Preliminary Report

- Group Number: 13
- Group Members: Elyse Levine, Kunal Mehta, Mandeep Singh
- Dataset used: CT Accidental Drug Related Deaths 2012-June 2017

Progress:

Our chief objective for the past week has been studying the dataset more intently and planning how to clean and organize the data to best suit the questions we have, and precisely what actions we'd have to perform on the dataset to answer these. We have also tried to design a detailed star/snowflake schema that we will base our database on. Our dataset has 32 variables with a total of 3583 records, and in general is somewhat noisy and unclean.

Some of the conclusions we have reached regarding the data are as follows:

- Some of the dimensions are essentially useless since they have observations for very few records, for instance, the AmendedMannerOfDeath and the DescriptionOfInjury variables. Therefore, as of now we don't intend to include them in any of our operations.
- Some of these seem to be logically needless, for instance the variable Death State. Given that the dataset contains observations of drug-related deaths in Connecticut, it is a must that for each record, the Death State will be CT. Therefore, we intend to fill all the empty values of Death State as CT. Another example of such a variable is the MannerOfDeath variable. Given that the dataset is that of Accidental deaths, it is of no surprise that essentially all the records have Accident as the Manner of Death, making this variable mostly useless.
- So far we have not come to any conclusion as to what we shall be doing regarding missing values in the Residence State and Residence County variables. One possible solution could be to determine Residence State via the Residence County name if available. We could automate this process by linking a geographical database and isolating the County and corresponding State names from it. As regards records that do not have any observations for both Residence State and Residence County, we have not reached a conclusion as to what to do regarding those records so far.
- From the DeathLoc variable, we intend to isolate the Death Town and make that a new separate dimension that will help derive more insightful inferences geographically.
- We currently have 13 dimensions that correspond to different categories of drugs. The observations in all of these are binary in nature, either a Y or an N. We intend to condense these into fewer, possibly one single dimension in the form of a bitset.

Future Plans:

The next immediate step will be to clean the data and construct the database as planned. This will provide us with a solid base on which to build our next step of the project on. Post completion of the dataset, we intend to visualize most of the queries that we outlined in the project description. The last component of our project will be to apply some data analysis techniques we've learnt this semester to the data, a couple of which are association rule mining and clustering. In addition, graphical analysis will be used in order to demonstrate basic trends and proportions. We also plan to learn exactly the format in which we are supposed to present the project, as that is something that can potentially hamstring our project. We feel the constraint of time will probably limit how many of our plans come to fruition and how many we abandon. Not being able to work side by side is seriously torpedoing our productivity and work speed, as is the truckload of assignments and exams that are characteristic of this period of the semester. That being said, we are optimistic that we will achieve what we intend to via this project, which we hope is a stepping stone to tackling problems of larger and tougher magnitudes.

Rough Snowflake Schema: