Project 2 Proposal

W200 Python 2018 Fall, Section 6, Group 3

High Level Summary

Names of Team Members: Chloe Wu, Hanna Rocks, and Sarah Iranpour

Name of github repository:

https://github.com/UCB-INFO-PYTHON/W200_Project2_RocksIranpourWu

Primary dataset: https://www.kaggle.com/usdot/nhtsa-traffic-fatalities

Location Variables

We intend to research some of the various location variables that are included in the accidents_2016 dataset. Relevant fields include, but are not limited to: *land_use, milepoint, latitude, longitude, special_jurisdiction, work_zone, city, state_name, county, and national_highway_system*.

Figure 1 describes the number of traffic fatalities in 2016 by state, sorted from the state with the most fatalities (California) to the state with the least (District of Columbia). Further analysis will be performed to identify a way to standardize these data. For example, the number of roads within the state, or square miles of the state may be used to make these data comparable.

To gain a more detailed understanding of traffic fatality locations, we will use the latitude and longitude to map the various accidents to identify “hot zones” for traffic fatalities. These locations may be viewed by season to identify any variation depending on the time of year. States with the highest occurrences of traffic fatalities will be analyzed for additional trends.

Below is a list of questions we will further explore:

1. Which states had the highest number of traffic fatalities per roadway and per square mile?
    a. Within these states, which cities and/or counties experience the greatest number of traffic fatalities?
    b. Do these results vary by year? By season?
2. What percentage of fatalities occur in work zones?
    a. Are work zones where fatalities occur properly indicated with appropriate signage?

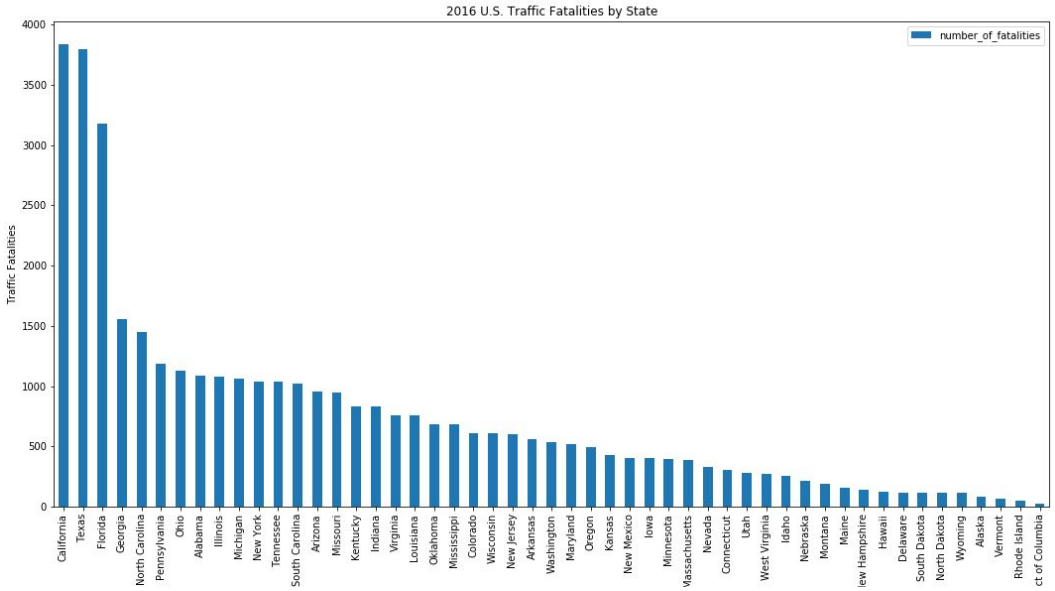

Figure 1: 2016 U.S. Traffic Fatalities by State

Time of the Year Variables

We would like to look at the time aspect of the datasets, specifically the following data groups and data elements:

- **Time of the accident**: day_of_crash, month_of_crash, year_of_crash, day_of_week, hour_of_crash, minute_of_crash Year of crash, month of crash, day of the week, hour of the crash
- **Atmospheric condition**: atmospheric_conditions_1, atmospheric_conditions_1_name, atmospheric_conditions_2, atmospheric_conditions_2_name, atmospheric_conditions
- **Light condition**: light_condition, light_condition_name
- **Road condition**: to investigate its correlation with atmospheric condition, roadway_surface_condition, roadway_surface_condition_name
- **Miscellaneous**: bus_use, special_use, school_bus_involved, driver_drinking, critical_event_precrash, pre_impact_stability

Our preliminary analysis on the time aspect is presented in **Figure 2**. It describes the number of traffic accidents in 2016 by month. The data shows gradual increase from the beginning of the year and reaching peak value in October, and gradual decrease in November to December. It would be interesting to see the monthly distribution by state, by weather / light and various conditions. Furthermore, it would be also interesting to refer to datasets from other years to see if the trend is consistent among different years.

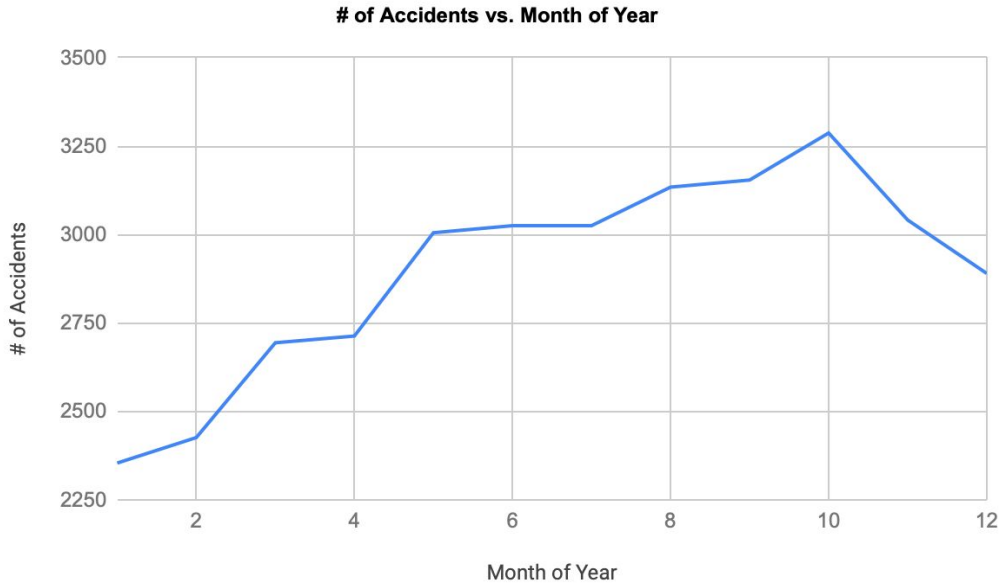

Figure 2: 2016 U.S. Traffic Accidents by Month

Therefore, in our detailed report, we would like to dive deeper into the data to understand the following questions:

1. When accidents and fatalities occurred more frequently?
    a. During certain months, days of the week, and time of the day?
    b. What are the accident / fatality rate during holiday season? People tend to drink more for celebration during holidays. Will this have an impact?
    c. Are drunken drivers involved more frequently in certain time of the year?
    d. What are the accident / fatality rate during normal work week days?
    e. What about normal work week nights vs. weekend nights?
2. Are accidents and fatalities connected to weather conditions?
    a. What type of weather condition tends to have higher accident / fatality rate?
    b. What is the relation between weather condition and light condition? Under which light condition does the data shows higher accident / fatality rate?
    c. What is the relation between weather and road condition?
3. It would be interesting to refer to datasets from other years to see if the trend is consistent among different years. The 2015 dataset is available on the website as our supplementary dataset.

Vehicles/Persons Involved Variables

We would like to look into some of the details of the accidents, such as the types of vehicles and the people involved. Some variables that may be relevant are, number of motor vehicles, number of parked vehicles, number of persons not in motor vehicle, number of persons in

vehicle, vision_obstruction, age, gender, etc. **Figure 3** shows the number of accidents that occured based on gender. Males were involved in more accidents than women in 2016.

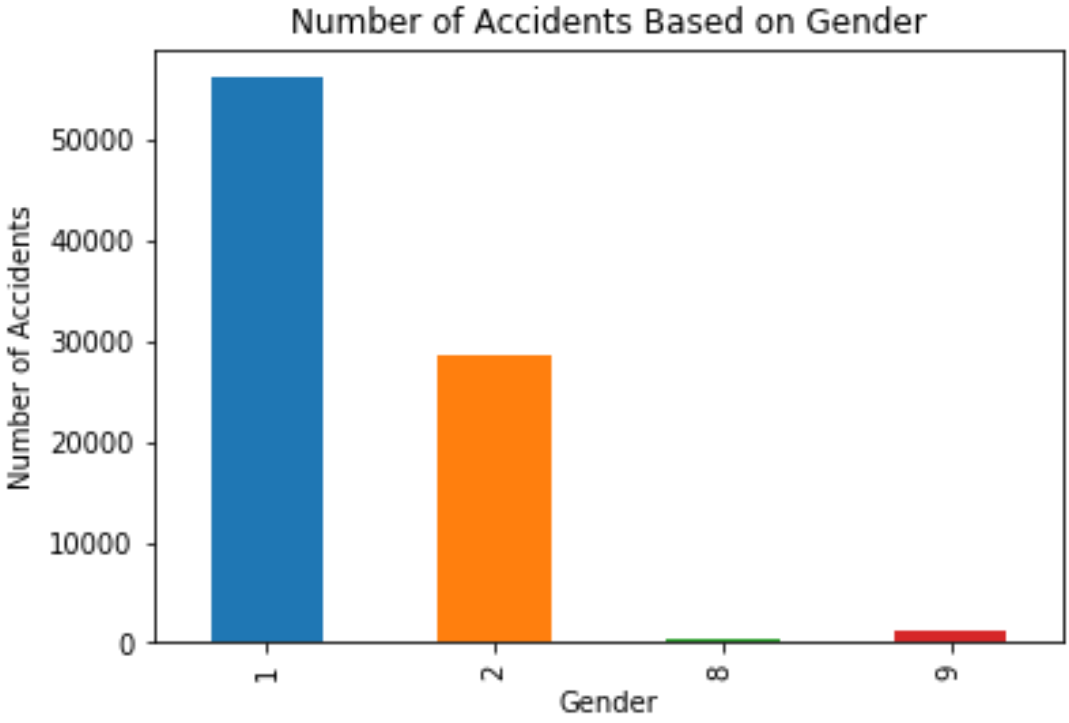

Figure 3: 2016 US Traffic Accidents by Gender

Here are some further questions we will explore:

1. What types of vehicles had the most fatalities? Motorcycles? Cars?
    a. Were both cars in motion or was one non-moving
    b. How many vehicles were involved?
    c. What parts of the vehicle were damaged? (Damaged_2016 dataset)
2. Is there high involvement of bikes? Pedestrians?
3. Which states have the most hit and runs? (Vehicle dataset)
4. How many people were involved? Were they in cars or mostly pedestrians? (People_2016 dataset)
    a. What age has the most fatalities?
    b. What gender has the most fatalities
5. What was the state of the driver? Were they distracted? (Distract_2016 dataset)
    a. Did the driver try to avoid the accident? (Manuever_2016 dataset)
    b. Was the driver's vision obstructed? (Vision_2016 dataset)
    a. Fatal injuries

Supplemental Datasets

- Explore datasets from other years to compare time series trend

- Number of highways/roadways in each states
- Square mileage in each state