Data Science Capstone Project - Milestone Report

Accidental Drug-Related Deaths in Connecticut 2012-2018

***Executive Summary***

The subject data is accidental drug-related death in the state of Connecticut from 2012-2018. The goal of this capstone project is to develop a data product using an R programming language that will allow us to find the cross-correlation between age, sex, race, the cause of death, death place, and death person place of living. Based on raw data we have a hypothesis that there is should be a strong relationship between age and cause of death, race and age and sex and age. The data has a large amount of information on accidental drug-related death by race, sex, age, Death County, the cause of death and geographical location of the incidents. The data analysis can help the health department and safety authorities have a deep understanding of the situation and make necessary conclusions to prevent, stop and improve the drug-related deaths in the state of Connecticut.

***Capstone Project Goals***

*Identifying areas of Interest*

Experiential Goals -- to contribute to solving a community health problem in a meaningful, effective, and culturally sensitive fashion, specifically by:

Working on data available on reliable sources and working closely with other agencies of the governance find the best solution to minimize the accidental drug-related deaths.

* Finding and applying evidence-based solutions to a defined community problem.
* Based on data analysis find the pattern and correlations between sex, race, and age which can help emphasize our attention to a particular group that is more vulnerable to this issue.
* Help health department, police department and government’s all other related agencies work together to find the complex solution to this matter.

***Getting and Loading Data***

The data was obtained from the official site of the state of Connecticut ([<https://data.ct.gov/Health-and-Human-Services>](https://data.ct.gov/)) in excel format which been converted into CSV format in order to upload into R studio for further data wrangling and visualization. These data files were downloaded and saved in the working directory in R.

***About the Dataset***

The dataset has 5105 observations and 41 variables for the state of Connecticut in drug-related deaths from 2012 to 2018 including the variables like sex, race, Death County, death city with geographical coordinates, cause of death, injury place, and location of the death, etc.

* “Date” column shows the date of the accident in the format of m/d/Y/ H/m.
* “Age” column includes all ages from 14 to 87 years old.
* “Sex” column has three levels: male, female and 7 observation of unknown.
* “Race” column includes 7 levels (white, Hispanic white, black, Hispanic black, Asian other, other, and unknown) which shows the dominance of white people in the state.
* “Death County” column has all state’s 8 counties observations included in our dataset
* The “Location” column has 7 different levels that describe the location of a person after the accident.
* The “Description of Injury” column includes 306 different types of drug usages.

In data set also included the cause of death, type of drugs (17 different kinds) been used, DeathCityGeo and ResidenceCityGeo where in last two columns we have cities names and geographical coordinates aka latitudes and longitudes.

***Cleaning and Wrangling the Data***

After downloading the data into the R studio, we need to take some steps of cleansing activities to ensure that unwanted and unnecessary data is removed from the Corpus data. We named our dataset accid2 and by exploring the structure of the data by using glimpse function, we can discover that we have a factor, integer, and character variables in our data. Using the pipe command (%>%) we will remove the duplicate rows and check the number of rows. There are 11 columns ("ID", "DateType", "ResidenceCity", "ResidenceState", "LocationifOther", "InjuryPlace", "InjuryCity", "InjuryCounty", "InjuryState", "OtherSignificant", "InjuryCityGeo") that is not valuable to our objectives and for that reason, we will remove them from our dataset by using the select function. The date format in our dataset would be more useful if we will break it into three different columns converting them as a double.

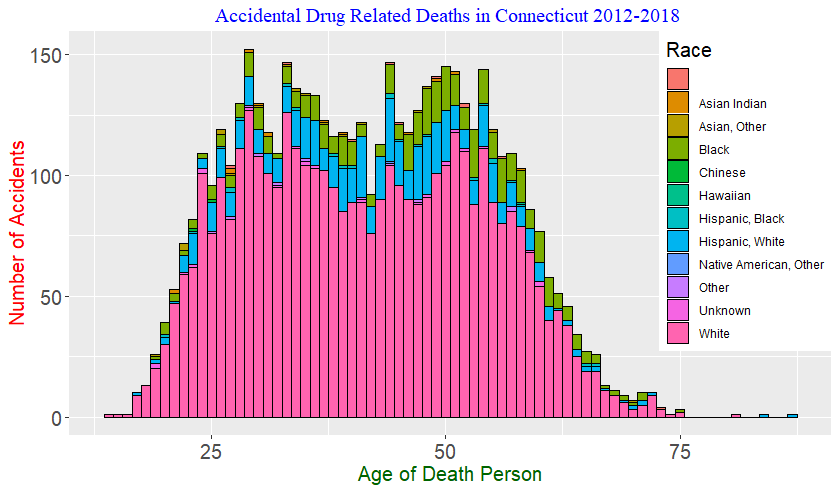
After wrangling we start to explore data trying to find real associations and correlations between our data variables. By filtering some levels of our variables (for example: in race column, we have white, black, Hispanic\_white, and etc, in sex column we have male and female) we can notice that we are going to need to add more columns for grouping and creating the bins for that groups. We will create age\_groups, sex\_groups, and race \_groups which will help us after in the linear regression model and chi-squared test calculations.

***Plotting and Mapping***

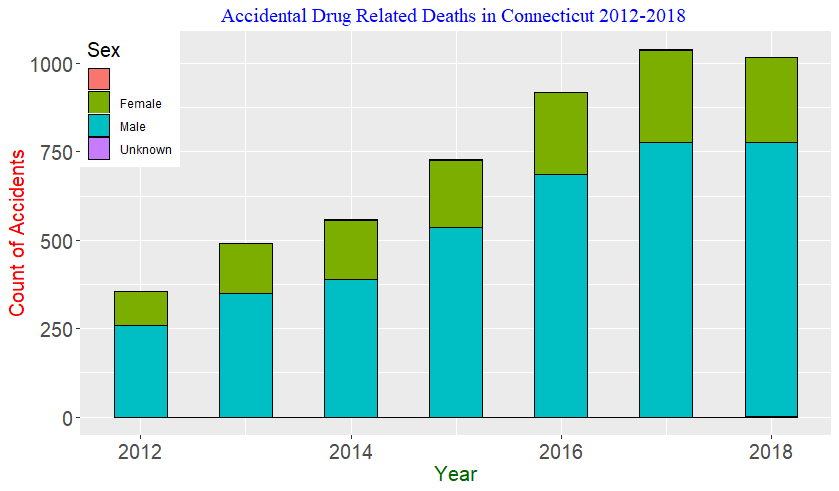
Visualizing the year and number of accidents correlation we can notice that since 2012 the accidental drug-related deaths increased drastically until 2017 and slightly dropped in 2018 which can be a strong concern for state population and for the entire nation as well.

Our goal is to find any type of association or correlation between the variables that we have in our dataset and for that reason we can use the R programming language in order to create plots, tests, and maps to have more global understanding of how big this issue is and how we can interact with this epidemic problem.

In this plot we can see how strong the correlation between age and race is, and to have more precise results to this hypothesis we can run the chi-square test



which will give us p-value that will describe the correlation level of age and race. (X-squared = 113.41, df = NA, p-value = 0.03448). A smaller p-value means that there is stronger evidence in favor of the hypothesis that we have a strong relation to age and race and specifically to the age group of 35-50 years old. Using the concept we can plot the sex relation to the age of the person and as we can see in a plot that we have the same strong relationship between these two variables as well.



The chi-squared test result can sustain this hypothesis (p-value = 0.03498). Similar chi-squared test for race and death-county and sex and death-county maintain the hypothesis that there is a relation between race and sex to the state counties which gives us more valuable information that we have state issue by counties. In our dataset, we have a column “deathcitygeo” where we have geographical coordinates of the cities that by using the library(stringr) and extract and regex functions we can create new columns identifying the longitudes and latitudes. The results show that the existing issue is more emphasized in the larger counties. With library(leaflet) we were able to create maps where we could visualize the geographical locations of the different type of drugs that been used at the time of accidents.