Open- Ended Capstone Step 4: Data Exploration

1. Is the data homogenous in each column?

   Each column in the dataset adds a level of detail for each incident. For each incident there is more than adequate data for the purposes of this project.

2. How do you anticipate this data will be used by data analysts and scientists downstream?

   The data can be used to plot, using geolocation, the incidents onto a map. Each incident will be able to filter certain criteria such as critical vehicle information which can be used for purposes in determining car insurance rates. The data can be filtered by region, race of driver, reason for incident (description), and much more. Other information such as agency and location could potentially be used to determine critical intersections/regions where traffic flow could be improved to prevent incidents.

3. Does your answer to the last question give you an indication of how you can store the data for optimal querying speed and storage file compression?

   Upon examine of the data we notice that multiple citations could be given to the same traffic stop. That means that much of the data for the indecent is also the same. Examples of this are vehicle details, geolocation, agency information, and Incident details.

4. What cleaning steps do you need to perform to make your dataset ready for consumption?

   In order to consolidate the information, we must split the date into smaller table. Each table contains attributes directly associated with the entity. Information will be divided into seven tables for easier querying. Each table below will have a primary key id which will correspond to the table described above.

   Table Driver

   - *Race*
   - *Gender*
   - *Driver City*
   - *Driver State*
   - *DL State*

   Table Vehicle

   - *VehicleType*
   - *Year*
   - *Make*
   - *Model*
   - *Color*

Table Agency

- *Agency*

Table SubAgency

- *SubAgency*

Table Report

- *Accident*
- *Belts*
- *Personal Injury*
- *Property Damage*
- *Fatal*
- *Commercial License*
- *HAZMAT*

Table Incident

- *Arrest Type*
- *Description*
- *State*
- *Violation Type*
- *Charge*
- *Article*
- *Contributed To Accident*

Table TimeStamp

- *Date Of Stop*
- *Time Of Stop*
- *Location*
- *Latitude*
- *Longitude*
- *Geolocation*

5. What wrangling steps do you need to perform to enrich your dataset with additional information?

   We must ensure that a per incident we can properly query the necessary information. We must first establish each unique incident based on time. This is the primary table for the entity tree. This table will be named time; of which, we will populate with each unique incident. Next, each incident has three sub entities; A report, an Agency, and Driver(s). It is important to note than an incident may involve multiple drivers. Each driver has a vehicle table, while each agency has a sub-agency that may have involved on the incident.