ETL Project

By: Mike Whitmore, Ernesto Beltran

**Extraction:**

We used two different data sources from data.world, which are listed below. The Gun-deaths data source has death count by county and state dated from 1999 to 2016. Police-deaths data source has police deaths by name, date, cause of death, and by the police department dated from 1971 to 2016. The Gun-death data was in a .tsv file format, and Police department deaths in csv file format.

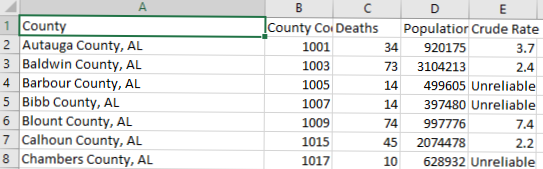
1. Police Deaths, url: <https://data.world/fivethirtyeight/police-deaths>
2. Gun Deaths, url: <https://data.world/nkrishnaswami/gun-deaths-by-county>

**Transformation:**

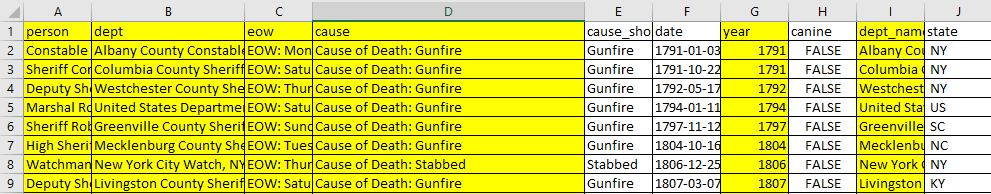
The first step taken was separating the state from the county in the gun-death data source because the state is the unique identifier needed to join the two data sources.

Next, we renamed the columns in the gun-death data source by replacing spaces with an underscore. Also, changing all column titles to be lower case letters to minimize potential errors in PostgreSQL.

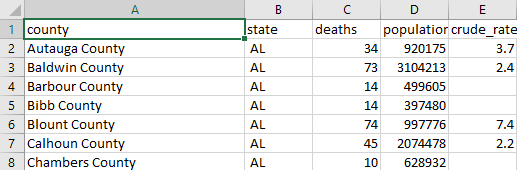
We removed columns in the police-deaths data sources, columns removed are person, ”dept”, “eow”, “cause”, “year”, and “dept name”. In the gun-death data source, we replaced "unreliable" values with "null" on the "crude\_rate" column. Then replaced "null" with blanks because we encountered a problem when uploading to PostgreSQL.

Before: 

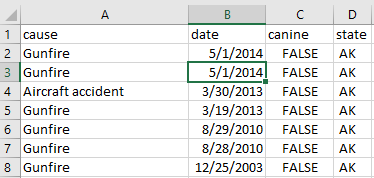
*Gun deaths*

*Police deaths*

After:



*Gun deaths*

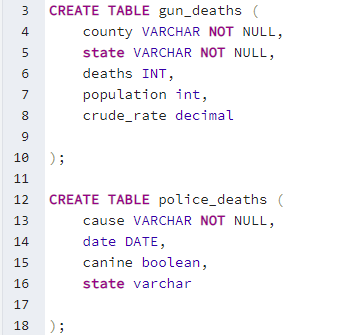


*Police deaths*

**Loading**

Using PostgreSQL, we created a new database, "etl\_database." We created 2 tables “gun\_deaths” table and “police\_deaths” table with matching column headers to the csv files.

“county”, “state”, and “cause” columns are varchar data types. Population and deaths are integer data types. “crude\_rate” is a decimal data type, date column is a date data type, and “canine” a Boolean data type.



Then we uploaded the csv files to each table.

We created two queries and assigned each of them an alias. The first query was to get the count of each death type by state and assigned it to alias “pd” on the “police\_deaths” tables. The second query we selected all the count columns for type of death from the “police\_deaths” table, then summed the deaths by state in the “gun\_deaths” table and attached it to “nopd” alias, and joined the “pd” table to “nopd” table on column state. To finalize our database and to hold our query we create a view and named it “police\_death\_summary.”

