Analysis of Drug Overdose related Deaths

Kumari Nishu (kn2492)
Neelam Patodia (np2723)
Arusha Kelkar (ak4432)
Tanvi Gautam Pareek (tgp2108)

2019-12-12

Contents

1 Introduction

Deaths due to drug overdose is a serious issue in USA. The problem of drug abuse in USA has a long history of 48 years. The federal budget to bring the situation under control has been rising year on year with the allocation in 2018 being $27.7 billion [1].Over 63,600[2] poeple have died in 2016 due to drug overdoses which is greater than the deaths attributed to motor vehicle accidents, homicides and suicides. According to the Centres for Disease Control and Prevention, the 2016 Connecticut age-adjusted rate for drug induced mortality was 25.1 per 100,000 population compared to the 2016 national rate of 17.1[3]. This increase is mostly attributed to misuse of prescription medication, making it an issue of public health concern in Connecticut. As part of this research project we wanted to investigate the prevailing issue of drug abuse in Connecticut.

1.1 Objective and Approach

Our objective is to find out the reasons and patterns contributing to high drug abuse in Connecticut by utilizing the publicly available data on this issue. We have thus considered a three fold approach to unfold this problem. These have been described below.

1) **Demographic attributes in Drug Abuse:** We will explore all the dimensions related to demographic aspect of drug abuse. As per our initial hypothesis, drug abuse is highly correlated with the demographic attributes of population such as age, gender, race, and area. There should be an underlying pattern between drug consumption and these demographical attributes.

2) **Temporal Aspect of Drug Abuse:** This will be an orthogonal aspect of drug abuse. We will explore the temporal patterns in the data. Specifically we are looking to answer questions such as:

- What is the year by year trend in overall number of drug abuse cases from 2012 to 2018?
- Is there any seasonality in the number of cases particularly related to different seasons as winter/summer?
- Are people more likely to do drugs on certain time of the years, if so what is distribution from that angle?

3) **Causal Diagnosis for Drug Abuse:** Lastly, we investigated the underlying root cause in the form of chemicals consumed and type of injuries people suffered from.

- From this investigation, we are trying to find the most common chemicals found in the drug abuse cases and their co-consumption.
- We inquired the broader classes of drug abuse types such as ingested pills, alcohol, abuse of medication, substance abuse etc. And we will try to attribute the categories contributing to the 80% cases following the pareto principle.

By following this three fold approach, we expect to unravel significant insights bringing us closer to our quest of understanding drug abuse in Connecticut.

1.2 Individual Responsibility

1) **Member 1:** This member explored the dataset in terms of sparsity and missing values. Description of each column of the dataset and how these are associated with our broader objective of the project. She also worked on the sampling procedure and on the interactive visualization.

2) **Member 2:** This member looked into the relation of the number of drug abuse cases to the time when it occured/reported. She explored the relation between the year, the time of the year when the death happened and the number of deaths during the duration.

3) **Member 3:** This member explored the relation between the region where the person died due to drug abuse and the number of deaths that occured in the regions. Also explored the different demographic attributes as age, race, gender associated with the drug abuse cases.

4) **Member 4:** The member explored the exact cause of death i.e. the drug responsible for death, the description of the injury and the number of deaths that occured due to that particular reason.

2 Data Exploration

2.1 Data Source

We have procured the data on accidental drug related deaths for the years 2012-2018 from the United States Government's official data repository. Data is derived from an investigation by the Office of the Chief Medical Examiner which includes the toxicity report, death certificate, as well as a scene investigation. (data source is: https://catalog.data.gov/dataset/accidental-drug-related-deaths-january-2012-sept-2015)

2.2 Data Variables and Description :

The data includes various parameters and we have associated these parameters to the 3 fold approach we highlighed above in order to understand which fields could be utilized for the respective analysis. For the sake of making variable names easily distinguishable, we have put the variable names in quotes.

1) **Temporal Fields**

- 'Date': Date when the accidental death happened or was reported.

2) **Demographics Fields**

- 'Sex', 'Age', 'Race': Sex, age, race of the drug addict person.
- 'ResidenceCity', 'ResidenceState', 'ResidenceCounty': This dentotes the city/state/county of residence place of the drug addict person. We have similar fields for the 'death' city/state/county for each case.
- 'DeathCityGeo', 'ResidenceCityGeo', 'InjuryCityGeo': These fields provide the latitude and longitudinal of the death case.
- 'Location': This field denotes the place of death such as hospital or residence.

3) **Causal Fields**

- 'COD' which denotes cause of death and has all the chemicals consumed by the victim (separated by commas)
- A Y/N column for each of the chemicals causing death of the person "Heroin", "Cocaine", "Fentanyl" and other such 14 chemicals in total
- 'AnyOpioid' to denote if opiod was consumed or not.
- 'DescriptionOfInjury' which denotes how the drug abuse took place. Eg. via substance abuse or injection etc.
- 'MannerofDeath' denotes the manner in which death occured

The data spans over a period of 7 years from 2012-2018 and has 41 variables at our disposal. There are 5,105 observations in our dataset which provided sufficient scope to proceed with our investigation.

```
##
##  Number of rows in the original dataframe:  5105
##  Number of columns in the original dataframe:  41
##
```

2.3 Data Challenges and Resolution

1) We wanted to identify the type of injury and the initial driver in the case of drug abuse as this information would be very crucial for the overall analysis of drug abuse and the root cause underlying the same. Data contained a column "DescriptionOfInjury" but this had freely written text in plain english language which was challenging to categorize in fixed set of classes.

In order to overcome this challenge, we parsed the freely written text in the form of tokens. Later on we derived the token frequency from all drug abuse cases. This token frequency helped us to compute the most frequent underlying root cause and type of injury associated with that.

2) "DateType" in the dataset informed us when did the drug abuse case occur. However, it led to some ambiguity because it had two values : "dateReported" and "dateOfDeath".

This led to some gap in both dates for each case. Hence we aggregated the numbers for different time range and it was thus an approximate value near the boundary time range.

2.4 Data Cleaning

The data had missing values which were treated in a pair-wise manner for each of our analysis. For free textual data, we resorted to natural language processing techniques to clean the data. In particular we observed the following anomaly:

- The column "DeathCounty" had "USA" as one of entries, which we cleaned during data cleaning.
- Their is a number entry for column "DeathCity" in one of the drug abuse cases

2.5 Missing Data Exploration

We began by exploring the missing values in the dataset. This was crucial as it guided us towards assessing the usability of the data i.e if a column had too many missing values it would be better to not use it. It further guided us on building the interaction variables i.e we assured that the columns involved in the analysis have sufficient values.

Observations : We observed the following:

1) Data related to demographics such as "Age", "Sex", "Location", "InjuryPlace", "MannerofDeath" have less than 1% of missing values and can be used for the purpose of our analysis.

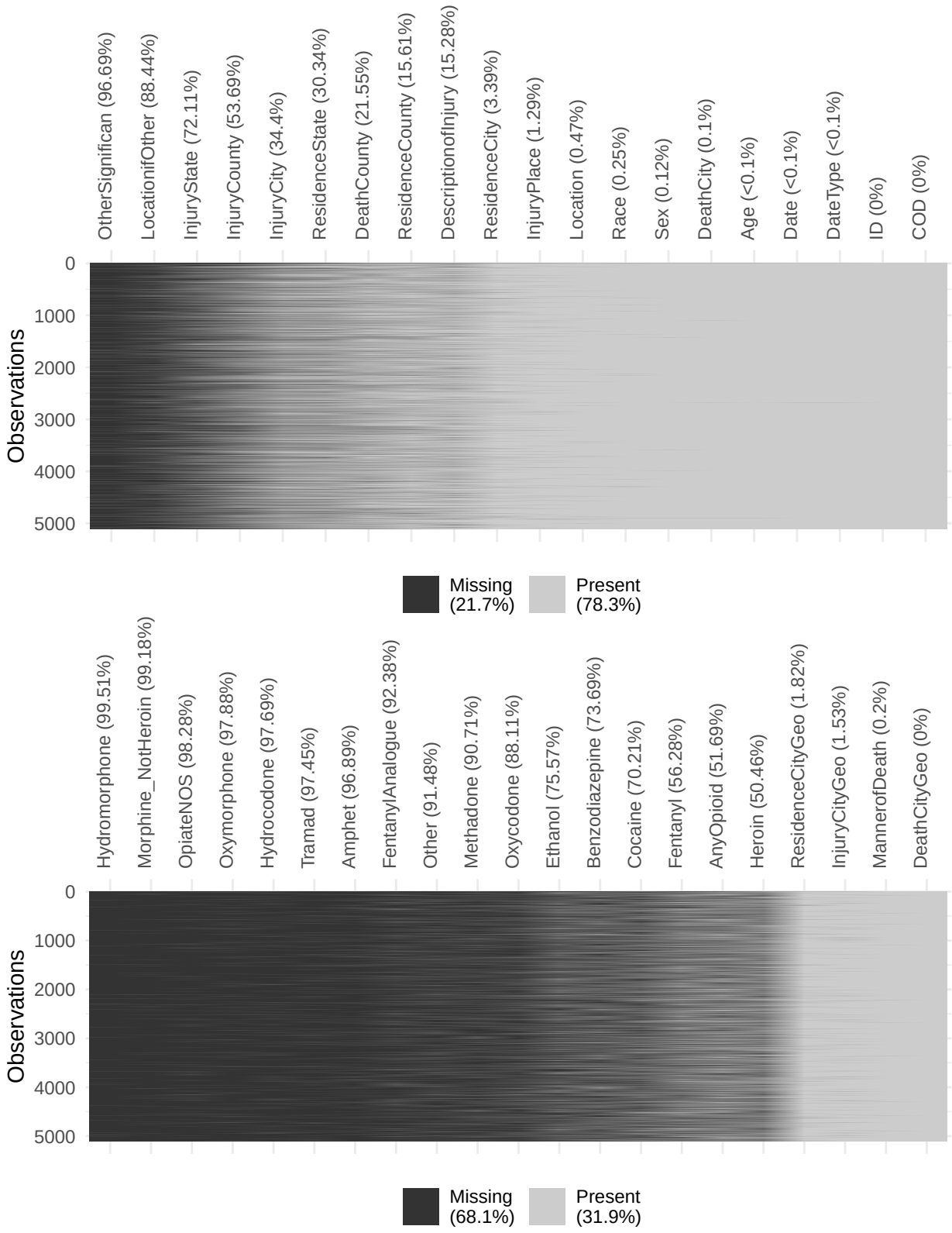
2) Time stamps i.e "Date" and "DateType" have been reported for nearly all candidates thus enabling us to perform trend analysis

3) Spatial Data parameters such as "InjuryState", "InjuryCounty", "ResidenceState", "ResidenceCounty" seem to have more than 15% missing data and goes as high as 72%. We could thus utilize other data parameter: 'DeathCityGeo', 'InjuryCityGeo' and 'ResidenceCityGeo' which have less than 2% missing data.

4) The column 'Description of Injury' has ~15% missing data but could still be utilized as it conveyed important information corresponding to scenarios of death due to drug overdose. We thus conducted an independent analysis of this column using a word cloud and the data seemed sufficient.

5) Though there seemed to be a lot of missing values corresponding to the individual drugs, it was completely okay as it signified whether a candidate consumed that particular drug or not. A blank value here had been assumed to signify that drug was not taken.

The data thus seemed to be usable for our analysis.



3 Demographical Aspect of Drug Abuse

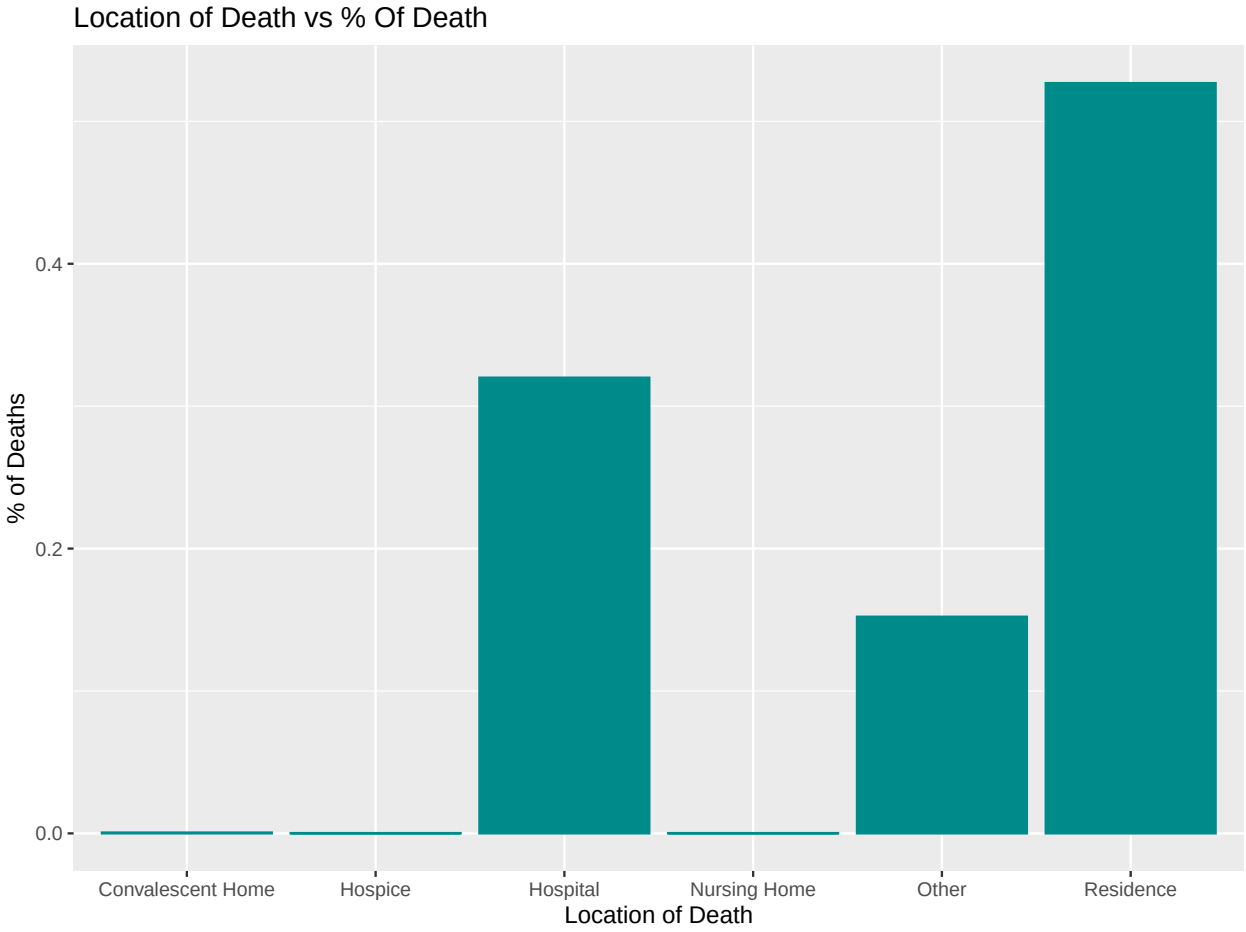
3.1 Assessing the distribution of fields

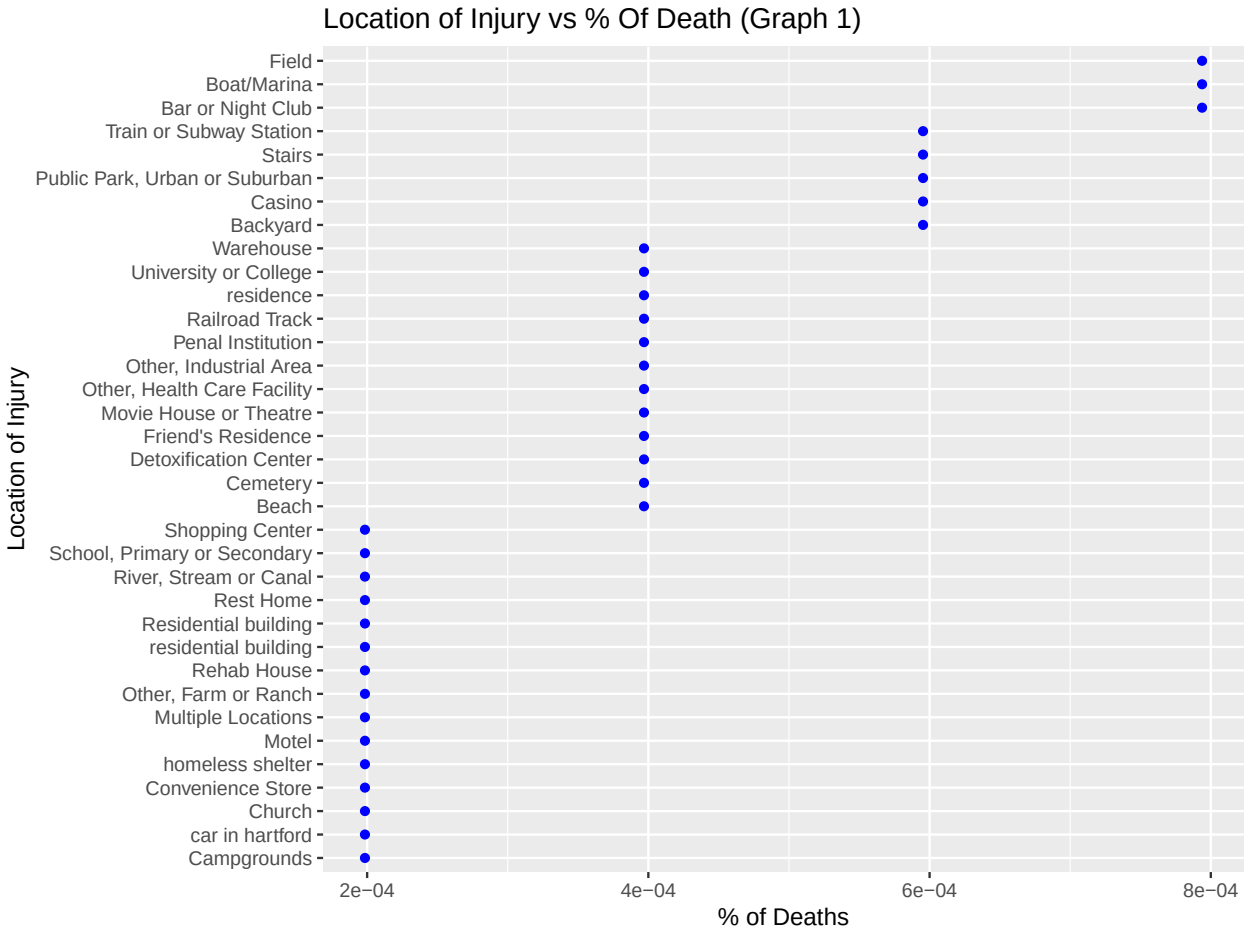
Once we identified the parameters containing sufficient data, we assessed the distribution of each of these variables i.e we visualized the spread of numeric data, the frequency of the categories within a column, the proportion of death along various axes etc. This paved the way in identifying the parameters which needed to be further explored for truly unravelling drug abuse situation here.

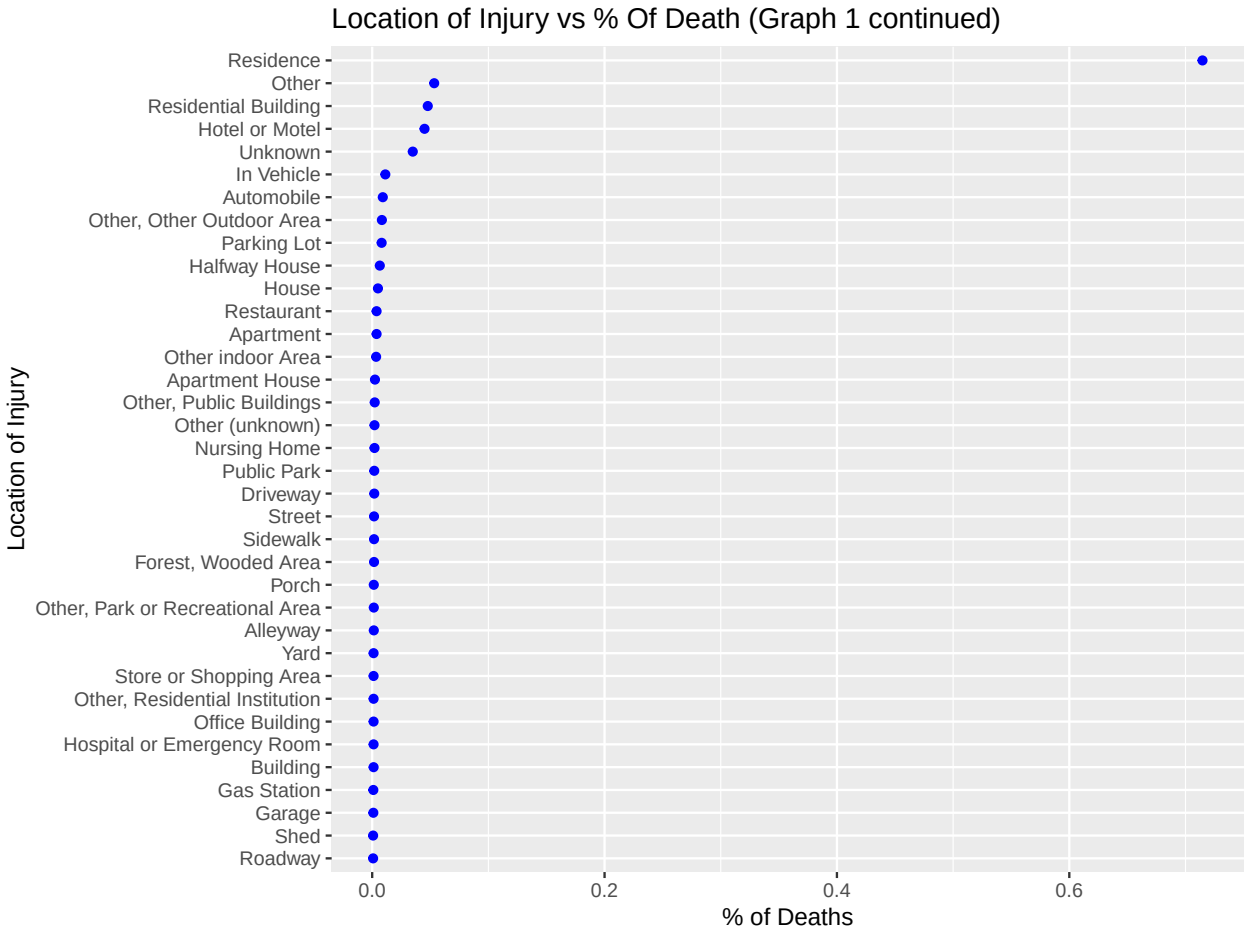
1) **Location of Injury and Location of Death**

The data consisted of information pertaining to both Location of Injury and Location of Death. We wanted to understand if the location of injury was same as that of death, which could have potentially suggested that post drug abuse death might be instantaneous as people did not have time to relocate. Here, we observed that majority of death occured in residence followed by that in "hospitals". However, majority of injuries also occured at "Residence" potentially suggesting that most injuries and deaths occured at Residence. This probed us to further investigate:

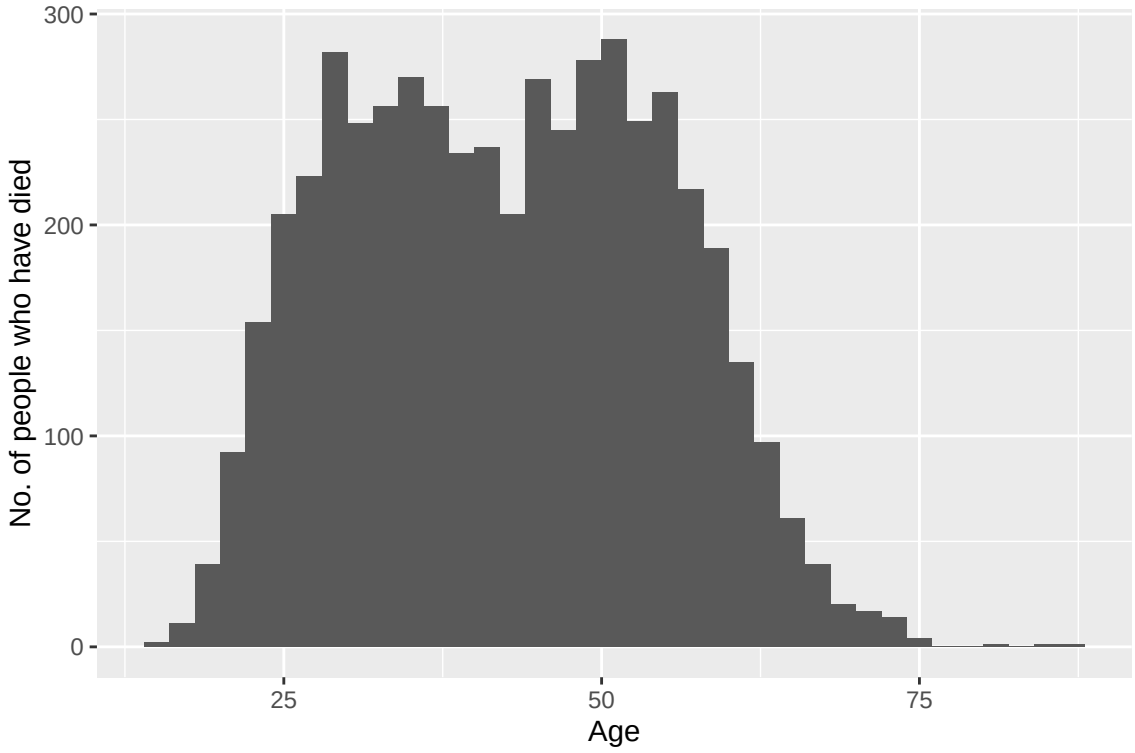
- Which cities/states have more "Residence" reporting death due to drugs?




Location of Injury vs % Of Death (Graph 1)
Location of Injury
Field
Boat/Marina
Bar or Night Club
Train or Subway Station
Stairs
Public Park, Urban or Suburban
Casino
Backyard
Warehouse
University or College
residence
Railroad Track
Penal Institution
Other, Industrial Area
Other, Health Care Facility
Movie House or Theatre
Friend's Residence
Detoxification Center
Cemetery
Beach
Shopping Center
School, Primary or Secondary
River, Stream or Canal
Rest Home
Residential building
residential building
Rehab House
Other, Farm or Ranch
Multiple Locations
Motel
homeless shelter
Convenience Store
Church
car in hartford
Campgrounds
2e−04
4e−04
6e−04
8e−04
% of Deaths



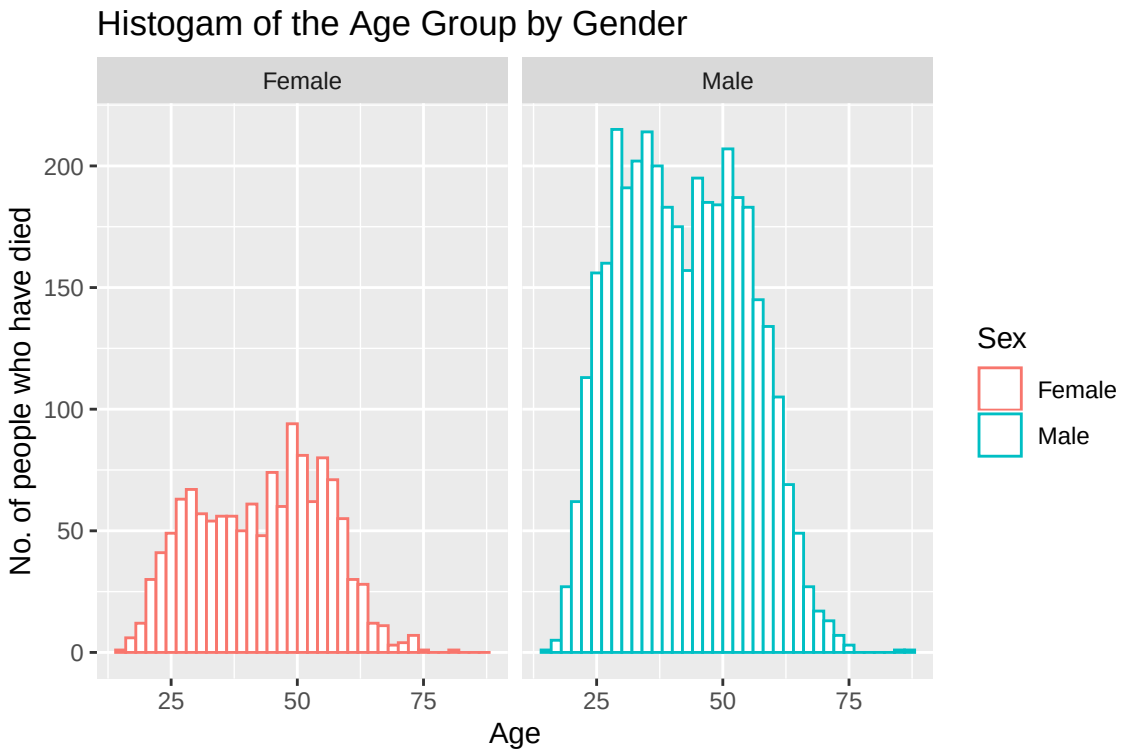
2) **Age** : Since age was the only continuous variable available here, we proceeded by analyzing its distribution across other categorial variables.The overall data appeared to be bimodal at the ages ~28-30 and ~50 years. The plot appeared to be skewed to the right




```
## $title
## [1] "Histogam of the Age Group"
##
## attr(,"class")
## [1] "labels"
```

3) **Sex**

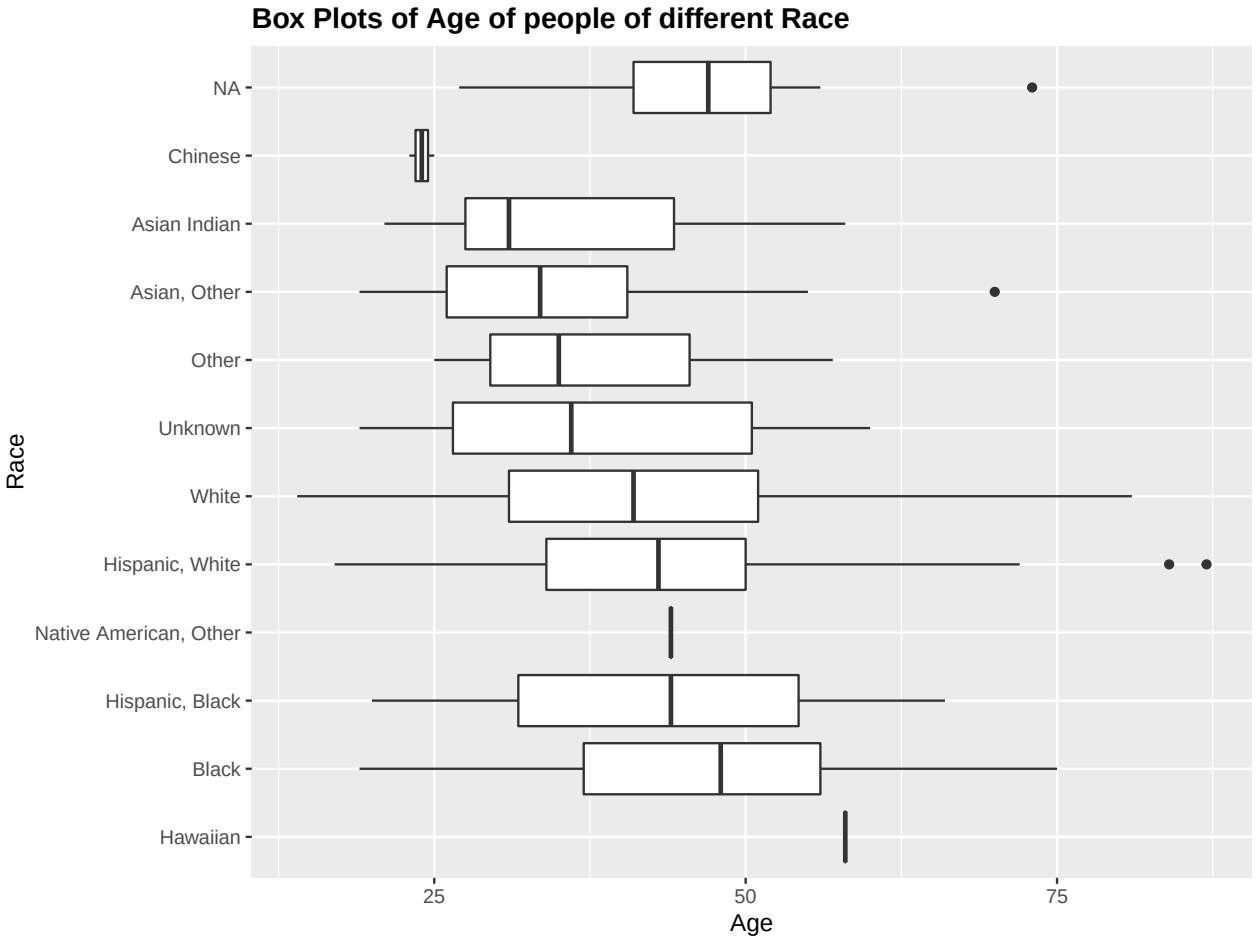
- The number of males who died due to drug overdose was more than twice that of females.
- The median age of females who died due to drug abuse was more than males

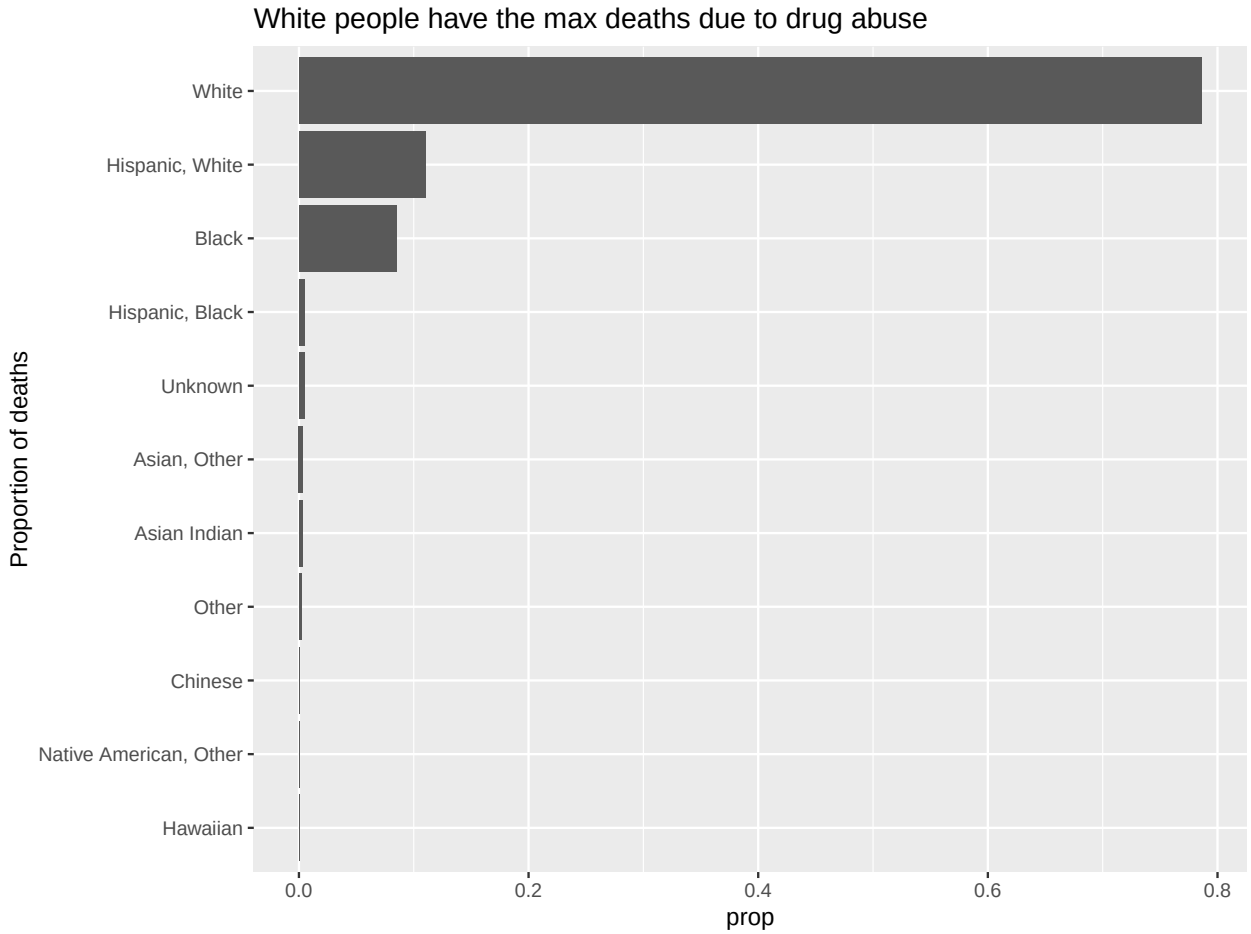

Histogam of the Age Group by Gender
Female
Male
No. of people who have died
200
150
100
50
0
25
50
75
Age
Sex
Female
Male


Box Plots of Age of people by Gender
Race
Male
Female
25
50
75
Age

4) **Race**

- Most number of people who died due to drug abuse are "White" followed by "Hispanic, White" and "Black".
- The median age of death amongst "Chinese" was the lowest while that of "Black" is the highest.



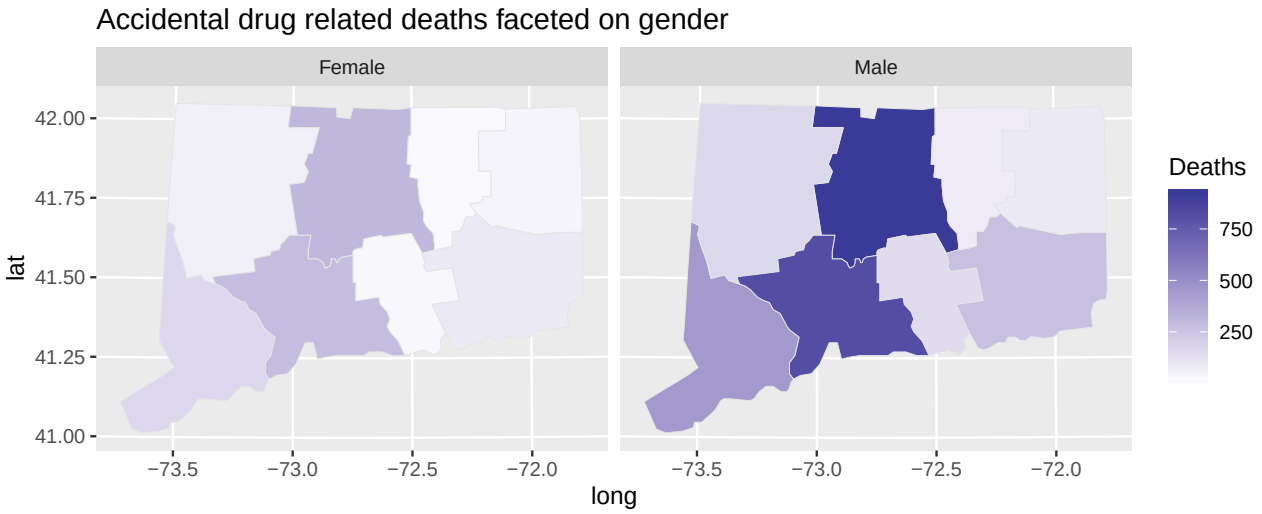


3.2 Spatial Analysis

1) **Sex** Gender based drug abuse has always been a topic of discussion when it comes to analysis regarding drugs. To find the corresponding insights for our dataset and county based distribution of the same lead us to plot choropleth plots for our analysis.

The choropleth plot below represents the number of accidental deaths due to drug abuse in the state of Connecticut according to its various counties.

- The number of males who died due to drug abuse are on the whole more than the number of females who died due to drug abuse for the years 2012 to 2018 which can be inferred from the fact that the choropleth plot for males has darker shades of purple than the choropleth plot for females.
- As it can be inferred from the plot, the county of Hartford has more number of deaths due to drug abuse as compared to other counties for males as well as females.
- The number of deaths in the counties of Middlesex and Tolland are very less as compared to other counties for males as well as females.

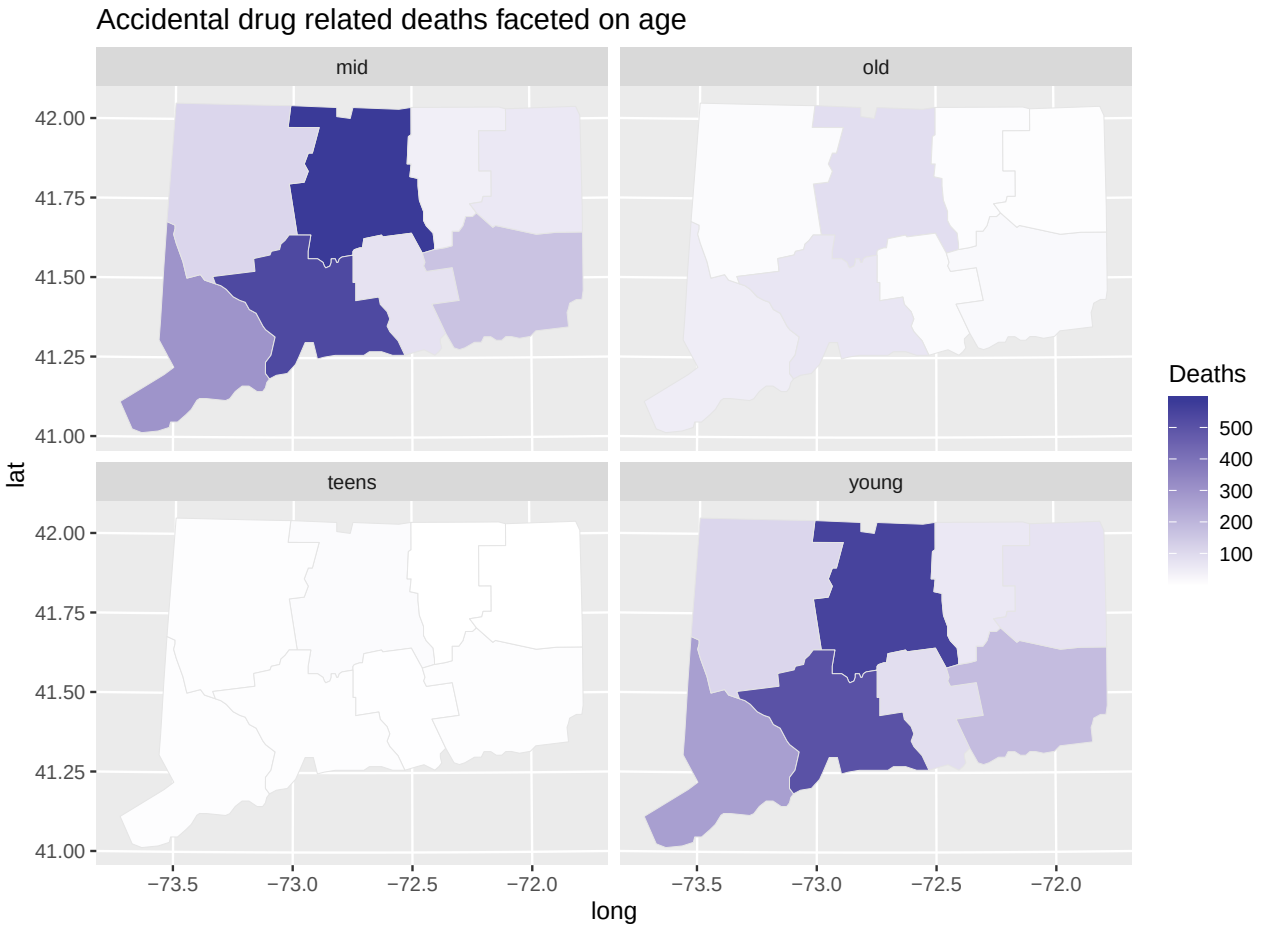


2) **Age**

Research regarding the age groups involved in drugs and their effects has been of interest since decades.Getting county wise insights regarding the age group of people who died due to drug abuse being our primary aim lead to us plotting choropleth plot faceted on age.

We wanted to explore the distribution of the number of deaths due to drug abuse across the various age groups over the different counties of the state of Connecticut. The below choropleth plot represents the number of deaths due to drug abuse for the different age groups with teens being people with age less than 19, young being people with age less than 40, mid being people with age less than 60 and old being people with age greater than 60.

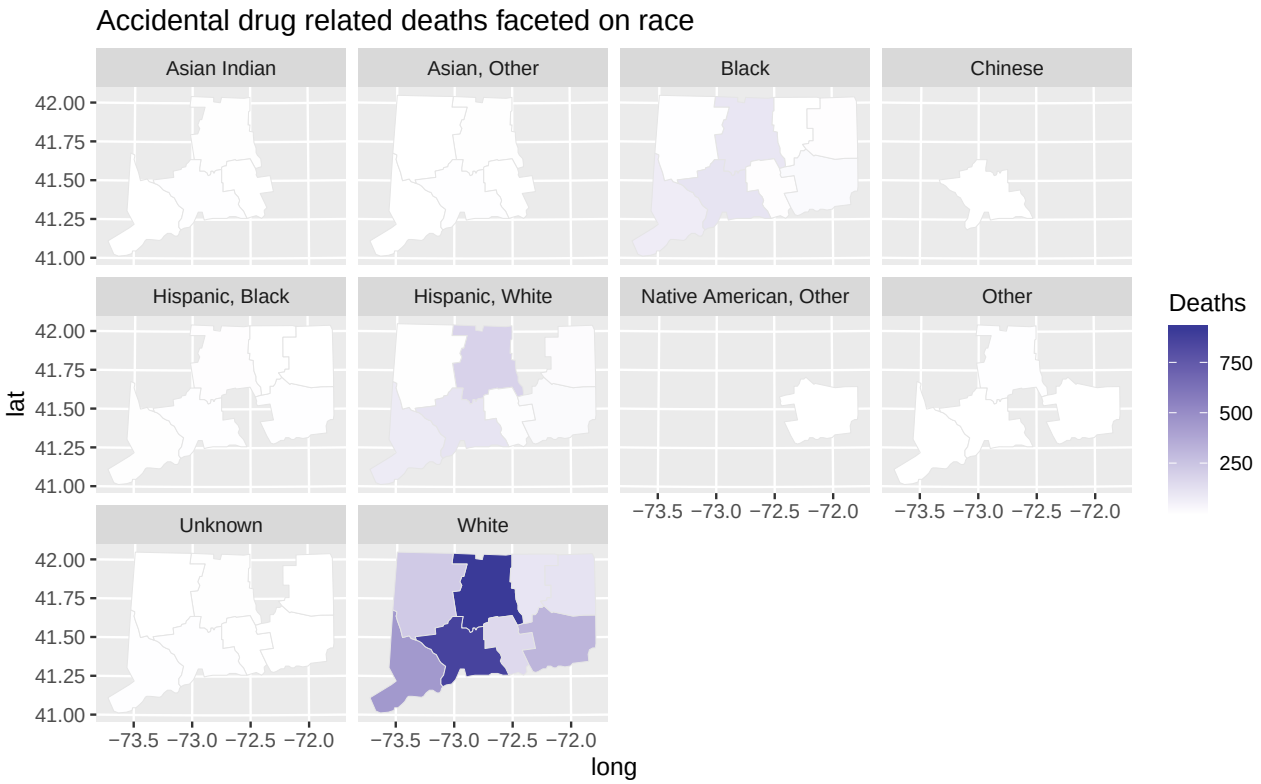
- There are very few teens who have died due to drug abuse in the state of connecticut.
- There are very few old people who have died of drug abuse in the state.
- Maximum number of people who died due to drug abuse were people between the age group of 20 to 60.
- Similarly as above, Hartford county has the maximum number of deaths due to drugs for mid and young aged people from the year 2012 to 2018.



3) **Race**

We wanted to identify the distribution of the race of the people who died due to drug abuse. We have plotted a choropleth plot for the death count for different ethnicity values.

- The ethnicity of the people who died due to drug abuse was mostly "white".
- There weren't any Chinese people who died due to drug abuse in the counties except in county of New Haven.
- The counties of Hartford, New Haven, Middlesex and Fairfield had hardly any Asian Indian people who died due to drug abuse.
- The counties of Hartford, New Haven and Fairfield had quite a few people of the black and hispanic white race who died due to drug abuse.

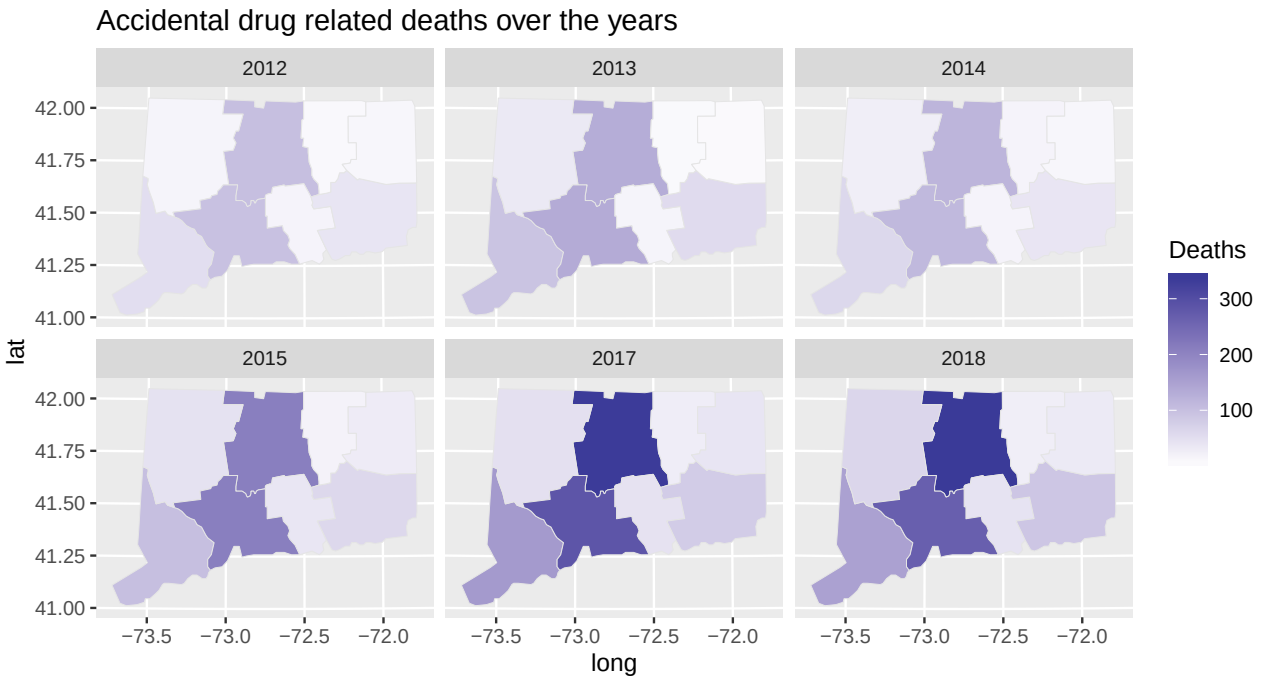


4) **Years 2012-2018**

We wanted to identify the trends of deaths due to drugs over the years in the different counties of Connecticut. So, choropleth plot representing the number of deaths over the years was plotted.

- After exploration, it can be inferred that the number of deaths due to drugs have increased over the years with the highest being in the years 2017 and 2018 across all the counties of Connecticut.

Note:The 'deathcounty' for the deaths in 2016 is not available

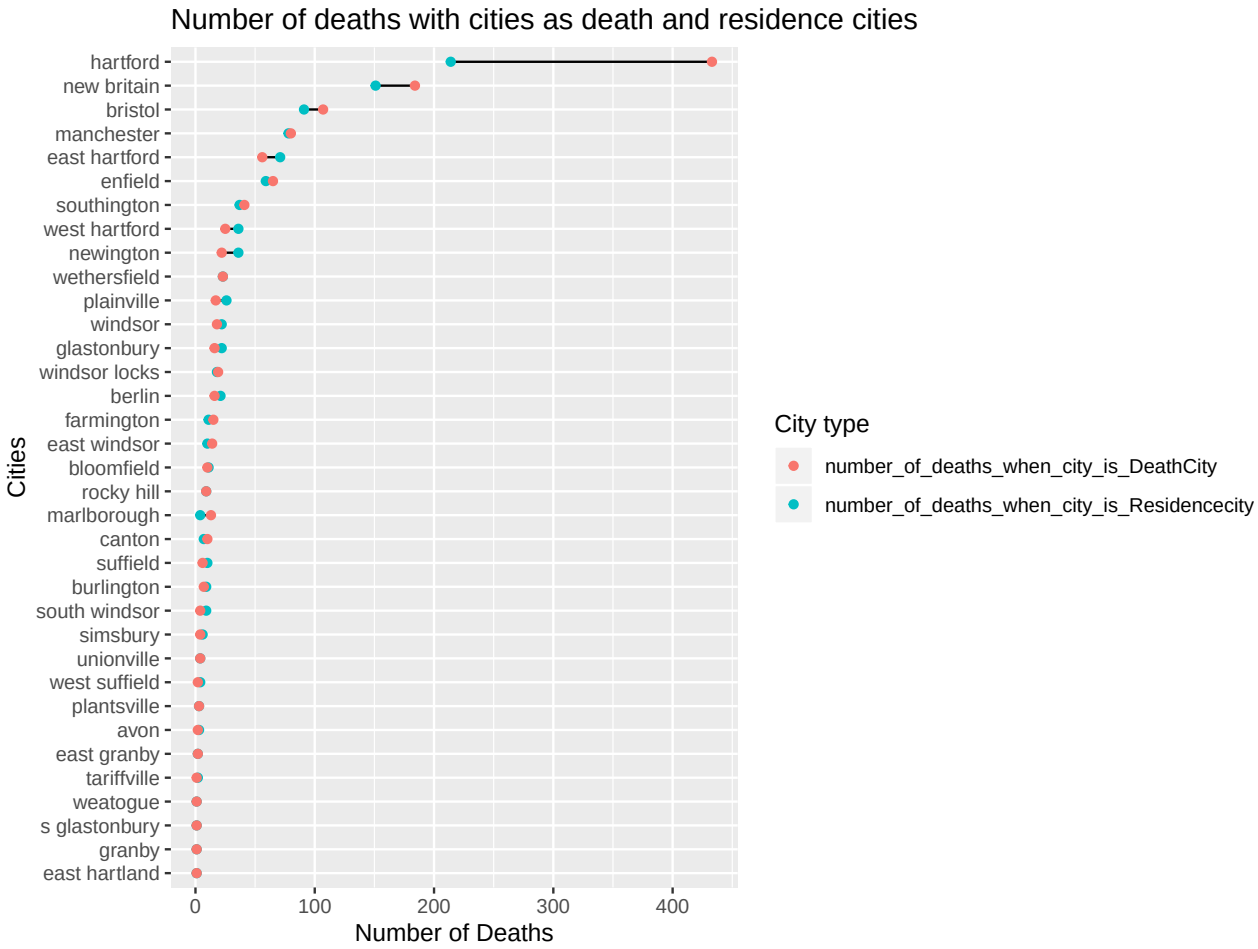


The "deathcounty" for the deaths in 2016 is not available

5) **Investigating Hartford**

As per our previous exploration, we found out that Hartford had the highest number of deaths due to drug abuse amongst all the counties of the state of Connecticut. Therefore,we decided to explore the deaths in the cities of Hartford due to drugs. The below cleveland plot shows the number of deaths in the cities with the death city being that city and the residence of the person who died being that city.

- The city of Hartford in the Hartford county had the highest number of deaths with Hartford being the death city and the highest number of deaths with Hartford being the residence city of the person.
- For the city of Hartford, the number of deaths when the city is the death city is much more than the number of deaths when the person is a resident of Hartford city.
- Similarly for the cities of Great Britain and Bristol being the second and third ranked cities with the maximum number of deaths due to drugs in the county of Hartford.
- For the cities of East Hartford, West Hartford and Newington,the number of deaths with the cities being the resident cities of people was more than the deaths occuring in the cities,which implies that more people who were residents of these cities died in a different city.

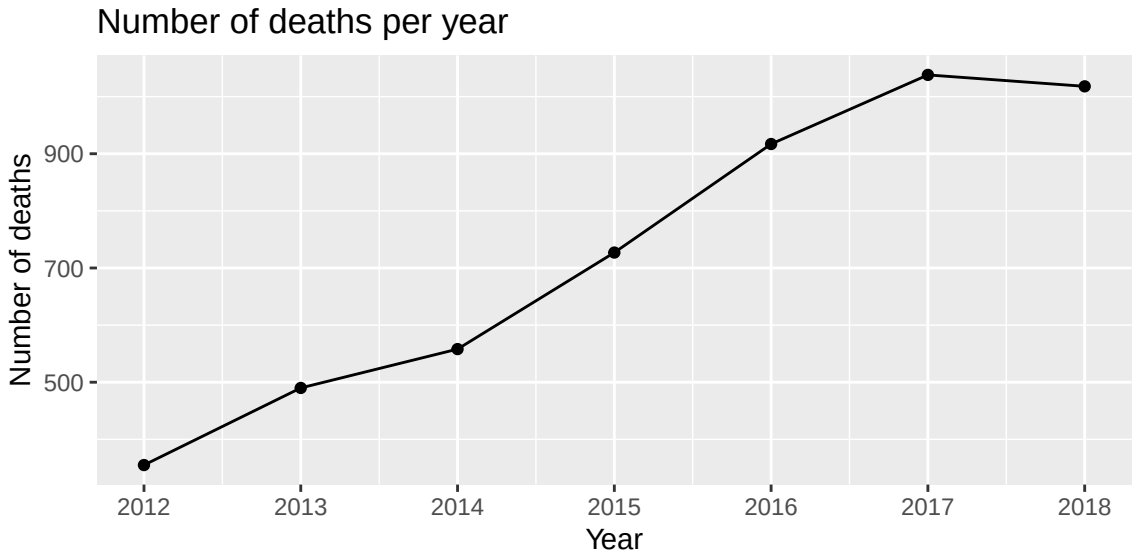


4 Temporal Aspect of Drug Abuse

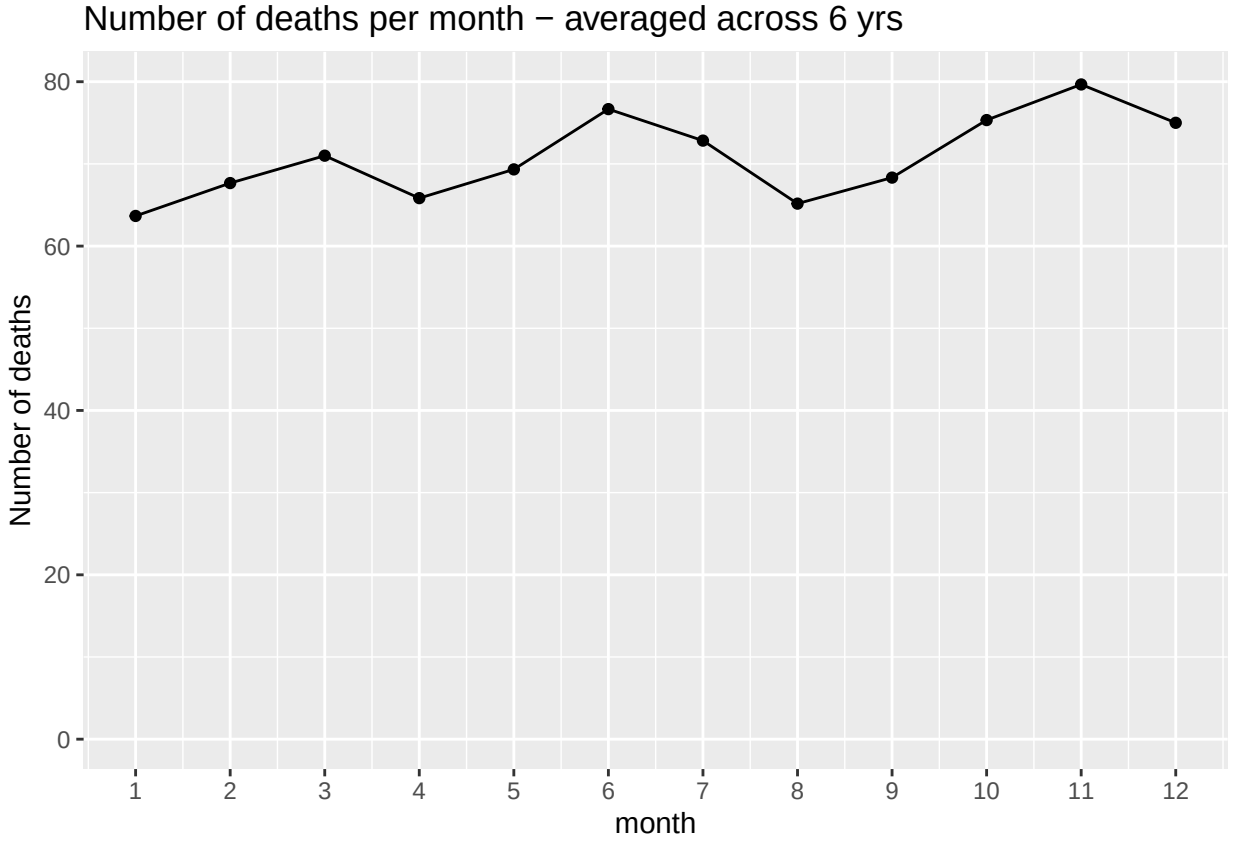
We explored the patterns of drug abuse with respect to time variables here.

4.1 Seasonality

We thus noted the number of deaths due to drug abuse reported each year from 2012-2018. We observed that in 2017 maximum number of deaths were reported, followed by in 2018.

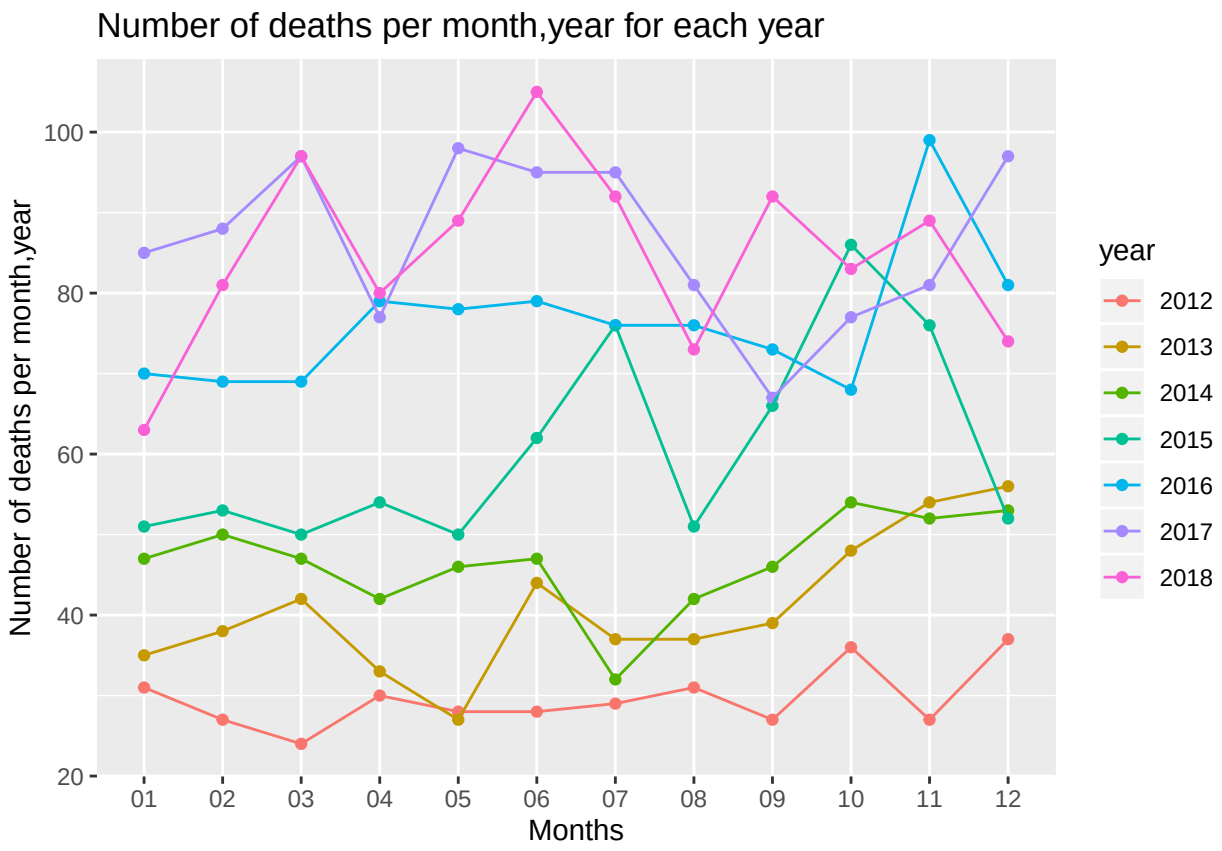


+ On plotting the average number of deaths each month for the last 6 years, we observed that on an average the number of deaths increased from January to March and it rose again in August through November.

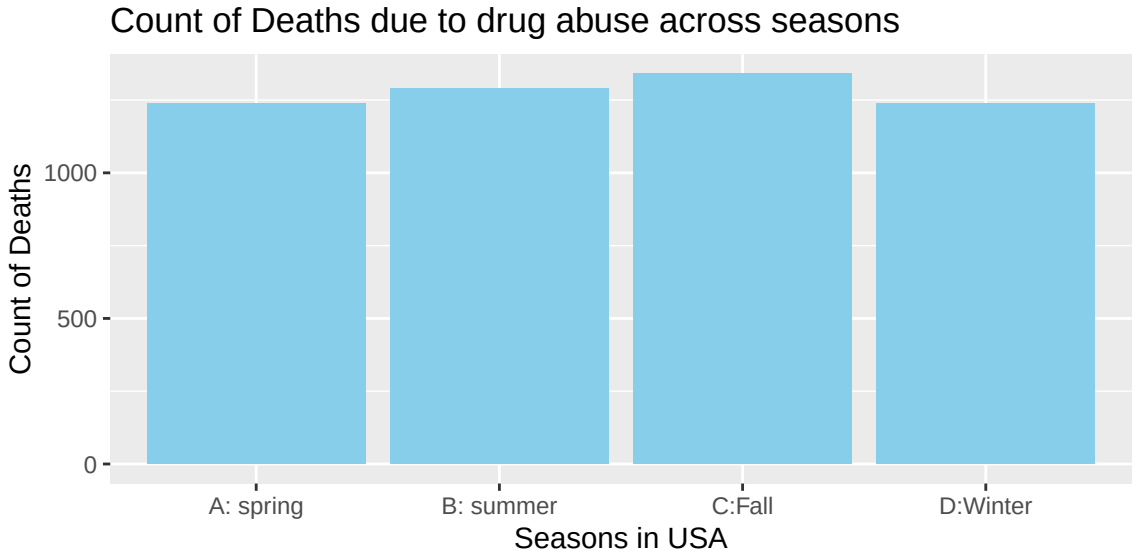


- Specifically in 2017, we observed that number of deaths have risen steeply since September till the end of the year and this rise continued for the 1st three months of 2018 as well. This suggested that during Fall 2017 and Winter 2018 deaths was on a rise which could be related to the cold weather conditions triggering people to consume more drugs and therefore resulted in more deaths. A similar pattern was also observed for the years 2013 and 2014. However, we do not have the required data variables in

this dataset to confirm our assumptions. Additionally, we observed that the highest death in 2018 had occured in June.



- The number of deaths are highest in Fall.

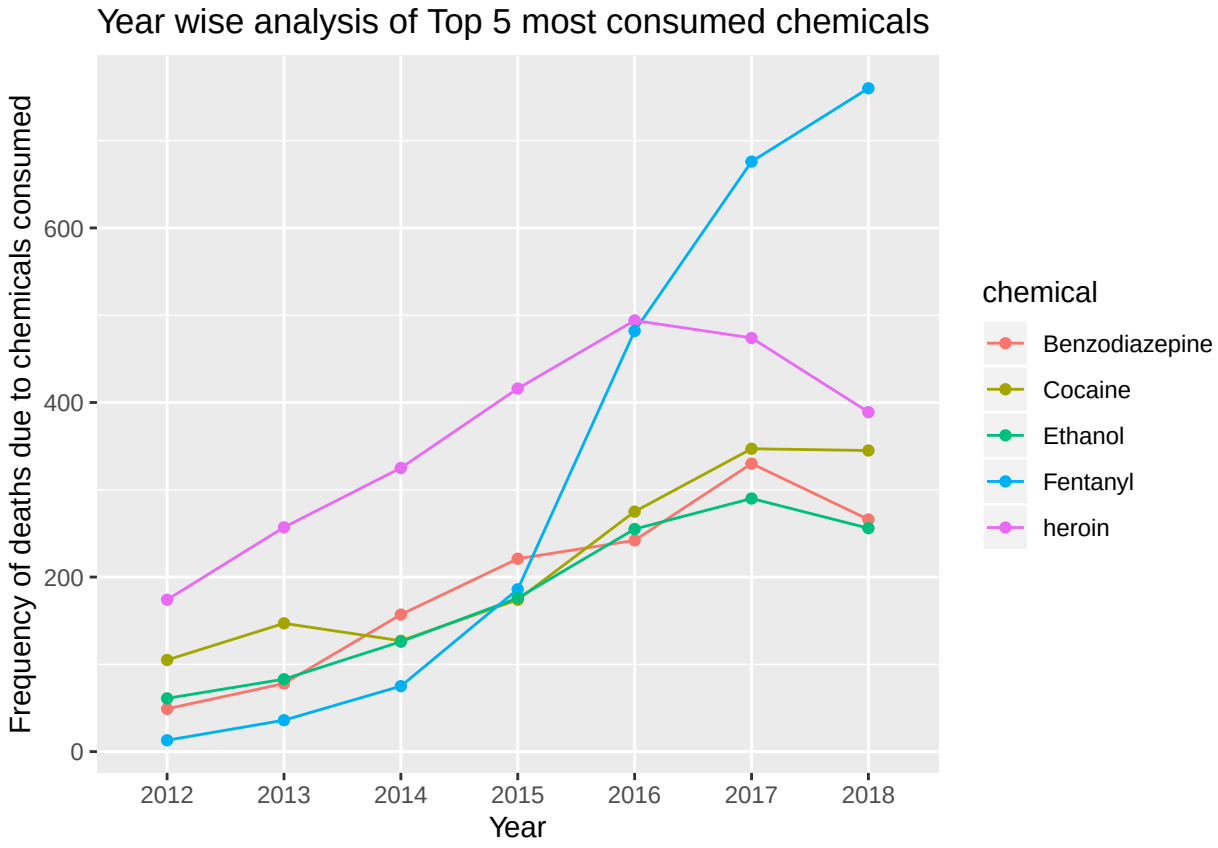


4.2 Patterns in Drug consumption over Years

In order to further understand the factors associated with increased deaths due to drugs year over year, we investigated the chemical compounds/drugs which were consumed during these years and were responsible for the deaths. We thus observe that:

- 'Fentanyl' which used to be the least consumed drug during 2012-2014, has been on a rise since 2015 onwards and surpassed the consumption of heroin in 2016. Rather 'Heroin' consumption has been decreasing post 2016 and it seems it is being canabalized by 'Fentanyl'.
- The consumption of other drugs has also been rising steadily.

```
##    all_chemicals Freq
## 7         Heroin 2529
## 5       Fentanyl 2232
## 3        Cocaine 1521
## 2 Benzodiazepine 1343
## 4        Ethanol 1247
```



5 Causal Diagnosis of Drug Abuse

We wanted to analyze the drug abuse cases in order to highlight the major contributing factors. We investigated the underlying root causes: major chemicals consumed, type of injuries people suffered from. From this investigation, we obtained the most common chemicals found in the drug abuse cases. Moreover,

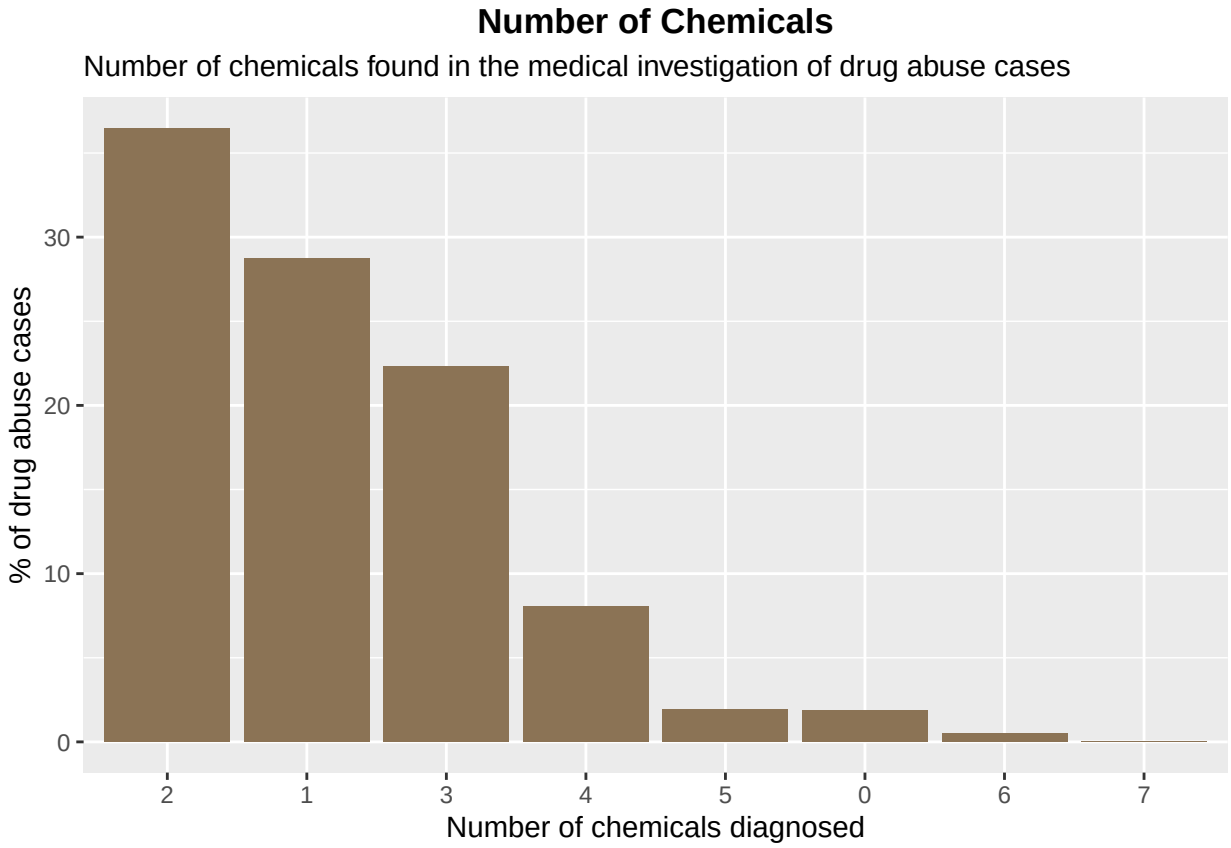
we inquired the forms of drug abuse such as ingested pills, alcohol, abuse of medication, substance abuse etc. And we also attributed the categories contributing to the most of the cases.

5.1 Derived Metric Calculation

In medical diagnosis of durg abuse cases, there are generally 14 types of chemicals found - Heroin, Cocaine, Fentanyl, FentanylAnalogue, Oxycodone, Oxymorphone, Ethanol, Hydrocodone, Benzodiazepine, Methadone, Amphet, Tramad, Morphine_NotHeroin, Hydromorphone. In this section we calculated the derived metrics as "diagnosed_chemicals" and "diagnosed_chemicals_count". Metric "diagnosed_chemicals" represents the list of chemicals found in each case of drug abuse, essentially in each row in the dataframe. Similarly, metric "diagnosed_chemicals_count" denoted the number of chemical found in each drug abuse case out of exhaustive list of 14 chemicals mentioned here. Both of these derived metrics will be explored further in the upcoming sections.

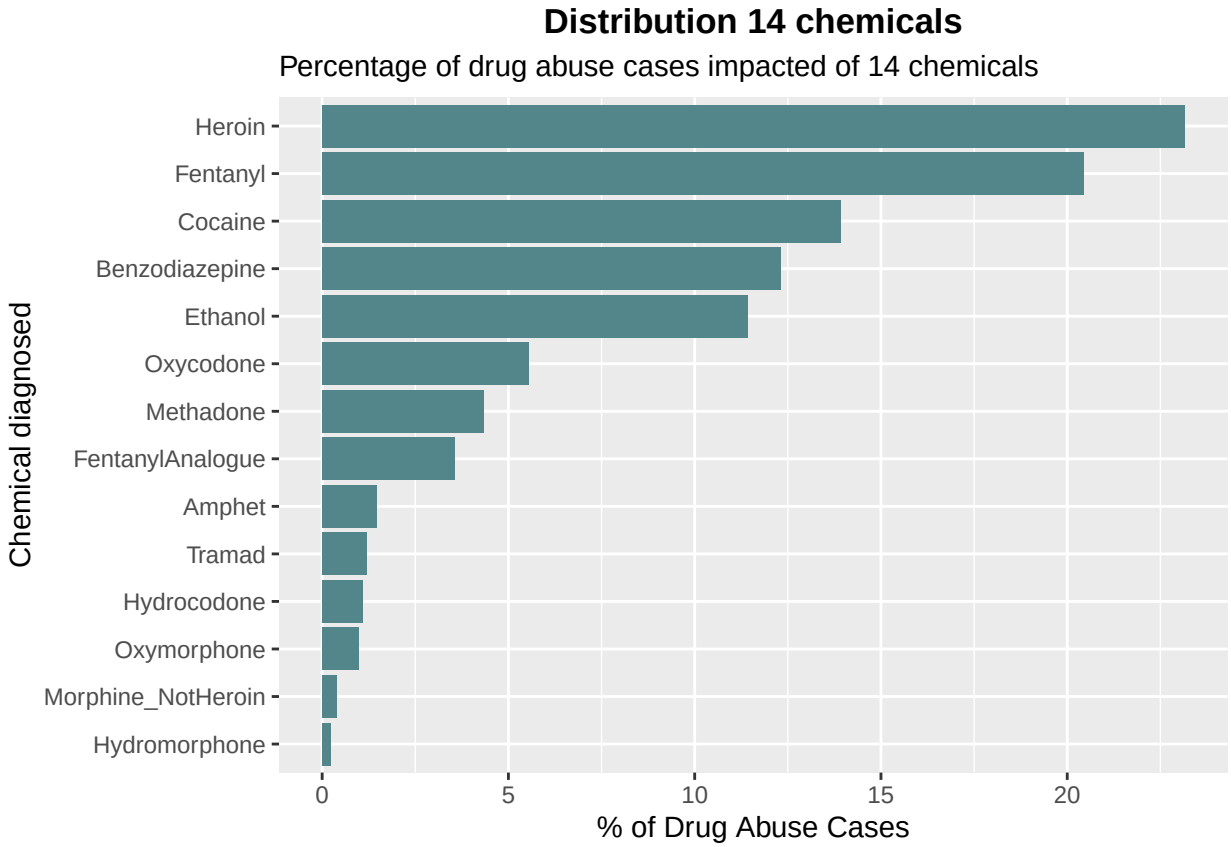
```
##                   Diagnosed_Chemicals diagnosed_chemicals_count
## 1 Fentanyl,Hydrocodone,Benzodiazepine                          3
## 2                              Cocaine                         1
## 3                       Heroin,Cocaine                         2
## 4                      Heroin,Fentanyl                         2
## 5                             Fentanyl                         1
## 6                               Heroin                         1
```

To further explore the derived metric "diagnosed_chemicals_count", we plotted the below histogram. This histogram represents the distribution of number of chemicals found in drug abuse cases. We noticed that generally there were 2 chemicals found in majority of drug abuse cases. In more than 37% cases, there were only 2 underlying causes.



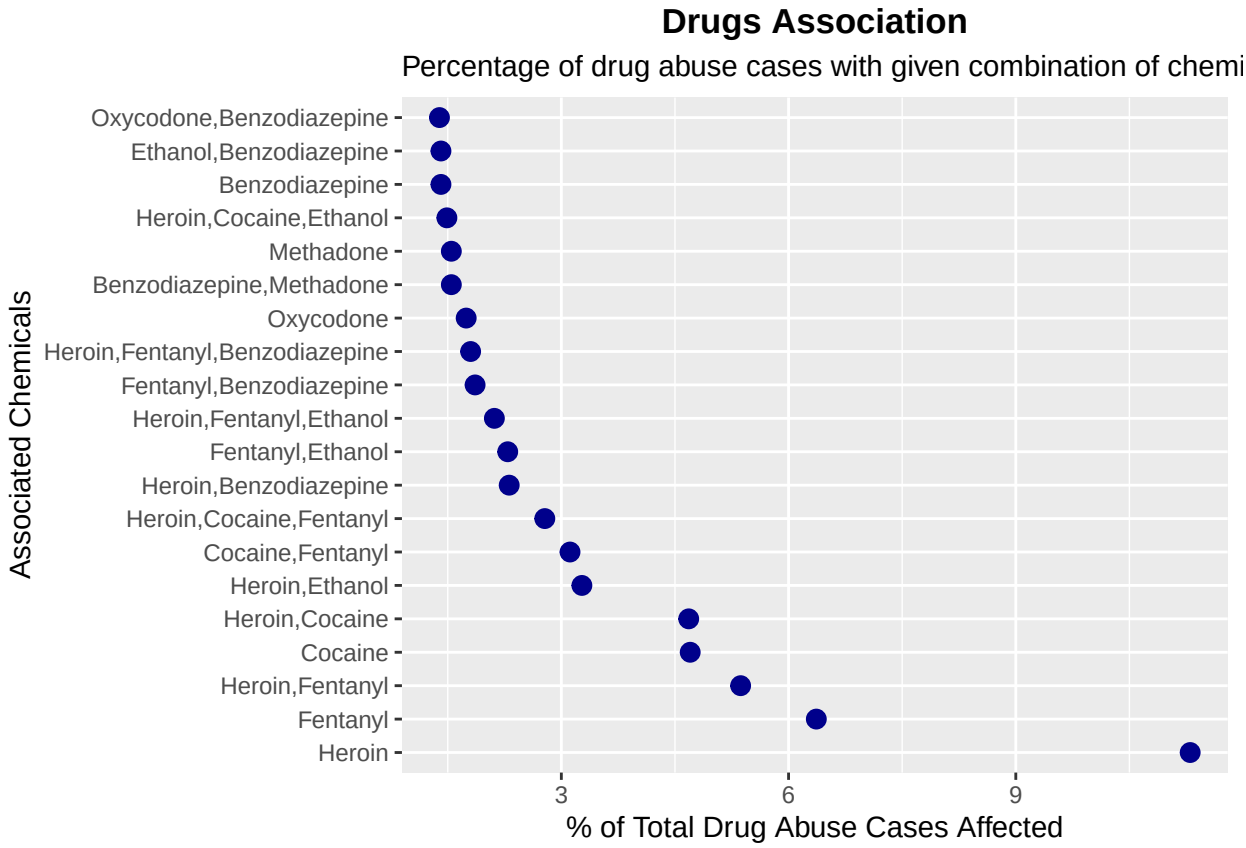
5.2 Distribution of 14 chemicals in Drug abuse Cases

Here we investigated the chemicals found in the medical investigation of drug abuse cases. We plotted the % of drug abuses cases where each of the chemical were diagnosed. Three major Chemicals - 'Heroin', 'Fentanyl', and 'Cocaine' occured in 47% of drug abuse cases. Rest of the underlying chemicals were relatively less frequent.



5.3 Association mining among 14 Chemicals

We wanted to further mine the association between these 14 chemicals - which combination of chemicals occured most frequently. Among all combinations that occured together, 'Heroin' with 'Fentanyl'(5%); 'Heroin' with 'Cocaine'(4.7%) and 'Heroin' with 'Ethanol'(3.3%) were the frequently found combinations. Hence these three were popular associations discovered. As Heroine was present in all the popular associations, it was concluded to be the most frequently consumed chemical.



5.4 Analysis of Injury Description - Text Mining using NLP

In this section, we explored the injury description for all drug abuse cases. The injury description("DescriptionofInjury") was basically the broader categorization of the mode/form of drug consumed in that particular drug abuse case. This field contains the freely written text hence we neeed to perform the basic natural language processing.

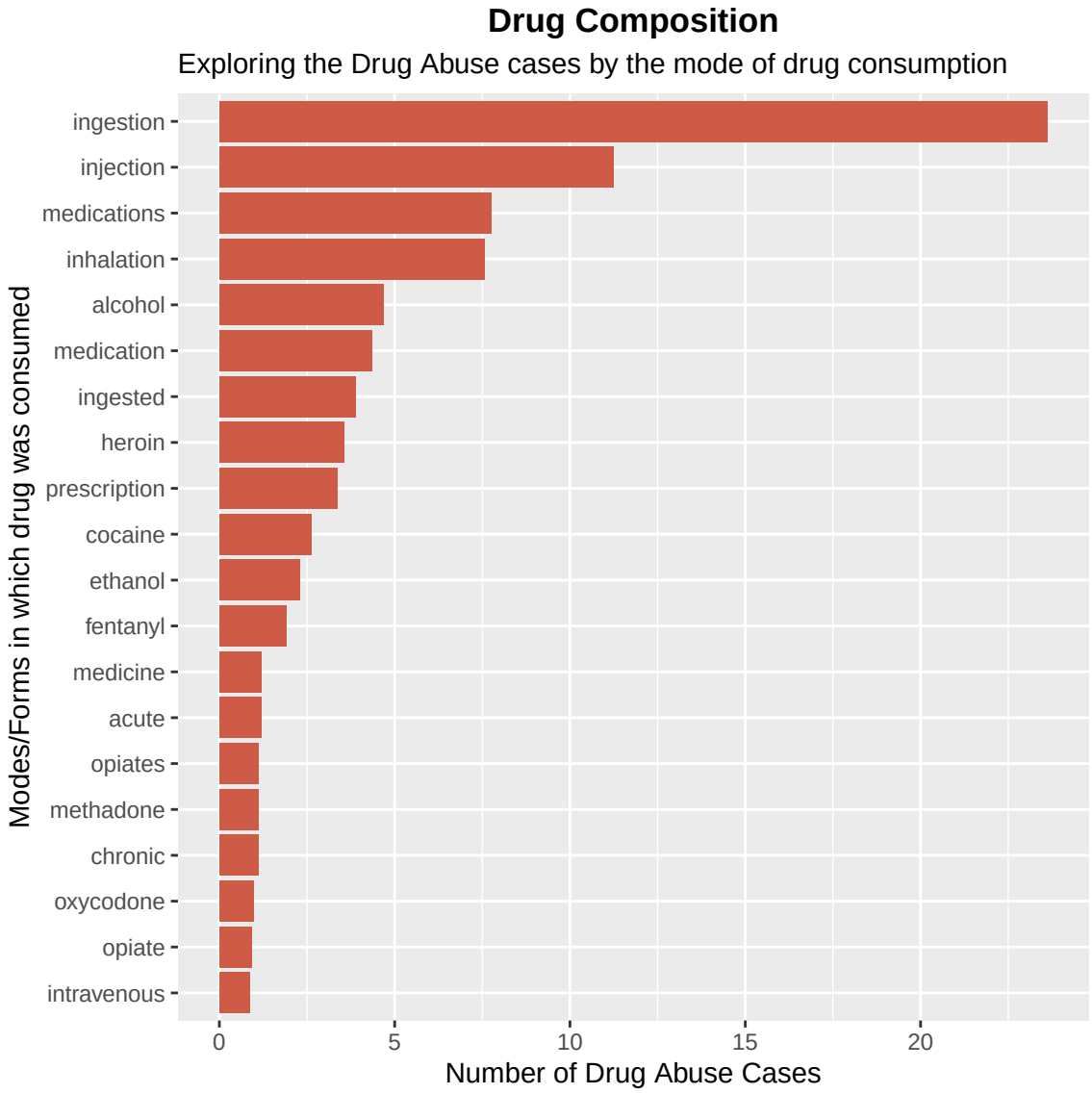
Below are the following steps performed by us:

1) Tokenization: We are first breaking the description of every drug abuse case in tokens.

2) Stop Word removal : From these tokens we majorly looked for the modes and forms in which drugs were consumed. Hence, tokens with noun and verb POS tag (Parts of Speech) appeared here, were our candidate tokens. In order to extract these candidate tokens, we removed the stop words that appeared in the token list.

First we used the stop word removal step to remove all the stop words and the tokens with verbs and nouns POS tag from the token list. We called these filtered tokens as drug consumption modes. We calculated the frequency of these modes.

Insights: Ingestion, Injection and medications were the most frequent modes of drug consumption: Ingestion in 23% drug abuse cases, injection with 11% cases and medications in 7.7% cases. Together these three modes accounted for the 42% of total drug abuse cases.



Further we built the word cloud of the description of injuries.

6 Interactive Component

The objective of the interactive component is to see the number of deaths due to drug abuse in the state of Connecticut across various parameters **year,months,race,age and gender**.

6.1 Tools functionalities

- Year is a slider which takes values from 2012 to 2018.
- Months is a slider which takes values from 1 to 12.
- Race is a dropdown which takes one of the values from Black, Chinese,White,Asian Indian,Asian other,Hispanic Black,Hispanic White,Native American other,other and unknown.
- Age is a dropdown which takes one of the values from young,teens,mid and old.
- Gender is a dropdown which takes one of the values from male, female and unknown.

6.2 Visualizations

We are showing a total of 4 outputs in our interactive component:

- Linechart showing number of deaths over the years selected by the user.
- Multiple linechart showing monthwise number of deaths over the years according to the selections made by the user.
- Barchart showing frequency of deaths due to a particular chemical/chemicals according to the selections made by the user.
- Finally a table is displayed of all the filtered drug abuse death cases so that the user can see all the information for that particular selection at one place.

You can also hover over the line as well as the bar chart to see the exact values for that particular year or chemical respectively.

The introduction of the interactive component in our project makes it easier for any type of user to get information as desired by him/her across the above mentioned input parameters. The variation of information is better represented by the interactive component than the static graphs. The static visualizations are restricted only to a single parameter which is not the case in interactive component.

6.3 Link to RShiny Application

Please visit the below link to experiment with our RShiny application pertinent to this analysis.

https://tanvipareek.shinyapps.io/interactive/

7 Conclusion and Summary

We analyzed data regarding deaths due to drug overdoses in the state of Connecticut from 2012 to 2018. We followed a three-fold approach to gather insights and find patterns in the drug abuse cases and further answer our initial questions.

- Are drug abuse cases correlated with demographic aspects of population?
- What is the trend in the overall number of drug abuse cases from 2012 to 2018?
- What are the most common drug types and type of injuries people suffered from in the drug abuse cases?

The hypothesis for our demographic aspects was supported by our analysis and supporting graphs that drug abuse is highly correlated with the demographic attributes of population such as age, gender, race, and area. Deaths due to drug abuse have been on the rise over the years, thereby answering our initial questions regarding the temporal aspect of the analysis. Finally, the causal diagnosis helped us gather information about the chemicals and their co-consumption which are responsible for the deaths in Connecticut.

All the analysis and the inferences from the graphs are concurrent to the findings from the various online news articles that we came across regarding the same.

Overall, drug abuse has been a grave problem not only in Connecticut but all over the world and needs to be paid heed to immediately, so that the lives of many innocent people can be saved.

News Link related to Drug Overdose Deaths :

https://www.sciencedaily.com/releases/2019/10/191029182501.htm

8 Github Repository

All our files and analysis can be found in the below Github repository.

https://github.com/arushakelkar/EDAV_Final_Project_Group31

9 References

[1] https://wallethub.com/edu/drug-use-by-state/35150/

[2] https://www.pewresearch.org/fact-tank/2018/05/30/as-fatal-overdoses-rise-many-americans-see-drug-addiction-as-a-major

[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/2019/10/191029182501.htm