□space{3.5in}Analysis of Accidental Drug Related Deaths

Kumari Nishu (kn2492)

Neelam Patodia (np2723)

Arusha Kelkar (ak4432)

Tanvi Gautam Pareek (tgp2108)

2019-12-10

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 CDC, 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 misue of prescription medication, making it an issue of public health concern in Connecticut. As part of this research project we want to investigate the prevailing issue of drug abuse in Connecticut.

Objective and Approach

Our objective is to find out the reasons and patterns contributing to high drug abuse in Connecticut by utilizing the publically 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 season as winter/summer?
- Are people more likely to do drugs on weekend compared to weekdays if so what is distribution from that angle?

3. **Causal Diagnosis for Drug Abuse:** Lastly, we will investigate 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 will inquire 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.

Individual Responsibility

1. **Member 1:** This member is exploring the dataset in terms of sparsity and missing values. Description of each column of the dataset and how these can be associated with our broader objective of the project. She is also working on the sampling procedure. She will also be working on the interactive visualization.
2. **Member 2:** This member is looking into the relation of the number of drug abuse cases to the time when it occured/reported. She is exploring 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 is exploring the relation between the region where the person died due to drug abuse and the number of deaths that occured in the regions. Also exploring the different demographic attributes as age, race, gender associated with the drug abuse cases.
4. **Member 4:** The member is exploring 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.

Data Exploration

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. (ata source is: https://catalog.data.gov/dataset/accidental-drug-related-deaths-january-2012-sept-2015 (https://catalog.data.gov/dataset/accidental-drug-related-deaths-january-2012-sept-2015))

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.

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/country for each death 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 15 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 6 years from 2012-2018 and has 41 variables at our disposal. There are 5,105 observations in our dataset which provides sufficient scope to proceed with our investigation.

```
# Loading all required packages
library(wordcloud)
library(memisc)
library(tidyverse)
library(downloader)
library(naniar)
library(openintro)
library(ggplot2)
library(dplyr)
library(choroplethr)
library(maps)
library(scales)
library(tidyr)
library(forcats)
```

```
# Reading the dataset
df_main <- read.csv("Accidental_Drug_Related_Deaths_2012-2018.csv",na.strings=c("","NA"))
cat(paste("\n Number of rows in the original dataframe: ", nrow(df_main), "\n",
          "Number of columns in the original dataframe: ", ncol(df_main), "\n \n" ), sep="")
```

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

```
print( head(df_main, 2))
```

```
##        ID                   Date     DateType Age  Sex  Race ResidenceCity
## 1 14-0273 06/28/2014 12:00:00 AM DateReported  NA <NA>  <NA>          <NA>
## 2 13-0102 03/21/2013 12:00:00 AM  DateofDeath  48 Male Black       NORWALK
##   ResidenceCounty ResidenceState DeathCity DeathCounty Location
## 1            <NA>           <NA>      <NA>        <NA>     <NA>
## 2            <NA>           <NA>   NORWALK   FAIRFIELD Hospital
##   LocationifOther DescriptionofInjury InjuryPlace InjuryCity InjuryCounty
## 1            <NA>           substance        <NA>       <NA>         <NA>
## 2            <NA>                <NA>        <NA>       <NA>         <NA>
##   InjuryState                                 COD OtherSignifican Heroin
## 1        <NA> Acute fent, hydrocod, benzodiazepine            <NA>   <NA>
## 2        <NA>                 Cocaine Intoxication            <NA>   <NA>
##   Cocaine Fentanyl FentanylAnalogue Oxycodone Oxymorphone Ethanol
## 1    <NA>        Y             <NA>      <NA>        <NA>    <NA>
## 2       Y     <NA>             <NA>      <NA>        <NA>    <NA>
##   Hydrocodone Benzodiazepine Methadone Amphet Tramad Morphine_NotHeroin
## 1           Y              Y      <NA>   <NA>   <NA>               <NA>
## 2        <NA>           <NA>      <NA>   <NA>   <NA>               <NA>
##   Hydromorphone Other OpiateNOS AnyOpioid MannerofDeath
## 1          <NA>  <NA>      <NA>      <NA>      Accident
## 2          <NA>  <NA>      <NA>      <NA>      Accident
##                           DeathCityGeo                   ResidenceCityGeo
## 1         CT\n(41.575155, -72.738288)         CT\n(41.575155, -72.738288)
## 2 Norwalk, CT\n(41.11805, -73.412906) NORWALK, CT\n(41.11805, -73.412906)
##                  InjuryCityGeo
## 1 CT\n(41.575155, -72.738288)
## 2 CT\n(41.575155, -72.738288)
```

Data Challenges and Resolution

1. We want to identify the type of injury and the initial driver in the case of drug abuse as this information will be very crucial for the overall analysis of drug abuse and the root cause underlying the same. Data contains a column “DescriptionOfInjury” but this has freely written text in plain english language which is challenging to categorize in fixed set of classes.

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

2. “DateType” in the dataset which informs us when did the drug abuse case occur. However, it leads to some ambiguity because it has two values as “dateReported” and “dateOfDeath”.

This could lead to some gap in both dates for each case. Hence we will be aggregating the numbers for different time range and it would thus be an approximate values near the boundary time range.

Data Cleaning

The data has missing data which are 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” has “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

Missing Data Exploration

We began by exploring the missing values in the dataset. This is crucial as it guides us towards assessing the usability of the data i.e if a column has too many missing values it might be better to not use it. It further guided us on building the interaction variables i.e we assure 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

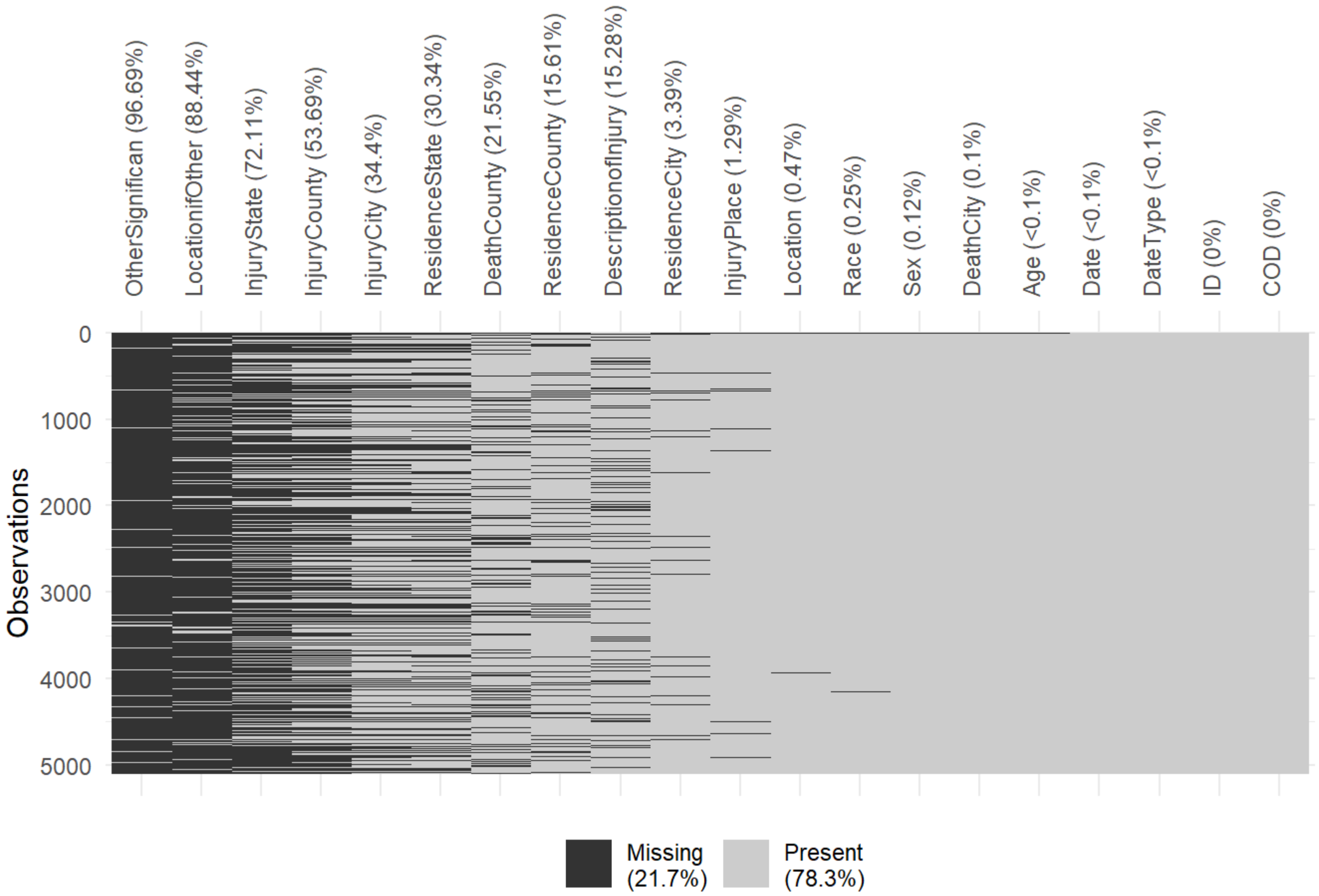
3. Spatial Data parameters such as “InjuryState”, “InjuryCounty”, “ResidenceState”, “ResidenceCounty” seem to have more than 15% missing data and goes as high as 72%. We could thus utilize other data parameter: ‘DeathCityGeo’, ‘InjuryCityGeo’ and ‘ResidenceCityGeo’ which have less than 2% missing data.

4. The column ‘Decsription of Injury’ has ~15% missing data but could still be utilized as it conveys important information corresponding to scenarios of death/drug overdose. In case of an independant analysis of the reasons around a word cloud, the data seems to be sufficient.

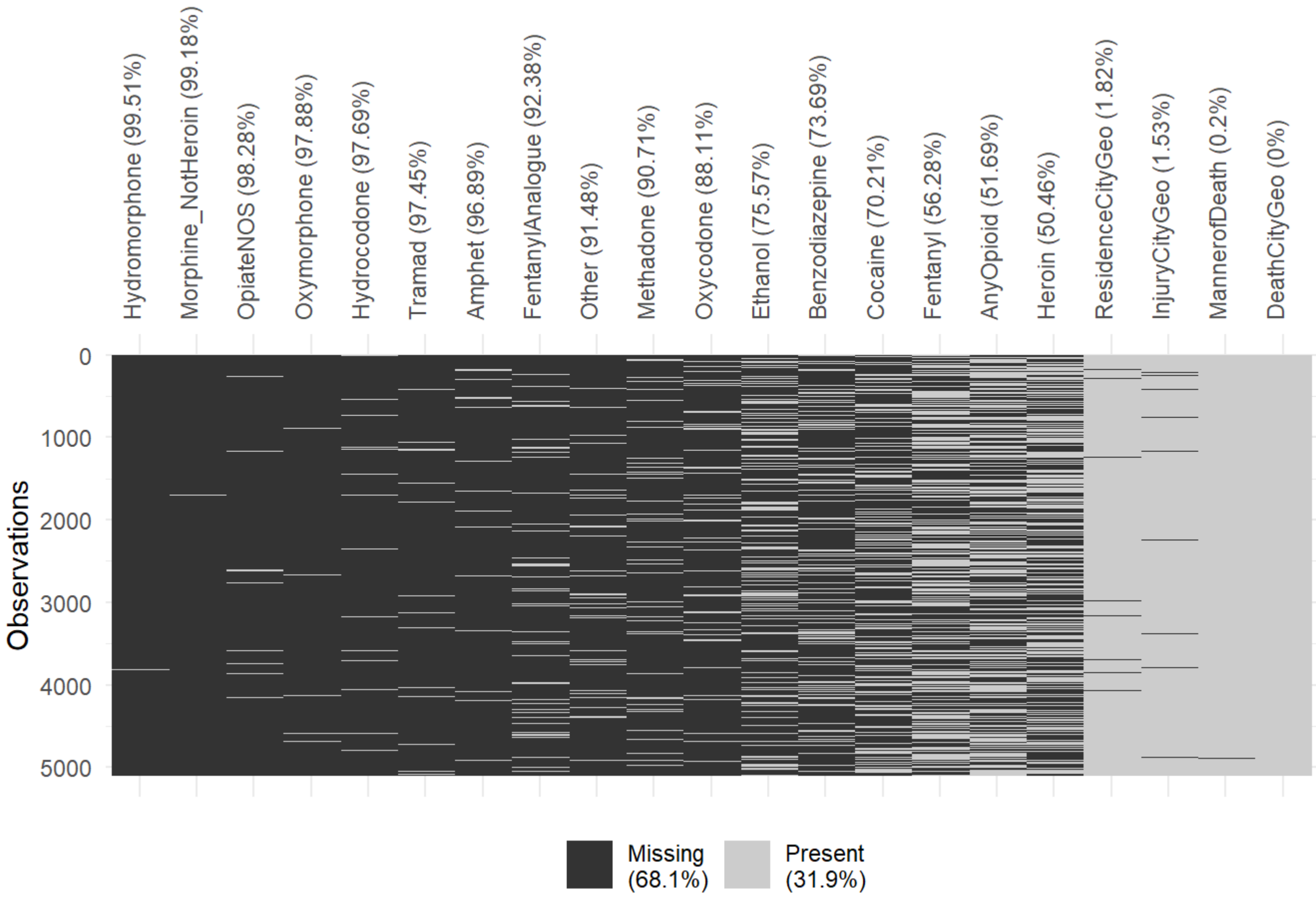
5. Though there seems to be a lot of missing values corresponding to the individual drugs, it is completely okay as it signifies whether a candidate consumed that particular drug or not. A blank here has been assumed to signify- drug not taken.

The data thus seems to be usable for our analysis.

```
#Replacing blank values with NA
drug_acc <- df_main
#1st 20 parameters
vis_miss(drug_acc [,c(1:20)], sort_miss = TRUE) +
 theme(axis.text.x = element_text(angle = 90))
```



```
#2nd 21 parameters
vis_miss(drug_acc [,c(21:41)], sort_miss = TRUE) +
 theme(axis.text.x = element_text(angle = 90))
```



Demographical Aspect of Drug Abuse

Assessing the distribution of fields

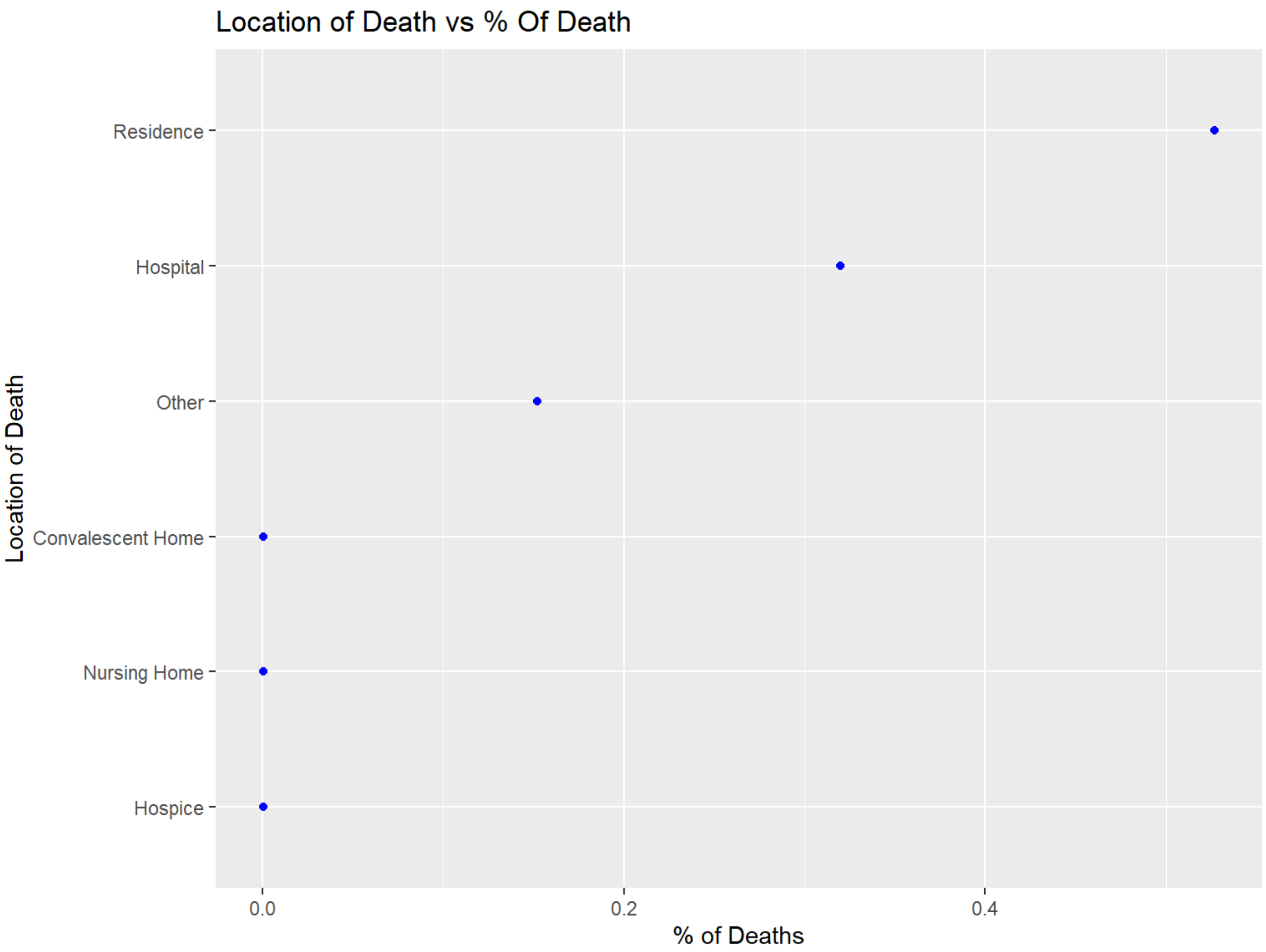
Once we have identified the parameters containing sufficient data, we now assess the distribtion of each of these variables i.e we visualize the spread of numeric data, the frequency of the categories with a column, the proportion of death along various axis etc. This paves the way in identifying the parameters which need to be further explored for truly unravelling drug abuse situation here.

1. **Location of Injury and Location of Death**

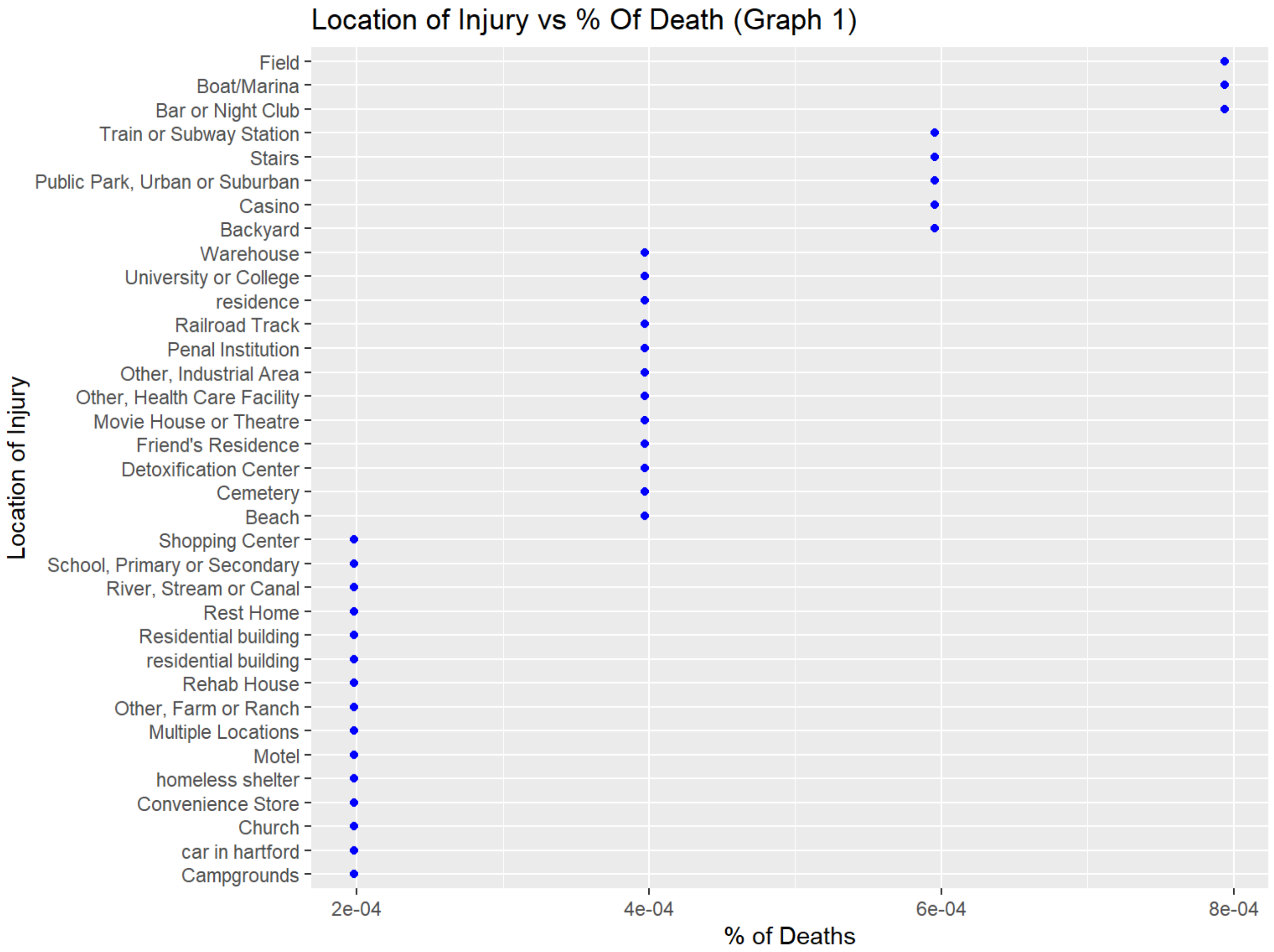
The data consists of information pertaining to both Location of Injury and Location of Death. We wanted to understand if the location of injury is same as that of death, which could have potentially suggested that post drug abuse death might be instantaneous as people did not have time to relocate. Here,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

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

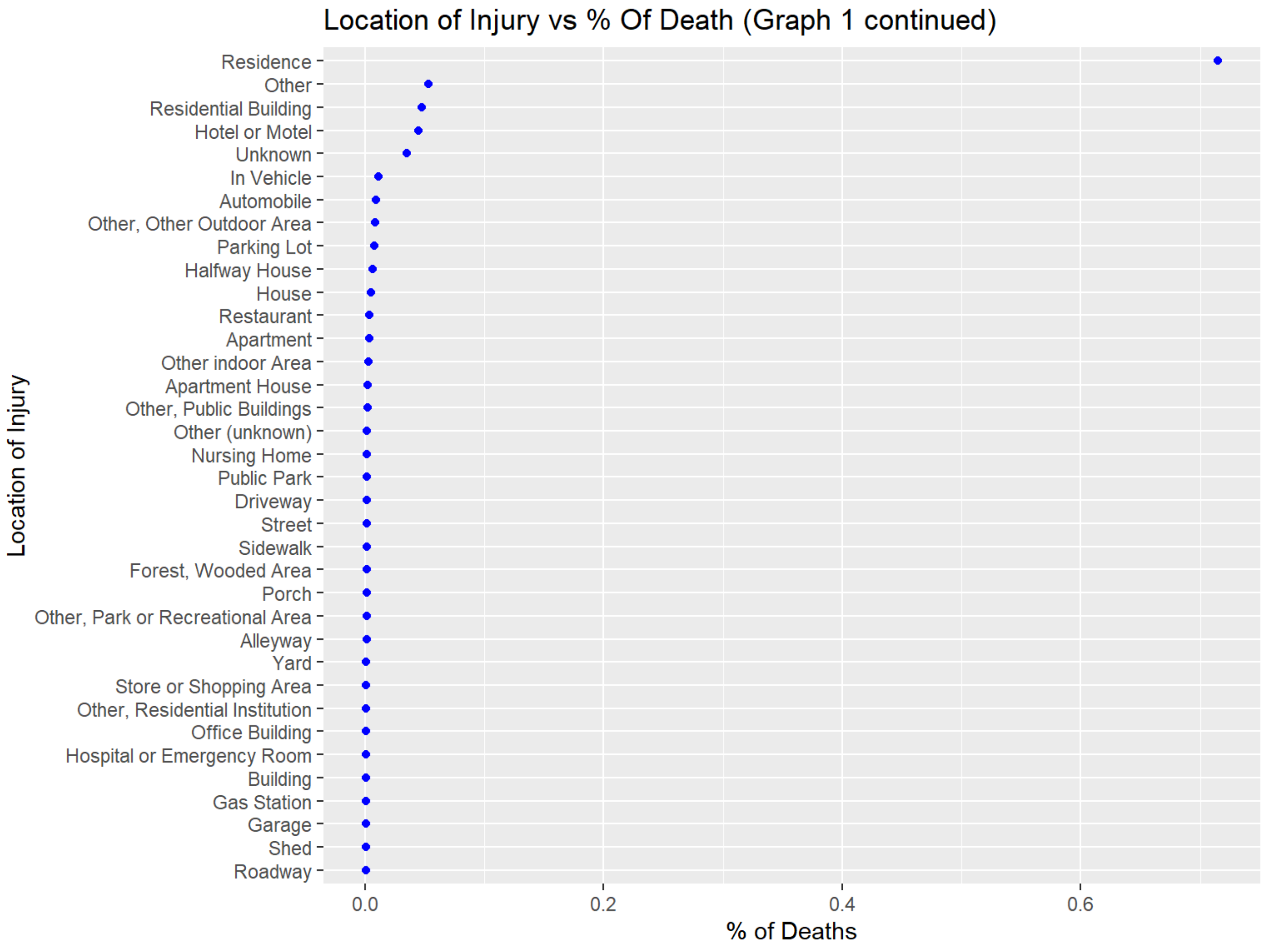
```
# By visualizing the location of death
x<- df_main %>%
    select(Location) %>%
    na.omit()

death_by_loc <- x %>%
                group_by(Location=x$Location) %>%
                summarise(Freq=n()) %>%
                mutate(perc_Death=Freq/nrow(x))
#Graph plot
ggplot(death_by_loc,aes(x = perc_Death, y = reorder(Location, perc_Death))) +
    geom_point(color = "blue")+
  ggtitle("Location of Death vs % Of Death")+
  xlab('% of Deaths') +
  ylab ('Location of Death')
```



```
# Visualizing the location of injury
x<- df_main %>%
    select(InjuryPlace) %>%
    na.omit()
injury_by_loc <- x %>%
                group_by(InjuryPlace=x$InjuryPlace) %>%
                summarise(Freq=n()) %>%
                mutate(perc_Injury=Freq/nrow(x))
injury_by_loc <- injury_by_loc[order(injury_by_loc$perc_Injury),]

#Graph plot
ggplot(injury_by_loc[c(1:35),],aes(x = perc_Injury, y = reorder(InjuryPlace, perc_Injury))) +
    geom_point(color = "blue")+
  ggtitle("Location of Injury vs % Of Death (Graph 1)")+
   xlab('% of Deaths') +
  ylab ('Location of Injury')
```

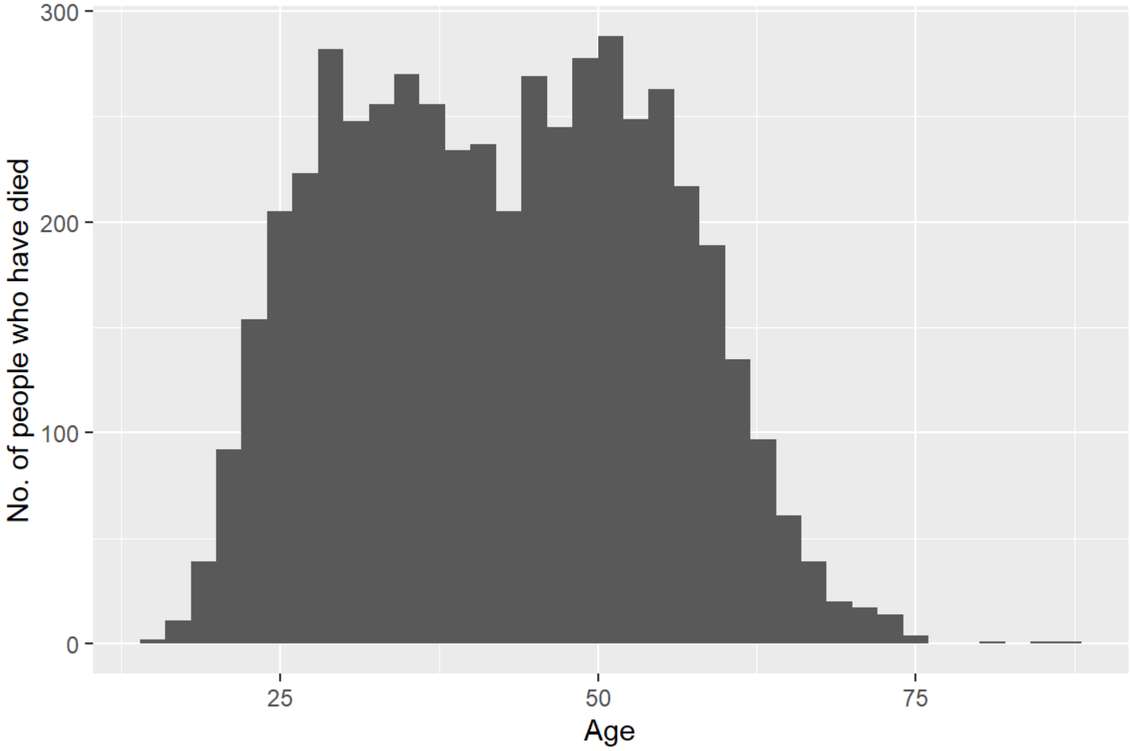


```
ggplot(injury_by_loc[c(36:71),],aes(x = perc_Injury, y = reorder(InjuryPlace, perc_Injury))) +
    geom_point(color = "blue")+
  ggtitle("Location of Injury vs % Of Death (Graph 1 continued) ")+
   xlab('% of Deaths') +
  ylab ('Location of Injury')
```



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

```
#Age
ggplot(drug_acc, aes(x=Age)) +
  geom_histogram(binwidth = 2, boundary = 0, closed = "left") +
  ylab('No. of people who have died')
```



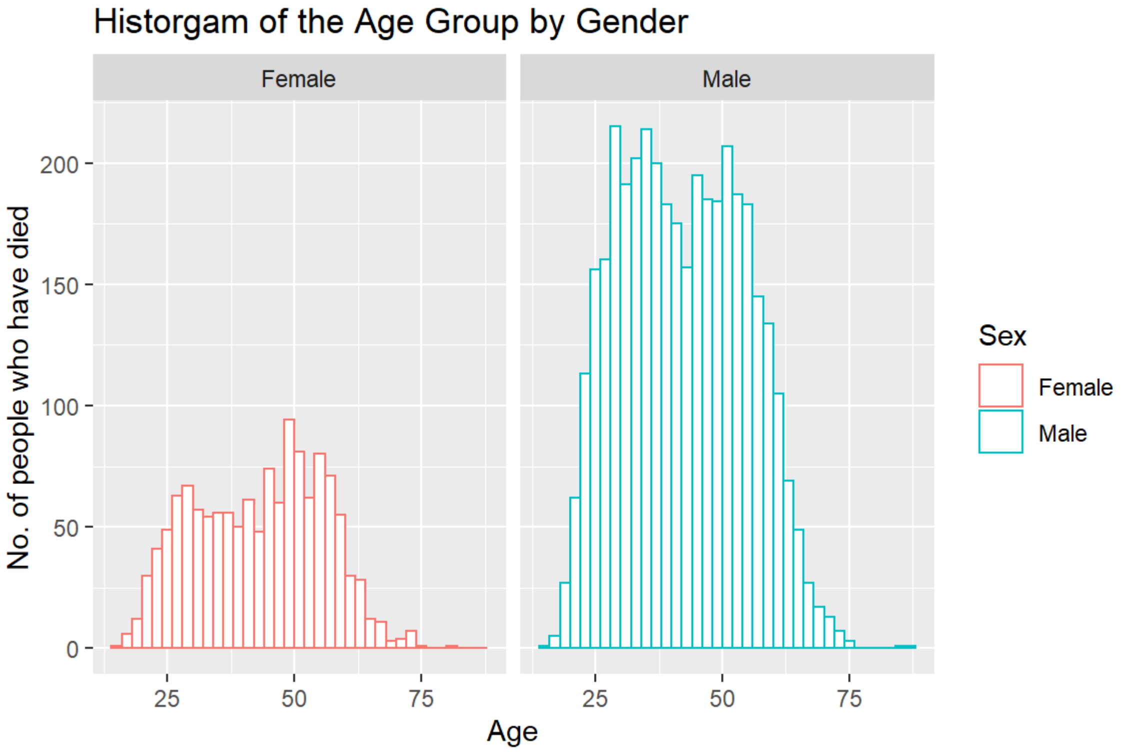
```
  ggtitle("Historgam of the Age Group")
```

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

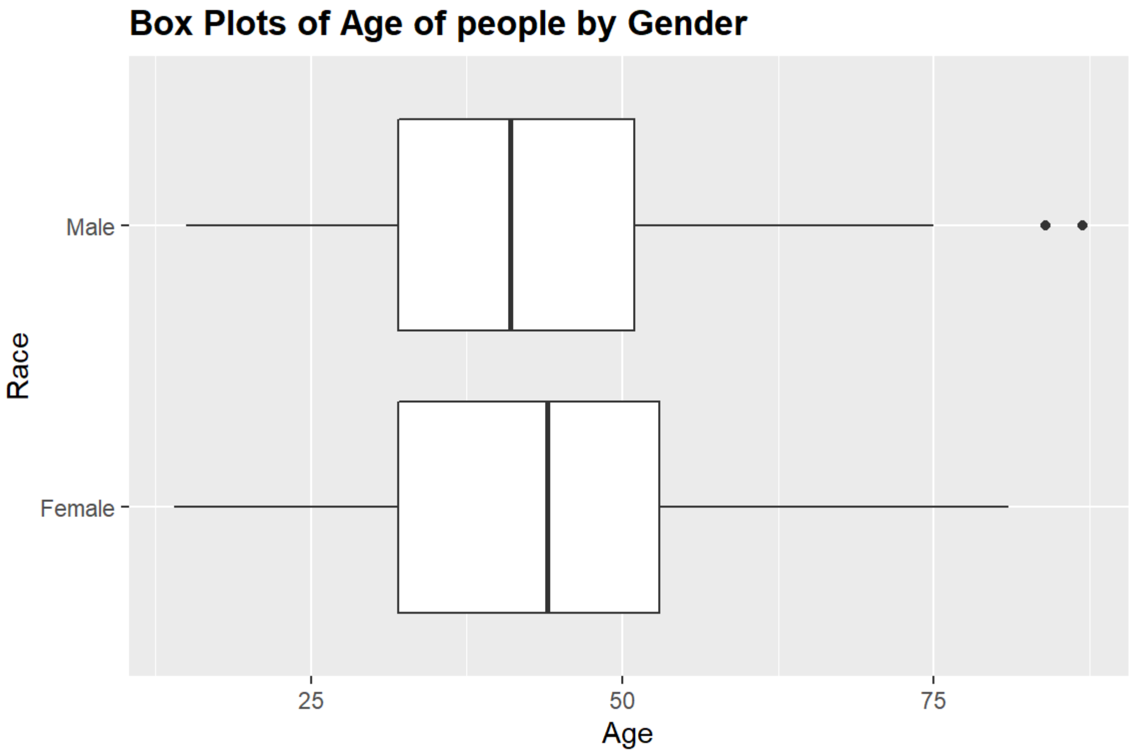
3. **Sex**

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

```
#Sex
ggplot(filter(drug_acc, Sex == "Female" | Sex=='Male'), aes(x=Age,color=Sex)) +
  geom_histogram(binwidth = 2, boundary = 0, closed = "left",fill='white') +
  ggtitle("Historgam of the Age Group by Gender") +
  ylab('No. of people who have died')+
  facet_wrap(~Sex)
```



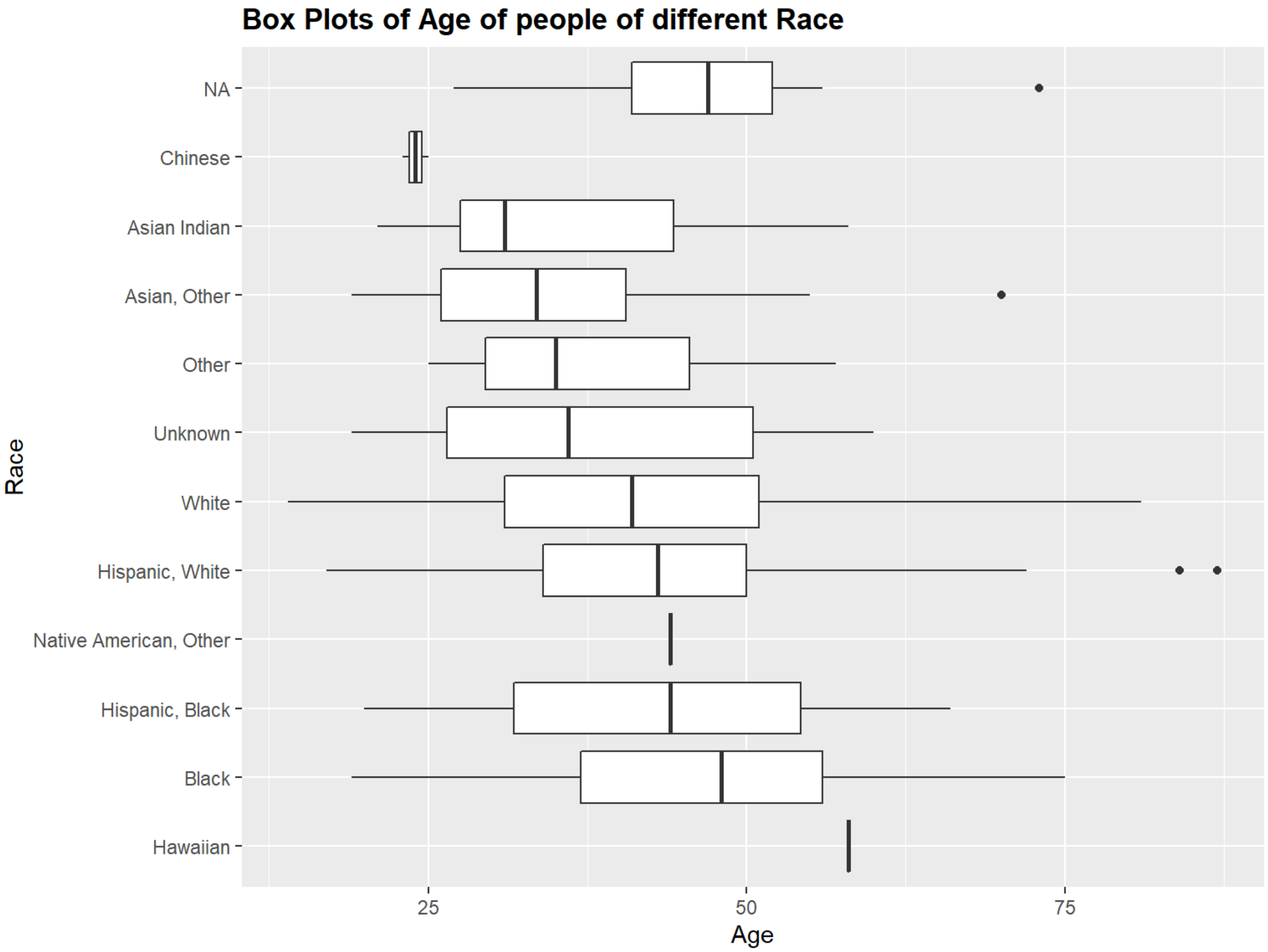
```
#Sex and Age
ggplot(data=filter(drug_acc, Sex == "Female" |  Sex=='Male'),aes(x=reorder(Sex,-Age,FUN=median),
y=Age)) +
  geom_boxplot() +
  ggtitle("Box Plots of Age of people by Gender") +
  xlab("Race") +
  ylab("Age") +
  theme(plot.title = element_text(face = "bold"))+
  coord_flip()
```



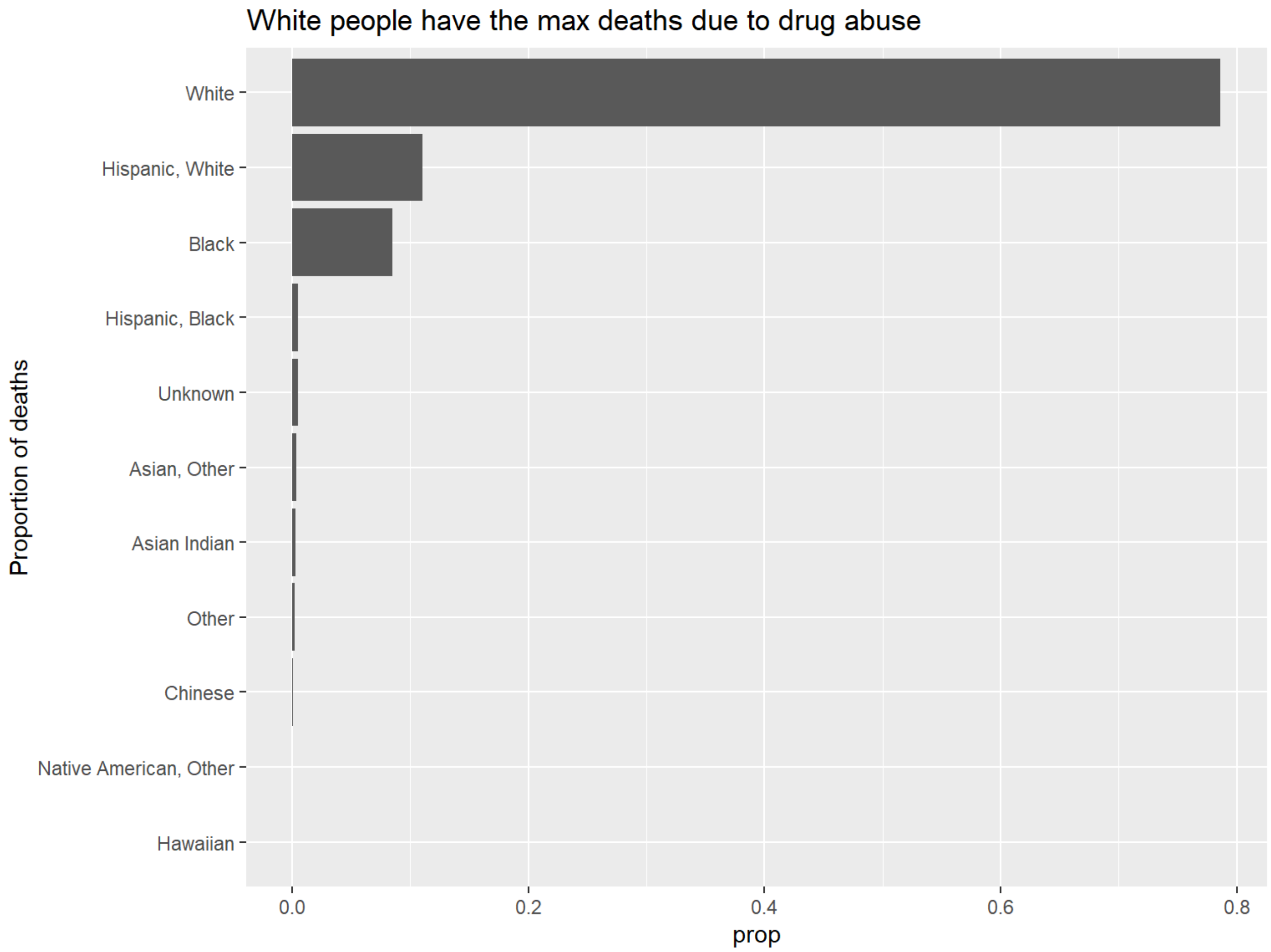
4. **Race**

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

```
#Race and Age
ggplot(data=drug_acc,aes(x=reorder(Race,-Age,FUN=median), y=Age)) +
  geom_boxplot() +
  ggtitle("Box Plots of Age of people of different Race") +
  xlab("Race") +
  ylab("Age") +
  theme(plot.title = element_text(face = "bold"))+
  coord_flip()
```



```
race_plot<-drug_acc %>%
  filter(!is.na(Race))%>%
  group_by(Race)%>%
  summarize(Freq=n()) %>%
  mutate(prop=Freq/sum(Freq)) %>%
  ggplot()+
  geom_bar(aes(x=reorder(Race,prop), y=prop),stat='identity')+
  coord_flip()+
  ggtitle('White people have the max deaths due to drug abuse')+
  xlab('Proportion of deaths')

race_plot
```



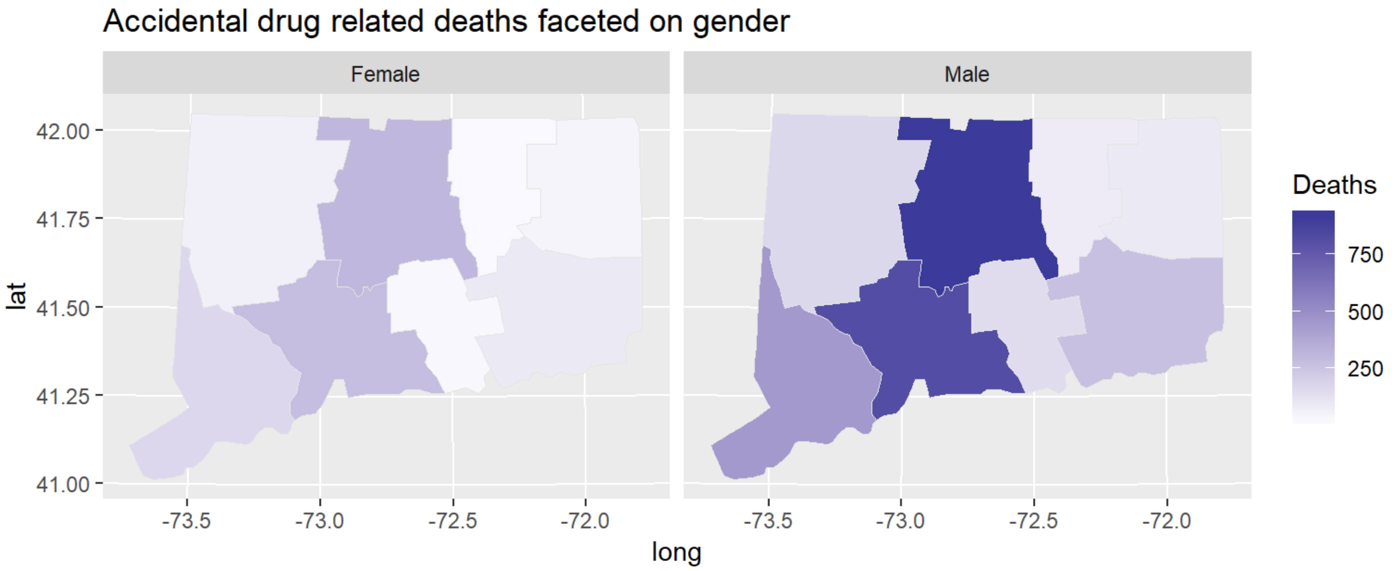
Spatial Analysis

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.

```
#Reading the dataset
#Replacing blank values with NA
demo_df <- drug_acc

#adding a new column age group
demo_df$age_grp<-demo_df$Age
demo_df$age_grp <- ifelse((demo_df$Age>=14 & demo_df$Age<=19) , 'teens',demo_df$age_grp)
demo_df$age_grp <- ifelse((demo_df$Age>=20 & demo_df$Age<=40) , 'young',demo_df$age_grp)
demo_df$age_grp <- ifelse((demo_df$Age>=41 & demo_df$Age<=60) , 'mid',demo_df$age_grp)
demo_df$age_grp <- ifelse((demo_df$Age>=61) , 'old',demo_df$age_grp)

demo_df$statename <- abbr2state(demo_df$ResidenceState)
demo_df$DeathCity <- tolower(demo_df$DeathCity)
demo_df <- demo_df %>% filter(!is.na(demo_df$DeathCounty ))
#unique(demo_df$DeathCity)
k <- demo_df %>%
  count(demo_df$DeathCounty,demo_df$Sex)

names(k)[names(k) == "demo_df$DeathCounty"] <- "subregion"
names(k)[names(k) == "demo_df$Sex"] <- "gender"
k <- k %>% filter(gender != "Unknown")
k <- k %>% filter(!is.na(subregion ))
us_county <- map_data("county")
us_county <- us_county[us_county$region == "connecticut", ]
k$subregion <- tolower(k$subregion)
deathdf <- left_join(us_county, k)
#deathdf
q <- ggplot(data = deathdf,
            aes(x = long, y = lat,
                group = group, fill = deathdf$n))

q + geom_polygon(color = "gray90", size = 0.1) +
  coord_map(projection = "albers", lat0 = 39, lat1 = 45) +
  labs(title = "Accidental drug related deaths faceted on gender", fill="Deaths") +
  scale_fill_gradient2()+
  facet_wrap(vars(deathdf$gender))
```



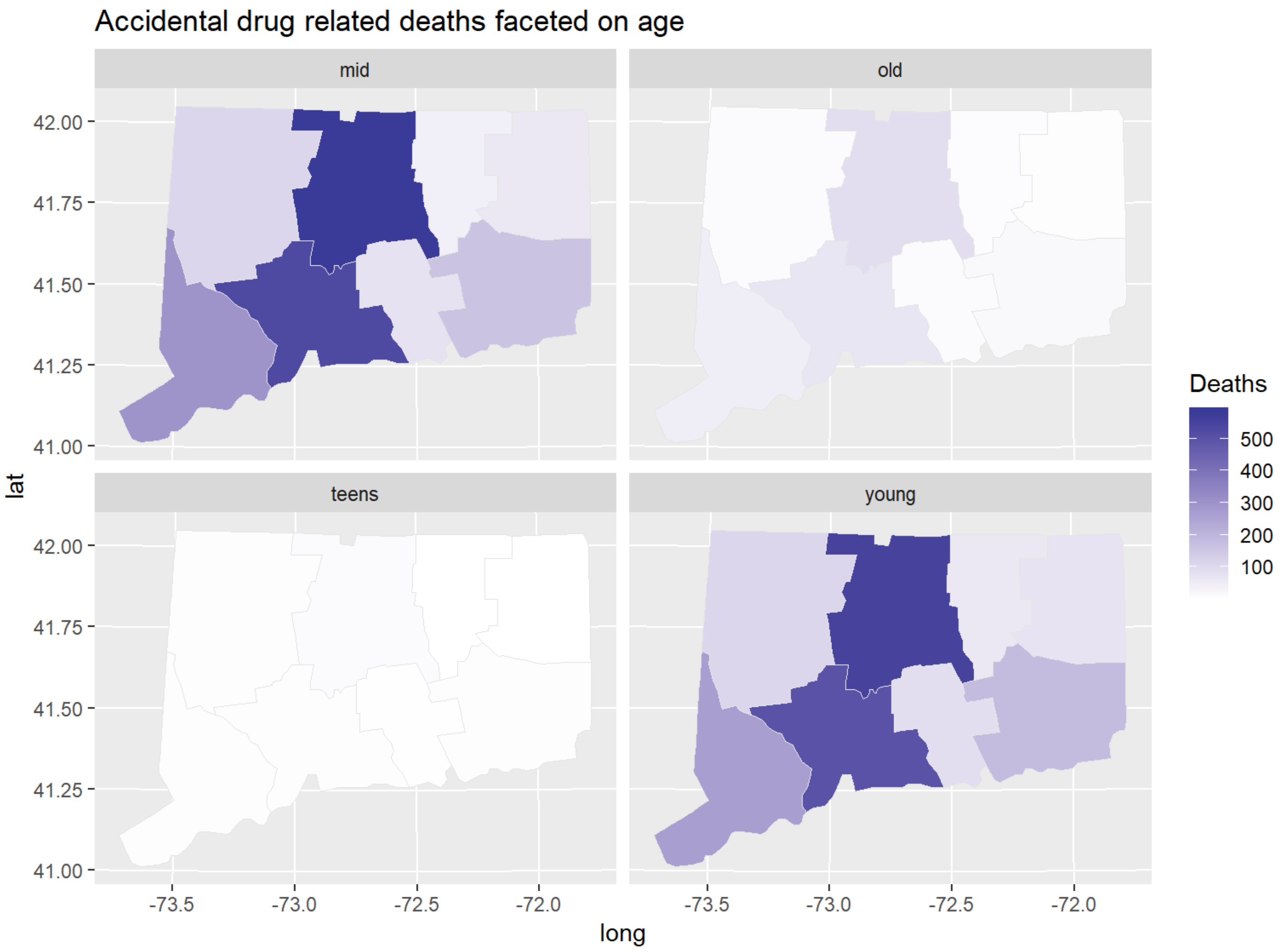
2. **Age**

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.

```
k_age <- demo_df %>%
  count(demo_df$DeathCounty,demo_df$age_grp)
#%>%
names(k_age)[names(k_age) == "demo_df$DeathCounty"] <- "subregion"
names(k_age)[names(k_age) == "demo_df$age_grp"] <- "age"

#k <- k %>% filter(gender != "Unknown")
k_age <- k_age %>% filter(!is.na(subregion ))

us_county <- map_data("county")
us_county <- us_county[us_county$region == "connecticut", ]
k_age$subregion <- tolower(k_age$subregion)
deathdf_age <- left_join(us_county, k_age)
#deathdf_age
q <- ggplot(data = deathdf_age,
            aes(x = long, y = lat,
                group = group, fill = deathdf_age$n))

q + geom_polygon(color = "gray90", size = 0.1) +
  coord_map(projection = "albers", lat0 = 39, lat1 = 45) +
  labs(title = "Accidental drug related deaths faceted on age", fill="Deaths") +
  scale_fill_gradient2()+
  facet_wrap(vars(deathdf_age$age))
```

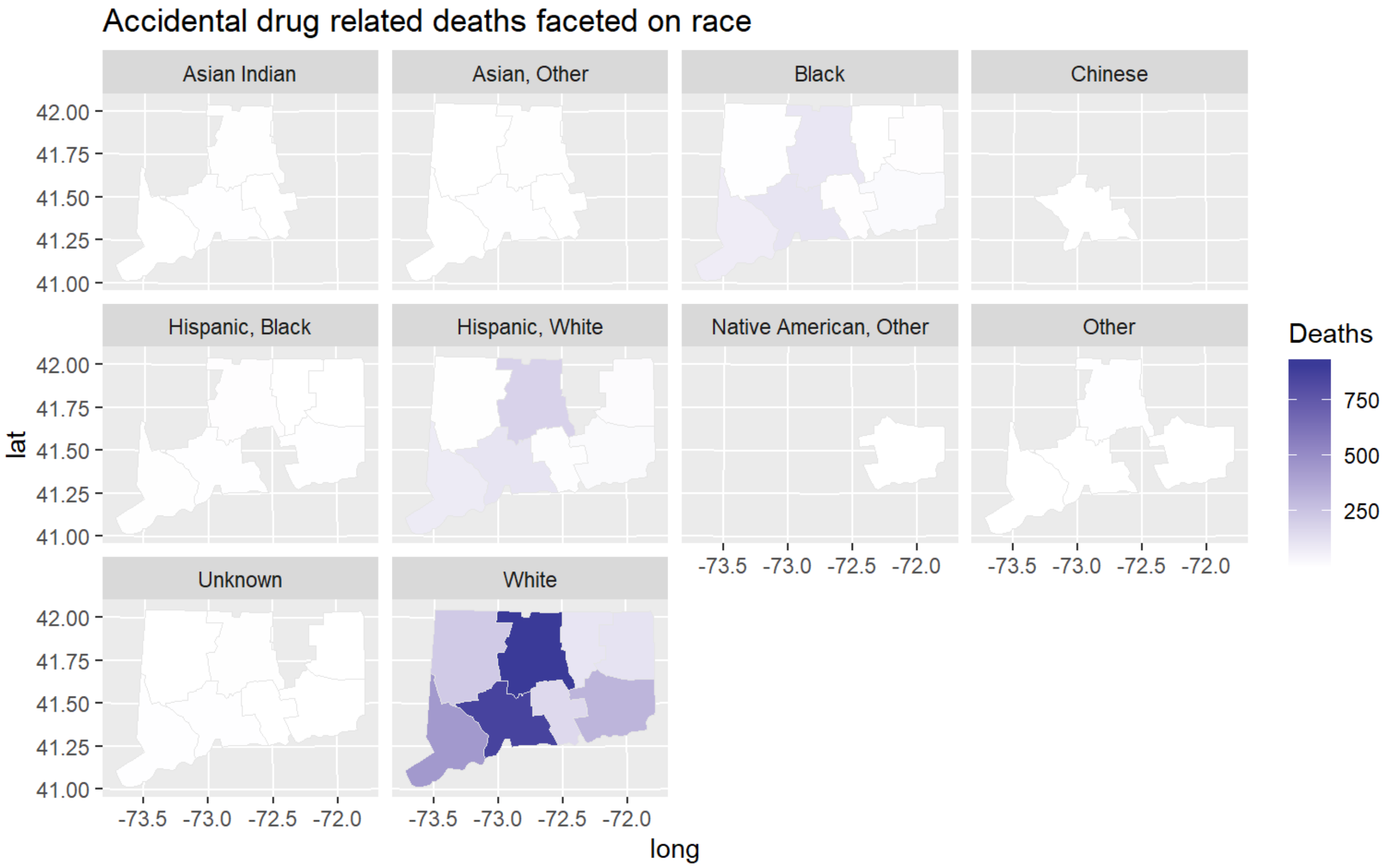


3. **Race**

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 valuess.

- 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.

```
k_race <- demo_df %>%
  count(demo_df$DeathCounty,demo_df$Race)
#%>%
names(k_race)[names(k_race) == "demo_df$DeathCounty"] <- "subregion"
names(k_race)[names(k_race) == "demo_df$Race"] <- "race"
k_race <- k_race %>% filter(!is.na(race))
k_race <- k_race %>% filter(!is.na(subregion ))
us_county <- map_data("county")
us_county <- us_county[us_county$region == "connecticut", ]
k_race$subregion <- tolower(k_race$subregion)

deathdf_race <- left_join(us_county, k_race)

q <- ggplot(data = deathdf_race,
            aes(x = long, y = lat,
                group = group, fill = deathdf_race$n))

q + geom_polygon(color = "gray90", size = 0.1) +
  coord_map(projection = "albers", lat0 = 39, lat1 = 45) +
  labs(title = "Accidental drug related deaths faceted on race", fill="Deaths") +
  scale_fill_gradient2()+
  facet_wrap(vars(deathdf_race$race))
```



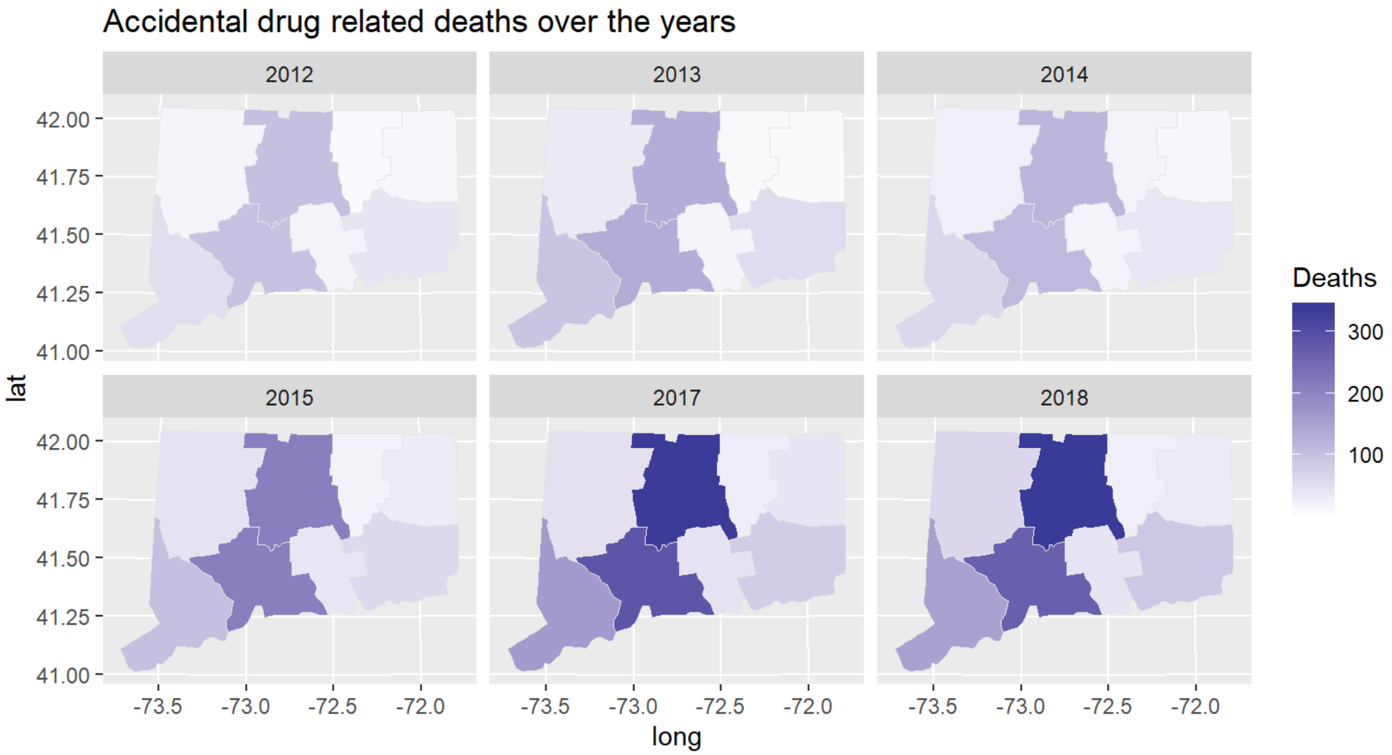
4. **Years 2012-2018**

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

```
demo_df$year <- substr(demo_df$Date, 7, 10)

k_year <- demo_df %>%
  count(demo_df$DeathCounty,demo_df$year)
#%>%
names(k_year)[names(k_year) == "demo_df$DeathCounty"] <- "subregion"
names(k_year)[names(k_year) == "demo_df$year"] <- "year"
k_year <- k_year %>% filter(!is.na(year))
k_year <- k_year %>% filter(!is.na(subregion ))
us_county <- map_data("county")
us_county <- us_county[us_county$region == "connecticut", ]
k_year$subregion <- tolower(k_year$subregion)

deathdf_year <- left_join(us_county, k_year)

q <- ggplot(data = deathdf_year,
            aes(x = long, y = lat,
                group = group, fill = deathdf_year$n))

q + geom_polygon(color = "gray90", size = 0.1) +
  coord_map(projection = "albers", lat0 = 39, lat1 = 45) +
  labs(title = "Accidental drug related deaths over the years", fill="Deaths") +
  scale_fill_gradient2()+
  facet_wrap(vars(deathdf_year$year))+
  labs(caption = 'The "deathcounty" for the deaths in 2016 is not available')+
  theme(plot.caption = element_text(face ='italic',hjust=0,size =10,color='red'))
```



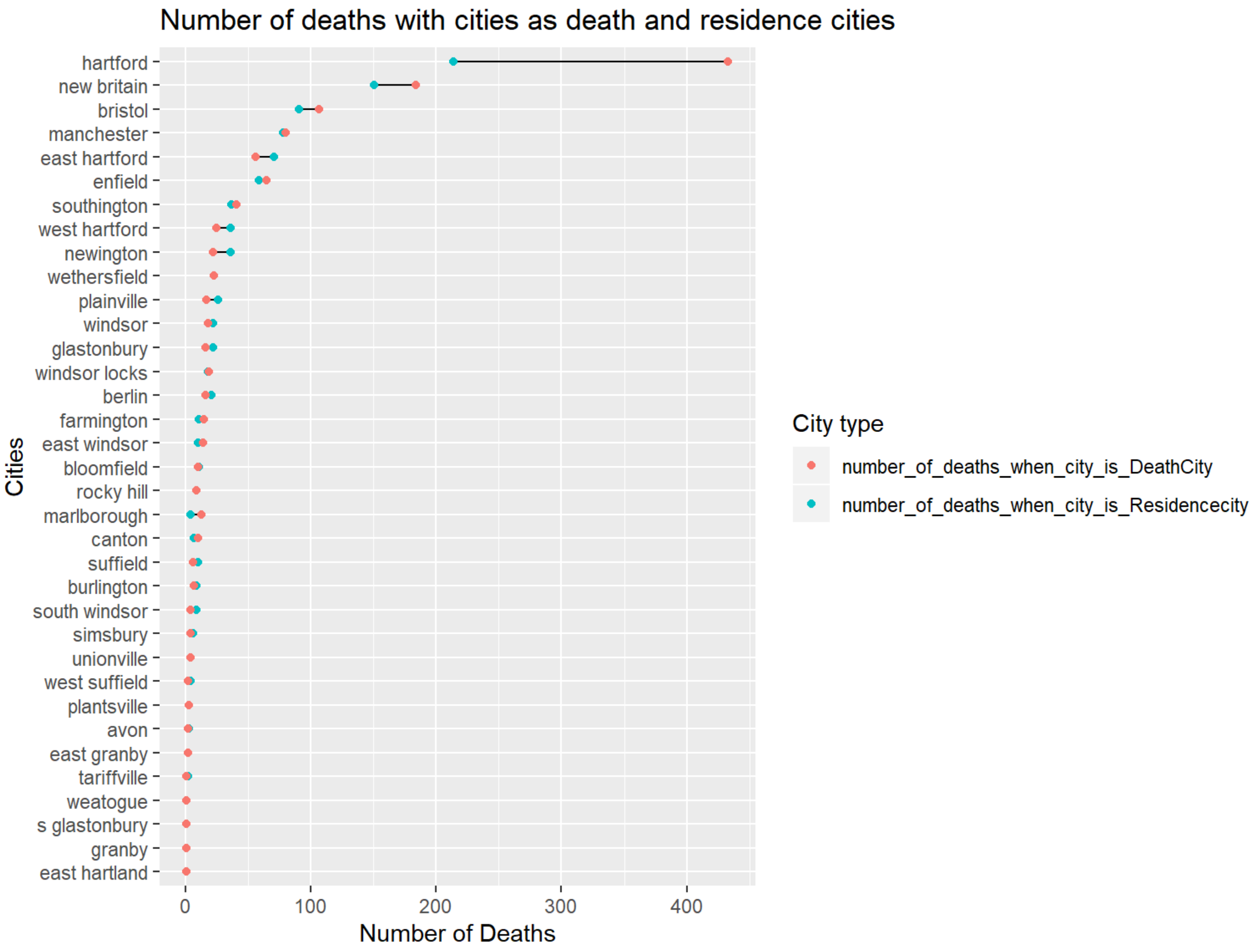
The "deathcounty" for the deaths in 2016 is not available

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 Hartdford 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 than these cities.

```
tempdf <- demo_df[demo_df$DeathCounty == "HARTFORD", ]
tempdf <- tempdf %>% filter(!is.na(tempdf$DeathCounty))

x<- tempdf %>%
    count(tempdf$ResidenceCity)
#x$var <- "residencecity"

names(x)[names(x) == "tempdf$ResidenceCity"] <- "city"
names(x)[names(x) == "n"] <- "number_of_deaths_when_city_is_Residencecity"
x$city <- tolower(x$city)
y<- tempdf %>%
    count(tempdf$DeathCity)
#y$var <- "deathcity"

names(y)[names(y) == "tempdf$DeathCity"] <- "city"
names(y)[names(y) == "n"] <- "number_of_deaths_when_city_is_DeathCity"

x <- merge(x,y,by="city")
x <- x %>% filter(!is.na(city))
x <- as.data.frame(x)

h<- x %>% gather(citytype, value, number_of_deaths_when_city_is_Residencecity:number_of_deaths_w
hen_city_is_DeathCity)

ggplot(h, aes(y = reorder(city, value),x = value))+
  geom_line(aes(group = city))+
  geom_point(aes(color = factor(citytype)))+
  labs(x="Number of Deaths", y="Cities", color = "City type")+
  ggtitle("Number of deaths with cities as death and residence cities")
```



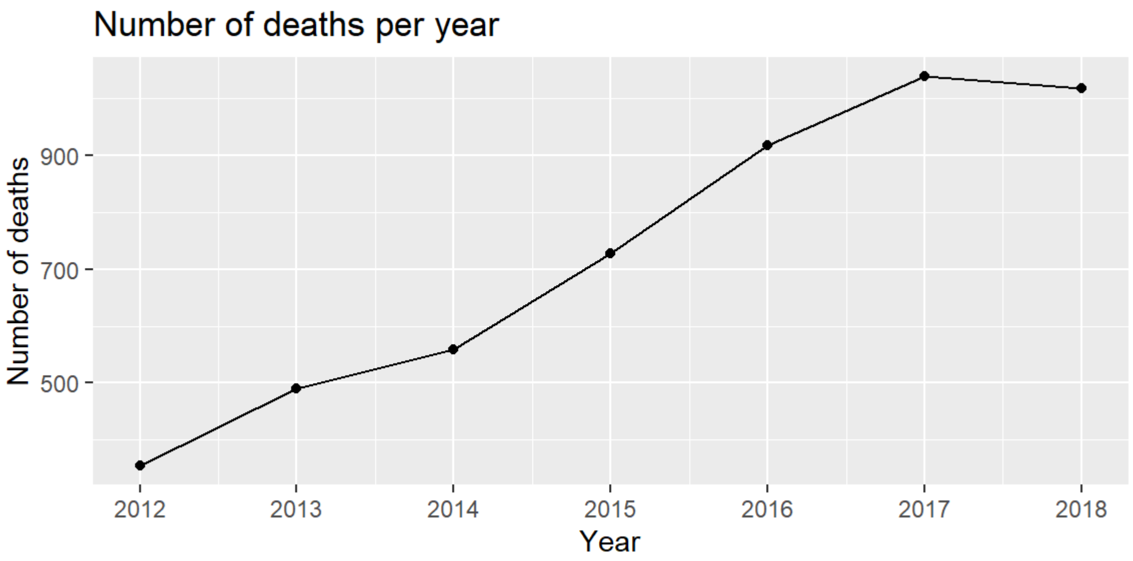
Temporal Aspect of Drug Abuse

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

Seasonality

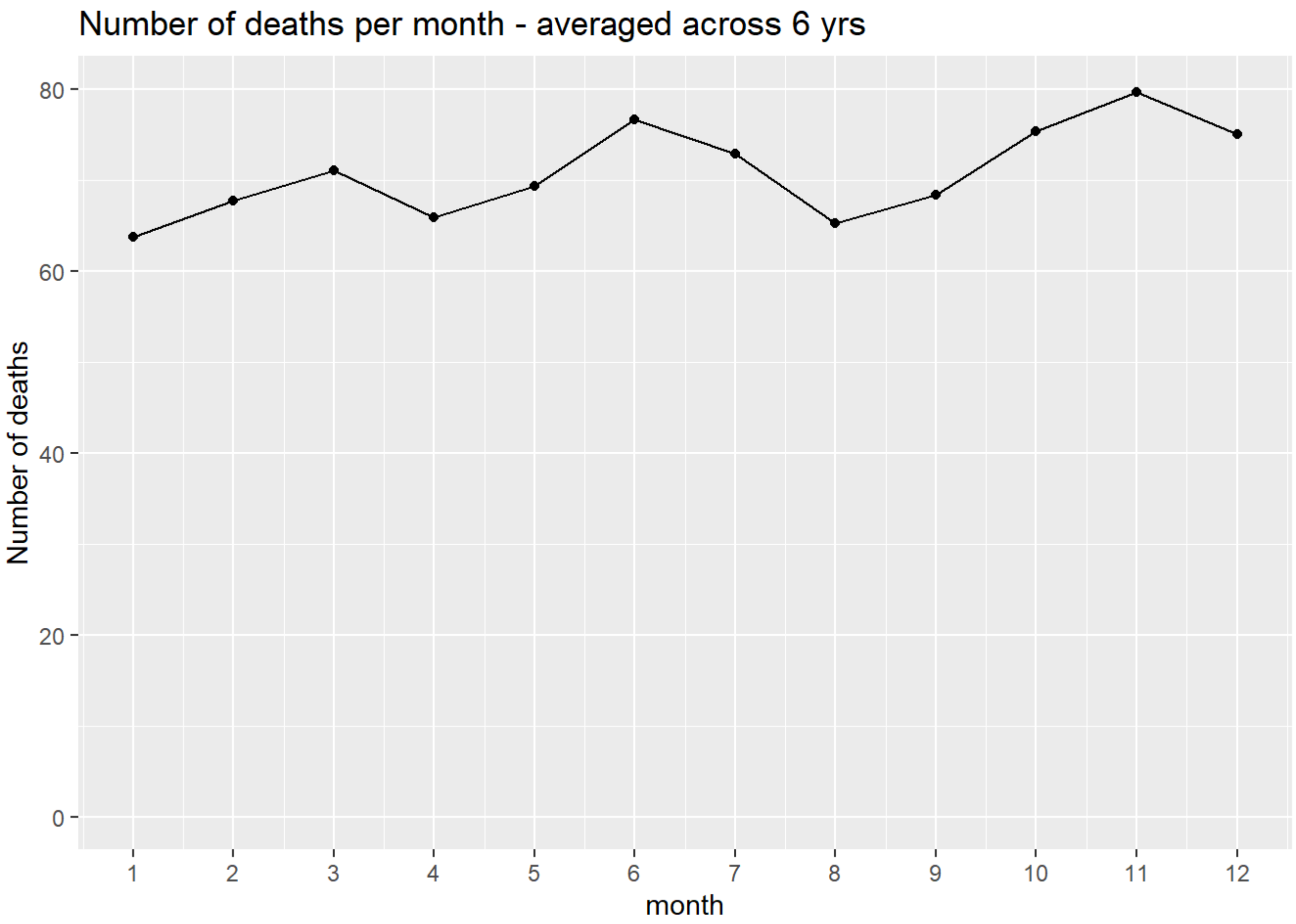
We thus begin by noting the number of deaths due to drug abuse reported each year from 2012-2018. We observed that in 2017 maximum number of deaths were reported, followed by in 2018.

```
#Converting Date to Date variable
temp<-df_main
temp$Date<-as.Date(temp$Date, format = "%m/%d/%Y")

temp_yr<-temp%>%
  group_by(yr=as.numeric(format(temp$Date,"%Y")))%>%
  summarise(cnt=n())

temp_yr%>%
  filter(!is.na(yr))%>%
  ggplot()+
  geom_line(aes(x=yr,y=cnt))+
  geom_point(aes(x=yr,y=cnt))+
  xlab("Year")+
  ylab("Number of deaths")+
  ggtitle("Number of deaths per year")+
  scale_x_continuous(labels = as.character(temp_yr$yr), breaks = temp_yr$yr)
```



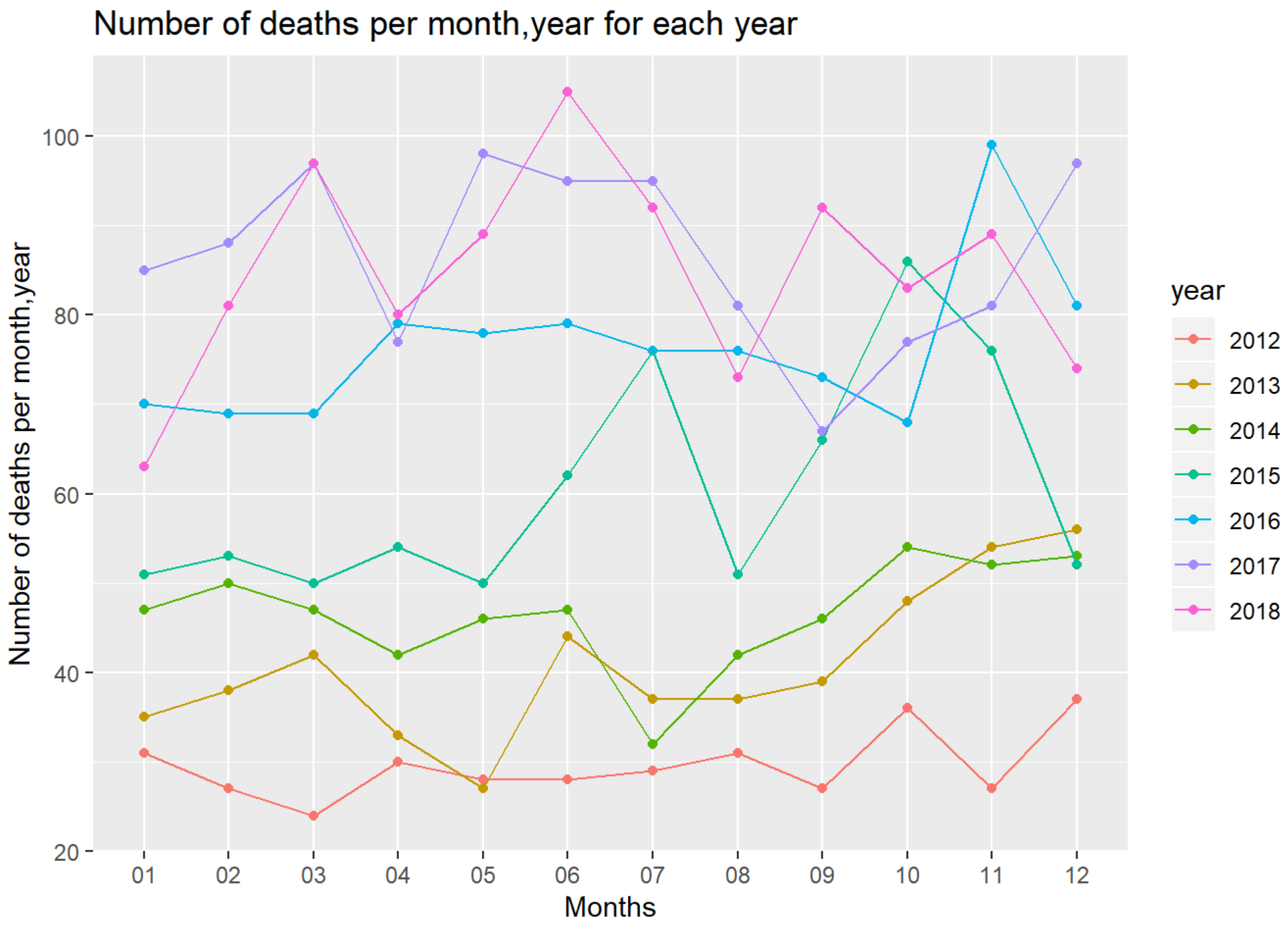
+ On plotting the average number of deaths each month for the last 6 years, we observe that on an average the number of deaths increase from January to March and it keeps rising again in August through November.

```
temp_month <- temp %>%
 group_by(month=as.numeric(format(temp$Date,"%m"))) %>%
 summarise(Freq = n()/6)
temp_month%>%ggplot(aes(x=month,y=Freq))+geom_point()+geom_line()+xlab("month")+ylab("Number of
 deaths")+ggtitle("Number of deaths per month - averaged across 6 yrs")+scale_x_continuous(label
s = as.character(temp_month$month), breaks = temp_month$month)
```



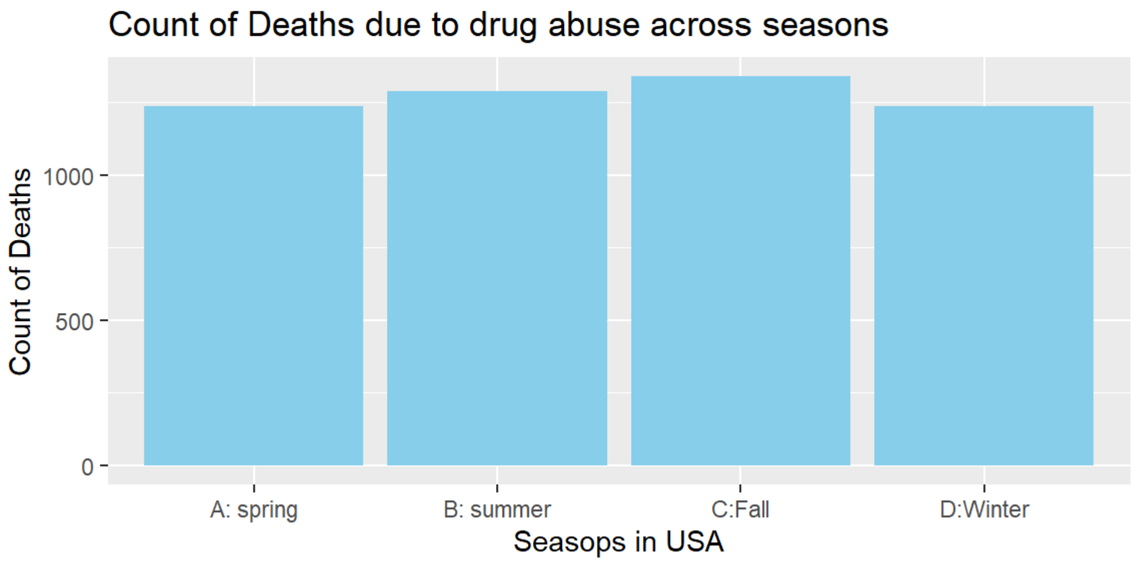
- Specifically in 2017, we observe that deaths have been rising steeply since September till the end of the year and this rise continues for the 1st three months of 2018 as well. This suggests that during Fall 2017 and Winter 2018 deaths is on a rise which could be related to the cold weather conditions triggering people to consume more drugs and therefore result in more deaths. A similar pattern is also observed for the years 2013 and 2014. However, we do not have the required data variables in this dataset to test our assumptions. Additionally we have observed that the highest death in 2018 has occured in June.

```
#Deaths per month for each year
temp$month <- factor(format(temp$Date,"%m"))
temp$year <- factor(format(temp$Date,"%Y"))
a<-temp%>%
  group_by(month,year)%>%
  summarise(freq=n())
a<-gather(a, key, value, -month, -year)

a%>%
  filter(!is.na(month))%>%
  ggplot(aes(x = month, y = value,color=year,group=year))+geom_point()+
      geom_line()+
    xlab("Months")+
    ylab("Number of deaths per month,year")+
    ggtitle("Number of deaths per month,year for each year")+
    scale_x_discrete(labels = as.character(a$month), breaks = a$month)
```



- The number of deaths are highest in Fall.

```
x<- temp %>%
    select(month) %>%
    na.omit()

x$month<-as.numeric(x$month)
x$season<-x$month
x$season <- ifelse((x$month>=3 & x$month<=5) , 'A: spring',x$season)
x$season <- ifelse((x$month>=6 & x$month<=8) , 'B: summer',x$season)
x$season <- ifelse((x$month>=9 & x$month<=11) , 'C:Fall',x$season)
x$season <- ifelse((x$month==12 | x$month==1 | x$month==2) , 'D:Winter',x$season)

a<-x%>%group_by(season)%>%summarise(freq=n())
a%>%ggplot(aes(x=season,y=freq,group=1))+
  geom_histogram(stat = "identity",fill="skyblue")+
  xlab('Seasops in USA')+
  ylab('Count of Deaths')+
  ggtitle("Count of Deaths due to drug abuse across seasons")
```



Patterns in Drug consumption over Years

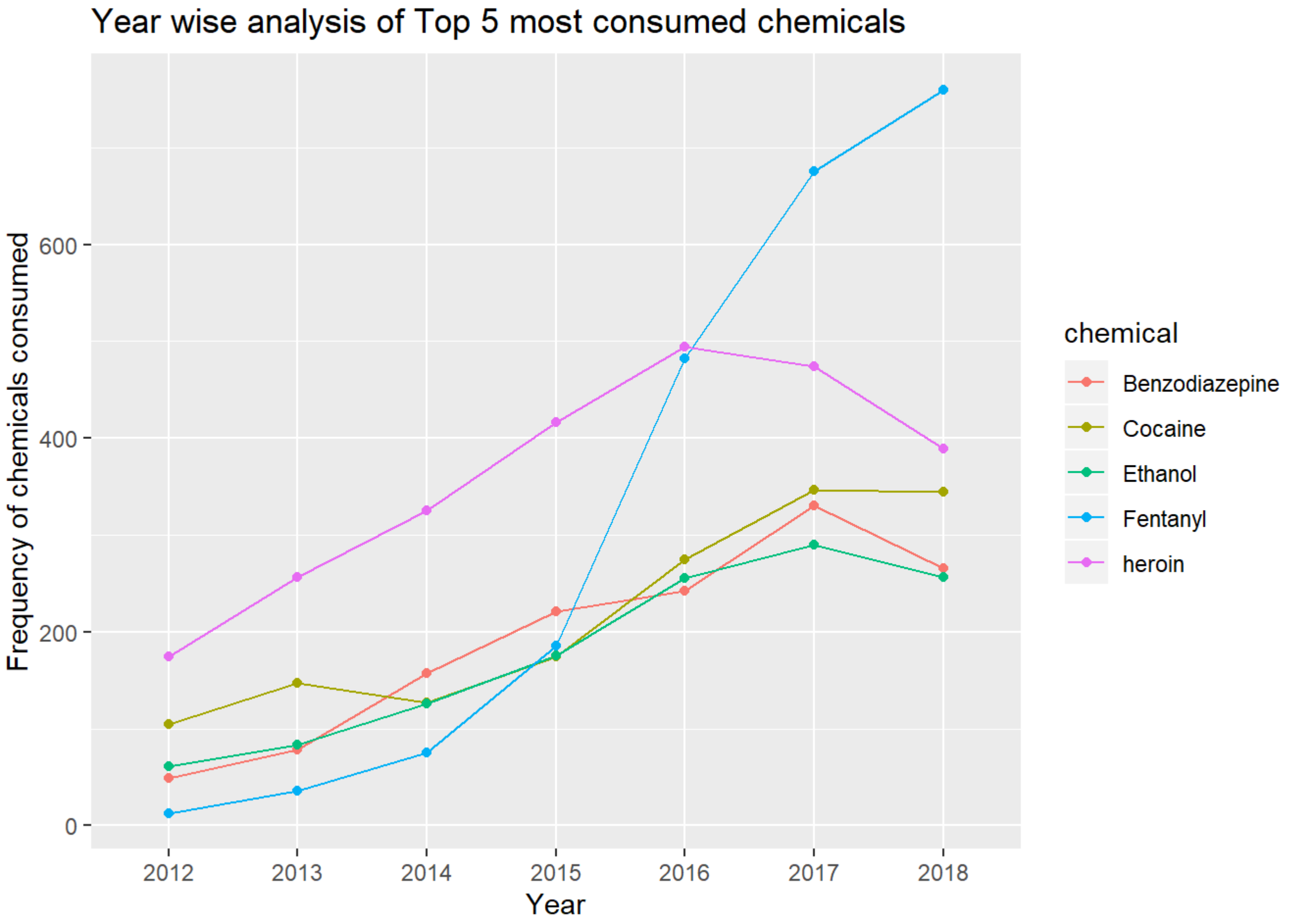
In order to further understand the factors associated with increased deaths due to drugs year over year, we investigate the chemical compounds/drugs which were consumed during these years and were responsible for the deaths. We thus observe that: + 'Fentanyl' which used to be the least consumed drug 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.

```
#only plotting top 5 chemicals
#df_chem <- read.csv("data.csv",na.strings=c("","NA"))
df_chem<-df_main
chemical_cols <- c("Heroin","Cocaine","Fentanyl" ,"FentanylAnalogue", "Oxycodone" ,"Oxymorphone"
,"Ethanol","Hydrocodone","Benzodiazepine","Methadone", "Amphet","Tramad","Morphine_NotHeroin","H
ydromorphone")

    df_rca <- subset(df_chem, select=c('Date','DescriptionofInjury', chemical_cols))
    all_chemicals <- c()

    for (row in 1:nrow(df_rca)) {
      chemical_list <-  ''
      chemical_count <- 0
      for (chemical in chemical_cols){
        val = df_rca[row, chemical]
        if (!is.na(val)){
          chemical_list <- paste(chemical_list, chemical, sep=',')
          chemical_count <- chemical_count+1
          all_chemicals <- c(all_chemicals, chemical)
        }
      }
    }
df_chem<- as.data.frame(table(all_chemicals))
df_chem<-df_chem[with(df_chem, order(-df_chem$Freq)), ]
head(df_chem,5)
```

	all_chemicals <fctr>	Freq <int>
7	Heroin	2529
5	Fentanyl	2232
3	Cocaine	1521
2	Benzodiazepine	1343
4	Ethanol	1247

5 rows

```
he<-temp%>%group_by(yr=fct_explicit_na(year),Heroin)%>%summarise(cnt=n())%>%filter(Heroin=='Y')
he$Heroin<-"heroin"
he<-he%>%rename(chemical=Heroin)
co<-temp%>%group_by(yr=fct_explicit_na(year),Cocaine)%>%summarise(cnt=n())%>%filter(Cocaine=='Y'
)
co<-co[1:7,]
co$Cocaine<-"Cocaine"
co<-co%>%rename(chemical=Cocaine)
fe<-temp%>%group_by(yr=fct_explicit_na(year),Fentanyl)%>%summarise(cnt=n())%>%filter(Fentanyl==
'Y')
fe$Fentanyl<-"Fentanyl"
fe<-fe%>%rename(chemical=Fentanyl)
be<-temp%>%group_by(yr=fct_explicit_na(year),Benzodiazepine)%>%summarise(cnt=n())%>%filter(Benzo
diazepine=='Y')
be$Benzodiazepine<-"Benzodiazepine"
be<-be%>%rename(chemical=Benzodiazepine)
eth<-temp%>%group_by(yr=fct_explicit_na(year),Ethanol)%>%summarise(cnt=n())%>%filter(Ethanol==
'Y')
eth$Ethanol<-"Ethanol"
eth<-eth%>%rename(chemical=Ethanol)
df1<-dplyr::bind_rows(he,fe,co,be,eth)
df1%>%ggplot(aes(x=yr,y=cnt,color=chemical,group=chemical))+geom_line()+geom_point()+xlab("Year"
)+ylab("Frequency of chemicals consumed")+ggtitle("Year wise analysis of Top 5 most consumed che
micals")
```



Causal Diagnosis of Drug Abuse

We want to analyze the drug abuse cases in order to highlight the major contributing factors. We will investigate the underlying root causes: major chemicals consumed, type of injuries people suffered from. From this investigation, we are trying to find the most common checmicals found in the drug abuse cases. Moreover, we will inquire the forms of drug abuse such as ingested pills, alcohol, abuse of medication, substance abuse etc. And we will try to attribute the categories contributing to the most of the cases.

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 thsi section we are calculating 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, eesentially in each row in the dataframe. Similarly, metric “diagnosed_chemicals_count” denotes the number of checmical found in each drug abuse case out of exaustive list of 14 chemicals mentioned here. Both of these derived metrics will be explored further in the below sections.

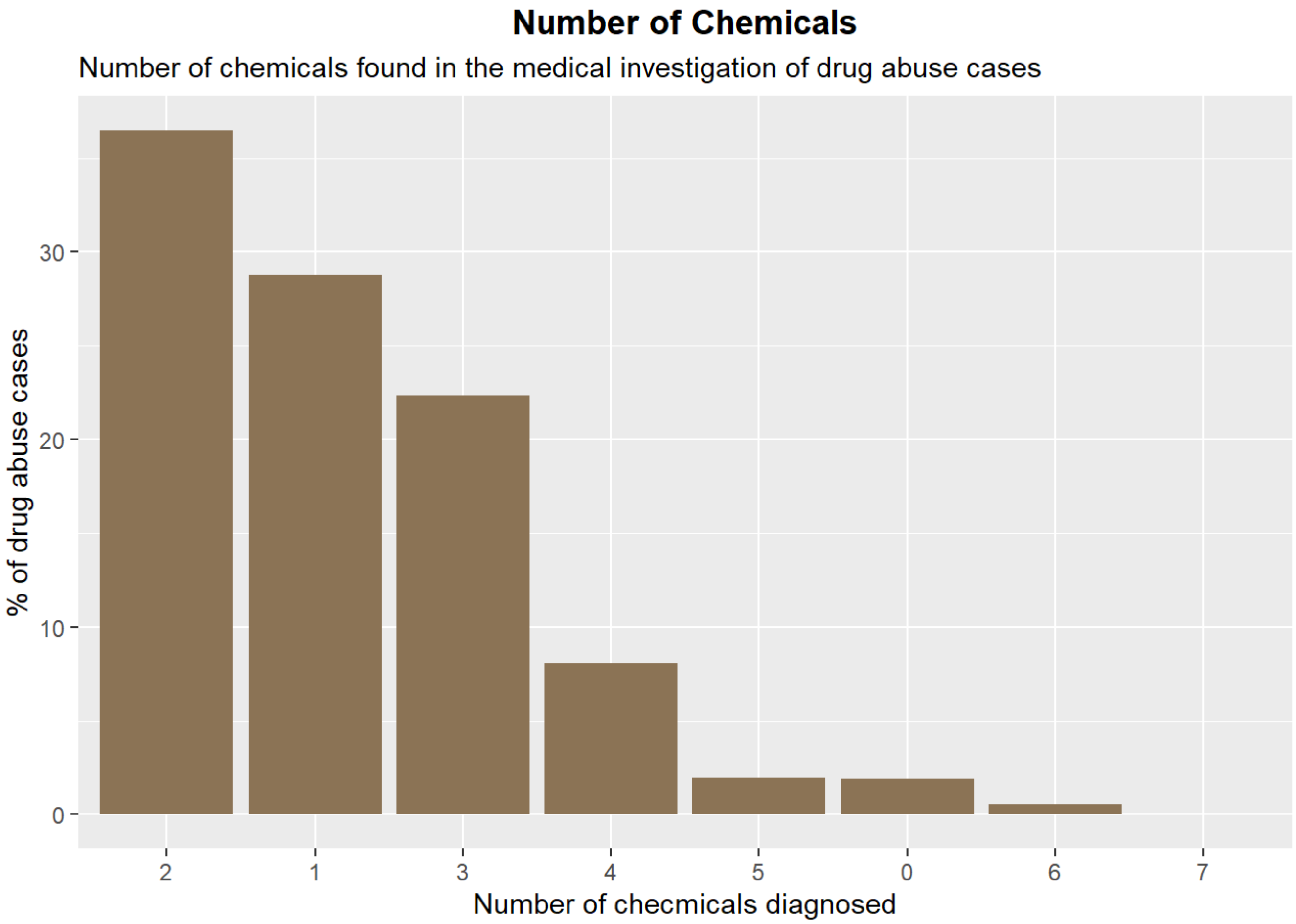
```
#df_main <- read.csv("Accidental_Drug_Related_Deaths_2012-2018.csv",na.strings=c("","NA"))

# List of chemicals
chemical_cols <- c("Heroin","Cocaine","Fentanyl" ,"FentanylAnalogue", "Oxycodone" ,"Oxymorphone"
,"Ethanol","Hydrocodone","Benzodiazepine","Methadone", "Amphet","Tramad","Morphine_NotHeroin","H
ydromorphone")

# Subset relevant columns for root cause analysis
df_rca <- subset(df_main, select=c('Date','DescriptionofInjury', chemical_cols))

# Adding diagnosed chemical list for each drug abuse case
diagnosed_chemicals = c()
count_diagnosed_chemicals = c()
all_chemicals <- c()

for (row in 1:nrow(df_rca)) {
  chemical_list <-  ''
  chemical_count <- 0
  for (chemical in chemical_cols){
    val = df_rca[row, chemical]
    if (!is.na(val)){
      chemical_list <- paste(chemical_list, chemical, sep=',')
      chemical_count <- chemical_count+1
      all_chemicals <- c(all_chemicals, chemical)
    }
  }
  diagnosed_chemicals <- c(diagnosed_chemicals, substr(chemical_list, 2, nchar(chemical_list)))
  count_diagnosed_chemicals <- c(count_diagnosed_chemicals, chemical_count)
}
df_rca$Diagnosed_Chemicals <- diagnosed_chemicals
df_rca$diagnosed_chemicals_count <- count_diagnosed_chemicals

head(df_rca[, c('Diagnosed_Chemicals', 'diagnosed_chemicals_count')])
```

	Diagnosed_Chemicals <chr>	diagnosed_chemicals_count <dbl>
1	Fentanyl,Hydrocodone,Benzodiazepine	3
2	Cocaine	1
3	Heroin,Cocaine	2
4	Heroin,Fentanyl	2
5	Fentanyl	1
6	Heroin	1

6 rows

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

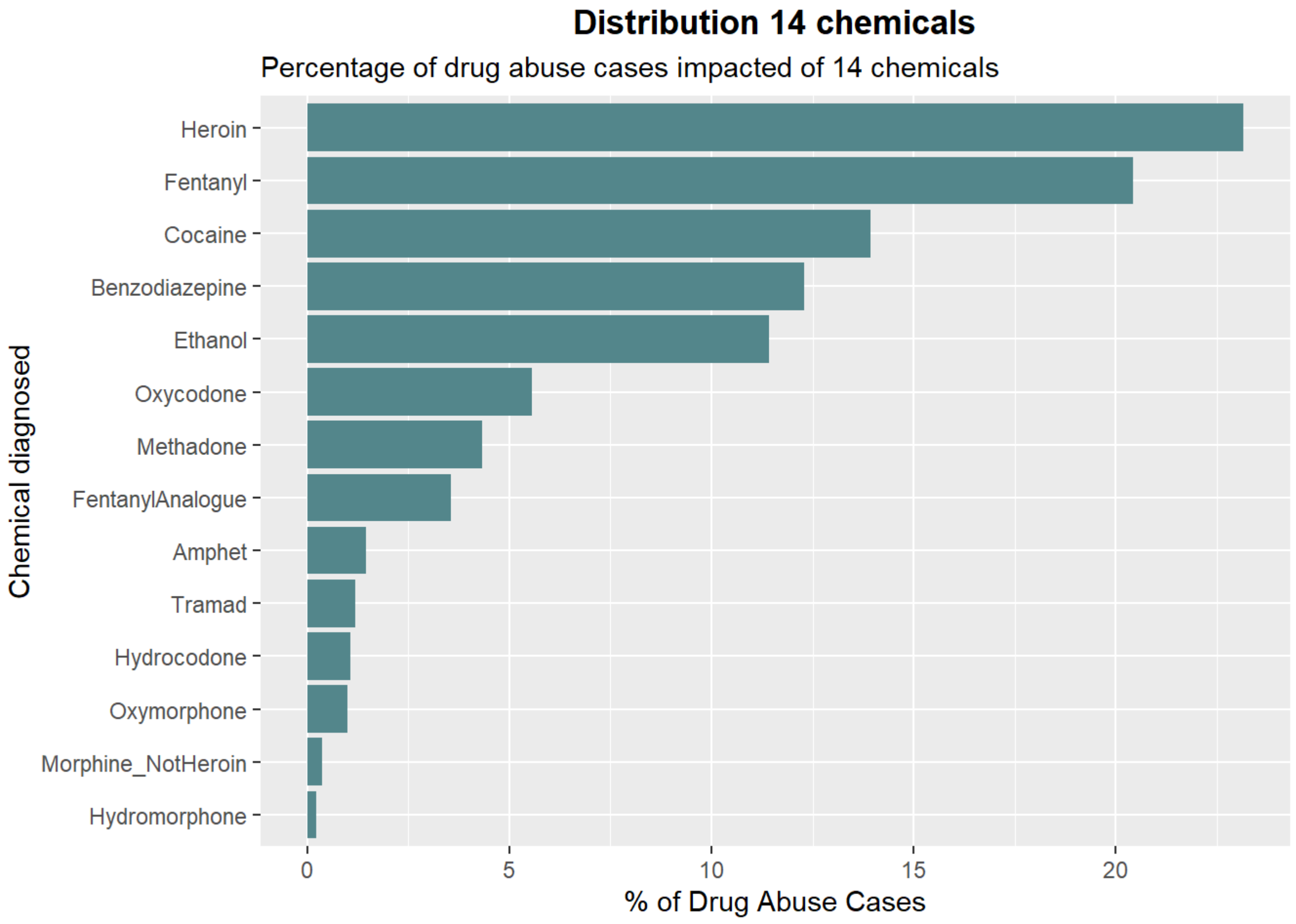
```
# histogram of number of chemicals in different drug abuse caess
total = nrow(df_rca)
temp_df <- df_rca %>% group_by(diagnosed_chemicals_count) %>% summarise(Count = n()*100/total)
ggplot(temp_df, aes(x = reorder(diagnosed_chemicals_count, -Count), y=Count)) +
  geom_histogram(stat="identity", fill = "burlywood4") +
  ggtitle("Number of Chemicals",
          subtitle = "Number of chemicals found in the medical investigation of drug abuse case
s") +
  labs(x = "Number of checmicals diagnosed", y = "% of drug abuse cases") +
  theme(plot.title = element_text(face = "bold"))  +
  theme(plot.caption = element_text(color = "grey68"))+
  theme(plot.title = element_text(hjust = 0.5))
```



Distribution of 14 chemicals in Drug abuse Cases

Here we investigate the chemicals found in the medical investigation of drug abuse cases. We have plooted the % of drug abuses cases where each of the chemical were diagnosed. Three major Chemicals - Heroin, Fentanyl, and Cocaine account occur in 47% of drug abuse cases. Rest of the underlying chemicals are relatively less frequent.

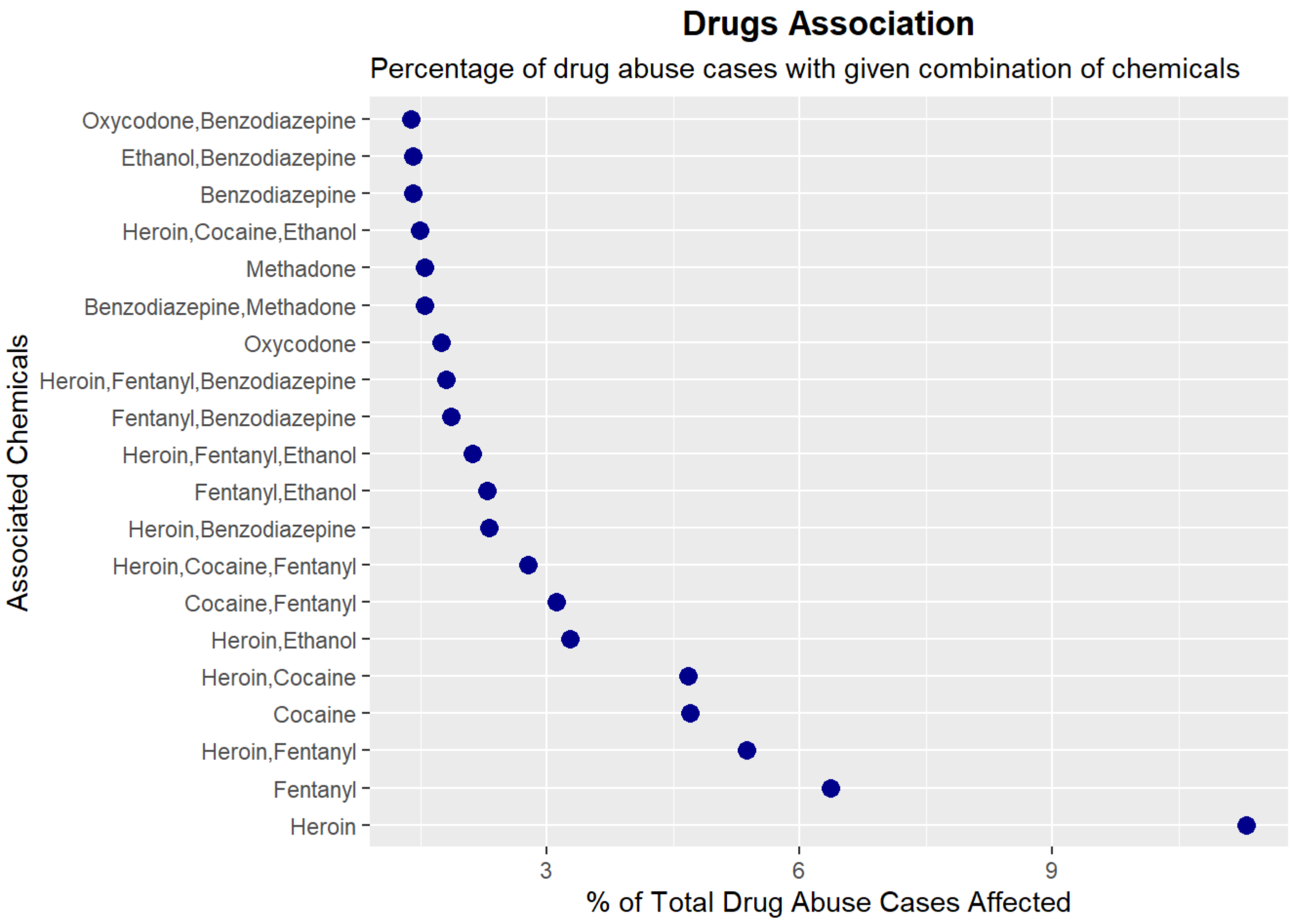
```
# Contribution of chemicals in different cases
temp_df <- data.frame('Drug_Composition'=all_chemicals)
total = nrow(temp_df)
temp_df <- temp_df %>% group_by(Drug_Composition) %>% summarise(Count = n()*100/total)

ggplot(temp_df, aes(x = reorder(Drug_Composition, Count), y=Count)) +
  geom_histogram(stat="identity", fill = "cadetblue4") +
  coord_flip() +
  ggtitle("Distribution 14 chemicals",
          subtitle = "Percentage of drug abuse cases impacted of 14 chemicals") +
  labs(x = "Chemical diagnosed", y = "% of Drug Abuse Cases") +
  theme(plot.title = element_text(face = "bold"))  +
  theme(plot.caption = element_text(color = "grey68"))+
  theme(plot.title = element_text(hjust = 0.5))
```



Associatio mining aming 14 Chemicals

We want to further mine the association between these 14 chemicals - which combination of chemicals occur most frequently. Among all combinations occured together, Heroin with Fentanyl(5%) and Heroin with Cocaine(4.7%) and Heroin with Ethanol(3.3%) were the frequently found combinations. Hence these two are popular associations discovered. As we can notice here, Heroine is present in all the popular associations as Heroine is concluded to be most frequently found chemicals.

```
# Contribution of chemicals in different cases
temp_df <- df_rca[df_rca$Diagnosed_Chemicals>1,]
temp_df <- data.frame(temp_df  %>%
                group_by(Diagnosed_Chemicals) %>%
                summarise(Freq=n()) %>%
                mutate(Perc_Cases=100*(Freq/nrow(df_rca))))
temp_df <- temp_df[temp_df$Freq>1,]

#temp_df <- temp_df[,temp_df[order('Perc_Cases', 'Freq','diagnosed_chemicals')]]
temp_df <- temp_df[order(temp_df[,'Perc_Cases'] ),]

# Top 20 combinations occured together
temp_df <- tail(temp_df, 20)


ggplot(temp_df, aes(x = reorder(Diagnosed_Chemicals, -Freq), y=Perc_Cases)) +
  geom_point(color='darkblue', size=3) +
  ggtitle("Drugs Association",
          subtitle = "Percentage of drug abuse cases with given combination of chemicals") +
    coord_flip() +
  labs(x = "Associated Chemicals", y = "% of Total Drug Abuse Cases Affected") +
  theme(plot.title = element_text(face = "bold"))  +
  theme(plot.caption = element_text(color = "grey68"))+
  theme(plot.title = element_text(hjust = 0.5))
```



Analysis of Injury Description - Text Mining using NLP

In this section, we are exploring the injury description for all drug abuse cases. The injury description(“DescriptionofInjury”) is basically the broder categorization of the mode/form of drug consumed in that particular drug abuse case. This field contains the freely written text hence we need to perform the baisc natural language processing.

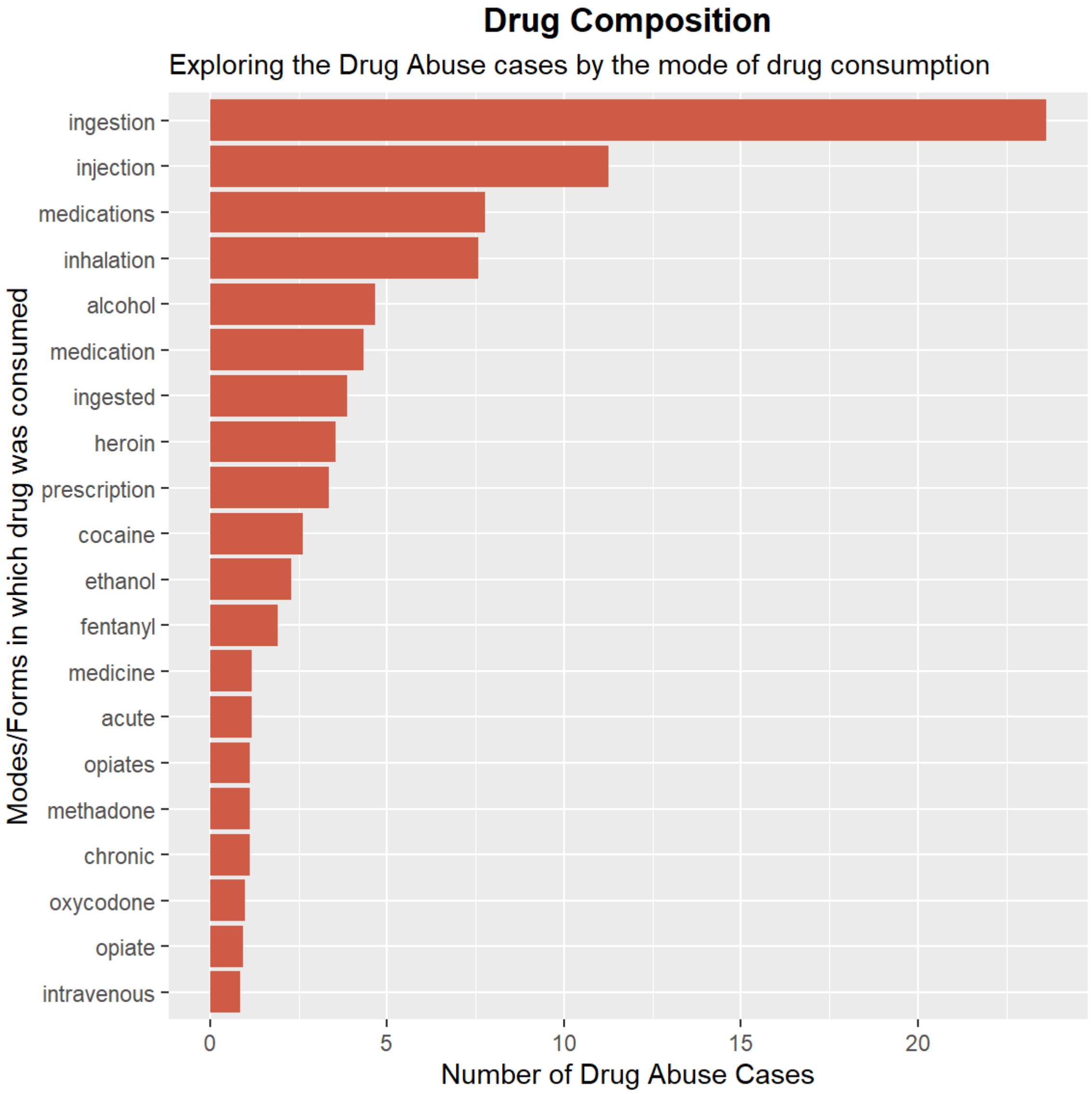
Below are the following steps performed by us:

1. Tokenization: We are first breaking the description of every drug abuse case in tokens.
2. Stop Word removal : From these tokens we are majorly looking for the modes and forms in which drugs were consumed.Hence, tokens with noun and verb POS tag appeared here will be our candidate tokens. In order to extract these candidate tokens, we are removing the stop words appeared in the token list.

First we are using the stop word removal step to remove all the stop words and the tokens with verbs and nouns POS tag from the token list. Lets us call these filtered tokens as drug consumption modes. We are calculating the frquency of these modesl.

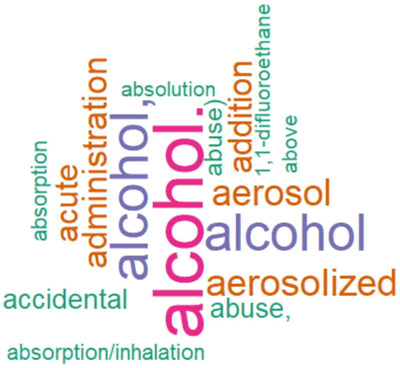
Insights: Ingestion, Injection and medications were the most frequent mode of drug consumption. Ingestion in 23% drug abuse cases, injection with 11% cases and medications in 7.7% cases. Together these three modes account for the 42% of total drug abuse cases.

```
stop_word_list = c('use','and', 'used', 'took', 'multiple', 'of', 'with', 'combined', 'misuse',
'including', 'consumed','in', 'abused', 'abuse','another', 'while','unknown', 'to','effects','us
age','own', 'others','drank', 'substance', 'drug')

remove_stop_words <- function(df){
  filter_df <-  df[!(df$Injury %in% stop_word_list),]
  return (filter_df)
}
all_injury <- c()

for (row in 1:nrow(df_rca)) {
  val = as.character(df_rca[row, "DescriptionofInjury"])
  split_val <- unlist(strsplit(val, ' '))
  all_injury <- c(all_injury, split_val)
}
all_injury <- tolower(all_injury)
temp_df <- data.frame('Injury'=all_injury)
temp_df <- temp_df %>% group_by(Injury) %>% summarise(Freq = n())

temp_df <- na.omit(temp_df)
temp_df <- as.data.frame(temp_df)

# stop word removal
temp_df <- remove_stop_words(temp_df)
total = sum(temp_df$Freq)
temp_df$Freq <- (temp_df$Freq)*100/total
df_word_cloud = temp_df

#temp_df <- temp_df[temp_df$Freq>1,]
temp_df <- temp_df[order(temp_df[,'Freq'] ),]

# plotting the top 20 major modes discovered
temp_df <- tail(temp_df, 20)
ggplot(temp_df, aes(x = reorder(Injury, Freq), y=Freq)) +
  geom_histogram(stat="identity", fill = "coral3") +
  coord_flip() +
  ggtitle("Drug Composition",
          subtitle = "Exploring the Drug Abuse cases by the mode of drug consumption") +
  labs(x = "Modes/Forms in which drug was consumed", y = "Number of Drug Abuse Cases") +
  theme(plot.title = element_text(face = "bold"))  +
  theme(plot.caption = element_text(color = "grey68"))+
  theme(plot.title = element_text(hjust = 0.5))
```



Further we are building the word cloud of the description of injuries.

```
# WordCloud of Injury Description
set.seed(1234)
wordcloud(words = df_word_cloud$Injury, freq = temp_df$Freq, min.freq = 1,
          max.words=500, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))
```



Interactive Component

Conclusion and Summary

Data regarding deaths due to drug overdose in the state of Connecticut from the year 2012 to 2018 was analysed. We followed the three fold approach to gather insights and find patterns in the drug abuse cases. The hypothesis for our demographic aspect 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 a 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 lives of many innocent people can be saved.

Github Link: # References [1] https://wallethub.com/edu/drug-use-by-state/35150/ (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-problem-in-their-community/ (https://www.pewresearch.org/fact-tank/2018/05/30/as-fatal-overdoses-rise-many-americans-see-drug-addiction-as-a-major-problem-in-their-community/)

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