Project 2 Final Report

Analysis of 2016 National Highway Traffic Safety Administration Accident Data

W200 Python Fall 2018, Section 6, Group 3

Group Members: Chloe Wu, Hanna Rocks, and Sarah Iranpour

Table of Contents

I. Introduction

For this project, we chose to examine data from the 2016 National Highway Traffic Safety Administration (NHTSA) Fatality Analysis Reporting System (FARS). NHTSA FARS was created to gather data on highway safety in the United States to allow for analysis and suggested solutions or improvements to highway systems across the country.

Records included in FARS "involve a motor vehicle traveling on a traffic way customarily open to the public and result in the death of a person (occupant of a vehicle or a non-occupant) within 30 days of the crash."[1]

Our primary dataset for analysis was the 2016 NHTSA FARS, obtained from Kaggle.com. A further description of Kaggle.com and the format of the source data is provided in **Section III.** To supplement parts of our analysis and provide additional insight, information was taken from the 2015 and 2017 NHTSA FARS reports.

II. Project Questions

For this project, we aimed to address several questions falling into the following categories:

1. Location Variables
    a. Which states had the highest number of accidents?
    b. Is there a significant difference between the number of fatalities occurring in rural areas versus urban areas?
2. Time Variables
    a. Was there a month of the year during which more accidents occurred?
    b. Was there an hour of the day during which more accidents occurred?
    c. Was there a difference between the number of accidents occurring on weekends versus weekdays?
3. Vehicles and/or Persons Involved Variables
    a. Were there certain types of drivers who were more susceptible to fatal accidents?
    b. Were there certain vehicle makes that were more frequently involved in fatal accidents?
    c. Which part of the vehicle was most frequently impacted in a fatal accident?

III. Data Preparation

The data for this project was downloaded from Kaggle.com, an online community that provides a platform for data scientists to download, publish, and explore datasets. Data shared through Kaggle has previously undergone cleaning procedures. As such, the data required very little cleaning before data exploration could begin. Although our primary data source was the 'accidents' dataframe, several supplemental data sources were also used. **Table 1** below gives a brief description of the columns that were used in our analysis.

Variable	Description
state_name	The state in which the crash occurred
land_use_name	Rural, urban, traffic way not in state inventory, not reported, or unknown
latitude	Identifies the location of the crash
longitude	Identifies the location of the crash
number_of_fatalities	The number of fatally injured people in the crash
MONTH / month_of_crash	The month in which the crash occurred.

[1] https://www.kaggle.com/usdot/nhtsa-traffic-fatalities/home. Accessed 12/8/18.

Variable	Description
DAY_WEEK / day_of_week	The day of the week on which the crash occurred. 1 Sunday 2 Monday 3 Tuesday 4 Wednesday 5 Thursday 6 Friday 7 Saturday -- Unknown
HOUR / hour_of_crash	The hour at which the crash occurred. 00-23 hour -- not applicable or not notified 99 unknown
DRUNK_DR	The number of drunk drivers involved in the crash. 00-99 number of drunk drivers involved in the fatal crash.
atmospheric_conditions_name	The prevailing atmospheric conditions that existed at the time of the crash as indicated in the case material
light_condition_name	The type/level of light that existed at the time of the crash as indicated in the case material
consecutive_number	The unique case number assigned to each crash
SEX	The gender of person involved in crash
MAKE	Vehicle make involved in the crash
AIR_BAG	Airbag deployment
MAN_COLL	Orientation of 2 vehicles in transport when they collide
MADREAS	Damaged areas on the vehicle as a result of accident
MDRDSTRD	Attribute which best describe driver's attention
MDRMANAV	Identify if driver tried to avoid accident

Table 1: Description of variables used in analysis

Before exploring these data, we performed a number of quality checks to ensure the accuracy of our analysis. These quality checks are described in detail for each section, below.

A. Location Analysis

The data source included the latitude and longitude for each recorded accident in 2015 and 2016 in the United States. We initially assumed that records without location information would have blank entries in the source table, however we soon discovered that this was not the case. By applying the function shown in **Figure 1**, we discovered 142 out of 32,538 records for 2015 and 129 out of 34,748 records for 2016 did not include valid geographic data. Both of these amounts represent less than 0.5% of the total records, and were therefore deemed insignificant and removed from geographical analysis.

```
def valid_latlon(x):
    '''Checks if latitude and longitude data are valid'''
    if (-90 < x.lat < 90) & (-180 < x.lon < 180):
        return True
    else:
        return False
def check_lat_lon(df):
    '''Creates a column to describe if the latitude and longitude data are valid (True)
        or invalid (False)'''
    df['valid_latlon'] = df.apply(lambda x: valid_latlon(x), axis=1)
```

Figure 1: Python Function to Identify Records with Invalid Latitude and/or Longitude

In addition to mapping the geographic location of each accident, the number of fatalities by state was examined for both rural and urban areas. While performing this analysis, we noted that the District of Columbia (DC) did not have any accident records on rural roadways. Thus, it was important to ensure that when merging rural and urban data, the DC records were not lost during the merge.

B. Time Analysis

For the time analysis, we first performed basic data sanity checks to describe the columns included in the dataset, the spread of the hours, and how many unique values were included. This allowed us to identify if there were any invalid records included in the dataset.

After performing this check, we made certain assumptions:

- If hour of crash is '99', it means there is no hour recorded for the accident. As our analysis is to compare the absolute accident amount under different conditions rather than the "accident rate", it is safe to assume that such records bear no value to our analysis. Thus, our analysis has removed records with hour_of_crash = 99 or HOUR = 99.
- For 2016 accident records, if their atmospheric condition is 'Unknown' or 'Not Reported', they are removed from our analysis. Please see **Figure 2**.
- For 2016 accident records, if their light condition is 'Unknown' or 'Not Reported', they are removed from our analysis. See **Figure 2**.

```
df2016_weather = accident2016_full[(accident2016_full['atmospheric_conditions_name'] != 'Unknown') &
                                   (accident2016_full['atmospheric_conditions_name'] != 'Not Reported') &
                                   (accident2016_full['light_condition_name'] != 'Not Reported') &
                                   (accident2016_full['light_condition_name'] != 'Unknown')]
```

Figure 2: Removing 'Unknown' and 'Not Reported' records from dataframe

C. Vehicles and/or Persons Involved Analysis

Similar to previous sections, we performed an initial sanity check on the variables used for this portion of the analysis. In most cases, the data was already clean and ready to analyze. However, as we analyzed time data with hours and minutes, we discovered a few issues. One issue was related to the military time formatting. For example, some accidents occurred right before midnight (hour 23) and the ambulance arrived at the scene right after midnight (hour 0). **Figure 3** shows an example of how we accounted for these situations. Another issue was potentially caused by human errors. For example, some accidents occurred at 1:30pm, but the ambulance arrival time was 1:20pm. In such cases, we removed the invalid values from our calculations.

```
# New Function to calculate the ambulance response time

def amb_diff(x):
    #change the unit to minutes
    total1 = (x.NOT_HOUR * 60) + x.NOT_MIN
    total2 = (x.ARR_HOUR * 60) + x.ARR_MIN

    #ambulance response time
    amb_resp_time = total2 - total1
    #error could be human error or the hour 23 changed to 0 hour
    if amb_resp_time < 0:
        if x.NOT_HOUR == 23 and x.ARR_HOUR == 0:
            amb_resp_time = 60 - x.NOT_MIN + x.ARR_MIN
        else:
            if x.NOT_HOUR == 23 and x.ARR_HOUR == 1:
                amb_resp_time = 60 - x.NOT_MIN + x.ARR_MIN + 60
    return amb_resp_time

def fix_amb_diff(df):
    # add response time column to the dataframe
    df['AMB_RESP_TIME'] = df.apply(lambda x: amb_diff(x), axis = 1)

fix_amb_diff(acc_df)
```

Figure 3: Python Function to Calculate and Fix Ambulance Response Times

IV. Our Data Stories

After completing necessary data cleaning procedures, we began analyzing the accident information. The following sections describe our findings related to accident locations, timing, and vehicles and/or persons involved.

A. Location Analysis

Data from 2015 and 2016 were analyzed to view locations within the United States that experienced a high number traffic accidents when compared to the rest of the country. **Figure 4** shows the geographic locations of recorded accidents in 2015 and 2016, respectively.[2]

The maps in **Figure 4** show that in both 2015 and 2016, there appeared to be a high concentration of fatal car accidents in Florida, Southern California, and along the Northeast Corridor of the U.S.

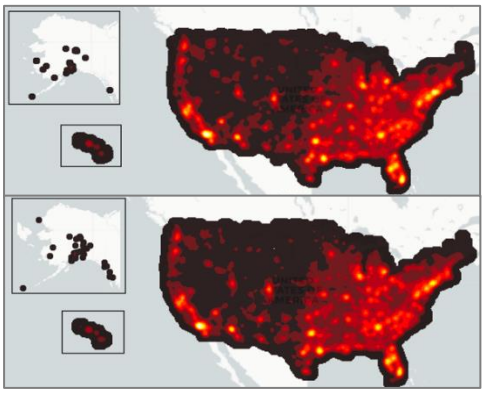

Figure 4: 2015 (top) and 2016 (bottom) Fatal Accidents in the United States

[2] Geographic visualization displays fatal accidents for which valid geographic data was available. See **Section III.A.** for details.

To confirm these results, we next examined specific accident and fatality data by state. The datasets included data for each of the 50 states and the District of Columbia.[3]

Accident Fatalities by State

As there did not appear to be a large variation in accident locations from 2015 to 2016, we moved forward with just analyzing the 2016 data by state. For a detailed analysis of the timing of fatal accidents throughout the year, see **Section B.**

To accurately compare the states to one another, data was obtained from the Federal Highway Administration (FHA) which reported the miles of roadway for each state and DC in 2016.[4] The report shows road miles designated as "rural" and "urban" for each state. With the help of this supplemental dataset, we were able to standardize the number of accident fatalities across the states by dividing the number of fatalities by the total road miles. **Figure 5** shows the accident fatalities per road mile across the United States in 2016.

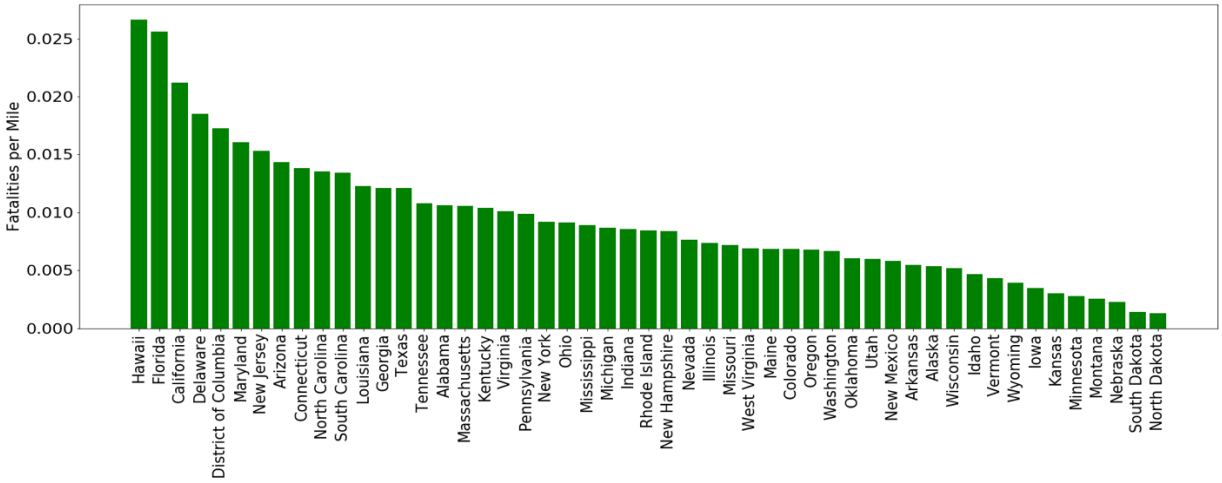

Figure 5: Accident Fatalities per Road Mile by State—2016

By standardizing the data, we found that some of the smallest states had the highest incidence of fatal traffic accidents. As described in **Figure 5**, the 5 states with the most traffic fatalities per road mile were: Hawaii, Florida, California, Delaware and DC. This is consistent with the observations from **Figure 4**, with the exception of Hawaii. Hawaii had only 1,509 reported road miles in 2016, which is approximately one-third the amount of road miles of the next greatest state, DC. This low number of road miles may explain Hawaii's top placement when examining fatalities per mile.

[3] For the purposes of this discussion, the District of Columbia shall be included as a "state" of the United States.

[4] Data obtained from: https://www.fhwa.dot.gov/policyinformation/statistics/2016/hm10.cfm. Accessed 12/3/18.

Fatal Accidents on Urban versus Rural Roadways

Next, we aimed to explore the differences between fatal accidents occurring on urban versus rural roadways in 2016. We targeted the top 5 states as determined by accident fatalities per mile in 2016. **Figure 6** shows traffic fatalities per road mile and geographic dispersion of those accidents for fatalities occurring on rural and urban roadways during 2016. Hawaii, Florida, and California had more fatalities per mile on urban roadways, while Delaware had more on rural roadways. DC has only urban roadways so a comparison is not available.

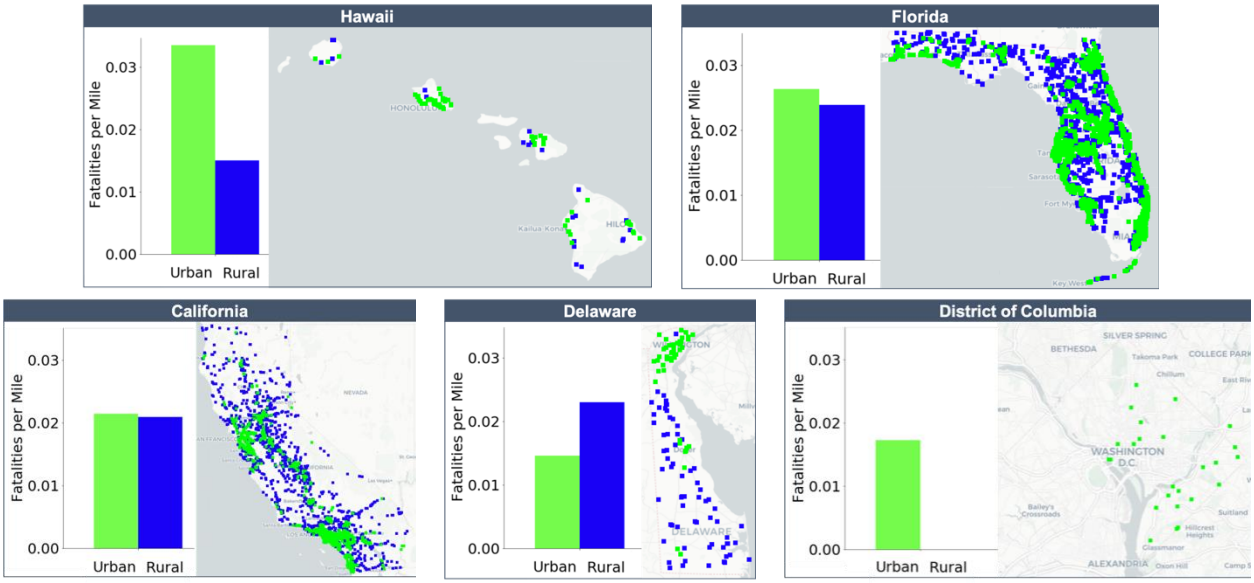

Figure 6: Fatalities per Road Mile and Geographic Dispersion of Fatalities, Urban and Rural Roadways—2016

B. Time Analysis

Accident Counts by Month and Day of Week

To conduct this analysis, we first load our downloaded datasets (in csv format) into accident2015, accident2016 and accident2017 dataframes. From there, we grouped the dataframes by Month or Day of Week, then calculated the accident counts in each month or Day of Week, and eventually combined 2015, 2016 and 2017 dataframes into one for plotting. **Figure 7** displays the results of this analysis.

We observed the following trends when reviewing accidents by month of the year:

- An increase in accidents from February to July;
- Number of accidents remain high from July to October.
- For 2015 and 2016, they reach peak value in October, and then decrease till December.

Additionally, when looking at accidents by day of the week, we observed the following:

- Drop in accidents from Sunday, and reaching lowest in Tuesday;
- Gradual increase from Tuesday all the way to Saturday.

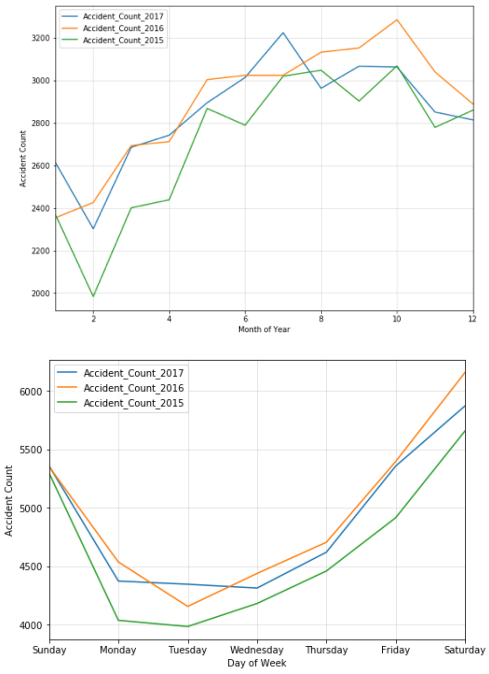

Figure 7: Accident distribution by month (top), and week (bottom)—2015-2017

Accident Counts by Hour of Day

Next, we analyzed the number of traffic accidents based on hours of the day from 2015 to 2017 (see **Figure 8**). Overall, we observed that the hourly trends year to year are remarkably similar.

- Decline in the accident count during night starting from midnight, and down to the lowest point around 4:00 in the morning.
- Increase in accident count after 5:00, and present a bump around 6:00 to 7:00, which could be interpreted as a result from morning traffic commute.
- Accident amount drops after around 7:00.
- Increase in accidents around 9:00, with a peak around 18:00, and goes down afterwards. The peak might be contributed by evening rush hour.

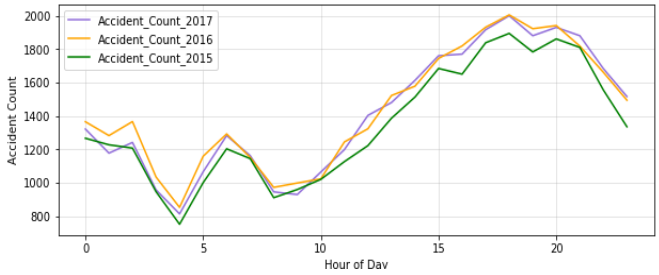

Figure 8: Accident distribution by hour of the day—2015-2017

Accident Counts by Weekday vs. Weekend

As we saw resembling patterns, we began to ask "what about the distribution in weekday vs. weekend?" If the small bump and peak in accident count could be interpreted as a result of morning and evening rush hours, the accident numbers in weekday would naturally follow similar pattern. With this question in mind, we segmented our 3-year accident datasets into weekend and weekday. We combined 2015, 2016 and 2017 weekday datasets, and calculated the average value of accident amount in every hour of weekday of these 3 years. Similarly, we merged 2015, 2016 and 2017 weekend datasets, and calculated the average value of accident amount in every hour of weekend of these 3 years. Using the formatted dataframe, we constructed a stacked bar chart shown in **Figure 9**.

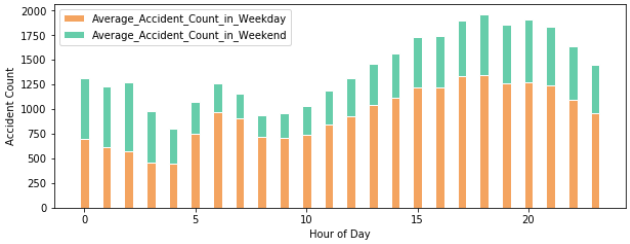

Figure 9: Average number of accidents of 2015-2017 by hour of day, categorized by weekday vs. weekend

From this analysis, we made the following observations:

- The weekday accident distribution by hour meets our expectation as mentioned above.
- The weekend distribution displays its uniqueness, in that,
  - Rather than decline after midnight, the accident counts stay high till 3:00.
  - The accident counts decrease afterwards and stay low till 10:00.
  - The number begins to grow afterwards all the way to approximately 20:00.
  - The uniqueness could be a result of people's activity hours are different between weekends and weekdays.

We further changed our perspective to explore alcohol involvement in accident during weekday vs. weekend. By categorizing 2016 accident dataset into "drunk-driver-involved" and "no-drunk-driver-involved" based off their DRUNK_DR value. If the value is greater than 0, it is considered as "drunk-driver-involved" ; otherwise, it is treated as "no-drunk-driver-involved", we were able to present the distribution by day in **Figure 10**. It presents noticeable hikes in value during Sunday and Saturday for drunk-driver-involved accidents.

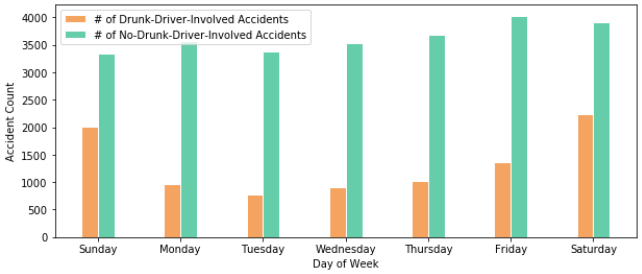

Figure 10: Alcohol involvement in accidents by day of week in 2016

Based on this observation, drunk drivers are involved more in accident during weekends than weekdays.

C. Vehicles and/or Persons Involved Analysis

In addition to analyzing accidents by location and time, we wanted to dive deeper into the actual crashes themselves. We were interested in gathering more details on the vehicles and persons involved in the accidents. To do so, we analyzed data from supplemental dataframes related to our primary accident table. These included person, vehicle, distract, and maneuver files.

Accident Fatalities by Persons

Which gender had a higher involvement in accidents?

We found that the amount of males involved in fatal accidents is almost double the amount of females (see **Figure 11**). However, one thing to note is that this does not necessarily mean that men are at higher risk than women to get in a fatal accident (it's possible that men drive more often than women).

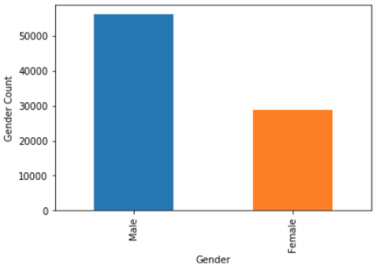

Figure 11: Number of accidents by gender

What was the state and behavior of the driver?

We conducted an analysis on the distract variable in the dataset and found that that the majority of drivers were not distracted or the distraction level was unknown or not reported. In this case, the number or unknowns and not reported were very high so we decided not to remove them from the dataframe to avoid skewing the data.

Approximately 61% of drivers were not distracted and about 31% were unknown or not reported. Another variable we looked as was whether or not the driver tried to avoid the accident. This is under the maneuver variable. This is another situation where we found that the majority, about 53%, of the data was 'not reported.' We did find that 33% of the drivers did not try to avoid the accident. For this portion, we may have had so many unknowns and not reported because the drivers and passengers of the vehicle may have passed away before they could be questioned.

After the time of the crash, it took (on average) approximately 7.8 minutes for someone to notify the authorities, 10.2 minutes for the ambulance to arrive at the scene of the accident, and 34.7 minutes to transport the person to the hospital.

Accident Fatalities by Vehicle

What are the top 10 vehicle makes involved in fatal traffic accidents?

Vehicle data from 2016 was analyzed to determine the vehicle makes that were involved in fatal accidents. **Figure 12** shows the top 10 vehicle makes that had a high number of involvement in crashes. Of the 98 vehicle types, we found that Ford, Chevrolet, and Toyota had the highest number of vehicles involved. It can be seen from **Figure 12** that Ford and Chevrolet's numbers are much greater than the rest of the vehicle types (approximately a difference of 5,000 – 10,000). One thing to note is that we do not have data on the number of people who own and drive each type of vehicle. Because of this, we were unable to normalize the data. This could be an area of further analysis for the two companies to determine the cause of the difference

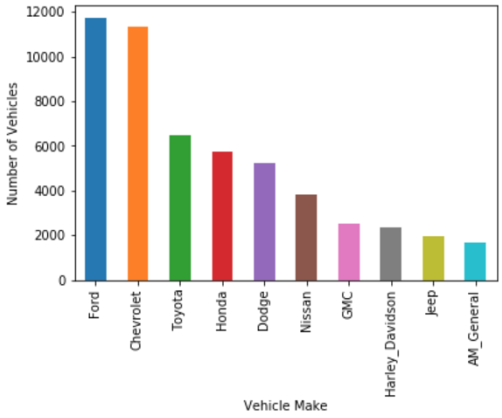

Figure 12: Top 10 vehicle types involved in fatal accidents—2016

(perhaps more people own and drive the type of car or there is a defect).

Which part of the vehicle was most often damaged that resulted in accident fatalities?

We also conducted some analysis on vehicle damage. **Figure 13** shows the various points of contact on a vehicle that were recorded in the data. The results can be seen in **Table 2**. The areas that resulted in the most damage were the front, front right, and front left of the cars. To validate these findings, we performed a check to investigate if the manner of collision matched the reported areas of damage. We found that the manner of collision is consistent with the damaged areas data. The cars usually collided front to front, at an angle, or front to rear.

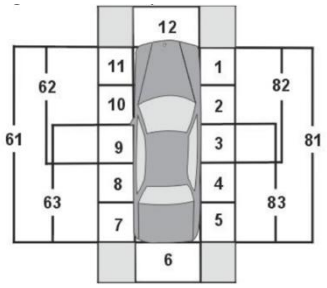

Figure 13: Areas of Impact—Initial Contact Point Element Values Diagram

Possible next steps related to this portion of the analysis would include a visualization of reported issues with the vehicle prior to the accident. Some possible maintenance issues recorded in the dataset were related to tires, the braking system, and headlights; however, these seem to be minor factors in traffic fatalities as initial analysis revealed the numbers are small. The majority of vehicles did not have any problems prior to the crash.

Damaged Area	Number of Vehicles	Damaged Area	Number of Vehicles	Damaged Area	Number of Vehicles
12	36,553	**2**	14,194	**6**	11,655
11	19,220	**13**	14,120	**5**	11,221
1	18,111	**3**	13,565	**4**	11,178
10	15,413	**8**	12,029	**14**	8,467
9	14,966	**7**	11,835	**15**	912

Table 2: Highest number of damaged areas on vehicles in accidents—2016

V. Conclusions

This concludes our analysis of 2016 National Highway Traffic Safety Administration Accident Data. With over twenty data tables, each containing between twenty and seventy variables, there is extensive opportunity to perform additional, deeper analysis in each of the three areas discussed above. It would also be interesting to perform an integrated analysis across the three areas. Further analysis would provide insight into traffic safety patterns and give regulators and traffic safety officials the resources required to make our roads safer throughout the United States.