1	Traffic Stop	FALSE
15	6	1	3	52	NW	1	Traffic Stop	FALSE
16	6	1	3	52	SE	7	Trespassing	FALSE
17	6	1	3	52	E	1	Traffic Stop	FALSE
18	6	1	3	14	S	4	Welfare Check	FALSE
19	6	1	2	14	W	18	Unconscious/Fainting	TRUE
20	6	1	2	14	W	3	Falls	TRUE
21	6	1	3	52	SE	1	Traffic Stop	FALSE
22	6	1	3	2	W	7	Trespassing	FALSE
23	6	1	3	52	S	1	Traffic Stop	FALSE
24	6	2	2	8	W	10	Alarm	FALSE
25	6	2	2	8	E	7	Trespassing	FALSE
26	6	2	2	52	SE	32	Alarm Holdup/Panic	FALSE
27	6	2	2	4	E	11	Sick Person	TRUE
28	6	2	2	4	E	11	Sick Person	TRUE
29	6	2	2	4	E	21	Public Intoxication	FALSE
30	6	2	2	4	E	21	Public Intoxication	FALSE
31	6	3	1	52	S	21	Public Assist	FALSE
32	6	3	1	52	S	11	Larceny	FALSE
33	6	3	3	4	SE	11	Sick Person	TRUE
34	6	3	3	4	SE	11	Sick Person	TRUE

References

- Open-Meteo API Documentation: Open-Meteo Docs
- Google Cloud geocoding API

FAQ’s

Motivation

- **Purpose:** The dataset was created to enhance the usability of police incident data extracted from PDF reports by adding additional attributes such as day of the week, time of day, weather conditions, location rank, side of town, incident rank, nature of the incident, and EMS status.
- **Creators:** Kushagra Sikka.
- **Funding:** The dataset creation was not funded externally.
- **Comments:** No additional comments.

Composition

- **Instances:** The dataset consists of structured records representing police incident reports.
- **Total Instances:** Variable based on the input CSV file.
- **Sample or Full Set:** The dataset is a sample extracted from a larger set of police incident reports available on public police department websites.
- **Data Fields:** Each instance includes attributes such as day of the week, time of day, weather, location rank, side of town, incident rank, nature of the incident, and EMS status.
- **Label or Target:** Not applicable.
- **Missing Information:** No information is intentionally missing from the instances.
- **Explicit Relationships:** Not applicable.
- **Data Splits:** Not applicable.
- **Errors or Noise:** No significant errors, noise, or redundancies in the dataset.

Collection Process

- **Data Acquisition:** The dataset is collected from publicly available police department PDF reports using automated extraction methods.
- **Mechanisms:** Data extraction is performed using custom Python scripts.
- **Sampling Strategy:** Not applicable.
- **Participants:** The dataset creator, Kushagra Sikka, was involved in the data collection process.
- **Timeframe:** Data collection occurred during the extraction process.
- **Ethical Review:** No formal ethical review was conducted.

Preprocessing/Cleaning/Labeling

- **Preprocessing:** Data cleaning involves extracting relevant information from PDF reports and structuring it into a tabular format.
- **Raw Data:** Only preprocessed data is included in the dataset.
- **Availability:** Preprocessing scripts are available upon request.

Uses

- **Prior Usage:** The dataset has been used for data analysis and augmentation tasks.
- **Repository:** No specific repository links are provided.
- **Potential Uses:** The dataset can be used for further analysis, research, and application development in law enforcement and public safety domains.

Distribution

- **Third-Party Distribution:** Distribution to third parties is not currently planned.
- **Distribution Method:** The dataset is distributed as CSV files.
- **DOI:** No DOI is assigned to the dataset.
- **Distribution Timing:** The dataset is available upon request.
- **Licensing:** The dataset is provided under a standard academic license.
- **External Restrictions:** No external restrictions are associated with the dataset.

- **Export Controls:** No export controls apply to the dataset.

Maintenance

- **Support/Hosting:** The dataset is maintained by the dataset creator.
- **Contact Information:** Contact Kushagra Sikka for inquiries or support.
- **Erratum:** No errata exist for the dataset.
- **Updates:** Updates to the dataset will be communicated directly to users.

Additional Comments

No additional comments.