US Vehicular Accidents

Kirsten Miller

5/15/2020

Table of Contents

## Introduction

I investigated vehicular accidents in the United States over a 21-year period (1996 to 2016). There are many potential ways to explore this dataset, so I tried to understand it broadly while also focusing on some more specific aspects and potential relationships. I first explored the data broadly by visualizing the number of people involved in vehicular accidents over varying timescales. I then looked at the breakdown of the number of people involved in accidents by many of the variables included to get a sense of the dataset. After this more general visualization, I attempted to uncover some of the relationships between specific variables and the number of people involved in accidents. In particular, I focused on injury severity, age, and alcohol involvement. I also focused on local states in portions of my analysis (Maine, New Hampshire, Vermont, and Massachusetts).

Some questions that guided this data exploration include:

* How has the number of vehicular accidents changed over different timescales (years, months, hours)?
* What is the breakdown of types of injury severity and has this changed over time?
* Are there significant differences between the ages of people involved in certain types of accidents?

While answering these questions, I also considered:

* How do vehicular accidents vary among regions of the country and local states and how does this compare to the United States as a whole?
* Could involvement of alcohol be a predictor for certain types of accidents?

## Methods

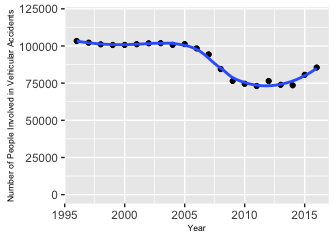
This dataset was collected by the National Highway Traffic Safety Administration (NHTSA) for the years 1996 through 2016. The data includes the following information for each person involved in a vehicular accident: state, county, month, day, year, hour, minute, manner of collision, number of vehicles involved, type of vehicle involved, number of people involved, age of driver, sex of driver, involvement of alcohol, and severity of injury.

An important note is that the data are broken down by individuals involved in vehicular accidents, not the unique accidents themselves. Therefore, my analysis is based on the number of individuals involved in vehicular accidents, not the number of vehicular accidents.

## Results

### Number of People Involved In Vehicular Accidents

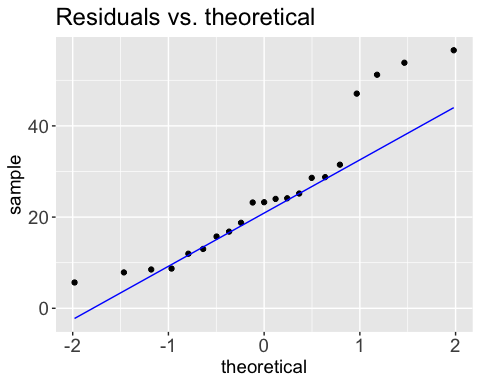
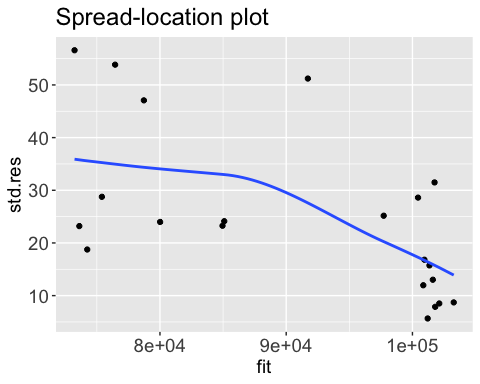
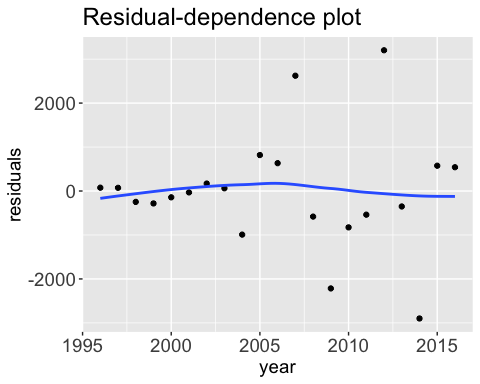
#### Number of people involved in vehicular accidents in the US over a 21-year period



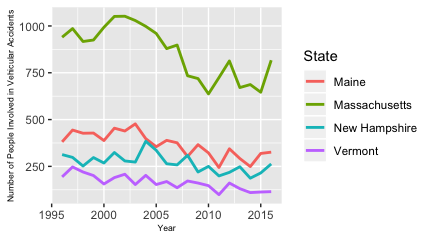
First, I examined the number of people involved in vehicular accidents in the US over the entire 21 year time period for the dataset (1996 - 2016). I used a loess fit and found that a second-degree loess fit was the best fit for this plot. The number of people involved in vehicular accidents was fairly constant (around 100,000 accidents) over the first 10 years of the dataset from 1996 to 2005. This was followed by a decrease in the number of people involved in accidents to a low number of 73073 in 2011, followed by a recent increase to 85496 accidents in 2016. The roughly linear trend prior to 2005 followed by a decrease in the number of people involved in accidents suggests that there could have been either a change in the method of data reporting or a change in regulations around 2005; however I was not able to find any concrete information to confirm or deny this.

##### Analytical plots for residuals

A second-degree loess was the best fit for the overall plot. The residual-dependence plot shows that the loess fit does approximate a horizontal line; however, there appears to be a potential fanning pattern of the residuals (the points become more scattered when moving from left to right) seen in the residual-dependence plot. The spread-location plot shows a decreasing spread of the original data, indicating that the variability decreases. I then checked the residuals for normality by comparing to the normal distribution (residuals vs. theoretical plot). The residuals align somewhat well with a theoretic distribution, although there may be some level of skew to the left.

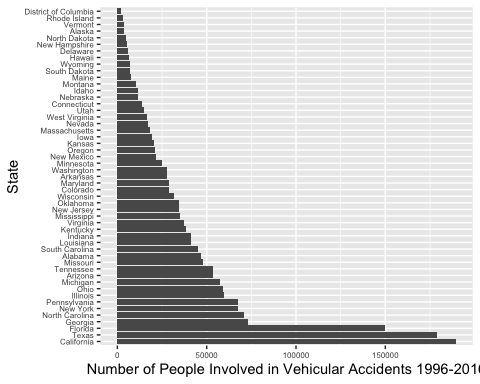


#### Number of people involved in accidents over time in local states



The number of people involved in accidents over time in local states appears to follow a similar pattern to the overall trend for the United States. Again, there is a change after 2005 (most pronouced in Massachusetts) which could indicate a change in the method of data reporting. In addition, with smaller total numbers of people involved in accidents in these states, the year-to-year variability is more apparent (compared to the number of people involved in accidents in the US as a whole). This also indicates the varying number of people involved in accidents by state, which is examined further below:

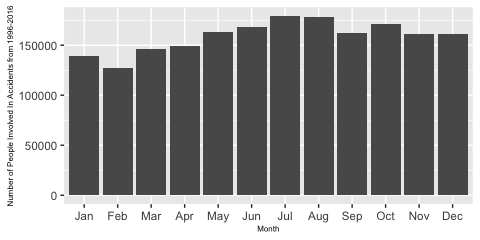
#### Cumulative number of people involved in accidents in each state from 1996 - 2016



The number of people involved in accidents in each state varies widely. The top 5 states with the greatest cumulative number of people involved in accidents over the 21-year period were California, Texas, Florida, Georgia, and North Carolina, while the 5 states with the least number of people involved in accidents were the District of Columbia, Rhode Island, Vermont, Alaska and North Dakota.

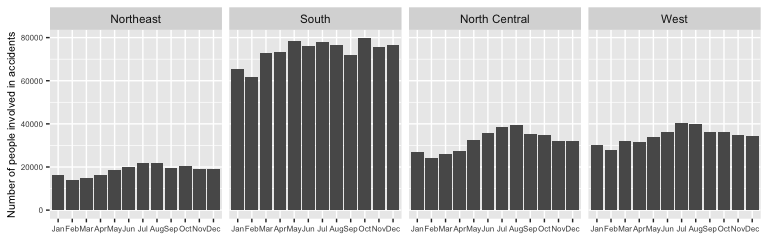
To further examine how the number of people involved in accidents changes over different timescales, I looked at the distribution of the total number of people involved in accidents over each month of the year.

#### Number of people involved in accidents per month of the year



I was surprised to see that the cumulative number of people involved in accidents was higher during the summer months (with the highest number occurring in July and the lowest number occurring in February). I would have predicted that there would be a greater number of accidents during the winter because of more weather events and difficult driving conditions. To examine this further, I looked at the number people involved in accidents each month by region (assuming that northern regions would experience more weather that could impact driving in the winter).

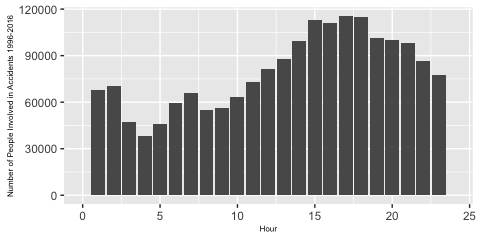
#### Number of people involved in accidents per month by region



Here, I was again surprised to see that the number of people involved in accidents was still higher in the summer in both the Northeast and North Central regions. However, the greater number people involved in accidents during the summer could be due to increased travel during these months. It was interesting to see that many more people were involved in accidents overall in the South Region (although this is likely because this region has a higher population).

Next, I looked at the distribution of the total number of people involved in accidents per hour of the day:

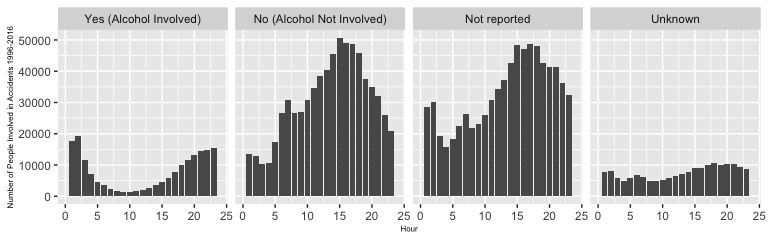
#### Number of people involved in accidents per hour of the day



This shows the most people were involved in accidents that occurred around 17:00, which is intuitive, as this is around the time when many people return home from work, as well as when it may be getting dark. It is also interesting to note that there is a small peak in the observed distribution at 6:00-7:00, which might be explained by the morning commute.

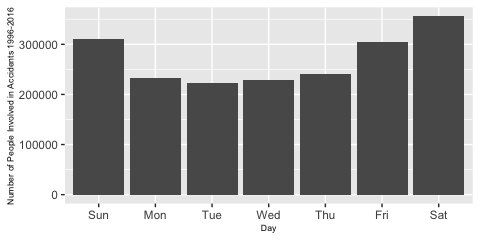
I was interested to see how alcohol involvement might be related to this distribution:

#### Number of people involved in accidents during each hour of the day, by alcohol involvement



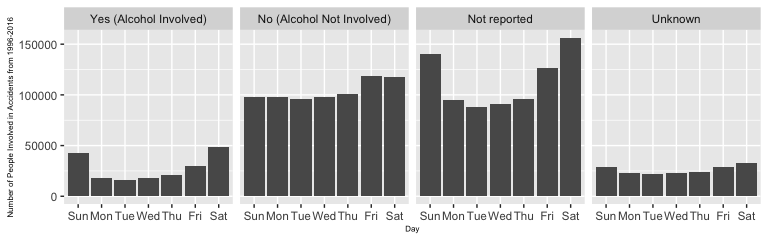
When alcohol was not involved, not reported, or unknown, the distribution of the number of people involved in accidents per hour is similar to the overall distribution. However, when alcohol is involved, the distribution is near opposite, with the peak of accidents occurring during the nighttime hours.

#### Number of people involved in vehicular accidents by day of the week



Finally, I examined the number of people involved in vehicular accidents by day of the week. The most people were involved in accidents that occurred on Friday and the weekend, while less people were involved in accidents that occurred during the week. The greatest number of people were involved in accidents on Saturdays (356068) whereas the least number of people were involved in accidents that occurred on Tuesdays (221969).

#### Number of people involved in vehicular accidents by day of the week, by alcohol involvement:

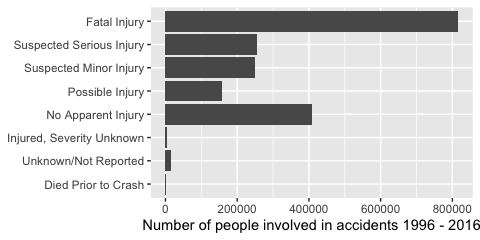


Again, I also looked at whether this breakdown by day of the week changed at all with alcohol involvement. When alcohol was involved (“Yes”), the distribution appears to be similar to the overall distribution, although the number of people involved in accidents on the weekend appears to be proportionally higher. When alcohol was not involved (“No”), the number of people involved in accidents is fairly consistent from Sunday through Thursday, and slightly larger on Friday and Saturday.

### Severity of Injury Sustained

After an initial investigation of some more of the dataset variables, (including vehicle type, manner of collision, and number of people involved), I decided that I was most interested in looking at injury severity. Below is the overall breakdown of injury severity for the cumulative number of people involved in accidents:

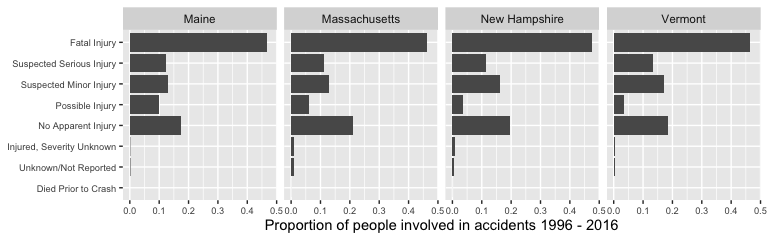
#### Injury Severity



Out of all the the people involved in accidents, fatal injuries makes up the largest category of injury level, with 815926 total fatalities recorded in this dataset. This is over twice as many as those who were reported to have no apparent injury. Those with possible, minor, or serious injuries also made up significant portions of the total number of people involved in accidents.

I also was curious to see if there were any evident differences in injury severity in local states, shown below:

#### Severity of injury by state



The breakdown of the categories of injury severity is fairly consistent across these states, with fatal injuries between 46 and 48 percent for all four states. The proportions are also fairly consistent with the overall breakdown for the United States above.

Next, I looked at injury severity by alcohol involvement:

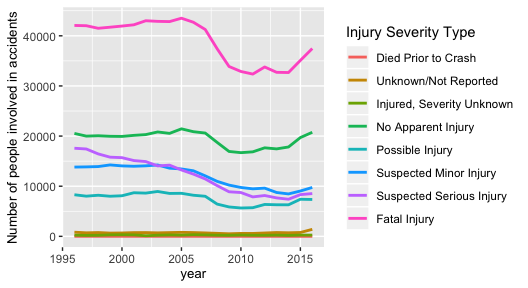
#### Severity of injury by alcohol involvment



When comparing the accidents in which it was known whether alcohol was involved (“Yes”) or not involved (“No”), it appears the proportion of fatal injuries was much higher when alcohol was involved. Fatal injuries made up 64.7% of accidents involving alcohol, while they only accounted for 36.6% of those accidents not involving alcohol.

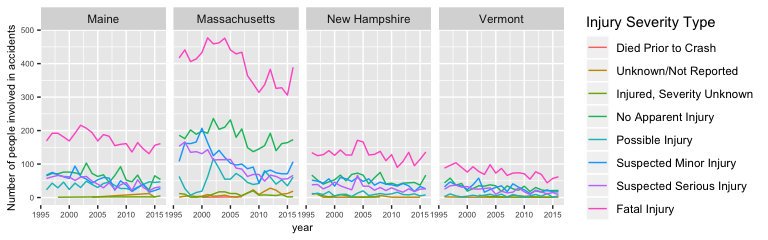
I also looked at how the different types of injury had changed over time:

#### Severity of injury over time



All types of injuries appear to have decreased over time, although after more substantial decreases between 2005 and 2010, injuries (as well as the no injury category) have increased again in the last 5 years of the dataset. Recalling the trend for overall number of people involved in accidents (the first figure in the results section), this decrease followed by a recent increase is very similar.

#### Severity of injury over time in local states

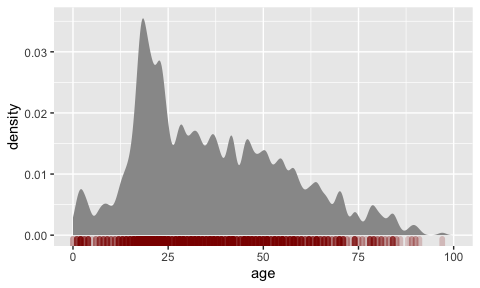


Again, the patterns over time in local states were relatively consistent and similar to the overall trend for the United States, with more year-to-year variability.

### Age of People Involved In Vehicular Accidents

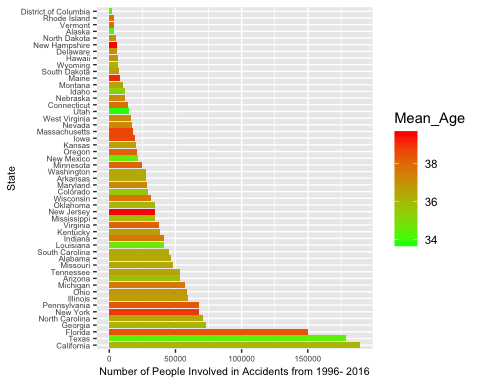
I was also interested in investigating the distribution of the age of people involved in vehicular accidents. I first looked at the distribution of the ages of people involved in vehicular accidents, and whether their average age has changed over time.

#### Density plot for age of people involved in vehicular accidents

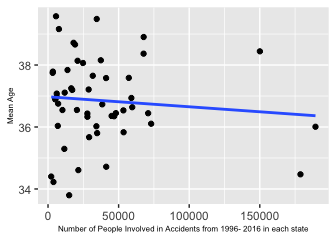


This density plot shows the distribution of the age of people involved in vehicular accidents. The median age of people involved in vehicular accidents over this time period was 33. However, looking at the density plot for age, the peak density occurs in the low 20’s. This is followed by a somewhat consistent density level from age 25 to age 50, followed by a decrease in density as age continues to increase.

#### Number of people involved in accidents for each state and mean age



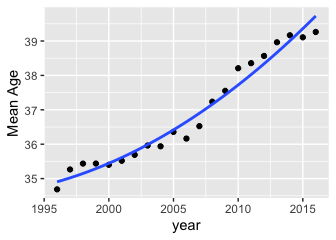
The plot above displays the number of people involved in accidents in each state, as well as the mean age in each state. The mean age of people involved in accidents does varies noticeably among the states. To further investigate this, I created a scatterplot comparing mean age and total number of accidents by state:



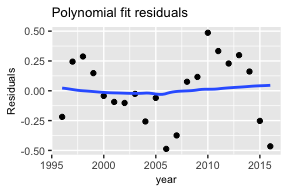
There is not much of a trend, and the three states with the highest number of people involved in accidents (California, Texas, and Florida) are noticeably separated from the rest of the data points. Based on this scatterplot, there does not appear to be a strong relationship between number of people involved in accidents and mean age in each state. The variation in mean age and number of people involved in acccidents between states could be influenced by the mean age and the total population overall in each state, so to further investigate it would be best to normalize both values based on the overall state mean age and population.

Next, I examined whether there has been a change in the age of people involved in vehicular accidents over time by plotting the mean age by year:

#### Mean Age of People Involved In Vehicular Accidents Over Time

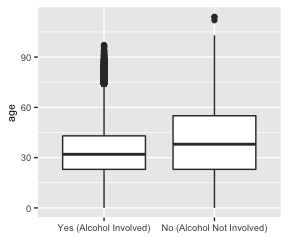


The mean age of people involved in vehicular accidents increased over time from 34.6877506 in 1996 to 39.2625147 in 2016. Although there was some inconsistency in the residuals, the data was best approximated by second degree polynomial fit, defining the relationship in the following equation:

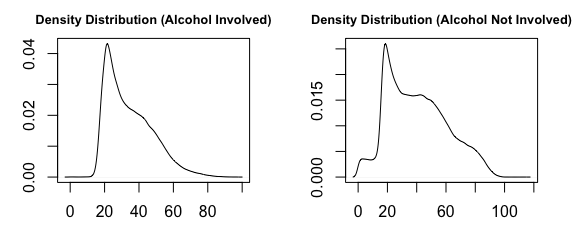


### Age By Group: Alcohol Involvement & Injury Severity

#### Age of people involved in accidents, by alcohol involvement

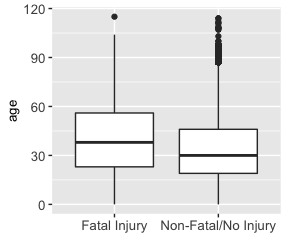


The median age of those involved in vehicle accidents when alcohol was involved was 32, and when alcohol was not involved the median age was 38. This suggests that the age of people involved in vehicle accidents with alcohol involvement could be lower than the age of people involved in accidents without alcohol involvement. Ideally, one would perform a parametric test to determine whether the difference in age between the two groups was significant or not. Below, I check the conditions required for performing a t-test. Unfortunately, the distributions are not normal and do not match each other so it is not possible to perform a t-test. After the peak around age 20 for both groups (the distributions are skewed to the right), there is a hump in the distribution at around age 45 which is even more pronounced for the alcohol not involved group. The alcohol-involved group also does not include many individuals under age 20, since people do not usually drive with their kids when drinking is involved, but this age group is represented in the alcohol-not-involved group.

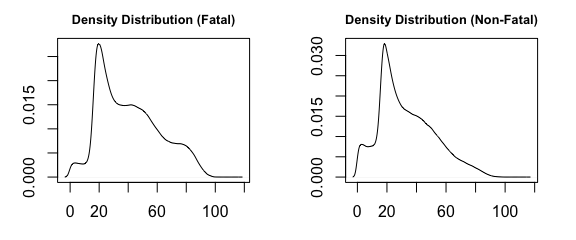


Next, I investigated whether there was a significant difference between the age of people involved in accidents who experienced fatal injuries and those who experienced nonfatal injuries.

#### Age of people who suffered fatal injuries versus non-fatal/no injuries



The median age of those who suffered fatal injuries was 38, compared to 30 for those who did not experience a fatal injury. This suggests that the age of those who experience a fatal injury may be older than those who experience a non-fatal or no injury. To test whether this difference is significant, a t-test could be performed. However, the conditions for a t-test are not met in this case because the distributions of both groups do not match the normal distribution (seen in the density plots below). Both groups are skewed to the right. In the group that experienced fatal injury, the density is higher in the middle age groups, with humps around 40 and 80, which are not as pronounced in the non-fatal group, which has a greater density for younger ages (below age 20).



##### Quantile - Quantile plot

The q-q plots below compare the values of the fatal group to those of the non-fatal group. They show that there is a systematic difference between the age of people who experienced fatal injuries and those who did not, with the age of those who experienced fatal injury being higher. In the plot on the right, this offset is quantified by adding 7 to the age of the non-fatal group. However, this q-q plot also reveals that the data should be divided into multiple groups for further analysis because the fit is not completely linear. The greatest discrepancy can be seen for below age 20, where the trend is most noticeably different from the line. There is also a lot of variability above age 90. Therefore, it would be best to determine the relationship between injury severity and age by dividing the data into two separate age groups: 0-20 and 20-90, because the relationship will likely be different for each group.



## Discussion

This exploratory analysis investigated various aspects of vehicular accidents in the United States over a 21-year period. The main areas of focus were change over varying timescales, injury severity, age of people involved, and alcohol involvement.

By analyzing this dataset, I found that the number of people involved in vehicle accidents, as well as the number of fatalities (in the US as a whole and in local states), has decreased over time. There are indications of a potential systematic decrease around 2005, although I did not find evidence to confirm this. However, both number of people involved in accidents and fatalities have increased in the last few years of the data examined. The number of accidents also varies by month of the year, hour of the day, and day of the week; accidents are higher during the summer months, evening hours, and weekend days.

Fatalities make up a significant component of the people who experienced vehicular accidents recorded in this dataset, and proportions of fatal injury as well as change over time of fatal injury are relatively consistent in local states and close to 50% for this dataset.

The mean age of people involved in vehicular accidents has increased over time. The age of people who experience fatal injuries through vehicular accidents is typically higher than the age of those who do not.

Accidents involving alcohol are especially prevalent during nighttime hours and on weekend days. The proportion of fatal injuries is higher for people in accidents for whom alcohol was involved. These data also suggest that alcohol involvement is more common in younger people who are involved in accidents.

These results identify certain time periods during which more people are usually involved in accidents, which is helpful for understanding risk, as well as in consideration of how vehicle accidents can be managed and prevented. These findings also suggest that both young people and older people may be at risk for vehicle accidents due to differing reasons: young people may have more accidents related to alcohol use, while older people may be more at risk for fatal injuries if they are to be involved in a vehicular accident.

There are many ways that this dataset could be explored more to highlight these potential relationships, as well as to uncover further trends and patterns. Perhaps most pertinently, it will be important to understand why the number of people involved in accidents has again increased over the last 5 years of this dataset in order to reduce this number in the future. To further this analysis, I would suggest examining the trends over time while normalizing the data for overall population trends to see if some of the findings could be influenced by the characteristics of the general population. I would also suggest that more studies consider alcohol involvement in vehicle accidents, since this analysis found that it may be connected to fatalities as well as accidents involving young people.

## References

NHTSA Data: <https://www.nhtsa.gov>

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

Gimond, M. (2020). Exploratory Data Analysis in R. <https://mgimond.github.io/ES218/index.html>