Project Description

- Group Number: 13
- Group Members: Elyse Levine, Kunal Mehta, Mandeep Singh
- Dataset used: CT Accidental Drug Related Deaths 2012-June 2017
- Dataset link: https://www.kaggle.com/addynaik/ct-accidental-drug-related-deaths-2012june-2017

Dataset Description:

A listing of each accidental death associated with drug overdose in Connecticut from 2012 to 2018. The 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. Medical examiners on the scene have documented the demographic of each victim, along with their COD (cause of death). Twenty-nine variables are used to describe each of the 3583 cases. Importantly, there are many binary variables used to indicate the particular substance detected. Other variables included are integer variables, such as age, categorical variables such as gender, and race, and so on. On examining the dataset, it is clear from a cursory glance that this dataset requires a lot of cleaning, for instance, making the drug observations truly Boolean by substituting a TRUE for Y and a FALSE for no record, making sure the dates are all in MM/DD/YYYY format, pruning some trivial variables from the dataset.

Aims:

We as a group aim to examine the data behind accidental deaths associated with drug overdoses and come up with logical inferences regarding different demographics based on segregating factors like age, race, sex, geographic location, drugs involved and so on. For instance, some valuable inferences that we could derive from the dataset could be determining which age groups, or which races of people, or people of which sex were the most prone to accidental drug related deaths. Other inferences could include what geographical areas have higher densities of drug related deaths, how the rate of these deaths have grown over a certain period of time, what the future projection of these could be. Both individually and in combinations, we can also determine how significant a contributor to accidental deaths each drug is.

To each of these ends, we intend to use association rule mining, principal component analysis and some other data analysis methods to determine the inferences. Standard plots and other visualization methods will be key to our work. We hope to complete the project within the next 3 weeks, using packages taught in class as well as some that we may need to learn how to use on our own.