**Death Analysis by State ETL Project**

**Nancy Campos, Rachel Brinkley, Ryan Byrne**

When looking for trends in outbreaks in the United States we chose the National Outbreak Reporting System (NORS) and the National Center for Health Statistics Mortality Surveillance System, which is a database maintained by the Center for Disease Control (CDC), to pull data from. Both datasets were in CSV format. The data obtained from the CDC had size restrictions, so we had to shrink our data by filtering our search down to adults (18 and above) in major metropolitan areas from 2009 to 2017. We named the NORS dataset “outbreak” table and named the CDC dataset “death” table. We then loaded our CSVs into Juypter Notebook.

Once the tables were loaded into pandas, we dropped columns that did not provide useful information. Then, we renamed the columns to be more intuitive. We used pgAdmin4 as our relational SQL database and created tables to load our CSVs into.

Once the data was loaded into pgAdmin4, we used several cleaning methods. One of our tables had several blanks, so we renamed the “null”s to Unknown in the “etiology” column from the outbreak table to make the data useable. We then compared the state names in both tables and dropped any state name that was not listed in both tables. Additionally, we removed territories and items that were listed as “multistate.” We normalized all state names to match exactly and re-named “Washington D.C. to District of Columbia.”

We then created a table listing the populations for each state and did a left join on the outbreak and population tables so both datasets would include population numbers per state. Afterward, we used the “union” command to put the datasets into one table named “death\_illness\_cause\_population.” A comprehensive list of quires can be found in the “queries.sql” file within the ETL folder.

Some initial analysis was done by running queries to see the number of illnesses and deaths by state. Unsurprisingly, California had the highest amount of cases. Future data analysts are advised to run queries that take population into account such as number of cases per capita.

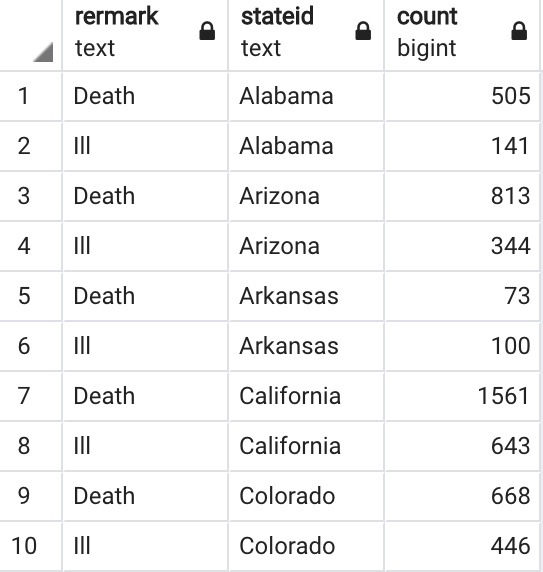
Image 1: Final table showing all fields

Image 2: Final number of cases, grouped by state