Analysis on Deaths Caused by Drug Overdose

Jayanth Dasamantharao

2023-08-08

Contents

INTRODUCTION

The problem of drug overdose has become a serious issue in the United States, with a history of over 50 years. To combat this issue, the federal budget allocated 42.5 billion USD in 2023, which is $3.2 billion more than the previous year[1]. In 2022, drug overdoses caused over 79,000 deaths, exceeding the number of deaths caused by motor vehicle accidents, homicides, and suicides. Although the estimated overdose deaths in the first nine months of 2022 declined from the same period in 2021, they were still 50% higher than pre-pandemic levels[2]. Misuse of prescription medication is a major contributing factor to this increase, making it a significant public health concern in Connecticut[3]. As part of our research project, we aim to investigate the issue of drug abuse prevailing in Connecticut.

OBJECTIVE & APPROACH

We aim to uncover the underlying causes and trends behind the high prevalence of drug abuse in Connecticut by analyzing publicly available data on this issue. To achieve this, we have devised a three-pronged strategy, which is outlined below.

Our goal is to gain insight into the patterns and causes of drug abuse in Connecticut by analyzing publicly available data. To achieve this, we have devised a three-pronged approach.

Firstly, we will examine the **demographic attributes** of drug abuse, such as age, gender, race, and area, as we hypothesize that these factors are highly correlated with drug consumption.

Secondly, we will explore the **temporal patterns** of drug abuse, including year-by-year trends, seasonal variations, and preferred times of drug use.

Finally, we will diagnose the **underlying causes** of drug abuse, such as the chemicals consumed and the type of injuries suffered. Our aim is to identify the most common chemicals found in drug abuse cases, the types of drugs most frequently abused, and the categories responsible for 80% of cases.

Data Exploration

Data Source:

The data pertaining to accidental drug-related deaths from 2012 to 2021 has been obtained from the official data repository of the United States Government. The Office of the Chief Medical Examiner conducted an investigation, which included a toxicity report, death certificate, and scene investigation, to derive this data. The source of this data is https://catalog.data.gov/dataset/accidental-drug-related-deaths-2012-2018.

Data Variables and Description:

The dataset we acquired contains various parameters that we have associated with the three-fold approach we outlined earlier, in order to determine which fields can be used for each analysis. To make the variable names distinguishable, we have enclosed them in quotes.

The dataset has three categories of fields: temporal, demographic, and causal. The temporal fields include "Date," which indicates the date when the accidental death occurred or was reported. The demographic fields include "Sex," "Age," and "Race," as well as fields indicating the city, state, and county of residence or death of the drug addict person. We also have fields providing the latitude and longitude of the death location and a field indicating the place of death.

The causal fields include "COD," which denotes the cause of death and lists all the chemicals consumed by the victim, separated by commas. We also have separate fields for each of the 14 chemicals causing death, indicating whether they were present or not. The field "Description.Of.Injury" denotes how the drug abuse took place, while "MannerofDeath" indicates the manner in which death occurred.

Overall, the dataset covers a period of ten years from 2012 to 2022 and contains 48 variables. With 9202 observations, we have sufficient data to conduct our investigation.

```
##
##  Number of rows:  9202
##  Number of columns:  48
##
```

Data Challenges and Resolutions

To understand the root cause and type of injury related to drug abuse, we analyzed the "Description.Of.Injury" column which contained free-form text. To categorize this data, we tokenized the text and calculated the frequency of each token across all drug abuse cases. This helped us determine the most common underlying causes and types of injuries.

The "Date.Type" column in the dataset provided information on when the drug abuse case occurred. However, the column had two values, "dateReported" and "dateOfDeath", which sometimes resulted in gaps between the two dates. To overcome this issue, we aggregated the data for different time ranges, providing an approximate value for the time of the event.

Data Cleaning:

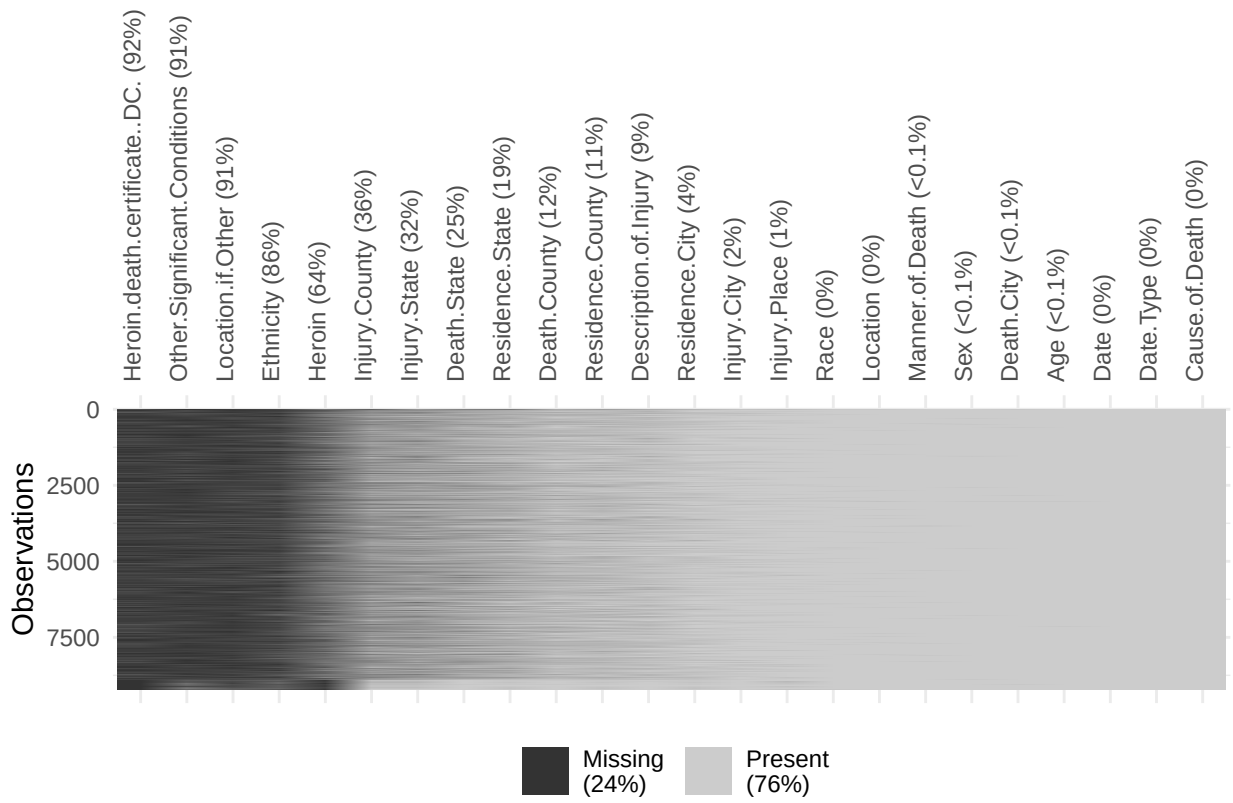
During our analysis, we encountered missing values in the dataset which were addressed using pairwise treatment. Additionally, we utilized natural language processing techniques to clean the free-text data. We also identified anomalies in the data during our cleaning process, such as the presence of "USA" in the "Death.County" column, which we resolved. Furthermore, there was a numerical entry in the "Death.City" column for one of the drug abuse cases.

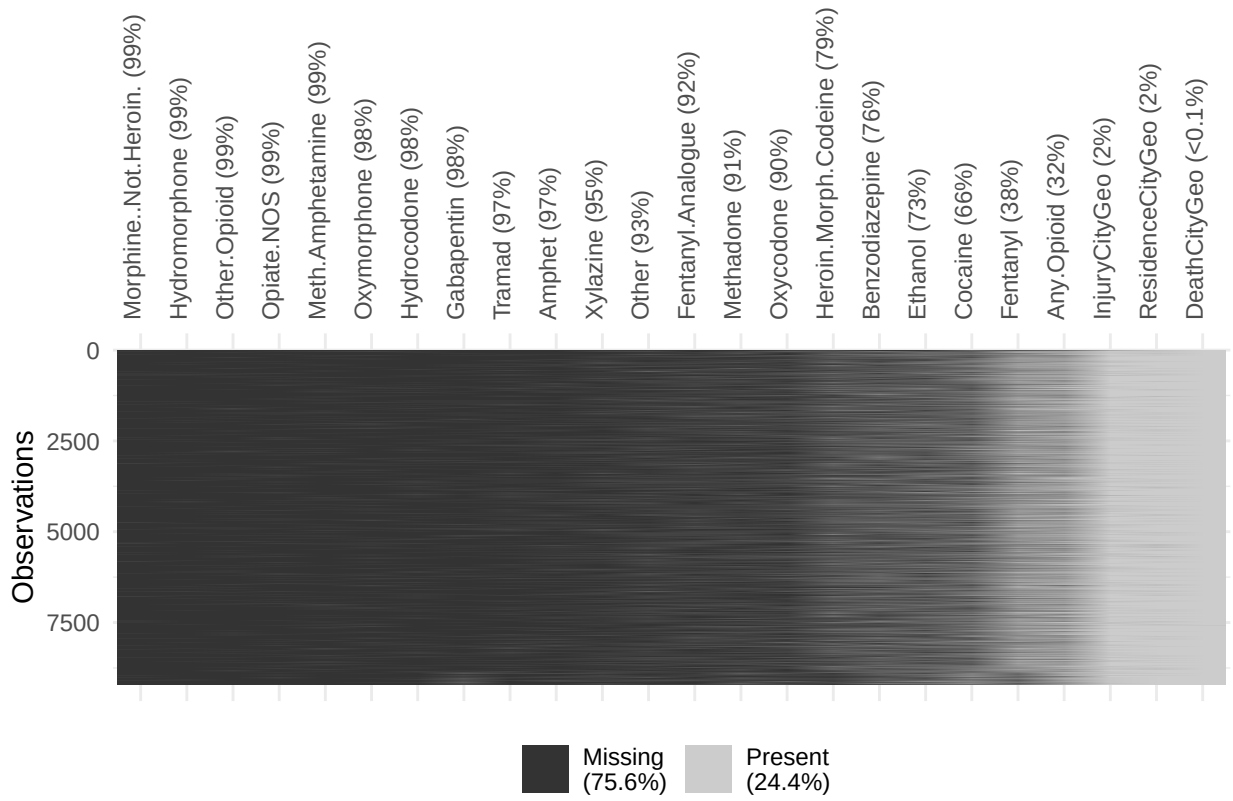
Missing Data Exploration

We started our analysis by examining the missing values present in the dataset. This was a crucial step as it helped us to evaluate the quality of the data and determine the suitability of the columns for our analysis. Based on our observations, we found the following:

1. Demographic information such as "Age", "Sex", "Location", "Injury.Place", "Manner.of.Death" had less than 1% missing values, which made them suitable for our analysis.
2. The temporal information provided by "Date" and "Date.Type" was available for almost all cases, making them useful for trend analysis.
3. Spatial data parameters like "Injury.State", "Injury.County", "Residence.State", "Residence.County" had more than 15% missing values, with some columns reaching as high as 72%. To mitigate this, we opted to use other parameters like 'DeathCityGeo', 'InjuryCityGeo', and 'ResidenceCityGeo' that had less than 2% missing values.
4. The "Description.of.Injury" column had about 15% missing values, but it was still valuable as it contained important information about the scenarios leading to drug overdose deaths. We utilized a word cloud analysis to examine this column and found the data to be sufficient.
5. The blank values in the drug-related columns indicated that the drug was not consumed by the candidate. Thus, the missing values in these columns did not pose any significant problem.

Overall, the data appeared to be suitable for our analysis, and we could proceed with our investigation.


Heroin.death.certificate..DC. (92%)
Other.Significant.Conditions (91%)
Location.if.Other (91%)
Ethnicity (86%)
Heroin (64%)
Injury.County (36%)
Injury.State (32%)
Death.State (25%)
Residence.State (19%)
Death.County (12%)
Residence.County (11%)
Description.of.Injury (9%)
Residence.City (4%)
Injury.City (2%)
Injury.Place (1%)
Race (0%)
Location (0%)
Manner.of.Death (<0.1%)
Sex (<0.1%)
Death.City (<0.1%)
Age (<0.1%)
Date (0%)
Date.Type (0%)
Cause.of.Death (0%)
Observations
0
2500
5000
7500
Missing (24%)
Present (76%)



Demographical Aspects of Drug Abuse

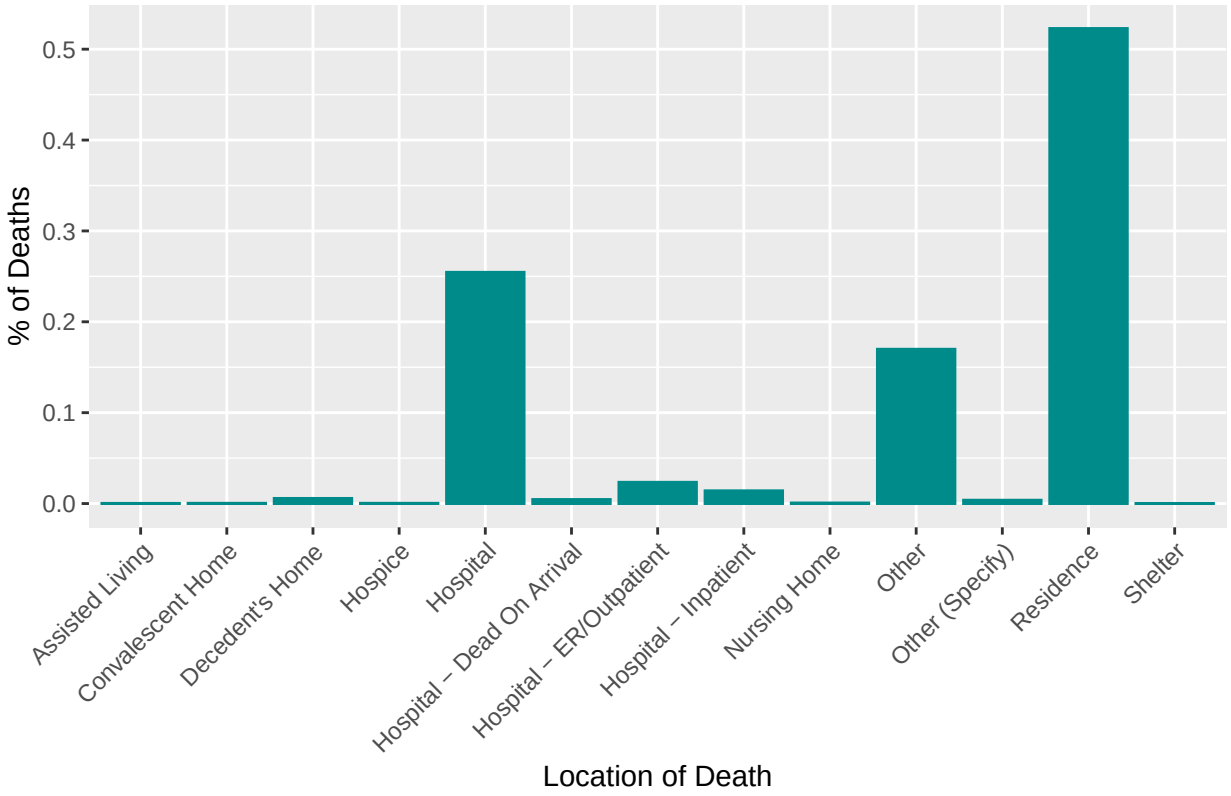
Assesing the distribution of the fields

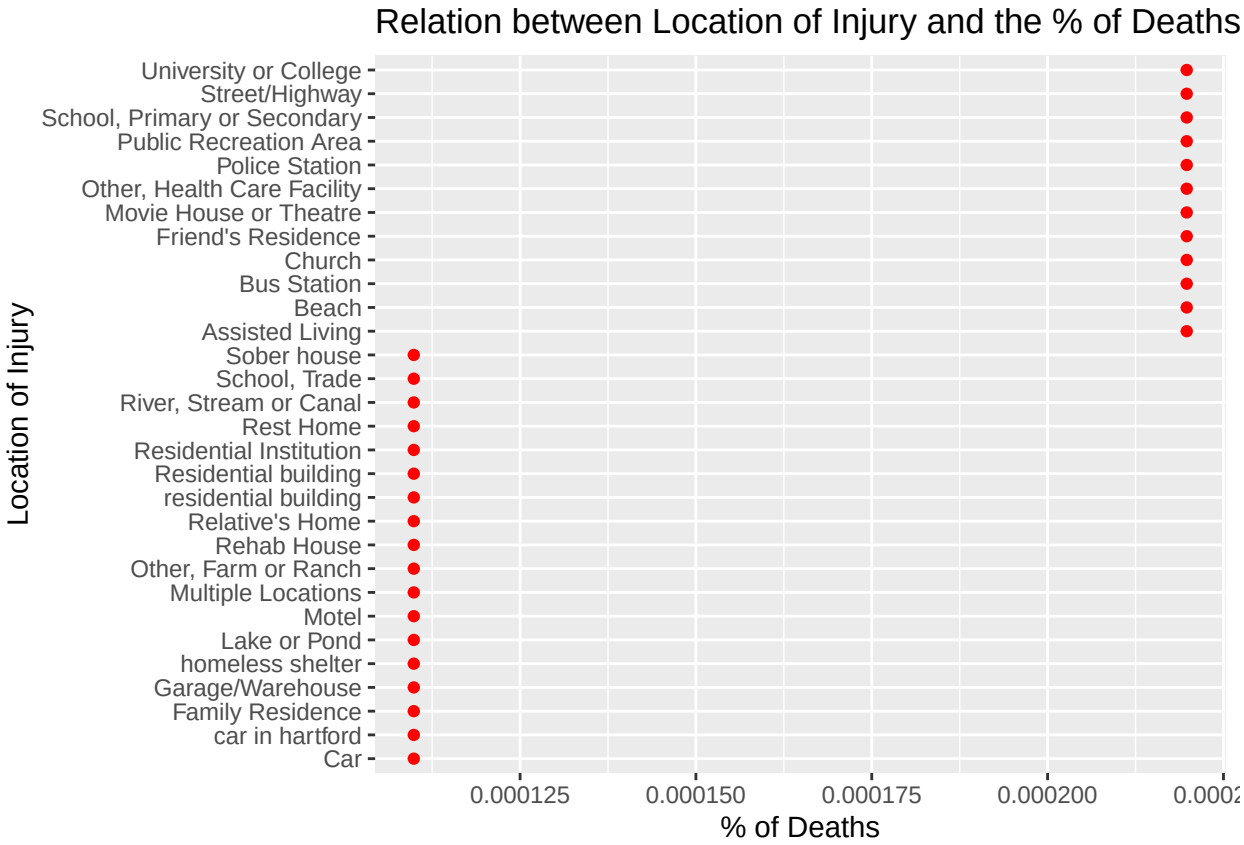
After identifying the parameters that had sufficient data, we examined the distribution of each of these variables. We created visualizations to observe the spread of numerical data, the frequency of categories within each column, and the proportion of deaths along various dimensions. This helped us identify which parameters needed to be further explored to gain a deeper understanding of the drug abuse situation.

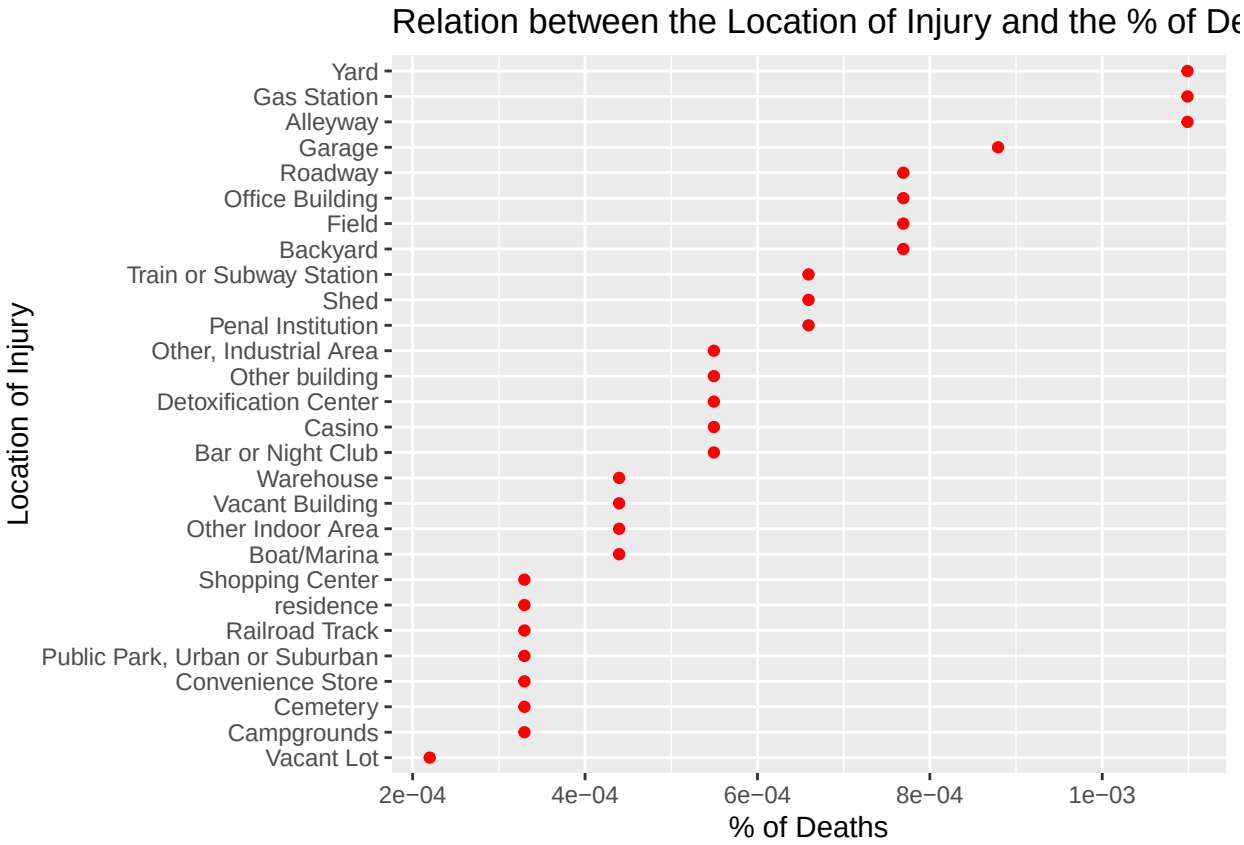
1) This section of analysis focuses on the comparison between the **Location of Injury and the Location of Death** in the dataset.

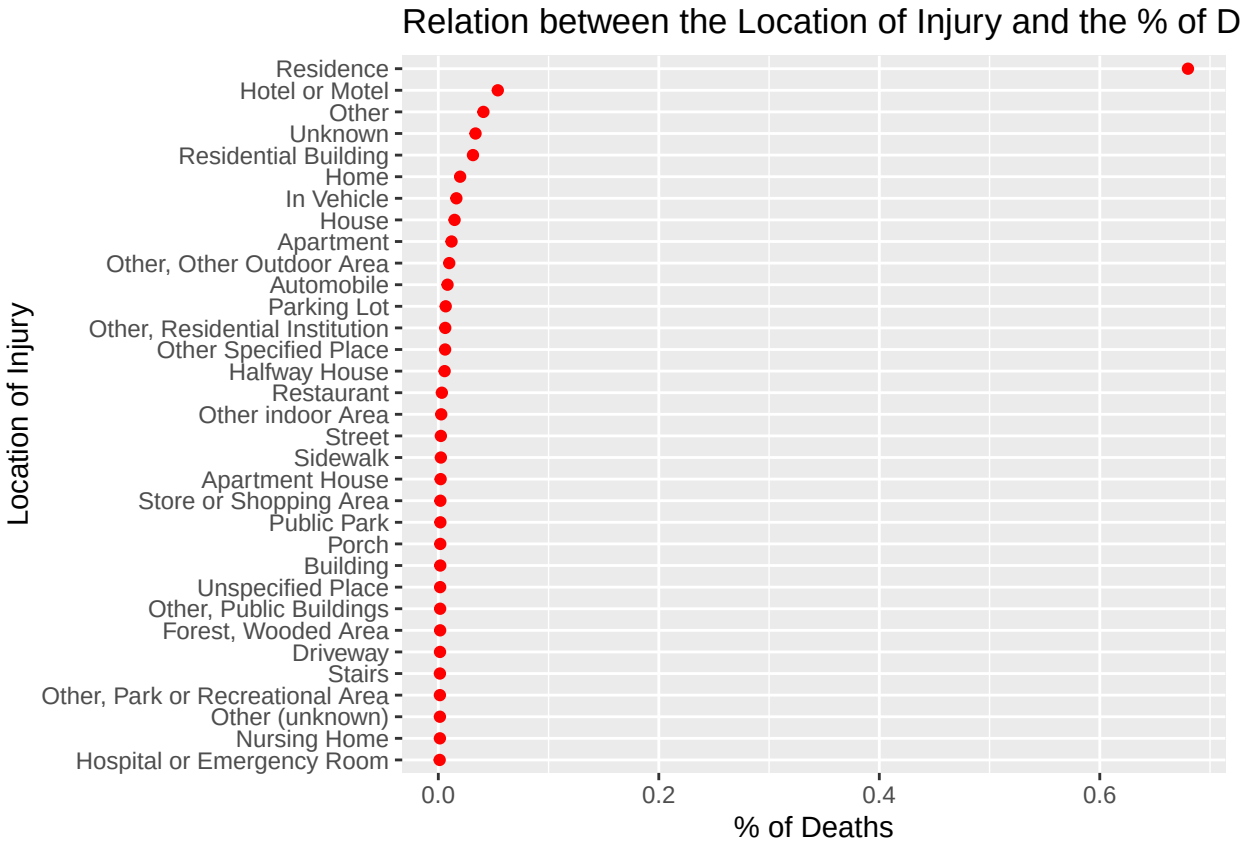
We analyzed the data related to the Location of Injury and Death to explore whether there was a correlation between the two. Specifically, we wanted to investigate if the location of injury was the same as the location of death, which could indicate that the death was instantaneous and occurred before the person could move to a different location. Upon analyzing the data, we found that a large number of deaths occurred in residences and hospitals. We also noticed that the majority of injuries occurred in residences, suggesting that most drug-related injuries and deaths happened at home. This prompted us to further investigate the cities and states where drug-related deaths were reported in residences.

Deaths and the % of deaths categorized by location.



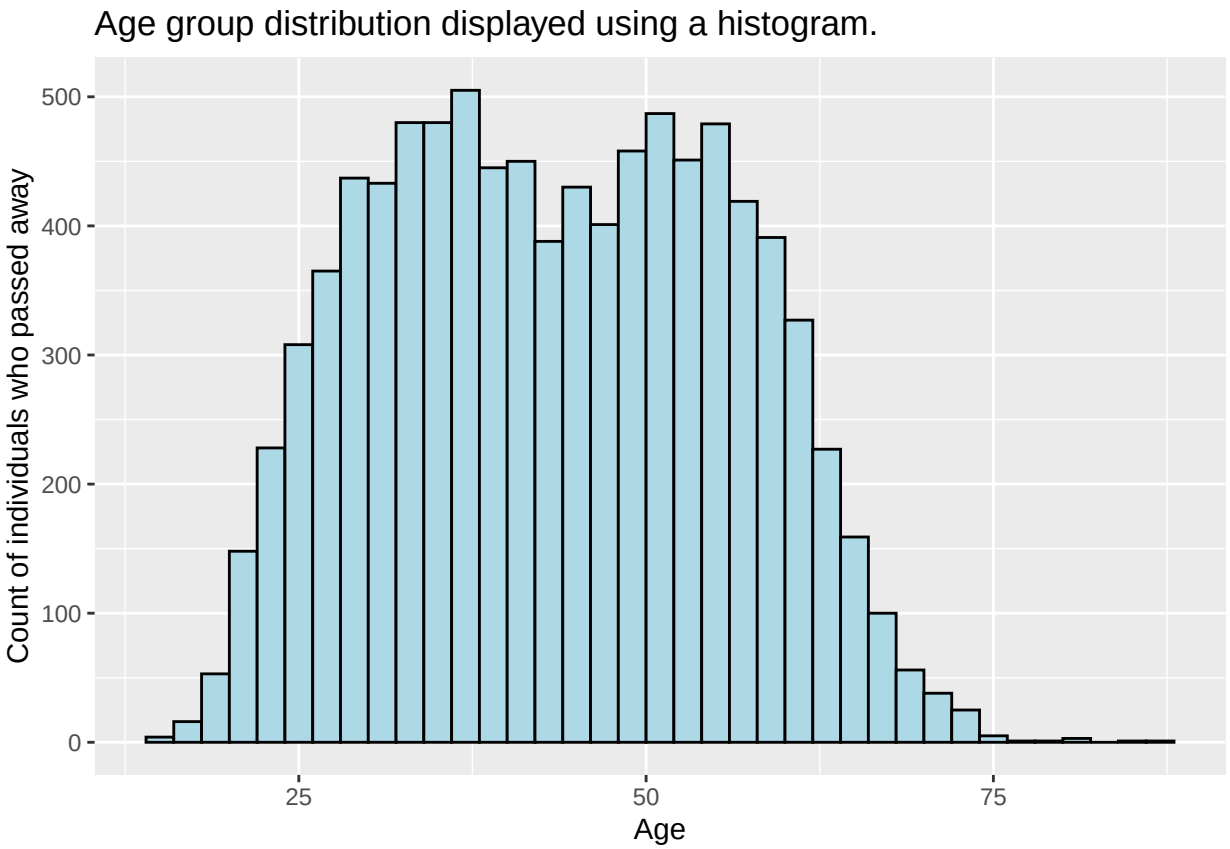

Relation between Location of Injury and the % of Deaths
Location of Injury
University or College
Street/Highway
School, Primary or Secondary
Public Recreation Area
Police Station
Other, Health Care Facility
Movie House or Theatre
Friend's Residence
Church
Bus Station
Beach
Assisted Living
Sober house
School, Trade
River, Stream or Canal
Rest Home
Residential Institution
Residential building
residential building
Relative's Home
Rehab House
Other, Farm or Ranch
Multiple Locations
Motel
Lake or Pond
homeless shelter
Garage/Warehouse
Family Residence
car in hartford
Car
0.000125
0.000150
0.000175
0.000200
% of Deaths


Relation between the Location of Injury and the % of De
Location of Injury
Yard
Gas Station
Alleyway
Garage
Roadway
Office Building
Field
Backyard
Train or Subway Station
Shed
Penal Institution
Other, Industrial Area
Other building
Detoxification Center
Casino
Bar or Night Club
Warehouse
Vacant Building
Other Indoor Area
Boat/Marina
Shopping Center
residence
Railroad Track
Public Park, Urban or Suburban
Convenience Store
Cemetery
Campgrounds
Vacant Lot
2e−04
4e−04
6e−04
8e−04
1e−03
% of Deaths



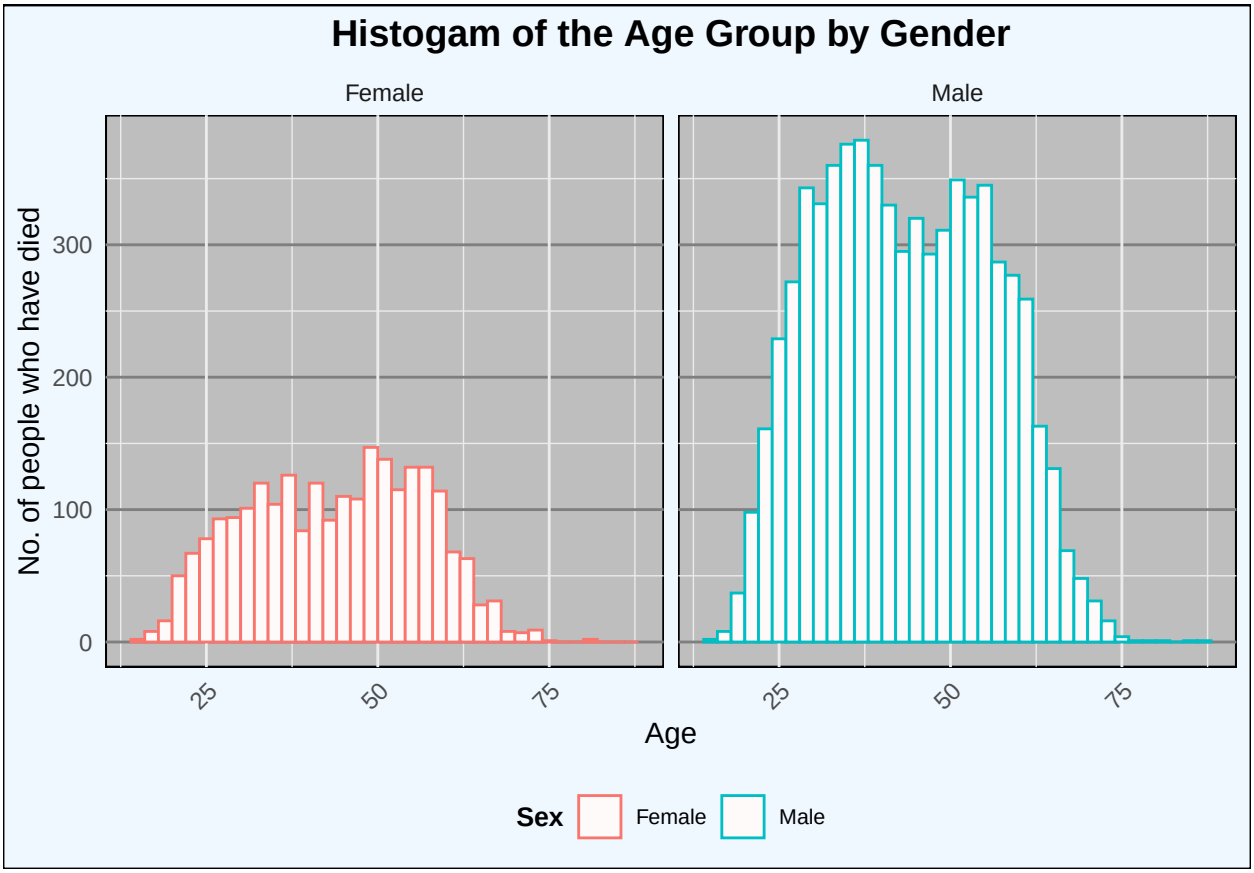
2) **Age**

As the dataset contained only one continuous variable, namely age, we analyzed its distribution across various categorical variables. The distribution of age across the entire dataset exhibited a bimodal shape with peaks at approximately 28-30 years and 50 years. Additionally, the distribution appeared to be skewed towards the right.



3) **Sex**

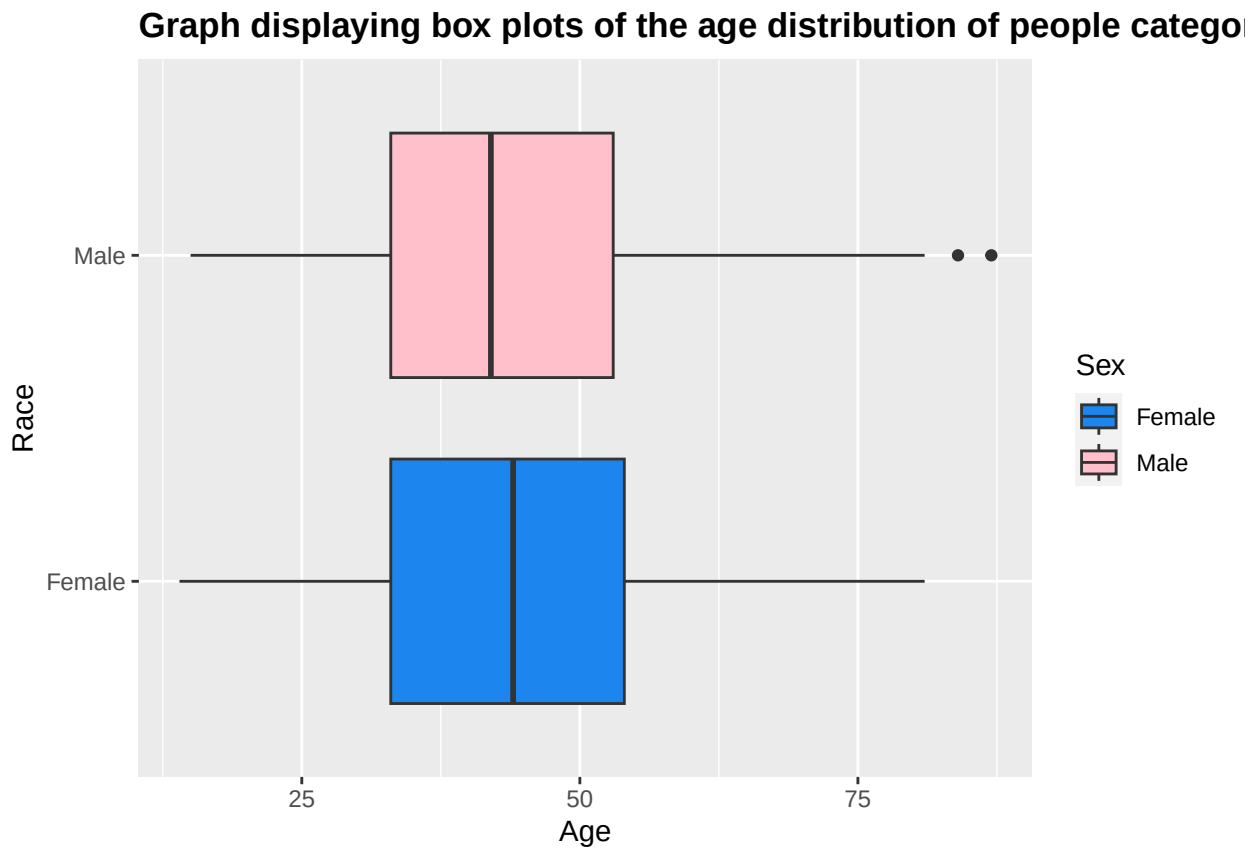
The count of male deaths due to drug overdose was over two times higher than that of females.




```
## # A tibble: 2 x 3
##   Sex    mean_age median_age
##   <chr>     <dbl>      <dbl>
## 1 Female     43.3         44
## 2 Male       42.9         42
```

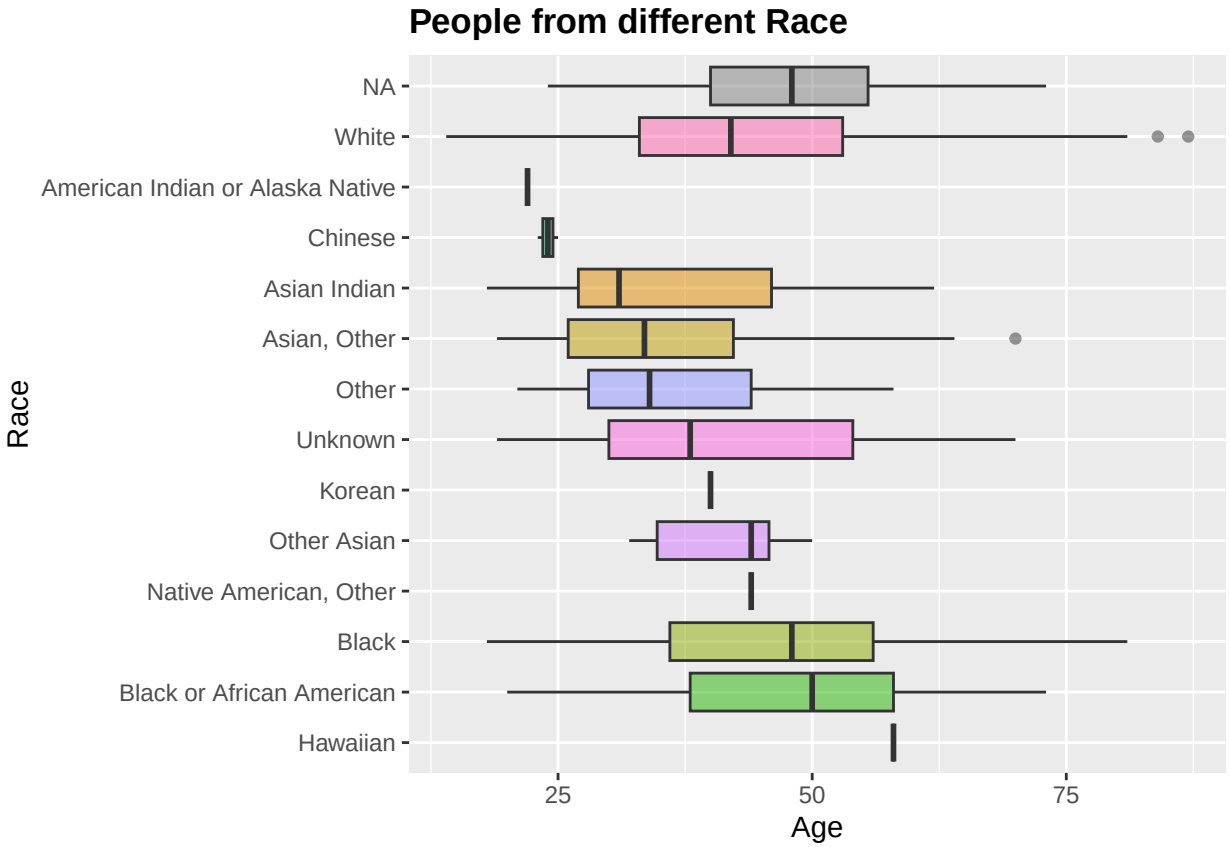
- Females who died due to drug abuse had a higher median age than males.
- The mean age for male who died due to drug abuse was lower than female with 42.94

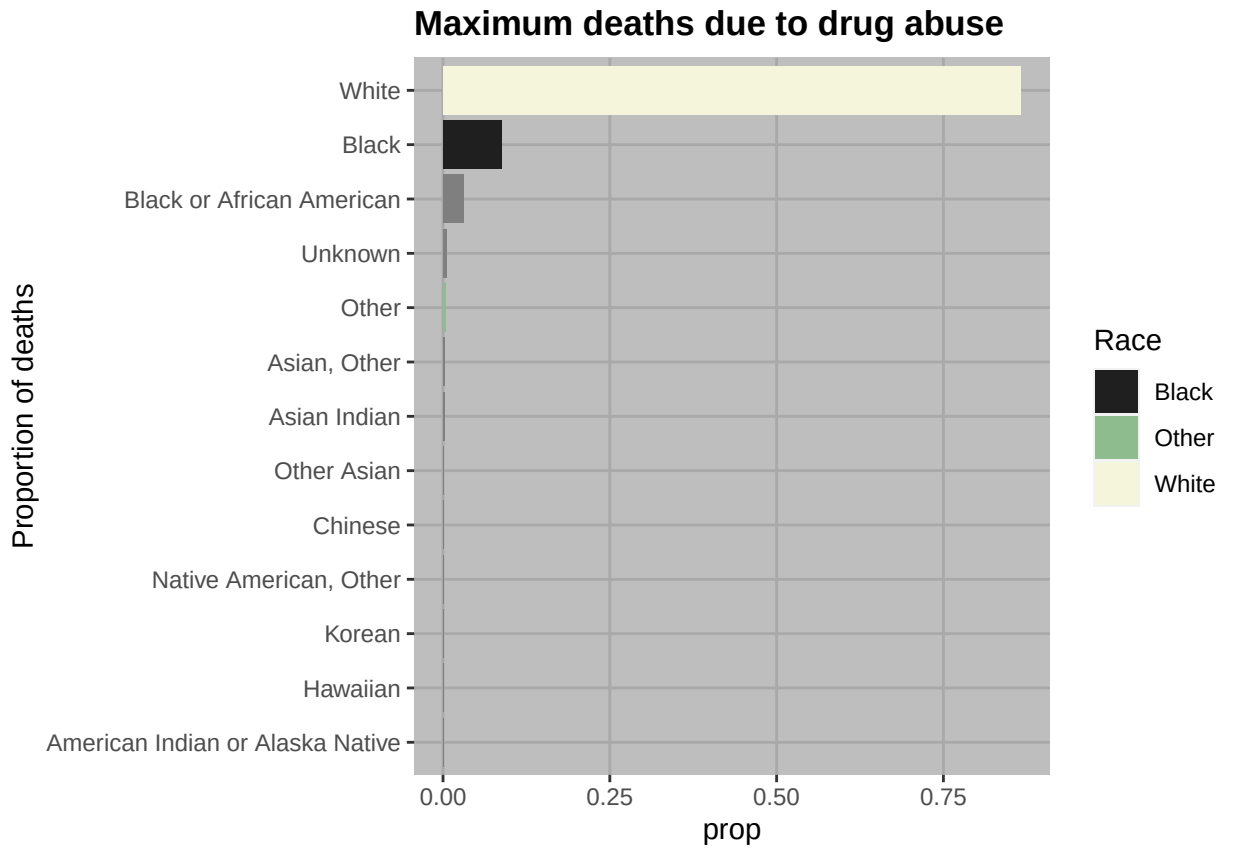
Sex and Age



4) **Race and Age**

The majority of drug-related deaths were among White individuals, followed by those of Hispanic, White, and Black ethnicity. Chinese individuals had the lowest median age of death due to drug abuse, while Black individuals had the highest median age


People from different Race
NA
White
American Indian or Alaska Native
Chinese
Asian Indian
Asian, Other
Other
Unknown
Korean
Other Asian
Native American, Other
Black
Black or African American
Hawaiian
Race
25
50
75
Age



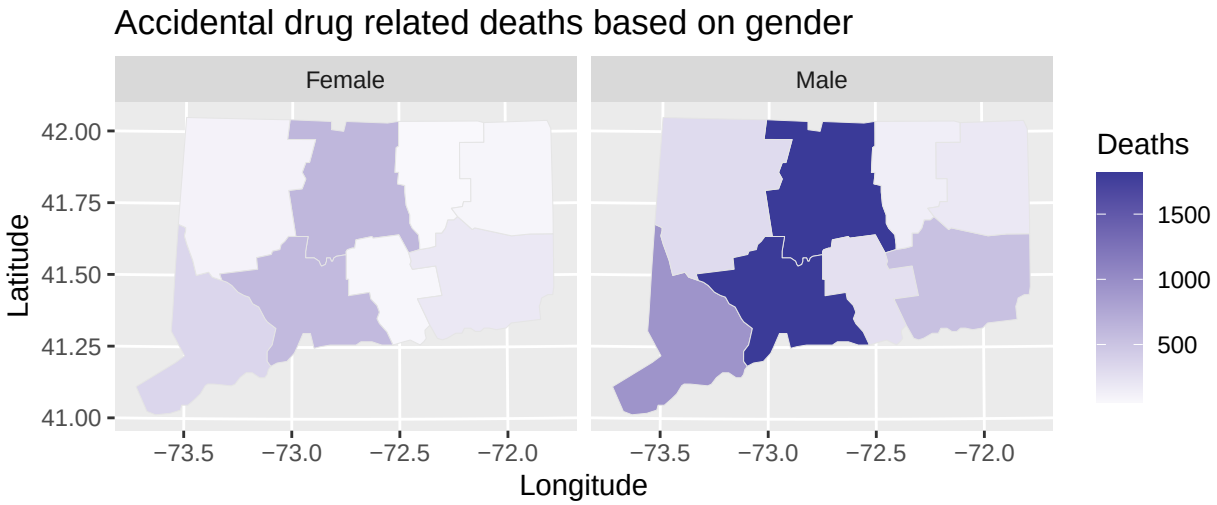
Spatial Analysis

Spatial analysis is a process of analyzing and understanding data related to places and geographic phenomena. It involves the use of specialized software and techniques to study and visualize patterns in spatial data, such as maps and GPS coordinates.

Spatial analysis can help identify patterns, trends, and relationships between different geographic features. It can also help in decision-making processes related to location wise issues and emergency response.

1) *Sex*

Gender-based drug abuse has always been a topic of discussion in drug analysis. To find insights into our data set and its county-based distribution, we created choropleth plots. The plot below shows the number of accidental deaths due to drug abuse in Connecticut's various counties.

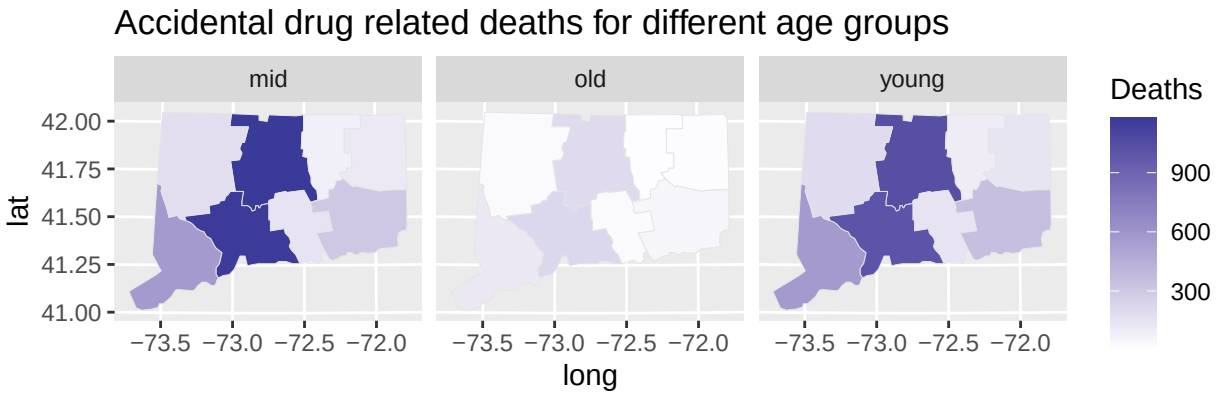


From the plot, we can infer that:

The number of males who died due to drug abuse is generally higher than females for the years 2012 to 2022. This is evident as the choropleth plot for males has darker shades of purple than that for females. The county of Hartford has the highest number of deaths due to drug abuse compared to other counties for both males and females. The number of deaths in the counties of Middlesex and Tolland are relatively low compared to other counties for males and females.

2) *Age*

The impact of drugs on different age groups has been an area of research for several years. Our primary objective was to gain county-level insights on the age distribution of individuals who died due to drug abuse, which led us to create a choropleth plot faceted on age.

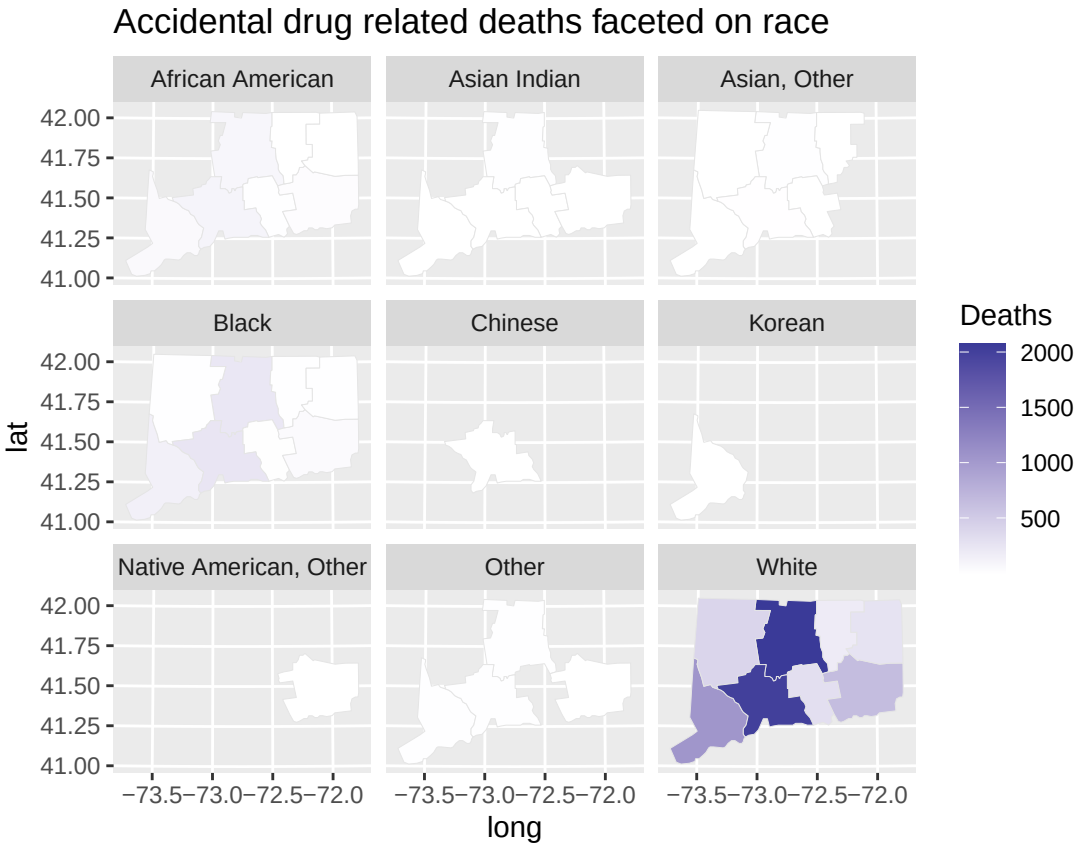


The choropleth plot shows the number of deaths due to drug abuse across the different age groups in the counties of Connecticut. The age groups are categorized as young (age 19 - 40), mid (age 40 - 60), and old (age $>$ 60).

There are very few cases of elderly individuals dying from drug abuse in the state. The highest number of deaths due to drug abuse was reported for individuals between the age of 20 to 60. Additionally, the county of Hartford had the highest number of deaths due to drugs for the mid and young age groups from 2012 to 2022.

3) *Race*

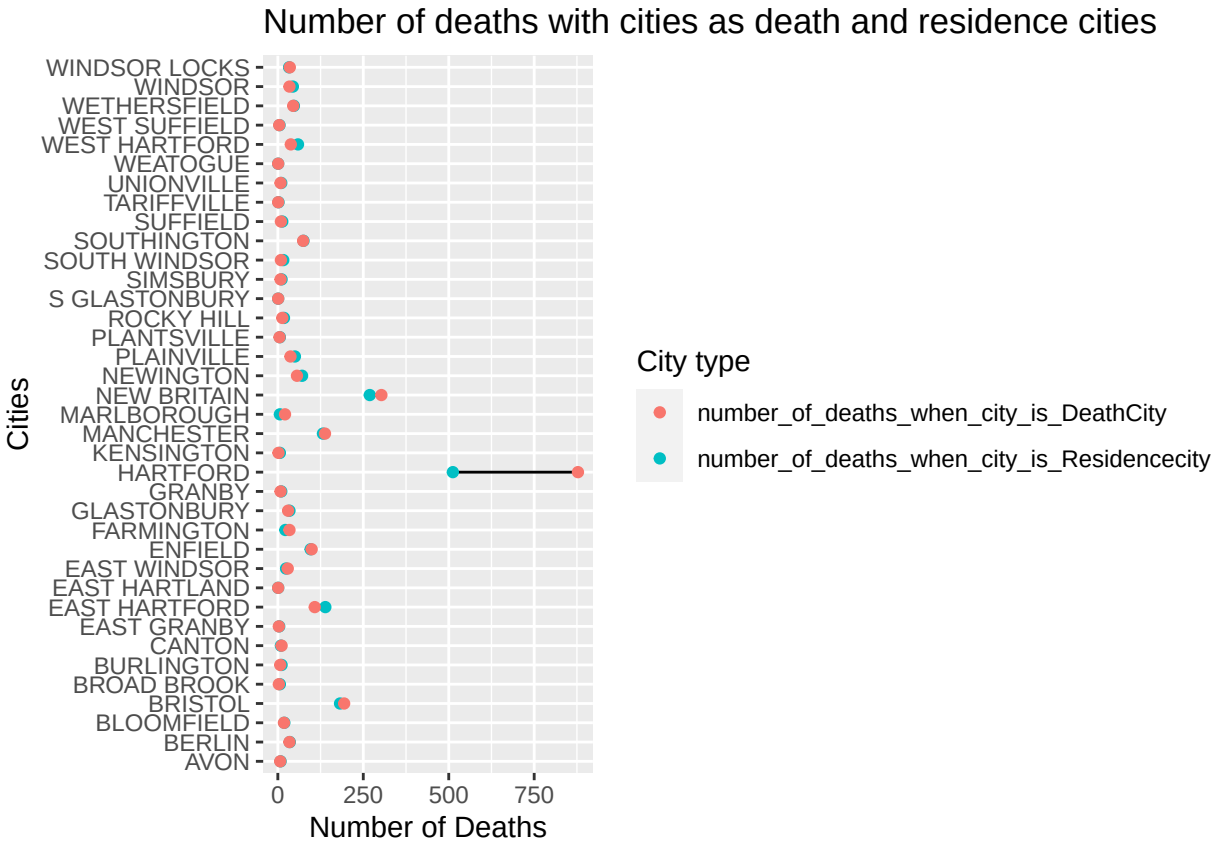
We conducted an analysis to determine the race distribution of individuals who died due to drug abuse. For this purpose, we created a choropleth plot to visualize the count of deaths for each ethnicity across different counties in Connecticut.



The majority of individuals who died due to drug abuse were of "white" ethnicity. Except for New Haven county, no Chinese individuals were reported to have died due to drug abuse in other counties. Asian Indian individuals who died due to drug abuse were relatively low in Hartford, New Haven, and Fairfield counties. A significant number of black and Hispanic white individuals were reported to have died due to drug abuse in Hartford, New Haven, and Fairfield counties.

4) *Understanding the Drug Epidemic in Hartford*

In our previous analysis, we discovered that Hartford had the highest count of deaths resulting from drug abuse compared to other counties in Connecticut. Hence, we decided to conduct a detailed study on the cities within Hartford county. The following Cleveland plot indicates the count of deaths in different cities, where the location of the demise and the victim's residential city are represented.



We observed the following trends from our analysis:

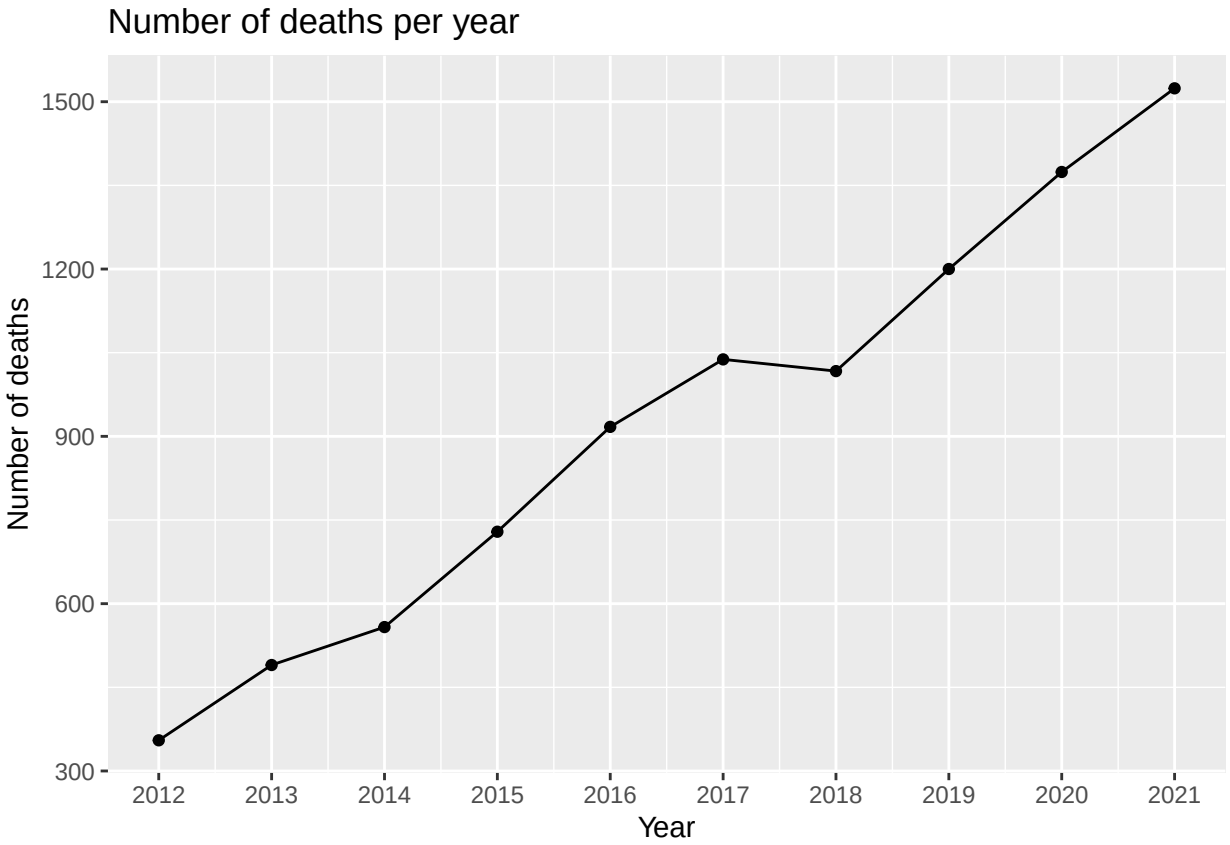
- Among all the cities in Hartford county, Hartford city had the highest count of deaths from drug abuse, both when it was the location of the demise and when it was the victim's residential city.
- The count of deaths in Hartford city was significantly higher when it was the location of the demise as opposed to the victim's residential city.
- Similarly, the cities of New Britain and Bristol were ranked second and third, respectively, in terms of the highest count of deaths resulting from drug abuse in Hartford county.
- Interestingly, for the cities of East Hartford, West Hartford, Newington and Plainville the count of deaths that occurred within these cities was lower than the count of deaths of residents from these cities who passed away elsewhere.

temporal Aspects of Drug Abuse

We conducted an analysis to examine the relationship between drug abuse and time variables.

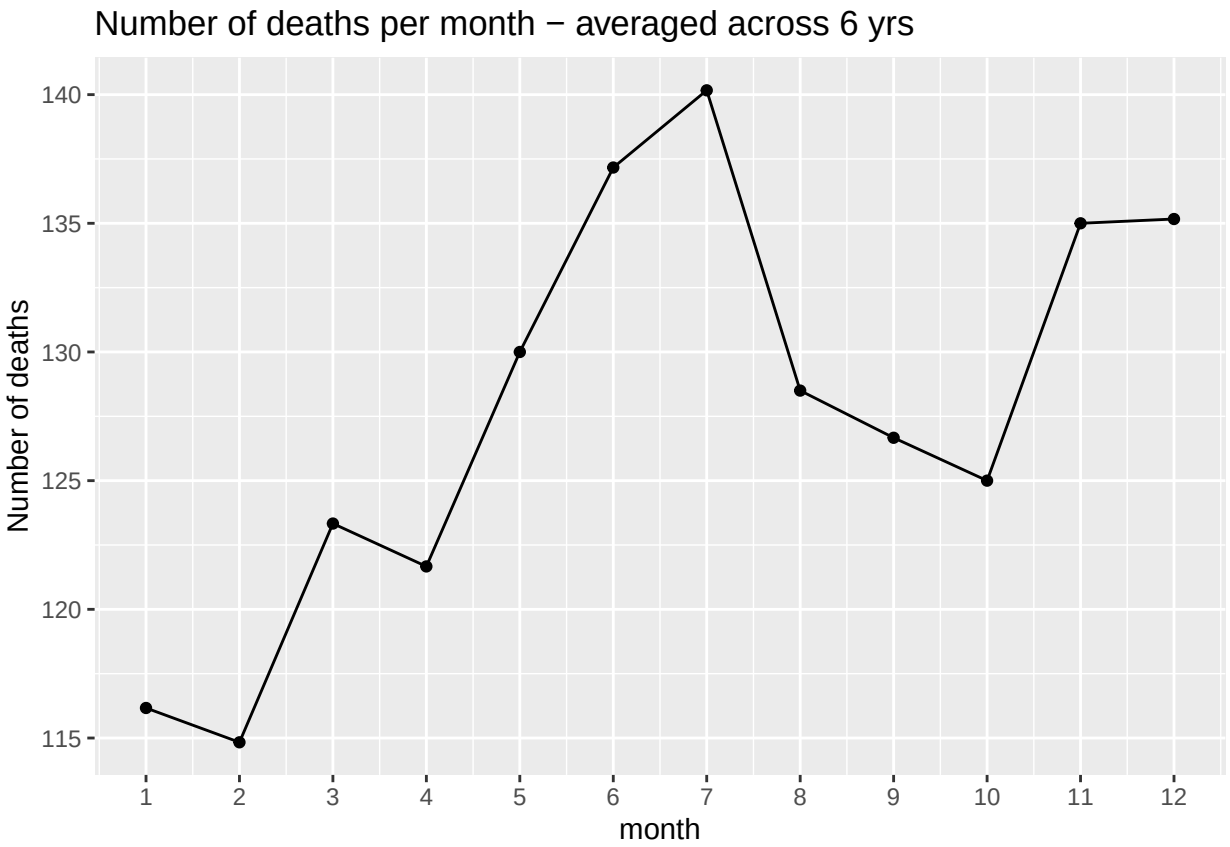
Seasonality

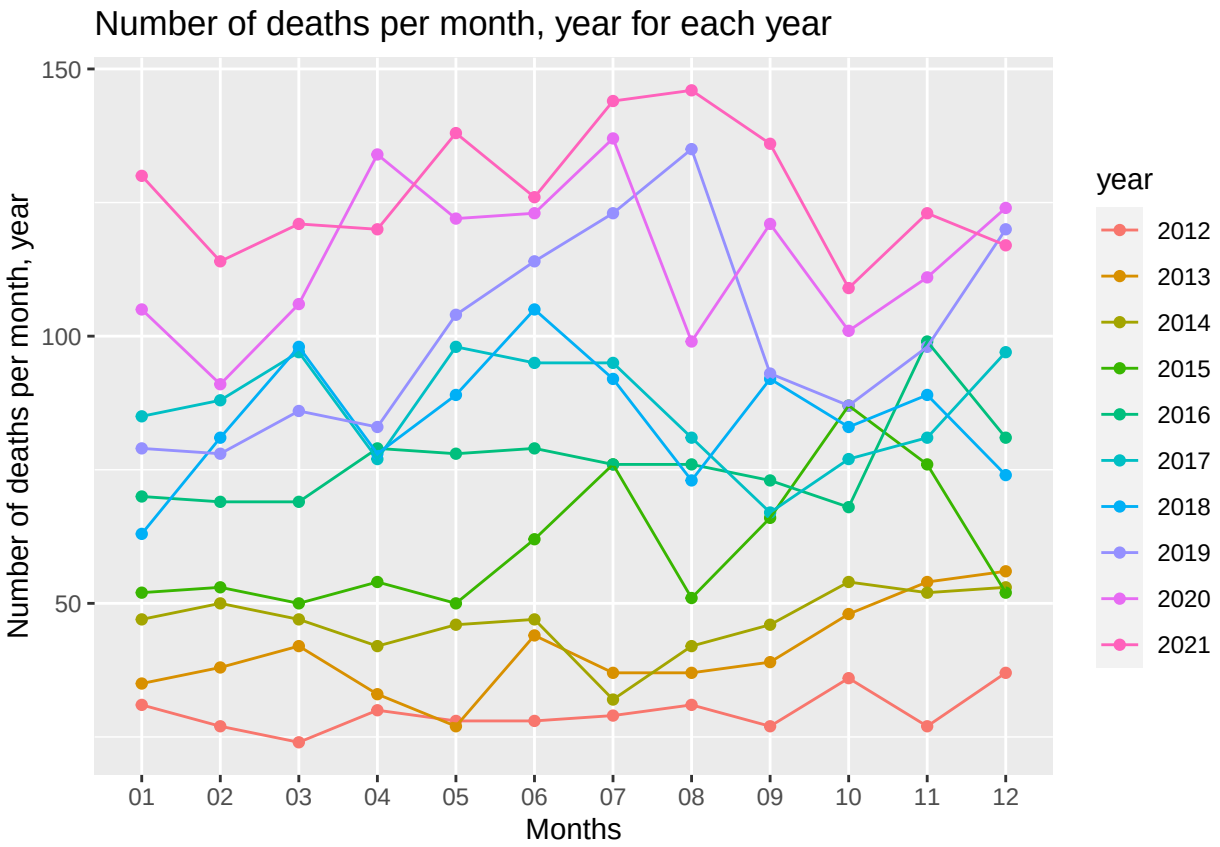
We examined the occurrences of drug abuse over time and analyzed the number of reported deaths each year between 2012 and 2022.



Our findings revealed that the highest number of deaths was reported in 2017, followed by 2018, with a consistent increase in deaths from 2018 on wards.

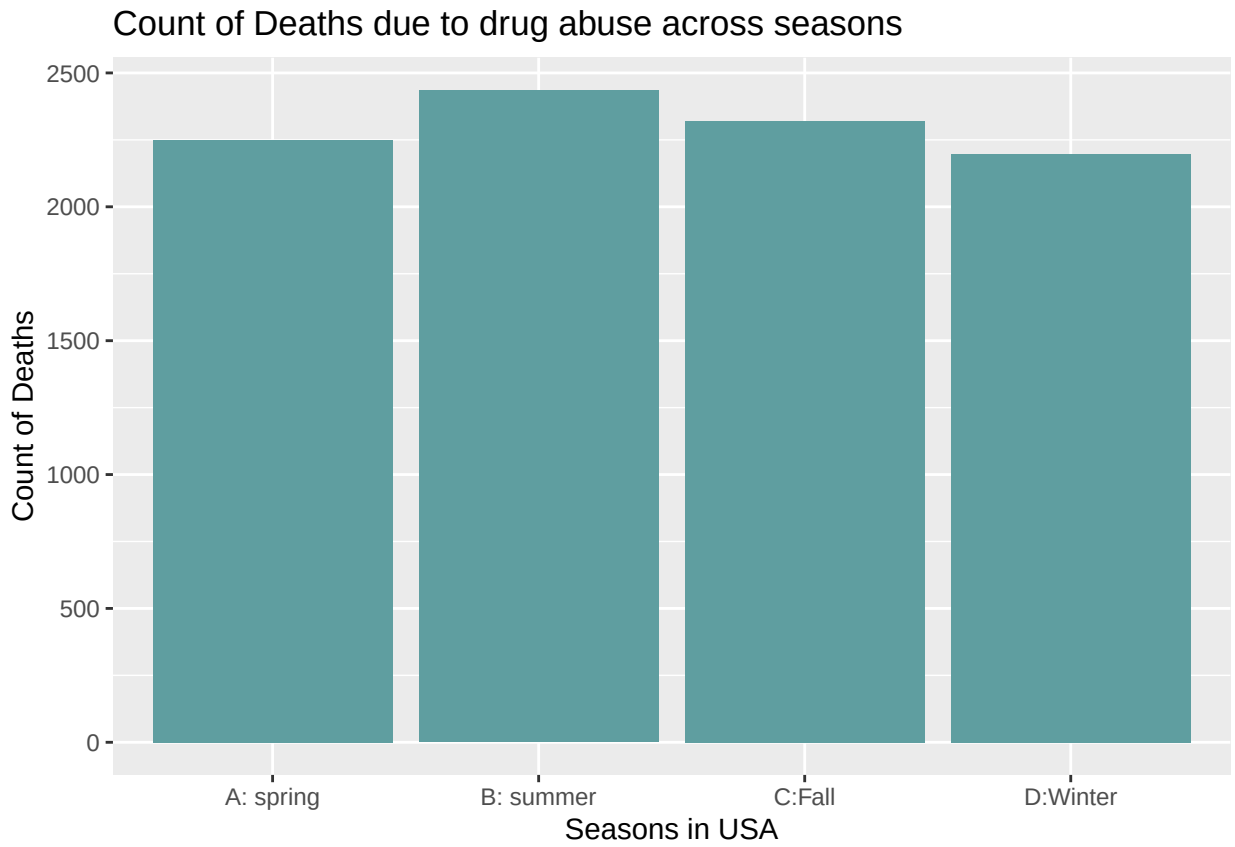
After plotting the average number of deaths per month for the past ten years, we noticed a pattern where the number of deaths tends to increase from April to July, and then again from October to November, on an average.


Number of deaths per month − averaged across 6 yrs
Number of deaths
140
135
130
125
120
115
1
2
3
4
5
6
7
8
9
10
11
12
month



In 2017, we noticed a significant increase in the number of deaths from September to the end of the year, and this trend continued into the first three months of 2018. This suggests that the rise in deaths during Fall 2017 and Winter 2018 may have been due to cold weather conditions that may have prompted people to consume more drugs, leading to more deaths. Similar patterns were also observed in 2013 and 2014, although we cannot confirm our assumptions as we lack the necessary data variables in this data set. Furthermore, we observed that the highest number of deaths in 2021 occurred in August, while the deaths getting better by the end of the year in comparison with 2020.

We continued analysing the data with patterns. Hence, we went ahead by working on different seasons.

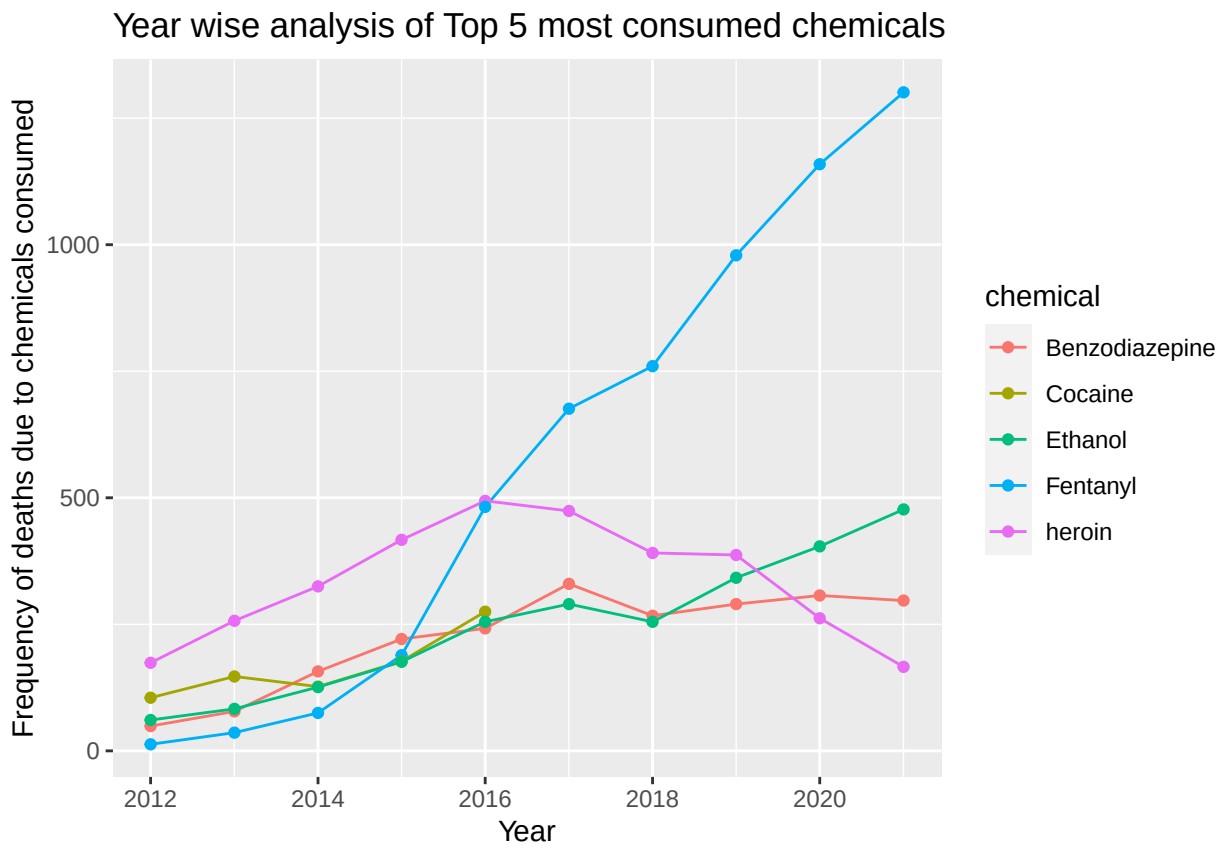


It is evident from the plot that the highest number of deaths occured were during Summer.

Patterns in Drug consumption over Years

To gain further insight into the factors contributing to the increase in drug-related deaths year after year, we analyzed the chemical compounds/drugs that were responsible for these deaths during the years under consideration.

```
##    all_chemicals Freq
## 5       Fentanyl 5673
## 8         Heroin 3348
## 3        Cocaine 3172
## 4        Ethanol 2471
## 2 Benzodiazepine 2239
```



Our findings indicate that:

'Fentanyl', which was the least consumed drug between 2012-2014, has been on the rise since 2015 and has surpassed 'Heroin' consumption in 2016. Conversely, 'Heroin' consumption has been decreasing since 2016 and it appears that 'Fentanyl' is cannibalizing it. In addition, the consumption of other drugs has also been increasing steadily over the years in contrast to 'Heroin'.

Drug Abuse Causal Diagnosis

Our aim was to investigate the factors responsible for drug abuse cases and identify their underlying root causes. We conducted an analysis to determine the major chemicals consumed and the types of injuries associated with drug abuse. Through this investigation, we were able to identify the most commonly used chemicals in drug abuse cases, as well as the various forms of drug abuse, including ingested pills, alcohol, and substance abuse. Furthermore, we categorized the factors that contributed to the majority of drug abuse cases.

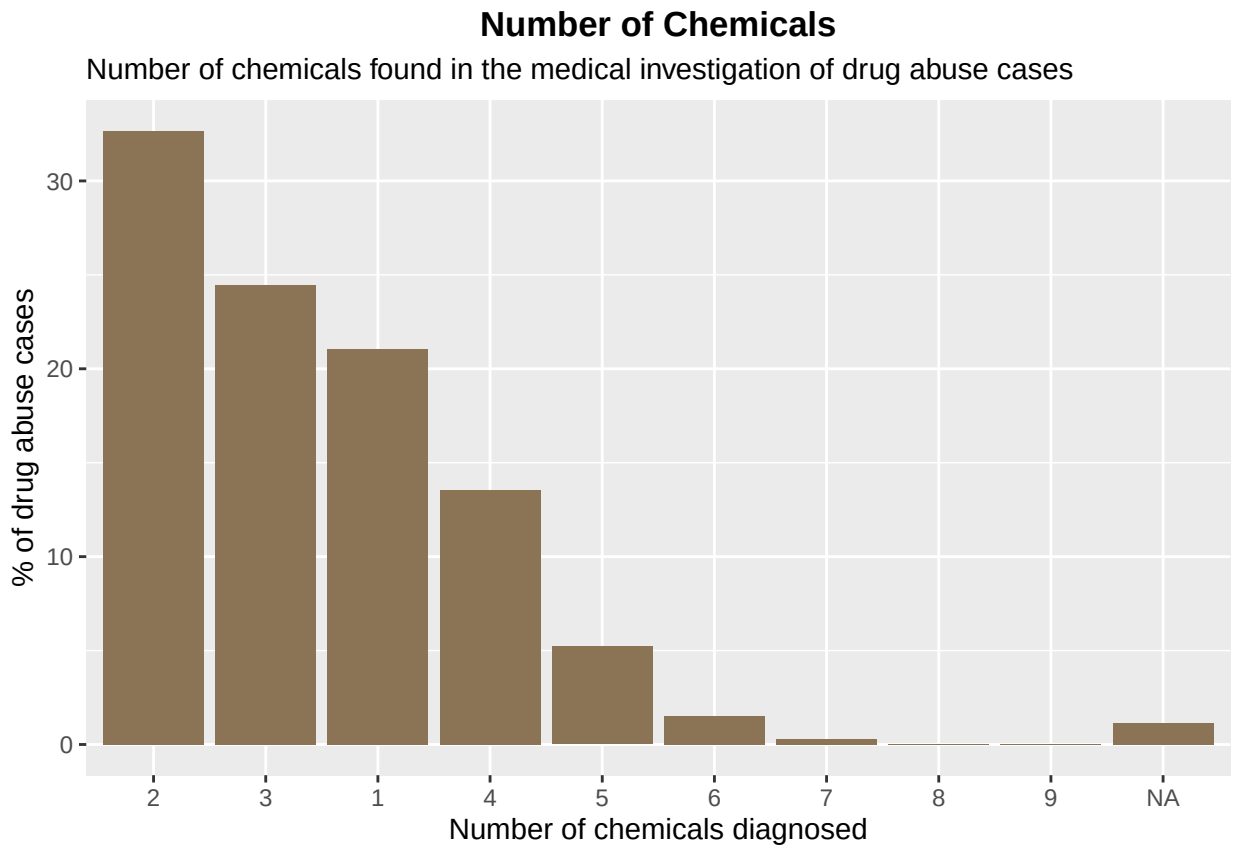
Derived Metric Calculation

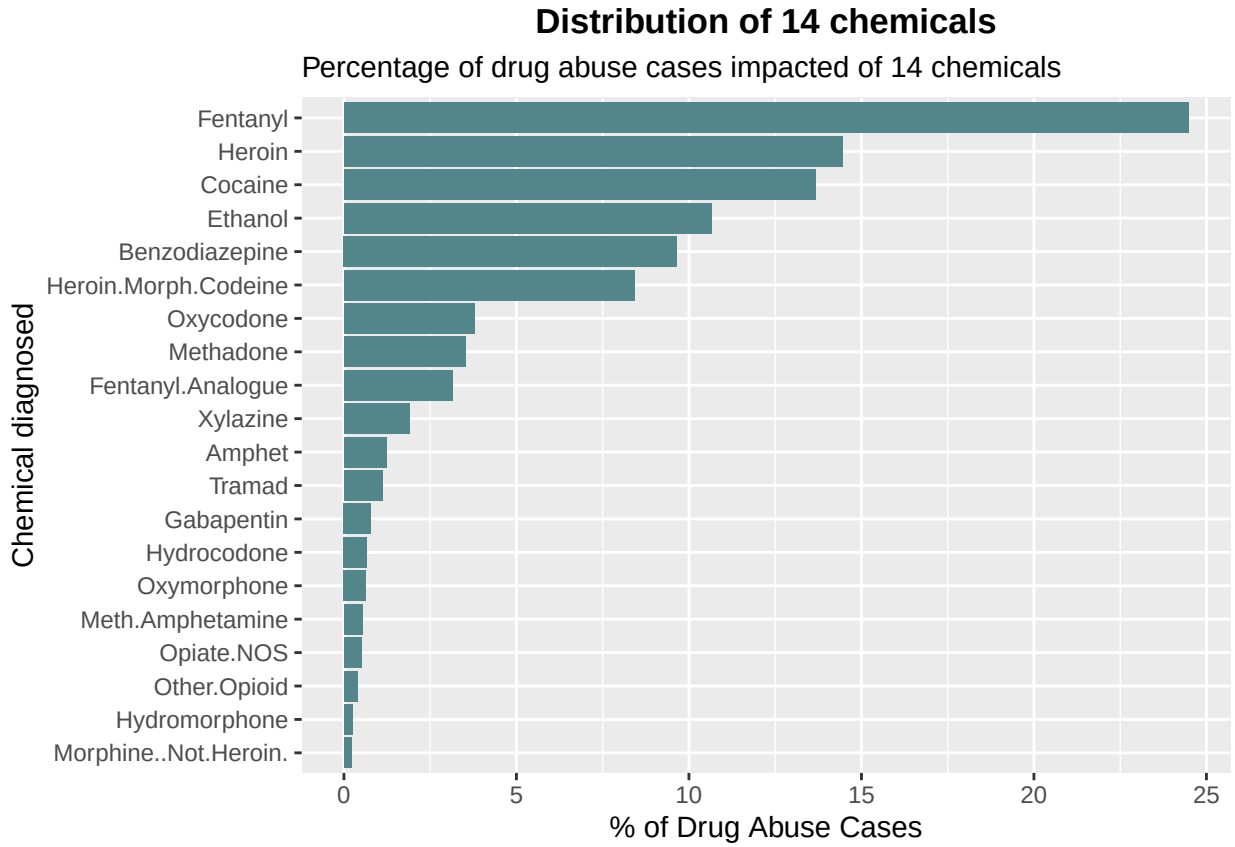
In the medical diagnosis of drug abuse cases, there are 14 types of chemicals commonly found. These include Heroin, Cocaine, Fentanyl, Fentanyl Analogue, Oxycodone, Oxymorphone, Ethanol, Hydrocodone, Benzodiazepine, Methadone, Amphet, Tramad, Morphine_NotHeroin, and Hydromorphone. We have computed two derived metrics: "Chemicals_Diagnosed" and "count_of_diagnosed_chemicals". The former represents the list of chemicals found in each case of drug abuse, i.e., in each row of the dataframe, while the latter denotes the number of chemicals found in each drug abuse case out of the total list of 14 chemicals mentioned above. We will further explore these metrics in the upcoming sections.

```
##                                          Chemicals_Diagnosed
## 9145 Oxycodone,Benzodiazepine,Gabapentin,Heroin.Morph.Codeine
## 9147                            Ethanol,Methadone,Gabapentin
```

```
## 9185                          Benzodiazepine,Amphet,Opiate.NOS
## 9189                                      Ethanol,Gabapentin
## 9194                                Cocaine,Fentanyl.Analogue
## 9196                      Fentanyl,Tramad,Xylazine,Gabapentin
##      count_of_diagnosed_chemicals
## 9145                            4
## 9147                            3
## 9185                            3
## 9189                            2
## 9194                            2
## 9196                            4
```

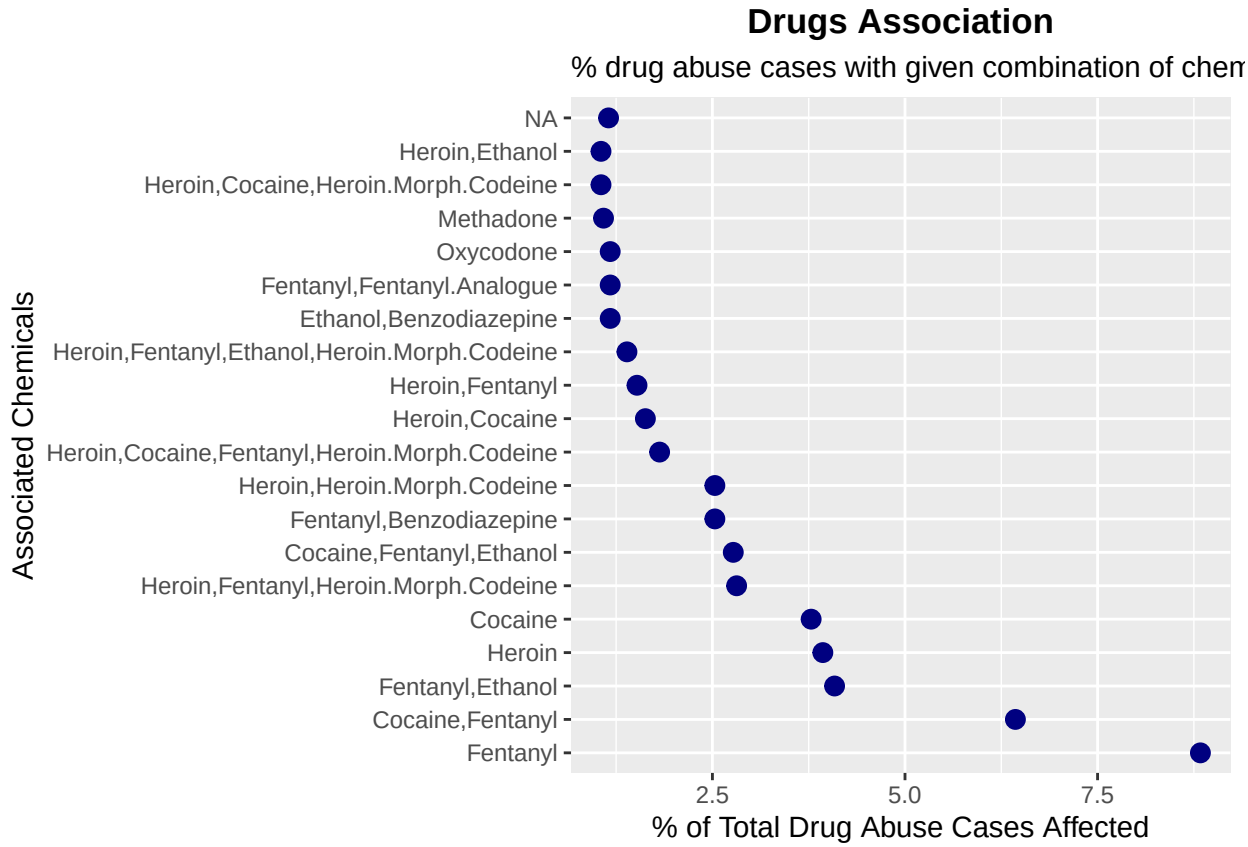
We analyzed the derived metric "count_of_diagnosed_chemicals" and created a histogram to represent the distribution of the number of chemicals found in drug abuse cases. The histogram showed that in most cases, there were only two chemicals found, and this was the case for more than 37% of the cases.





In our analysis of drug abuse cases, we examined the chemicals identified in medical investigations. We created a plot showing the percentage of drug abuse cases in which each chemical was diagnosed. The results indicated that 'Heroin', 'Fentanyl', and 'Cocaine' were the three most common chemicals found, occurring over 50% of the cases. The remaining chemicals were less frequently identified.

Association Mining among 14 chemicals



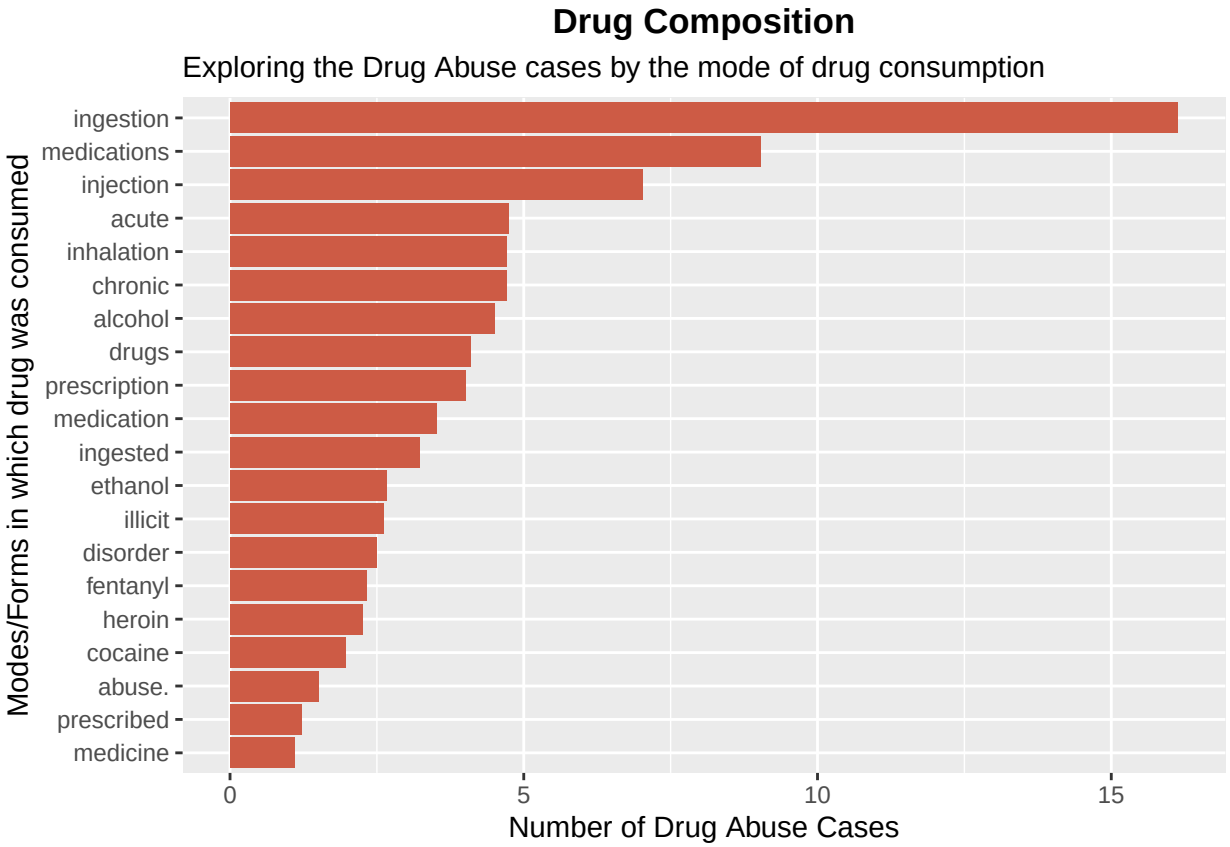
To investigate the co-occurrence of the 14 chemicals found in drug abuse cases, we analyzed the most frequently occurring combinations. We found that 'Heroin' was present in all the popular combinations. The top three most frequently occurring chemical combinations were 'Heroin' with 'Fentanyl' (5%), 'Heroin' with 'Cocaine' (4.7%), and 'Heroin' with 'Ethanol' (3.3%). This suggests that 'Heroin' is the most commonly consumed chemical among the drug abusers. Also, Fentanyl was observed to be the highest intake with over 8.2%.

Analysis of Injury Description - Text Mining using NLP

In this section, we examined the injury description for all drug abuse cases in order to categorize the mode of drug consumption. The "DescriptionofInjury" field contained free-form text, so we performed basic natural language processing on it.

We took the following steps:

1. Tokenization: We first split the description of each drug abuse case into tokens.
2. Stop word removal: We filtered out the stop words and extracted candidate tokens that had noun and verb POS tags, as these represented the modes and forms in which drugs were consumed. We then calculated the frequency of these drug consumption modes.



Our analysis revealed that ingestion, injection, and medication were the most frequent modes of drug consumption. Ingestion was the mode in 23% of drug abuse cases, followed by injection in 11% and medication in 7.7% of cases. Together, these three modes accounted for 42% of all drug abuse cases.


```
## Rows: 20
## Columns: 2
## $ Injury <chr> "medicine", "prescribed", "abuse.", "cocaine", "heroin", "fenta~
## $ Freq   <dbl> 1.102941, 1.225490, 1.511438, 1.960784, 2.246732, 2.328431, 2.4~
```

Conclusions and Summary

Our study focused on drug overdose deaths in Connecticut between 2012 and 2021. We took a three-pronged approach to uncover patterns and insights into drug abuse cases and to answer our initial questions.

Firstly, we investigated whether demographic factors such as age, gender, race, and location were linked to drug abuse cases. Our analysis and supporting graphs showed a strong correlation between these factors and drug abuse.

Secondly, we examined the overall trend in drug abuse cases from 2012 to 2021. Our findings revealed an increasing number of deaths due to drug abuse over time.

Lastly, we conducted a causal diagnosis to identify the chemicals and co-consumption responsible for drug overdose deaths in Connecticut. Our results provided valuable information about the types of drugs and injuries involved in these cases.

Our analysis and conclusions were consistent with similar reports found in online news articles [4]. The study highlighted the severity of the drug abuse problem in Connecticut and around the world and underscored the urgent need for action to save innocent lives.

References:

[1] https://www.whitehouse.gov/wp-content/uploads/2022/03/FY-2023-Budget-Highlights.pdf

[2] https://www.commonwealthfund.org/blog/2023/overdose-deaths-declined-remained-near-record-levels-during-first-nine-months-2022-states#:~:text=An%20estimated%2079%2C117%20Americans%20died,higher%20than%20pre

[3] https://portal.ct.gov/DPH/Health-Education-Management--Surveillance/The-Office-of-Injury-Prevention/Opioids-and-Prescription-Drug-Overdose-Prevention-Program

[4] https://www.sciencedaily.com/releases/2023/03/230330102329.htm

[5] https://r4ds.had.co.nz/

[6] https://r-graph-gallery.com/

[7] https://www.statmethods.net/advgraphs/index.html