# Analyzing Drug-Related Deaths in Connecticut from 2012-2023: A Comprehensive Data Exploration

# By: Holly Akati

# PUBH 1242

**Introduction**:

Accidental drug deaths have become a growing public health concern in recent times. The need for effective prevention and control measures requires comprehensive data analysis and insights. This project looks at a dataset that provides information on accidental drug-related deaths in Connecticut. It is designed to support public health research, policymaking, and prevention strategies by documenting detailed information about these fatalities. The data analysis was done using Python.

**About the dataset:**

The dataset contains records of accidental drug-related deaths in Connecticut from **2012 to 2023**. These cases include demographic details (such as age, gender, and race), the drugs involved, and the circumstances of each incident.

Data is available at the following link:

[https://data.ct.gov/Health-and-Human-Services/Accidental-Drug-Related-Deaths-2012-June](https://data.ct.gov/Health-and-Human-Services/Accidental-Drug-Related-Deaths-2012-June-2017/rybz-nyjw)-[2017/rybz-nyjw](https://data.ct.gov/Health-and-Human-Services/Accidental-Drug-Related-Deaths-2012-June-2017/rybz-nyjw)

 Connecticut Office of the Chief Medical Examiner

 Dataset owner: Tyler Kleykampii

**Objective:**

This project has two objectives:

1. To uncover trends and patterns to better understand the scale of the issue and inform possible interventions.
2. To apply NLP/ML methods to understand the unstructured free text information, and perhaps predict causes of death using the rest of the structured data.

**Data Exploration:**

The data analysis focuses on these specific variables:

* **Date of Death:** The specific date when the incident occurred.
* **Age, Gender, Race:** Demographic attributes of the deceased.
* **Drugs Mentioned:** The substances detected in the toxicology report.
* **Death Location:** Geographic details of where the death occurred (include exact coordinates like longitude and latitude of the location)
* **Description of Injury:** Describes how drugs were involved

**Demographic Data Analysis:**

The dataset contains a population of 11,981 deaths and contains 5 rows x 48 columns. Basic data analysis and visualization methods were used for the demographic analysis using the pandas, numpy and matplotlib libraries on python.

A graph with purple lines

Description automatically generated

**Figure 1:** This graph shows the age distribution, with a range of 13 to 87 years old and a median age of 44 years old, which is a higher range than the national average (25 to 34 years old) according to the National Center for Drug Abuse Statistics[[1]](#footnote-1).

A graph of a number of people

Description automatically generatedA graph with orange rectangular bars

Description automatically generated

**Figure 2**: Most of the deceased were male and most of them were white.

**Drug-Specific Data Analysis:**

This section analyzes the data, specifically in relation to the drugs mentioned and the death location (city, exact geographic location)

**A graph of different colored lines

Description automatically generated  
Figure 3:** This graph shows the prevalence of different drugs (e.g., fentanyl, heroin, or prescription opioids) over the years. There is a significant spike of Oxycodone in 2015 that falls right after.

**Geographic Distribution:**

A graph of a number of deaths

Description automatically generated

**Figure 4**: Shows the cities with the most deaths in Connecticut. The cities with the densest population are unsurprisingly on this list

***Advanced Techniques***

The dataset included exact geographical locations for each death. This data proved itself useful for some clustering analysis.

* Each geographic location was assigned to a cluster based on its proximity to cluster centers.
* kmeans.cluster\_centers\_ provides the central coordinates of each cluster (hotspots).

**Interactive Map**:

* Points are plotted with folium, color-coded by cluster.
* Cluster centers are marked to show the central hotspots.

A map with many colored dots

Description automatically generated

**Figure 5:** Interactive Map

The number of clusters, in this case nine (9), were chosen by using silhouette scores to identify the most optimal number of clusters. 9 had the highest silhouette score of 0.650.

**Data Cleaning for remaining textual data**

Real-world datasets often contain missing or inconsistent values, especially with detailed data describing medical events. To fully explore the data further, data cleaning using natural language processing (NLP) and support vector machines (SVM) were attempted.

The columns of interest were primarily **Description of Injury** and **Location + Other**.

A graph of a number of people

Description automatically generated with medium confidence

**Figure 6:** This graph shows the exploration within the **Location of Death** column. The information is interesting but leaves one to wonder what the “other locations” are.

A graph with numbers and symbols

Description automatically generated

**Figure 7:** When trying to look further into the **Other Locations,** the data shows major imperfections. “Friend’s House” comes up multiple times with different spelling and data cleaning is required. There was an attempt to simply combine similar words using NLP and SVM, specifically by putting phrases and words with similar enough words in the same vector, but this method requires more fine tuning before being deemed successful.

A graph of blue bars

Description automatically generated with medium confidence

**Figure 8:** the same SVM and NLP techniques were used for the **Description of Death** and the results combination worked, proving that these methods could also work with the **Locations** column with more fine tuning.

**Conclusion and Discussion**

So far, this project sets the foundation for potential robust data analysis on drug-related deaths in Connecticut. By systematically exploring the dataset, the analysis can provide critical insights into demographic disparities, geographic patterns, and trends over time. This information can eventually be very useful for policymakers, public health officials, and community organizations aiming to address the opioid crisis and prevent further loss of life.

The data could be explored further if the textual data can be thoroughly cleaned and used to find correlation between drug-overdose related public health initiatives and the number of deaths per year to assess how effective they are. The data could also be used to predict which drugs might reappear based on past trends.

Overall, this project didn’t fully achieve its goal of using NLP and ML to analyze the unstructured data, but it does show the beginning steps being taken to eventually reach that goal.

1. <https://drugabusestatistics.org/drug-overdose-deaths/#:~:text=Men%20aged%2025%20to%2034,%2C%20men%20are%20still%2075.64%25> [↑](#footnote-ref-1)