IISE DAIS Analytics Competition

Jack Francis[1] Christian Zamiela[1]

[1]Department of Industrial and Systems Engineering, Mississippi State University, Mississippi State, MS 39762

March 31 2020

Contents

0 Introduction

In 2018, 36,560 people died in vehicle crashes in the United States, which accounts for 1.2% of all deaths in the United States in 2018 [1]. Deaths due to vehicle crashes are grouped into the unintentional/accidental injury, because there is often not purposeful intent in these types of death. This is unique from many of the other top causes of death which have an unexplained combination of genetic and lifestyle factors, such as cancer. In total, unintentional accidents are the 3rd highest cause of death in the United States [1]. Unintentional accidents can generally be avoided through making different decisions that lead to the accident occurring. For example, unintentional deaths due to drunk driving can be completely mitigated by making it impossible to drive a vehicle while drunk. Draconian measures like this are often infeasible in the United States, due to political and social backlash, but policymakers often introduce large campaigns to "nudge" people to avoid negative behaviors [2]. For example, AT&T has promoted the "It can wait" campaign since 2010 to remind drivers to not text and drive. Mothers Against Drunk Driving (MADD) has campaigned since 1980 to help reduce drunk driving in the United States. While the immediate effectiveness of these campaigns is questioned [3, 4], both of these campaigns have led to new laws and policies, such as texting while driving bans in 48 of 50 states and over 1000 laws relating to drunk driving [5]. These laws have, over time, made Americans travelling on the road safer and led to a continued reduction in fatal crashes since 1996 [6]. However, there are still a large number of Americans dying each year in unintentional vehicle crashes, so it is vital to analyze fatal crashes to identify the most frequent causes so that new laws and guidelines can be introduced to further improve the safety of Americans on the roadways.

To identify which laws and guidelines would make roadways safer, a clear understanding of the causes of fatal crashes is necessary. Assuming that the causes are clear, policymakers can introduce campaigns and legislation to disincentivize these behaviors, thus making the roadways safer. Understanding all of the causes of fatal crashes is difficult for two reasons: 1) fatal crashes are exceedingly rare compared to the total number of vehicles and miles driven each year and 2) there are many causes that lead to fatal crashes. While both are important and should be considered, in this report we analyze the causes that lead to fatal crashes. There are a large number of causes, but we group these into four main types:

1. *Contributing Factors*: This group includes driver behaviors that lead to an increased risk of fatal crashes. Some examples include speeding, drunk driving, and drugged driving.

2. *Crash Patterns*: These are similarities among fatal crashes. Some examples include the time of the crash, type of road, and vehicle type

3. *Driver Profiling*: This group includes characteristics of drivers that may be correlated with fatal crashes. Some examples include driving history, age, and health condition.

4. *Vehicle Vulnerability*: Factors in this group directly relate to the reliability and quality of the vehicle involved in the fatal crash. An example is the type of damage for different vehicle types.

By analyzing these groups of causes of fatal accidents, we aim to provide recommendations to policymakers on which negative behaviors to disincentivize. By discouraging the most prevalent negative behaviors, American roadways will be made safer overall.

The remainder of the report is organized as follows. Each competition question is analyzed and results/recommendations are given in the corresponding section (i.e. Question 1 is analyzed in Section 1). Section 1 analyzes the relative safety of each state in the United States to find the safest and most dangerous states for vehicles. Sections 2-5 analyze and provide recommendations for each of the 4 groups of causes previously mentioned. The conclusion and major findings are given in Section 6. All source code is in the Appendix.

1 Question 1

1.1 Background

We begin our analysis by comparing the safety of each state in the United States to identify patterns within states that lead to more fatal crashes. For example, some contributing factors could be the number of miles of road, the population density, and the quality of the roads within each state. A higher number of accidents with fatalities could vary between states because of limited medical personnel and local law enforcement, hazardous speed limits, inconsistent road surfaces, or frequent dangerous weather [7]. In some studies, there is a correlation between lower gas prices and the number of people traveling [8]. Due to more people traveling, it is likely that more accidents will occur and lead to more deaths in these states. It is also common for people to get more frustrated and distracted with a higher traffic level [9]. So states with regions with higher traffic (e.g. New York City or Los Angeles) may have more fatal crashes due to driver frustration.

1.2 Analytical Approach and Assumptions

While comparing states by directly comparing the total number of deaths is straightforward,, this approach does not account for differences in states. For example this does not account for the varying population size, population densities, and miles of road in each state. We assume that states with larger populations will have more accidents with fatalities, because there are more opportunities for fatal crashes due to more people driving more miles over the course of a year. The number of large cities in each state could contribute to safety and danger factors. For this question, each state is evaluated by the number of fatalities in each accident. The number of fatal accidents is normalized in two ways 1) by the miles of road in the state and 2) the population density of the state. Locations within each state with a high number of fatalities are identified by examining the longitude and latitude of the accident. Further, we analyze the number of accidents that occur in rural and urban states, to examine whether rural or urban states are more likely to have fatal accidents.

1.2.1 Deaths

To investigate safety, we used the vehicle dataset to evaluate the state and the number of deaths that took place. Deaths is a numerical variable counting the number of deaths in each vehicle in the accident. Each accident case has a least 1 death, but there is not necessarily a death in each vehicle involved in the accident. The number of vehicles in an accident case ranges from 1 to 27. The number of deaths in a single vehicle accident ranges from 0 to 18. We note that there is no missing data for the number of deaths for each vehicle. We hypothesize that the most frequent number of deaths in each vehicle will be either 0 or 1 death, because fatal accidents are rare.

1.2.2 Longitude and Latitude

We use latitude and longitude of each accident to analyze the locations in the United States that have the most fatal accidents. The latitude and longitude are rounded the nearest whole number. We assume to see more deaths in locations with a higher number of people. The areas with the highest density of fatalities will likely surround large cities and highways with a large amount of traffic.

1.2.3 Rural Versus Urban

The classification for this variable is from the Federal Highway Administration. Out of all roads where accidents occurred, 1.77 percent do not have a rural or urban classification. The rest of the observations are classified as either rural or urban roads. As previously mentioned in latitude and longitude, we suspect that many of the deaths that occur in each state will be in more populated areas.

1.3 Tools and Data Sources Used

For identifying the factors contributing to accidents, we used the 2018 FARS dataset. Specifically, we used the variables state, deaths, longitude, latitude, and rural versus urban. Each table containing FARS data was downloaded as a CSV from the NHTSA website. From outside sources, we gathered tables that include total miles of public road in each state from the bureau of transportation [10]. Additionally, we collected the state population and square miles of land tables from the US Census Bureau to calculate population density [11]. All analysis was conducted using R in RMarkdown being used for generating the source code shown in Appendix 1.

1.4 Results

The number of deaths is used to identify which states are the most dangerous. Of all the vehicles involved in fatal accidents, 53 percent of the vehicles had a reported death. Showing that many fatal accidents are multi-car accidents. In terms of state safety, Rhode Island had the fewest deadly accidents, 56, and Texas had the most deadly accidents, 3305. Normalizing the number of fatal accidents by miles of road in a state and by the population of each state helps to show a clearer image of a state's relative safety. Table 1 highlights the States that are the most dangerous in each of the normalized categories.

We evaluate the effect of urban versus rural locations by using the geographic coordinates where the fatal accident occurred. We found that 56 percent of the incidents happen in urban

Table 1: Summary of State Fatal Accidents

Ranking	Number of Accidents	Accidents per 100 miles road	Accidents per 100,000 people
Top 1	Texas(3305)	Hawaii(2.48)	Mississippi(19.99)
Top 2	California(3259)	Florida(2.38)	South Carolina(19.07)
Top 3	Florida(2915)	California(1.86)	Alabama(17.92)
Top 4	Georgia(1407)	Delaware(1.62)	Wyoming(17.31)
Top 5	North Carolina(1321)	Maryland(1.46)	New Mexico(16.70)
Bottom 5	Wyoming(100)	Minnesota(0.25)	Minnesota(6.22)
Bottom 4	North Dakota(95)	Nebraska(0.22)	New Jersey(5.89)
Bottom 3	Alaska (69)	Montana(0.21)	Rhode Island(5.30)
Bottom 2	Vermont(60)	South Dakota(0.13)	Massachusetts(4.97)
Bottom 1	Rhode Island(56)	North Dakota(0.11)	New York(4.55)

locations. Urban areas such as large cities and busy highways between these large cities had a high number of deaths. In Figure 1, the DC, Virginia, Maryland (DMV), and nearby northern states have a high number of fatal accidents. The LA and Miami areas have a high density of fatal accidents as well. In general, areas with higher populations tend to have larger numbers of accidents. We identified that these areas are dangerous to drive, however only 56 percent of these fatal accidents occur in urban locations. Showing that more multi-death accidents occur in urban locations compared to rural locations.

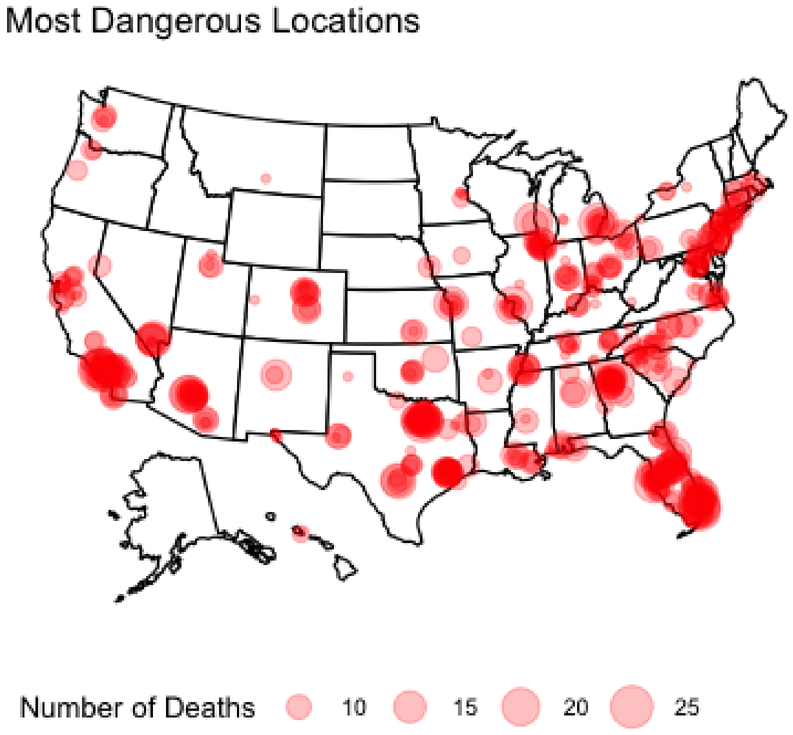


Figure 1: Most Dangerous Locations

Rural locations such as North Dakota and South Dakota have very few fatal accidents in one specific area (like New York) but have fatal accidents spread across the state. These states have some of the highest percentage of vehicle occupants that die. Thus, if a fatal accident occurs in a rural area, it is more likely that vehicle occupants will die. For example, 73 % of vehicle occupants involved in fatal crashes in Wyoming died. Similarly, South Dakota and North Dakota have vehicle occupant death percentage of 70 % and 62 %, respectively. More urban states, such as Hawaii and California have a lower percentage of vehicle occupant death at 42 percent and 45 percent, respectively. This finding again shows that urban areas are more likely to have multi-death accidents, while single-death accidents are more common in rural locations.

Also, comparing the number of miles of road to the population density per square mile of land can show why some states are more dangerous than others. We find that having a higher population density generally leads to a higher death rate per 100 miles when the population density is below 250. As you can see in Figure 2, there is a linear trend between population density and deaths per 100 miles. For states with a population density greater than 250, there is very little correlation between population density and deaths per 100 miles.

To identify the safest and most dangerous states, we use both of our normalized factors (i.e. accidents per 100 miles and accidents per 100k people). To do this, we first standardize both factors to have mean 0 and standard deviation 1. This is done because the range for accidents per 100k people is much larger than the accidents per 100 miles of road. Next, we sum the standard deviations for each variable to find the states that are the largest outliers. A positive standard deviation means that this state is more dangerous compared to average, while a negative standard deviation means that the given state is safer compared to the average state. For example, in Table 1, we show that Hawaii and Florida are the states with the highest rate of accidents per 100 miles of road. This is also shown when analyzing the standard deviation, as these are the only two states to have a standard deviation above 3 in this category. In this analysis, we are considering that both normalized factors are equally important, but this could be modified by weighting each factor. Our results for the top 5 most dangerous and safest states are shown in Table 2. We find that Florida, South Carolina, Hawaii, Mississippi, and Alabama are the most dangerous states. Whereas Minnesota, New York, Washington, Iowa, and Utah are the safest states.

Table 2: Summary of Most Dangerous and Safest States

Ranking	State	Std. Dev. Accidents per 100 miles	Std. Dev. Accidents per 100,000 people	Sum of Std. Dev.
Top 1	Florida	0.66	3.01	3.67
Top 2	South Carolina	2.10	1.22	3.32
Top 3	Hawaii	-0.92	3.20	2.28
Top 4	Mississippi	2.35	-0.08	2.27
Top 5	Alabama	1.79	0.04	1.83
Bottom 5	Utah	-0.99	-0.63	-1.62
Bottom 4	Iowa	-0.53	-1.13	-1.66
Bottom 3	Washington	-1.23	-0.45	-1.68
Bottom 2	New York	-1.77	-0.12	-1.89
Bottom 1	Minnesota	-1.33	-1.13	-2.46

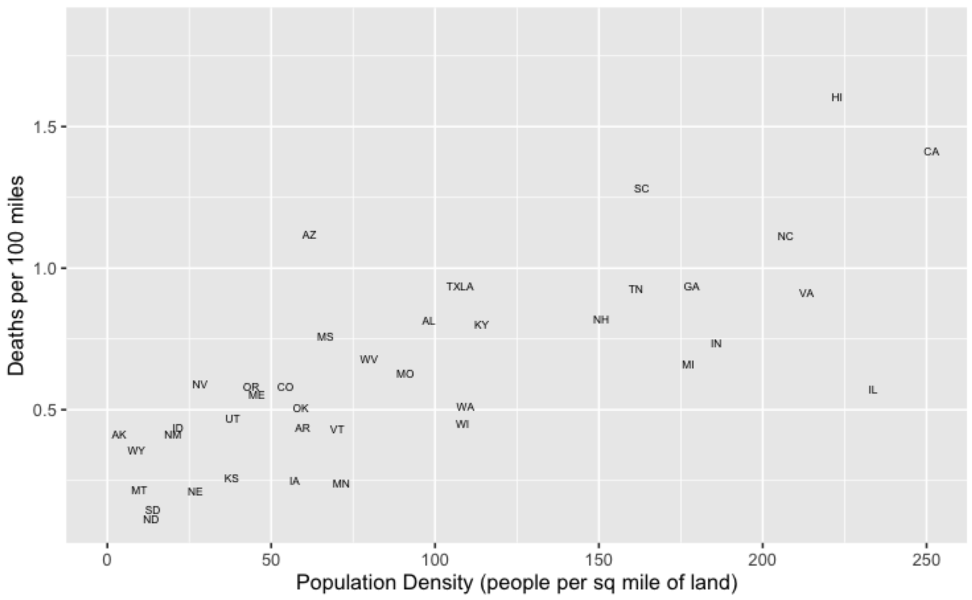


Figure 2: Relationship between population density and deaths per 100 miles for states with a population density below 250.

1.5 Discussion and Recommendations

We identified the states that are most dangerous by using 1) the number of accidents per 100 miles and 2) the number of accidents per 100k people. Locations with many miles of road and a high population density have an increased number of fatal accidents. Dense locations have higher numbers of total fatal accidents. Rural locations tend to have a high percentage of vehicle occupants in fatal crashes who died, but this may be due to more single car accidents on rural roads in these states. In general, we find that there is a balance between both of the factors investigated. In some states, such as Florida and Hawaii there are a large number of accidents relative to the amount of roadway, whereas in South Carolina, Alabama, and Mississippi there are a large number of accidents relative to the population.

2 Question 2

2.1 Background

For this question, the objective is to analyze various factors that contribute to fatal accidents recorded in the FARS dataset. We denote this group as "Contributing Factors" and are directly related to the driver's ability and include speeding, alcohol, drug use, distractions, and vision obstructions. Each of these factors negatively impact the driver's ability to drive at a high level. For example, drunk drivers believe they are capable of performing at similar levels to sober drivers and are often more willing to drive compared to drugged drivers [12]. There have also been many recent laws and guidelines to reduce distracted driving, including many bans on texting and driving. In fact, recent research has shown that infotainment systems, such as CarPlay and Android Auto lead to significantly reduced reaction times for drivers [13]. The reduction in reaction time is a common theme among each of these contributing factors. When speeding, there is less time to react to any obstruction in the road. Similarly for the other factors, each reduces the reaction time of the driver, which leads to accidents that may have been avoidable. By identifying the frequency of each factor present in fatal crashes, recommendations on public policy initiatives can be developed to encourage drivers to avoid these negative behaviors.

2.2 Analytical Approach and Assumptions

Within contributing factors, we analyze speeding, drunk driving, drugged driving, distracted driving, and driver's vision obstruction. For each of the investigated factors, we describe the analytical approach taken to process the data in the following subsections.

2.2.1 Speeding

To investigate speeding, we examined all vehicles involved in accidents where the police reported speeding was related. There are multiple types of speeding, which are shown in Table 3. For vehicle observations where it was unknown if speeding was related, we used other columns to fill in this data when possible. Specifically, we compared the vehicles current travel speed to the speed limit of the road on which the accident occurred. This method reduced the number of unknown cases from 2077 to 481, out of the total 51872 vehicles involved in fatal crashes. For the analysis, we examined how often each of the types of speeding occured among all vehicles in fatal accidents in 2018. In addition, we hypothesized that fatal speeding crashes are related to the travel speed of the car, the speed limit of the road, and the difference between these two. In general, we expected to see more fatal crashes when there were high travel speeds, due to the accident generating more force on the vehicle occupants. We expected to see most fatal crashes occur on roads with speed limits of 55, 65, and 70 mph, because these are the most common speed limits on highways and interstates. The higher speed limit also implies that on average, crashes occurring on these roads were at higher speeds compared to neighborhood and rural roads. Finally, we expected that the difference between travel speed and speed limit would show a strong relationship in the fatal crash data. If a vehicle is traveling significantly over or significantly under the speed limit, then a crash with another car would likely lead to a more severe outcome.

Table 3: Groups of speeding related accidents

Groups	Definition
1	No
2	Yes, Racing
3	Yes, Exceeded Speed Limit
4	Yes, Too Fast for Conditions
5	Yes, Specifics Unknown
6	No Driver Present
7	Unknown

2.2.2 Alcohol Use

To investigate alcohol use we examined all drivers that were either 1) identified as drunk or 2) above the federal limit of 0.8 Blood Alcohol Content (BAC). We used the federal limit, which is applicable in 49 out of 50 states. The exception to this rule is Utah, which has a limit of 0.5 BAC. Only 59 out of the 8644 drunk drivers had accidents in Utah, so we believe this assumption will not significantly affect the quality of the results. One of the most tragic results of drunk driving fatalities is sober people being negatively affected. So to investigate this, we examined how often sober people are involved in drunk driving accidents. In addition, we analyzed how drunk (i.e. BAC level) these drivers were. Finally, we look at how many accidents police thought alcohol was involved, but the driver was not legally drunk.

2.2.3 Drug Use

To examine drugged driving, we analyzed how many drivers had drugs in their system at the time of a fatal crash. The FARS dictionary has 8 groups of drugs shown in Table 4, which we use to find the most common type of drug used by drivers in fatal crashes. Drivers can have multiple drugs in their system at the time of a fatal crash. For analyzing how often a given drug group is used in fatal accidents, we count each occurrence of the drug present.

Table 4: Drug Groups in the FARS Dataset

Group Number	Drug Type
1	Narcotic
2	Depressant
3	Stimulant
4	Hallucinogen
5	Cannabinoid
6	Phencyclidine
7	Anabolic Steroid
8	Inhalant
9	Other Drugs

2.2.4 Distracted Driving and Vision Obstruction

The FARS dataset has over 20 different categories on distracted driving, but we group these based on common distraction types. Specifically, we examine cell phone, food, and in-car distractions to see the frequency of each type. Unlike the previous categories, distracted driving is often not reported (61 % of drivers), so analysis for this group is less conclusive compared to speeding, alcohol, and drug use. Similar to distracted driving, the FARS dataset has almost 20 different categories for vision obstruction. We group these based on common obstruction types present in multiple of the FARS categories. Our groups examine the frequency of weather-related, vehicle-related, and other external obstructions.

2.3 Tools and Data Sources Used

For identifying the factors contributing to accidents, we used the 2018 FARS dataset. Specifically, we used the vehicle, person, accident, drugs, distract, and vision tables. Similar to the previous question, each table was downloaded as a csv from the NHTSA website and was manipulated within R for analysis. All analysis was completed using R, with RMarkdown being used for generating the source code shown in Appendix (fill this in later).

2.4 Results

We find that speeding, alcohol, and drugs are frequently present in fatal crashes. Figure 3 summarizes our main findings of how often each of the 5 types of causes were present in fatal crashes. For speeding, we find that the most common crash scenario is a vehicle traveling 70 mph on a road with a 55 mph speed limit. Surprisingly, we find that most speeding related crashes occur on roads with speed limits of 35 mph, 45 mph, and 55 mph. These roads are often local roads or state highways with low visibility. Each of these types of roads had over twice as many fatal accidents as roadways with a speed limit of 70 mph. Our initial hypothesis was that high speeds would more frequently lead to fatal crashes, but in fact it appears that higher speeds on roads with low speed limits is the most likely scenario for a speeding related fatality. Officers typically rounded the speed of the vehicles to the nearest 5 mph, and we found a close to normal distribution of speeds among vehicles in accidents, with a mean of 70 mph. By examining the relationship between travel speed and speed limit for each speeding related accident, we find that 92% of vehicles involved in speeding related accidents were speeding. This shows that in speeding related accidents, it is much more frequent for two (or more) cars that are speeding to be in an accident than a speeding car and a non-speeding car.

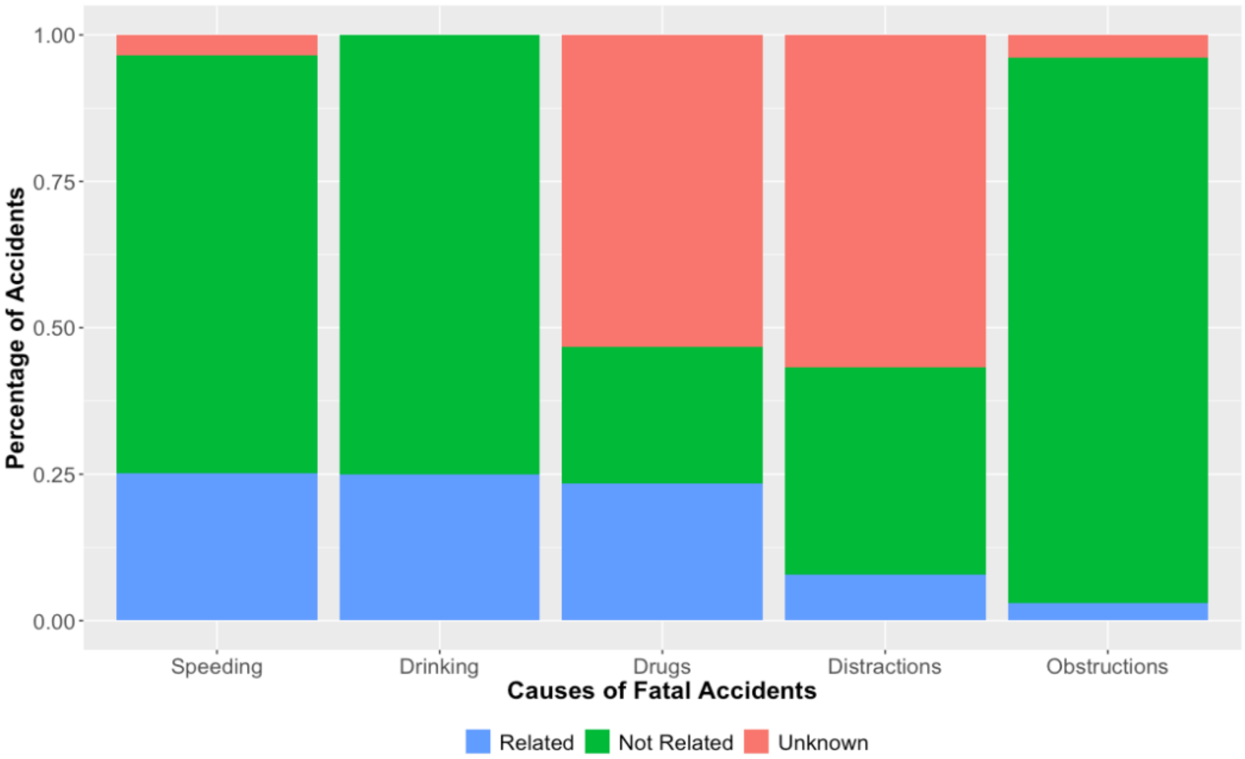


Figure 3: Percentage of crashes where each cause was involved. We aggregate all drivers for each accident, so if one driver out of multiple drivers involved in an accident was drinking, then the entire accident is coded as a drinking related accident.

As shown in Figure 3, alcohol is a significant factor in fatal accidents. Using the person table and filtering by people in the driver seat during accidents, we find that 6112 out of the 51872 drivers were suspected to be drunk by the police reporting the accident. However, only 5752 drivers were confirmed to be drunk by a BAC test. These 5752 drivers contributed to 5666 accidents, which shows that most accidents involving a drunk driver did not involve multiple drunk drivers. These accidents were either single car accidents, or involved other non-drunk drivers. We further analyze this group by examining the BAC of all drivers in drinking related crashes and find that for drunk drivers in fatal accidents the most likely BAC level is around 0.16-0.19. Out of all accidents involving drunk drivers, we find that 26.96% of drivers were above the legal limit, again showing that drunk drivers most often have accidents with sober drivers.

Drugged driving is much more difficult to analyze compared to speeding and alcohol because most drivers are not tested for drugs. Out of the 51872 drivers, over 65% were reported as unknown. Out of the drug groups which were present, we find cannabinoids, stimulants, and other drugs (i.e. not in one of the 8 mentioned groups) to be the most common drugs in a driver's system involved in a fatal crash. While cannabinoids are the most prevalent drug group, it is difficult to conclude that this drug is significantly related to fatal crashes. In many states cannabinoids are legal, and thus more drivers will have these in their system compared to many of the other drugs, which may be illegal. Further, only 15% of all drivers involved in fatal accidents were drugged.

Similar to drugged driving, distracted driving is also not reported very often. For 61% of all drivers in fatal crashes it was unknown if the driver was distracted. In total, only 5% of all drivers in fatal crashes had a distraction noted by the officer. Of these, 3.8% were reported as an unidentified distraction. Of the distraction groups we examined, cell-phone distractions were the most common, but only comprise 0.68% of all fatal crashes. Vision obstructed driving is quite rare in fatal crashes, over 93% of all fatal crashes had no visual obstruction. Of the groups we investigated, weather related obstructions were the most common, which comprised 1.37% of all fatal crashes.

2.5 Discussion and Recommendations

We find that alcohol use and speeding are the factors most prevalent in fatal accidents. Currently policy makers effectively portray the negative impact of drinking and driving, but we recommend that a similar campaign be used to combat speeding. Speeding on local roads and state highways is most likely to lead to fatal crashes, so one possible policy change is to more strictly enforce speed limits on these roads. From personal experience, we know that many of these local roads allow for up to 5 mph over the speed limit before an officer will stop a driver for speeding. By more strictly enforcing the speed limit on these roads, fatal crashes may be able to be reduced.

3 Question 3

3.1 Background

The objective of this question is to analyze fatal vehicle accidents to identify crash patterns. Vehicle crash patterns can be analyzed using a variety of related causes. For example, crash patterns can be found with regard to equipment safety, time, road condition, and type of vehicle to name a few. By analyzing crash patterns, the most common causes of crashes can be made public, so that drivers can avoid behaviors that commonly lead to crashes. Further, if a driver has to be on the road, knowing common crash patterns can help them drive safer and more attentively in scenarios where crashes are more likely. Similarly, understanding where and when fatal crashes are likely to occur is essential to many government entities. Government entities can introduce new laws to reduce the frequency of crash scenarios, leading to fewer overall crashes. Manufacturers would also benefit from understanding crash patterns, so that they now how to improve their vehicles and understand where they are more likely to see damage in fatal accidents. Safety is important to consumers purchasing a vehicle, so this provides these manufacturers with a competitive advantage [14]. A better understanding of the relationship between different vehicle models, crash types, and road conditions can help to reduce the number of fatal vehicle accidents.

3.2 Analytical Approach and Assumptions

Several variables could relate to crash patterns, and some are more significant than others. There are several factors in a crash that may show a pattern when a large number of crashes are aggregated. The first crash pattern factor we analyze is the time of the crash (i.e. month, day, hour). Day, time, and month may identify crash patterns because a large number of people have a similar regular day routine. Specifically, many people travel to work between 6-8 AM and leave work between 4-7 PM on the weekdays. Additionally, Next, road conditions are evaluated to show locations that a driver should avoid. There may be certain types of roads that are more dangerous than others, such as 2 or 4 lane roads compared to 1 lane roads. Finally, we evaluate crash type and vehicle model (e.g. automobile, truck, motorcycle) to see which types of accident a person is more likely to be involved in for varying car models.

3.2.1 Time, Day, and Month

The objective of analyzing this variable is to identify when fatal accidents occur and identify the underlying behavior that leads to these accidents. We hypothesize that more accidents will occur during morning rush hour from 6 am to 8 am, and evening rush hour from 5 pm to 7 pm. Also, we expect to see more accidents in the evening because people tend to drive to various hobbies, activities, and restaurants. We also look at how drunk driving is related to time of day. We hypothesize that drunk driving will occur most often at night.

3.2.2 Road Condition

Road conditions are evaluated using the number of lanes, weather, and road lighting. We expect raining to be significantly correlated with fatal accidents, because the roads will be slick. This increases the chance of hydroplaning and reduces the driver's vision. The number of lanes could provide insight into if more space on the road leads to less accidents. Analyzing whether the road was properly lit could emphasize the need for more lights along roads to help people with their vision, especially when the weather is not favorable.

3.2.3 Crash Type and Vehicle Model

We analyzed four of the most frequent models of vehicles: 1) automobiles, 2) trucks, 3) heavy trucks, and 4) motorcycles. We identify which vehicle models get in the highest number of fatal accidents. We also compare vehicle models to understand what types of accidents each model is more likely to be involved in. For example, we expect that motorcycles are more likely to be in an accident when they are not seen by the other driver. However, heavy trucks (i.e. 18-wheelers) are not expected to be involved in these types of accidents frequentlly. We hypothesize more automobiles will be in fatal accidents because they are more frequently driven and smaller than trucks.

3.3 Tools and Data Sources Used

For identifying the factors contributing to crash patterns, we used the 2018 FARS dataset. Specifically, we used the variables hour, day, month, number of lanes, weather, road lighting, accident type, and model. Each table containing FARS data was downloaded as a CSV from the NHTSA website. All analysis was conducted using R in RMarkdown being used for generating the source code shown in Appendix (fill this in later).

3.4 Results

Analyzing time, day, and month provide insight into when the most fatal accidents occur. Figure 4 shows the number of accidents each day and hour. For weekdays, there is a spike at 6 am, followed by a decrease until about 10 am. After 10 am, there is a slow increase until rush hour (5 - 7 pm), then a reduction in crashes until the next day at 6 am. Interestingly, evening rush hour (i.e 5-7 PM) accidents comprise 15 percent of accidents, while only being 12 percent of the total time during the week.

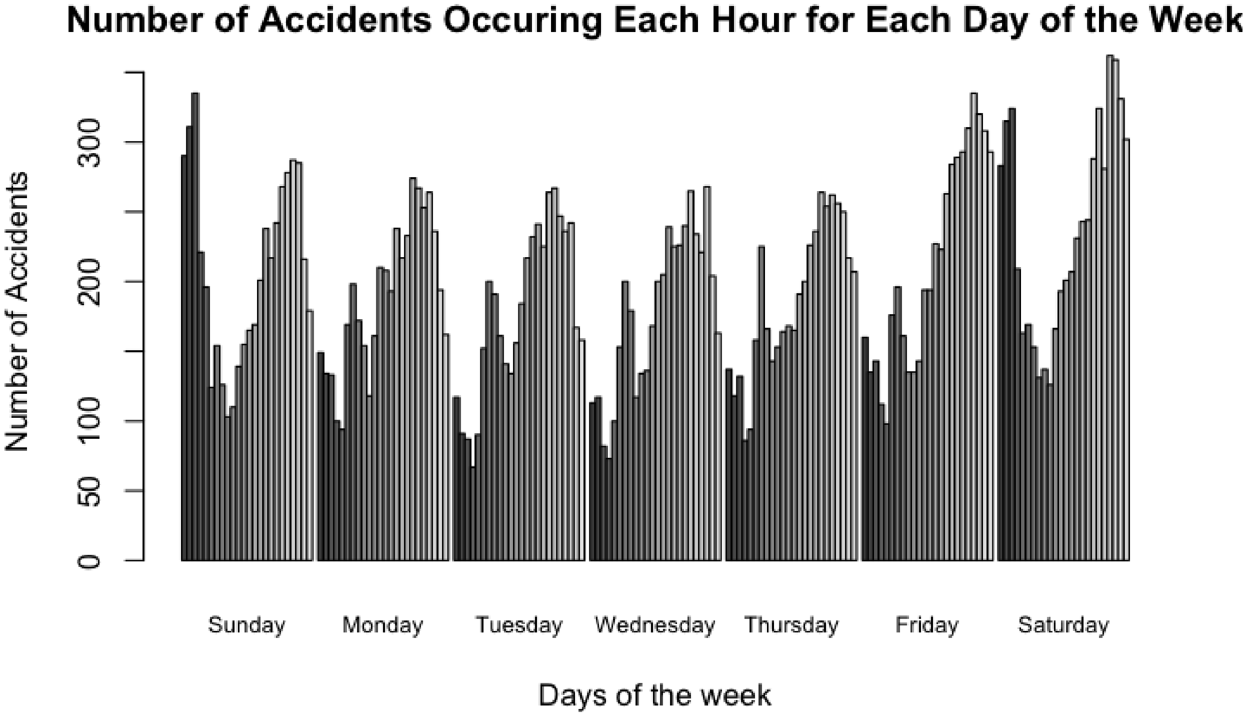


Figure 4: The frequency of fatal accidents on each day of the week at every hour of the day

There are significant peaks on weekend nights (Friday and Saturday night) from roughly 8 pm to 2 am. We hypothesized that number of drunk driving accident will increase at this time and find that this is a significant factor for this time period. Of all the drunk driving accident that happen between the hours of 12am and 3 am, 57 percent are on Saturdays and Sundays. Also, all the the drunk driving accidents that happen from 8pm to 12 pm, 41 percent are on Fridays and Saturdays. We do find that drunk driving accidents on Monday through Wednesday is significantly lower, which matches our intuition because more people tend to drink on the weekends. We analyzed when accidents occur during the year and find a higher number of fatal accidents during higher travel months in summer and fall. These seasons have considerably more fatal accidents than in spring and winter. We hypothesize that this is due to more travel in summer and fall due to summer vacation for schools and many sporting events in the summer and fall.

We analyzed the type of road fatal crashes most often occur and find that a majority of the accident happened on 2-lane roads. This could be due to their being no separation

Table 5: Crash Type Configuration Accident Attributes

Accident Configuration	Attributes
Road Departure	Intentionally driving off road, control loss, and avoid collusion
Single Driver	Collision with parked vehicle, object, pedestrian, and end departure
Forward Impact	Same direction, opposite direction, and rear end
Head On Collision	Direct impact on the front of vehicle
Angle Sideswipe	Lateral moves and same direction lane merges
Turn Across Path	Crossing the path going opposite direction and same direction
Turn Into Path	Coming into path going opposite direction and same direction
Straight Path	Directly striking left or right side of a vehicle
Other	Backing into, no impact, and Unknown crash type

between cars going in opposite directions. There are significantly fewer accidents as more lanes are added. The weather and lighting of the road also contribute to a large number of fatal accidents. Out of the weather conditions that generally are thought to negatively affect driver performance, rain was the most common, followed by snow, fog, smog, and smoke. We find that inclement weather was involved in 11 percent of fatal accidents.

Finally, we analyze the vehicle model and crash type configuration to identify what type of crash in each model of vehicle is likely to occur. The crash type configurations given by the FARS dataset are shown in Table 5. We find that trucks and automobiles have the highest number of accidents out of the vehicle models investigated. Figure 5 shows trucks and automobiles also have similar values for each crash type with little variation. Motorcycles have a significantly larger number of accidents that happen when turning across path likely due to other vehicles not looking for motorcycles. Motorcycles also have a higher number of the crash type turn into path, meaning people are less aware of Motorcycles when merging. Large trucks have a higher number of rear ends likely due to more blind spots and not being able to see every obstacle. Interestingly, we find that the most common accident crash type is road departures for all models except heavy trucks. Road departures, single drivers, and other configurations contribute to 50 percent of all accidents in automobiles, trucks, and motorcycles. This indicates that car mechanical performance is less likely to be the cause of many fatal accidents.

3.5 Discussion and Recommendations

As many previous studies suggest, rush hour and evenings when people are more likely to be traveling [15]. This likely correlates with the highest number of fatal accidents being during rush hour found in our study. We found a spike in drunk driving accidents at night time, especially on the weekends. We recommend that policy makers work with Uber and Lyft to reduce the price of late night rides, so that more drunk people will be driven home instead of driving themselves. It is important to note to policy makers that accidents with road departures, single drivers, and other are the most common accident type. More attention needs to go to staying on the road and other surrounding objects other than cars. It seems that people are less aware when there are fewer cars involved, and when pedestrians or

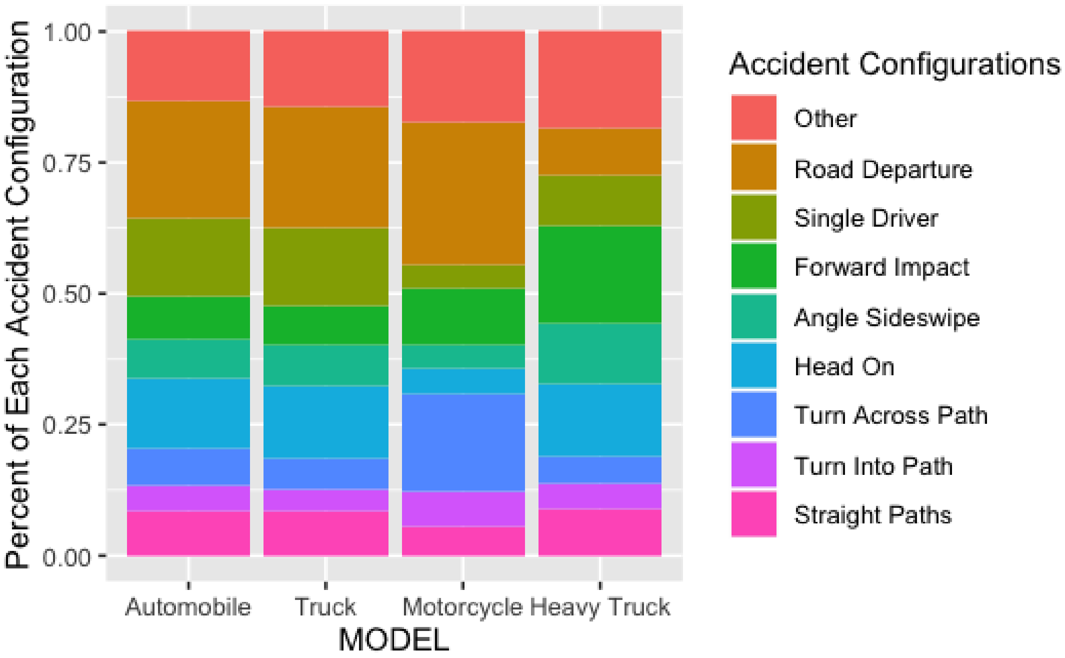


Figure 5: Percent of crash type for each model of vehicle. Atv, large vans, motor homes, and other vehicle are removed because they are in less than 1 percent of fatal accidents

animals are on the road.

4 Question 4

4.1 Background

The objective of this question is to profile drivers and identify the preexisting characteristics that make drivers more or less likely to be in fatal accidents. Some accidents occur as a result poor driving behavior and driver characteristics. Driver characteristics such as skill level, inexperience, and risk taking behavior are indicated as the key differences between younger drivers and older drivers. We hypothesize that the lack of experience and risk taking leads to more fatal crashes for younger people. Additionally, the number of violations tend be higher for younger drivers [16]. Studies also report that men are considered worse drivers due to behavior differences and as a result, men's insurance rate is often higher [17]. Health can also be a cause of fatalities in car accidents because it can affect a driver ability to perform. A study using the FARS dataset from 2010 [18] showed that BMI has a strong correlation with fatal car accidents. Compared to a healthy BMI of 18.5-25, drivers with a BMI between 35 and 40 have a 21 percent higher chance of death. Additionally, for a BMI over 40, there is up to an 80% increased risk of death.

4.2 Analytical Approach and Assumptions

The objective of this question is to identify high risk drivers using a combination of personal factors and driving behavior. Gender, age, and health conditions are the demographic variables investigated. While, we use driver's previous record and violations, such as speeding, DUI, and accident frequency, to identify the best and worst drivers. By identifying risky behavior, recommendations can be made on which drivers should have increased insurance rates and can also encourage behavior that exercises more precaution when driving. Each factor's analytical approach described in the following subsection.

4.2.1 Driver Health

Investigating a driver's health is a combination of a few different factors. Previous health conditions are not given in the dataset, but age and body mass index (BMI) serve as a method for identifying how a driver's health contributes to fatal accident. As age increases a driver could become worse at driving safely. This may be due to the general decrease in sight and hearing as people age. We hypothesis as age increase over 65 a driver ability to react and notice all danger signs decrease. Also, obesity is a common health disease among Americans. We hypothesize the number of deaths in accidents will increase as the body mass index increases, due to ill-fitting seatbelts and cars being tested with average size crash dummies.

4.2.2 Inexperience Errors

Inexperience is commonly associated with age. We hypothesize that new drivers and drivers under the age of 25 are at higher risk of being involved in fatal accidents because they have less experience. Inexperience can also relate to the type of crash. We also hypothesize that young drivers are likely to make similar mistakes. Mistakes of these drivers could be related to lack of awareness of their surroundings and the rules of the road. To analyze if young drivers had similar accident types, we use the accident configurations in the FARS dataset (Table 5 to see how the accident took place. For example, young drivers may be more prone to single car accidents, due to lack of following road signage.

4.2.3 Risk Taking

To analyze risk taking, we used the vehicle and person table to examine all previous driving violations, speeding, driver drinking, time since first altercation, time since last altercation, age, and sex. Sex and age along with the number of previous altercations and average time of altercations could show who is more likely to take risks. We assume younger drivers are more likely to take risks and previous studies show men are more likely to take risks due to their behavior [8].

4.3 Tools and Data Sources Used

For identifying the factors contributing to risky drivers, we used the 2018 FARS dataset. Previous drinking, accidents, suspensions, convictions, date of first and last, age, height, and weight, race, and accident type are the variables used. Each table was downloaded as a CSV from the NHTSA website and was manipulated within R for analysis. From outside sources, the percent of people in each BMI group was found in a study from the agency for healthcare and research [19]. All analysis was completed using R, with RMarkdown being used for generating the source code shown in Appendix 4.

4.4 Results

For driver health, BMI is evaluated to profile drivers. A BMI greater than 30 is considered obese. Figure 6 shows that from a BMI of 18 to 38, there is little change in the death percentage. There is a little curve downward for people at BMI from 20 to 32. A BMI of 14 to 17 shows a significant amount of variability because only 1.4 percent of Americans are underweight [19]. BMI of 40 and higher are grouped together because there are very few BMIs above 40.

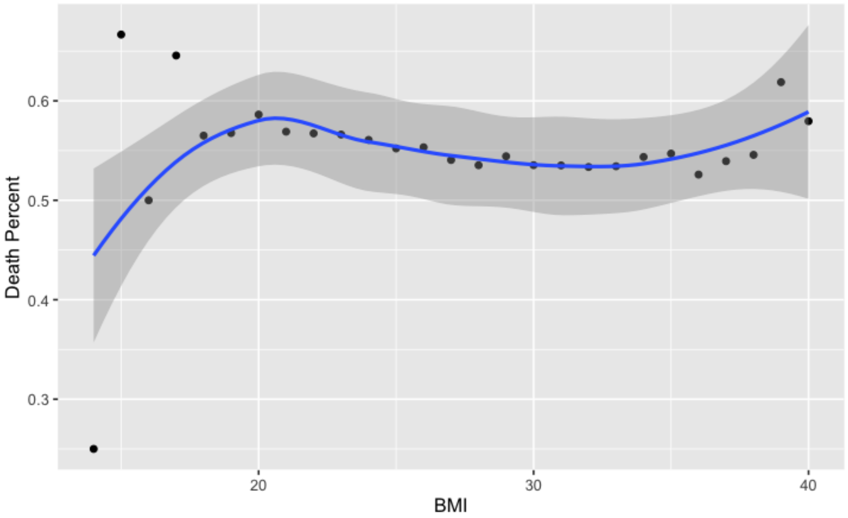


Figure 6: BMI is rounded to the nearest whole number and the average is plotted. BMI 40 and over is averaged into 1 point.

In terms of driver age, which we use as a proxy for driver experience, we find a significant correlation between age and the number of fatal accidents. The number of drivers in fatal accidents is higher at ages 15-25 and 26-35, accounting for over 40 percent of the observations. We see a decline in fatal accidents for the age groups 46-55 and 56-65, but see an increase for people over the age of 65. The crash type is also taken into consideration to analyze if an age group of drivers is more likely to get in preventable accidents. Table 5 from the FARS data dictionary shows the different crash type configurations.

In Figure 7, each crash type configuration percentage is given for each age group. Straight path, turn into path, and turn across path increases significantly for drivers over the age of 66. While single driver accidents drop significantly for drivers over the age of 66. For younger drivers from the age of 15-25, angle sideswipe, straight paths, and turn across frequency is very similar. Many of the crash type configurations are consistent across all the age groups.

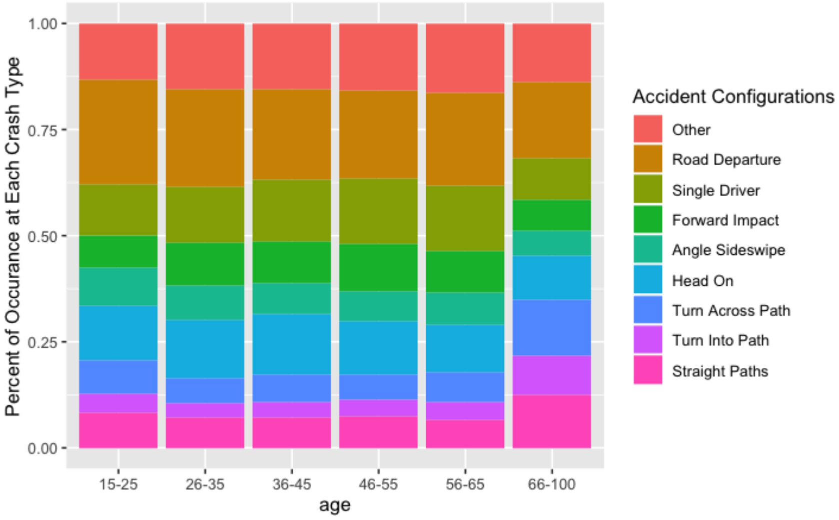


Figure 7: The percentage of crash type for each age group

We find an interesting fact when analyzing the frequency of violations for each driver. On average, the time between a driver's first violation and the driver's last violation (i.e. not including the fatal crash) is 195 days. However, the average time between a driver's last violation and the fatal crash is 465 days. This seems to indicate that as drivers commit violations, they are less likely to commit violations in the future. This could also be identifying the increase in driving skill of younger drivers, who make up a large proportion of total fatal accidents. In both cases, the time between first and last violation and the time between last violation and fatal crash are exponentially distributed, as expected.

We analyze the sex of the driver to identify which crash type configurations each sex is more likely to be involved in. Figure 8 shows that men and women generally have very similar results for each crash type configuration. The two categories that are significantly different are road departures for men and straight paths for women.

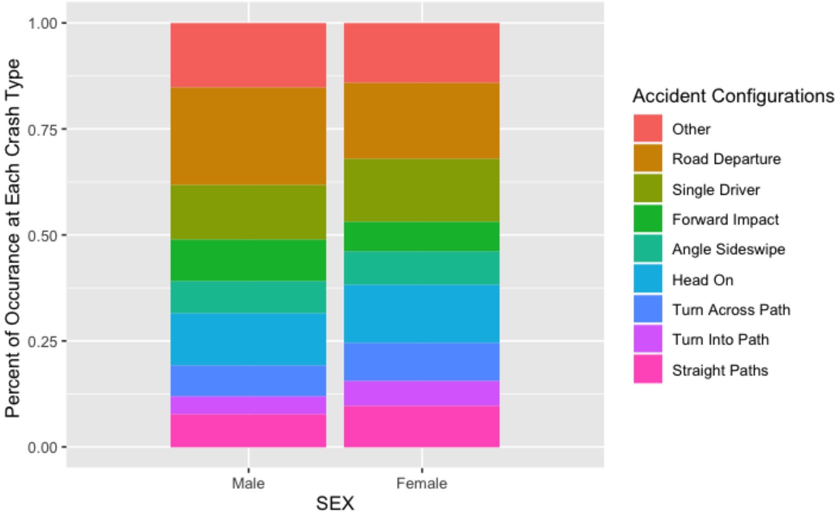


Figure 8: The percent of each crash type for different genders.

4.5 Discussion and Recommendations

Dissimilar to previous studies [18], it is interesting that we did not find much correlation between increased fatal accidents and obesity. One possible reason for this is that cars have become better manufactured to protect obese people since 2010. Further research and additional variables looking to driver's previous health conditions could provide more insight into high-risk drivers. We recommend that drivers be more careful after a previous violation or accident. Due to the exponential distribution of time between violations, drivers are more likely to be involved in another violation soon after the violation. By giving people driving lessons or teaching on safe driving practices, some of these accident may be avoidable. We would also advise people to be more careful around younger drivers from the age of 15 to 25 because they are in more accidents than any other age group studied. Men being in a significantly higher number of fatal accidents is likely due to different behaviors and more risk taking when driving. While men and women proportionally get in very similar crash type accidents, men get in more severe crashes, on average, than women. We recommend that policymakers provide additional teaching to young male drivers to further instill the necessity of safe driving.

5 Question 5

5.1 Background

Understanding how the quality and type of vehicles is related to fatal crashes is important for consumers and government agencies. For consumers, safety and reliability is more important than fuel economy and maintenance costs [14]. Government agencies can incentivize vehicle manufacturers to create vehicles that are safer by reducing the tax manufacturers would have to pay. Government agencies could also incentivize consumers to buy certain types of used vehicles, which are known to be safer through tax breaks as well. Thus, understanding the relationship between vehicle reliability and safety is critical in developing long term policy action for reducing fatal crashes.

5.2 Analytical Approach and Assumptions

For this question, we examined how three key vehicle attributes relate to rollover, fire/explosion, speeding related accidents, and the amount of damage. Specifically, we analyzed model year, body type, and vehicle make. We hypothesize that cars have gotten safer over time, so cars with a more recent model year should be safer compared to older cars in general. We also believe that the body type of the vehicle will be very important in the amount of damage and type of accident the vehicle is involved in (i.e. rollover, fire/explosion, speeding). The body types defined in the FARS dataset are shown in Table 6. Our initial hypothesis is that utility vehicles are more likely to rollover, due to the vehicle typically being top heavy. Similarly, we expect automobiles to be the most likely to be in speeding related accidents. We expect trucks and utility vehicles to be the safest overall, while motorcycles and buses are expected to be the most dangerous. Finally, we plan to explore a preliminary analysis into the make of the vehicle. Fatal crashes are fairly rare and with such a large number of vehicle makes, it may be difficult to make conclusive statements about vehicle makes that are not popular in the United States.

Table 6: Body Types in the FARS Dataset

Group Number	Body Type
1	Automobile
2	Automobile Derivative
3	Utility Vehicle
4	Van
5	Light Truck
6	Bus
7	Heavy Truck
8	Motor Home
9	Motorcycle/Moped
10	Other

5.3 Tools and Data Sources Used

For identifying the factors contributing to accidents, we used the 2018 FARS dataset. Specifically, we used the vehicle, damage, and accident tables. Similar to the previous question, each table was downloaded as a csv from the NHTSA website and was manipulated within R for analysis. All analysis was completed using R, with RMarkdown being used for generating the source code shown in Appendix (fill this in later).

5.4 Results

We first analyzed the body type of vehicles involved in all fatal crashes and the results are summarized in Table 7 and Figure 9. We find that over twice as many automobiles were involved in fatal accidents than any other body group. Automobile derivatives and motor homes are very rarely in fatal accidents, so for the remainder of the analysis we focus on the other 8 groups. As we initially hypothesized, utility vehicles and trucks were the most likely to rollover, due to the top-heaviness of these body types. Surprisingly, heavy trucks were the most likely to be involved in an accident with a fire/explosion. Finally, as expected motorcycles were the most likely to be involved in speeding related accidents. In terms of damage, buses were most likely to have a fatality while sustaining minimal damage. This is likely to be due to two factors 1) buses are quite large and so many other vehicles would not damage them in a crash and 2) buses typically transport large amounts of people/children without seatbelts so even small crashes could lead to a fatality. Motorcycles and automobiles were the most likely to sustain disabling damage, which makes sense because these are the two smallest vehicle groups by size.

For the vehicle make, we report the top 3 and bottom 3 of each make for the categories given in Table 8. There are a large number of makes, each with varying popularity in the United States, so we only consider makes which were involved in 500 accidents or more.

Table 7: Summary of Accidents by Body Type. The top 3 for each category are shown in bold

Body Type	Accidents	Rollover	Fire/Explosion	Speeding Related
Automobile	**20319**	12.1%	3.4%	**18.6%**
Automobile Derivative	17	11.8%	**11.8%**	**17.6%**
Utility Vehicle	**8827**	**22.3%**	3.2%	14.8%
Van	2081	12.1%	2.3%	10.4%
Light Truck	**8763**	**21.6%**	3.4%	15.6%
Bus	234	5.6%	3.0%	4.8%
Heavy Truck	4866	12.2%	**5.9%**	6.8%
Motor Home	38	**26.3%**	**5.3%**	13.8%
Motorcycle/Moped	5418	3.2%	2.0%	**32.5%**
Other	1209	7.9%	0.7%	10.4%

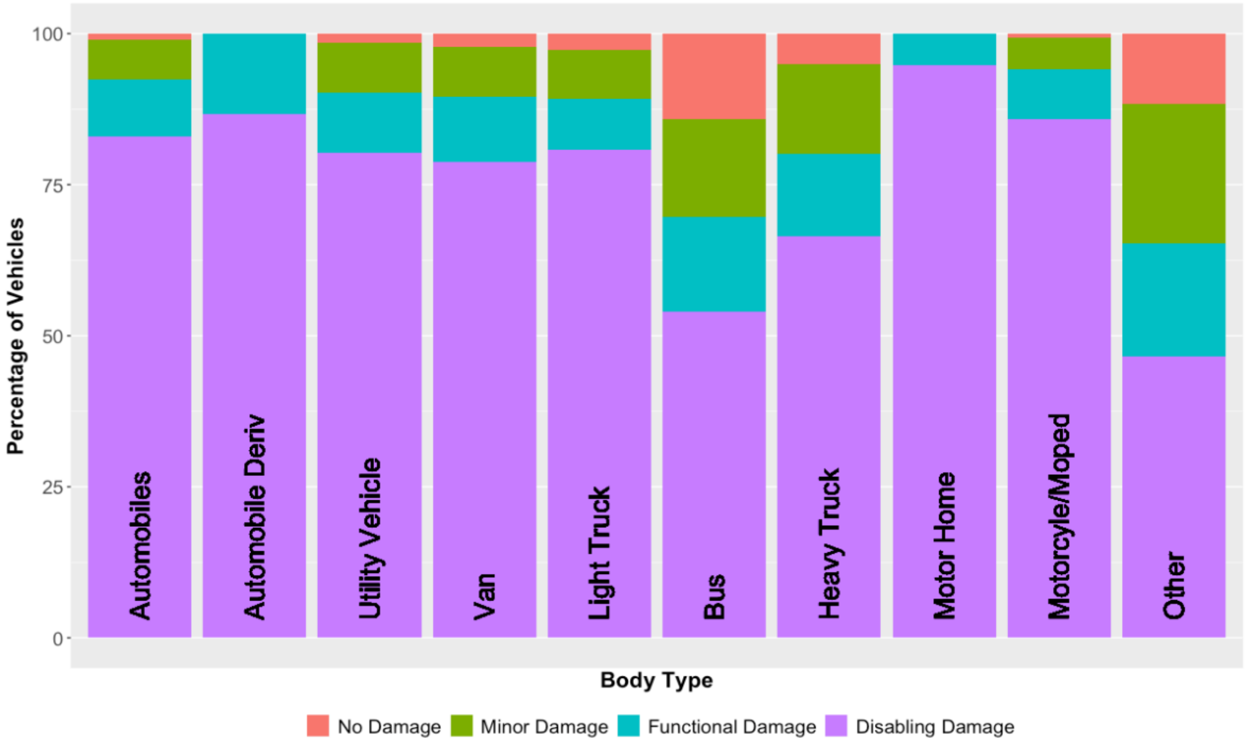


Figure 9: Percentage of vehicles of each body type that received each type of damage.

These makes combine for 90% of all vehicles in accidents, so we believe this is a good representative sample. Interestingly, the main conclusion from examining the top and bottom 3 for each accident type by make is that the body type is most important in determining how frequently each accident type occurs. For example, makes that have motorcycles (Suzuki, Kawasaki, Harley-Davidson, Honda, and Yamaha) have low rollover, low fire/explosion, but high speeding related accidents. Similarly, heavy truck makes (Freightliner, Volvo, Navistar, and Peterbilt) are more likely to be high fire/explosion but low on speeding related accidents. The makes that are most likely to be in accidents are also the most frequently purchased makes in the United States.

Table 8: The top 3 vehicle makes for each of the accident types investigated. The relevant statistic is given in parenthesis.

Ranking	Accidents	Rollover	Fire/Explosion	Speeding Related
Top 1	Ford (7234)	Jeep (25.4%)	Freightliner (7.9%)	Kawasaki (45.3%)
Top 2	Chevrolet (4304)	GMC (20.1%)	Volvo (7.1%)	Suzuki (42.6%)
Top 3	Toyota (3705)	Ford (19.3%)	Navistar (6.8%)	Yamaha (39.7%)
Bottom 3	Kensworth (528)	Suzuki (3.0%)	Honda (1.9%)	Freightline (6.2%)
Bottom 2	Volkswagen (512)	Kawasaki (2.2%)	Harley-Davidson (1.8%)	Navistar (5.8%)
Bottom 1	BMW (508)	Harley-Davidson (0.3%)	Yamaha (1.7%)	Peterbilt (5.7%)

Finally, we examine the model year of vehicles to see if vehicles have become safer overtime. We find that rollover and speeding related crashes decrease with newer cars. However, fire/explosion has gone up with newer cars. One possible explanation for this is the prevalence of electric vehicles, which may be more prone to fire/explosion due to the large batteries. The relationship of type of damage by model year from 1988-2018 is shown in Figure 10. We chose this time period, because 1988 is the first model year with over 100 accidents. We find that the percentage of vehicles with disabling damage has decreased over time. This shows that overall vehicles have gotten safer and are still at least somewhat operational after fatal crashes. Both minor damage and functional damage have increased since 1988, but these damage groups are much more favorable to disabling damage. One interesting finding from the data is a sharp decline in the number of vehicles of model year 2009-2011, which can be attributed to the recession leading to fewer vehicle sales overall.

5.5 Discussion and Recommendations

We find that to keep drivers safer, it is beneficial to recommend newer, large vehicles, such as trucks and utility vehicles. These larger vehicles are more prone to rollover, but overall are safer than automobiles, vans, and motorcycles. However, a very obvious tradeoff must be considered, because these larger vehicles are much worse for the environment compared to smaller vehicles. The data show that newer cars are also beneficial, so even within the more dangerous vehicle groups, it is beneficial to upgrade to the newer vehicle types. One public policy option to implement this strategy would be to give consumers a tax credit for buying cars that are newer and have high safety ratings.

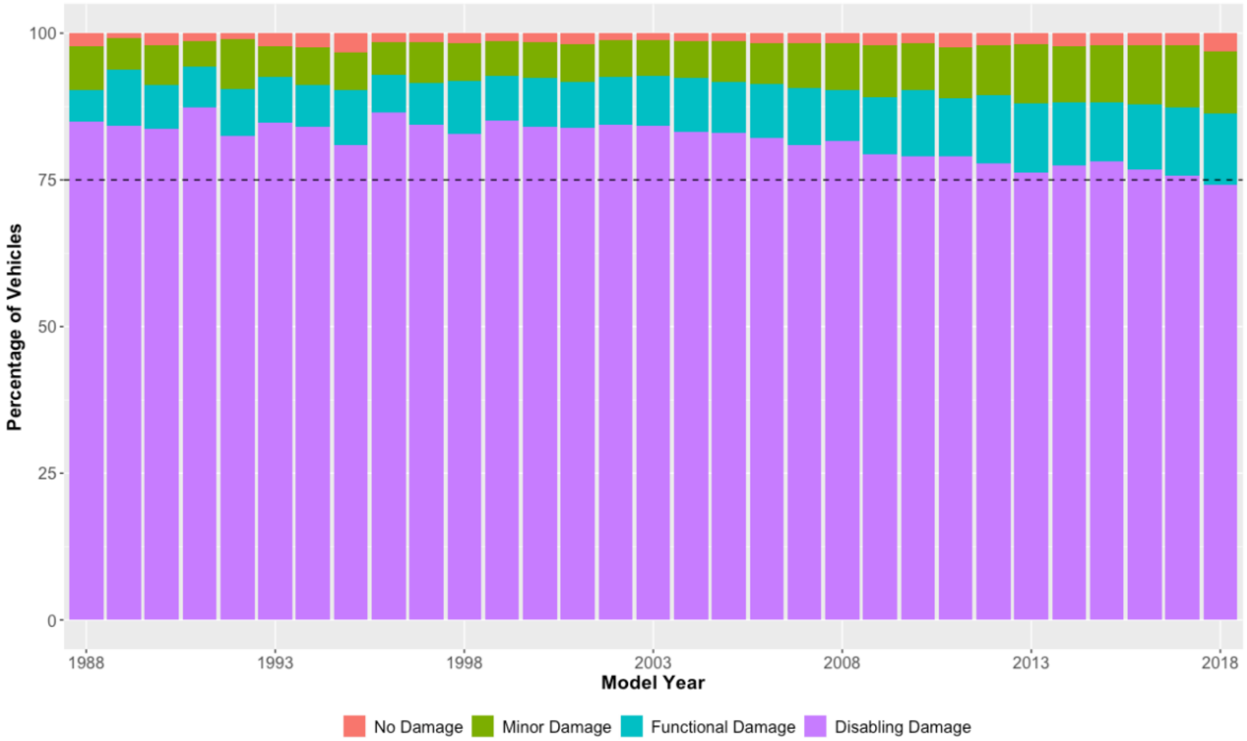


Figure 10: Evolution of damage types over the past 30 years (1988-2018). A dashed line at 75% of disabling damage highlights the difference between 1988 and 2018.

6 Conclusions and Recommendations

By analyzing the main causes of fatal crashes, we can provide recommendations to lawmakers, government officials, and the general public on behaviors to avoid while driving. Listed below are our major findings from the analysis:

1. We find that speeding is very common in fatal accidents. In fact, in 2018, speeding was related to more deaths in vehicles than drunk driving.

2. We find that one of the most common crash types for fatal accidents is motorcycles being hit while turning. While other vehicles had similar crash patterns, motorcycles were especially susceptible to this crash type.

3. We find that distracted driving is very rarely reported for fatal accidents. When distracted driving is reported, cell phone use is a very small percentage of fatal accidents.

4. We find that repeat offenders are much more likely to be in accidents or receive driving violations soon after accidents. Suggesting that a small portion of drivers account for a large percentage of accidents and violations.

5. We find that newer models are significantly safer than older model vehicles when involved in crashes. Newer models tend to have less damage when involved in accidents.

Based on these findings we have the following recommendations that we believe will help keep American roadways safer.

1. First, we recommend that officers more strictly monitor speeding related offenses. We see many campaigns for drunk driving and distracted driving, but our analysis shows that more people will be involved in speeding-related fatal accidents. Our recommendation for implementing this is to more strictly enforce speed limits, specifically on more rural roads.

2. Second, we recommend that government officials provide tax incentives to purchase newer vehicles. Even for used vehicles, there should be an incentive to buy more recently manufactured cars, because we find that these are significantly safer in fatal crashes.

3. Third, we recommend further driving school training for repeat offenders. We find that a small percentage of drivers account for a large amount of the accidents and driving violations. By targeting these drivers with further training, we can make the roadways safer.

References

[1] Gary Emerling and Katelyn Newman. The top 10 causes of death in america, Jan 2020. URL https://www.usnews.com/news/healthiest-communities/slideshows/top-10-causes-of-death-in-america?slide=9.

[2] Jia Shuo Yue, Chinmoy V Mandayam, Deepak Merugu, Hossein Karkeh Abadi, and Balaji Prabhakar. Reducing road congestion through incentives: a case study. In *Transportation Research Board 94th Annual Meeting, Washington, DC*, 2015.

[3] Daniel Eisenberg. Evaluating the effectiveness of policies related to drunk driving. *Journal of Policy Analysis and Management*, 22(2):249–274, 2003.

[4] John McDermott. At&t's anti-texting campaign: lots of impressions, zero sucess, Aug 2014. URL https://digiday.com/media/att-asks-twitter-whether-anti-texting-driving-campaign-working/.

[5] Mothers Against Drunk Driving. About us: Madd, 2019. URL https://www.madd.org/about-us/.

[6] IIHS. Fatality facts 2018: Yearly snapshot, Dec 2019. URL https://www.iihs.org/topics/fatality-statistics/detail/yearly-snapshot.

[7] Esurance insurance company. URL https://www.esurance.com/insights/dangerous-driving-study.

[8] Guangqing Chi, Willie Brown, Xiang Zhang, and Yanbing Zheng. Safer roads owing to higher gasoline prices: How long it takes. *American Journal of Public Health*, 105

(8):e119–e125, 2015. doi: 10.2105/AJPH.2015.302579. URL https://doi.org/10.2105/AJPH.2015.302579. PMID: 26066946.

[9] Norman Garrick, Carol Atkinson-Palombo, and Hamed Ahangari. Why america's roads are so much more dangerous than europe's, Nov 2016. URL https://www.vox.com/the-big-idea/2016/11/30/13784520/roads-deaths-increase-safety-traffic-us.

[10] Federal Highway Administration. Public road length, miles by ownership, 2013. URL https://www.bts.gov/content/public-road-length-miles-ownership.

[11] U.S. Census Bureau. U.s. states populations, land area, and population density, 2014. URL https://www.states101.com/populations.

[12] Nicholas Van Dyke and Mark T Fillmore. Alcohol effects on simulated driving performance and self-perceptions of impairment in dui offenders. *Experimental and Clinical Psychopharmacology*, 22(6):484, 2014.

[13] R Ranmath, N Kinnear, S Chowdhury, and T Hyatt. Interacting with android auto and apple carplay when driving: The effect on driver performance, Jan 2020. URL https://www.iamroadsmart.com/campaign-pages/end-customer-campaigns/infotainment.

[14] GFK. Us consumers cite car reliability, safety as more important than fuel economy, Apr 2019. URL https://www.gfk.com/en-us/insights/press-release/us-consumers-cite-car-reliability-safety-as-more-important-than-fuel-economy/.

[15] Carolina Tippett. 1 in 4 car accidents occur during rush hour, Jul 2019. URL https://www.ncdemography.org/2014/03/24/1-in-4-car-accidents-occur-during-rush-hour/.

[16] Jonathan J. Rolison, Shirley Regev, Salissou Moutari, and Aidan Feeney. What are the factors that contribute to road accidents? an assessment of law enforcement views, ordinary drivers' opinions, and road accident records. *Accident Analysis & Prevention*, 115:11 – 24, 2018. ISSN 0001-4575. doi: https://doi.org/10.1016/j.aap.2018.02.025. URL http://www.sciencedirect.com/science/article/pii/S0001457518300873.

[17] John Stossel. Are women worse drivers than men?, May 2009. URL https://abcnews.go.com/2020/story?id=3148281&page=1.

[18] Edmonds, Feb 2015. URL https://www.edmunds.com/car-safety/obese-drivers-face-higher-risks-in-cars.html.

[19] William Carrol and Jeffery Rhoades, Mar 2012. URL https://meps.ahrq.gov/data_files/publications/st364/stat364.shtml.

Appendix: Source Code

Setup and Library Import

```
library(pracma)
library(MASS)
library(plyr)
library(dplyr)
library(caret)
library(tidyverse)
library(lubridate)
library(ggplot2)
library(readxl)
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
opts_chunk$set(comment = NA)
```

Appendix 1: Question 1 Code (Safest States)

```
ACC_df <- read_csv("../FARS_Data/FARS2018NationalCSV/ACCIDENT.csv")
VEH_df <- read_csv("../FARS_Data/FARS2018NationalCSV/VEHICLE.csv")
PER_df <- read.csv("../FARS_Data/FARS2018NationalCSV/PERSON.csv")
VEH_df<-ACC_df%>%
  select(ST_CASE, LONGITUD, LATITUDE)%>%
  left_join(VEH_df, by=c("ST_CASE"="ST_CASE"))

miles_of_road <-
  read_excel("../Background_Information/2013 miles of road per state_abr.xlsx")
state_population <-
  read_excel("../Background_Information/2014 state population and total area.xlsx")
miles_of_road <- select(miles_of_road, State, Total,Abr.)

# Day of week, rural or urban, and Functional system are added to the Vehicle data frame
a<-select(ACC_df, ST_CASE, DAY_WEEK, RUR_URB, FUNC_SYS)
VEH_df<-left_join(VEH_df,a, by= c("ST_CASE"="ST_CASE"))
```

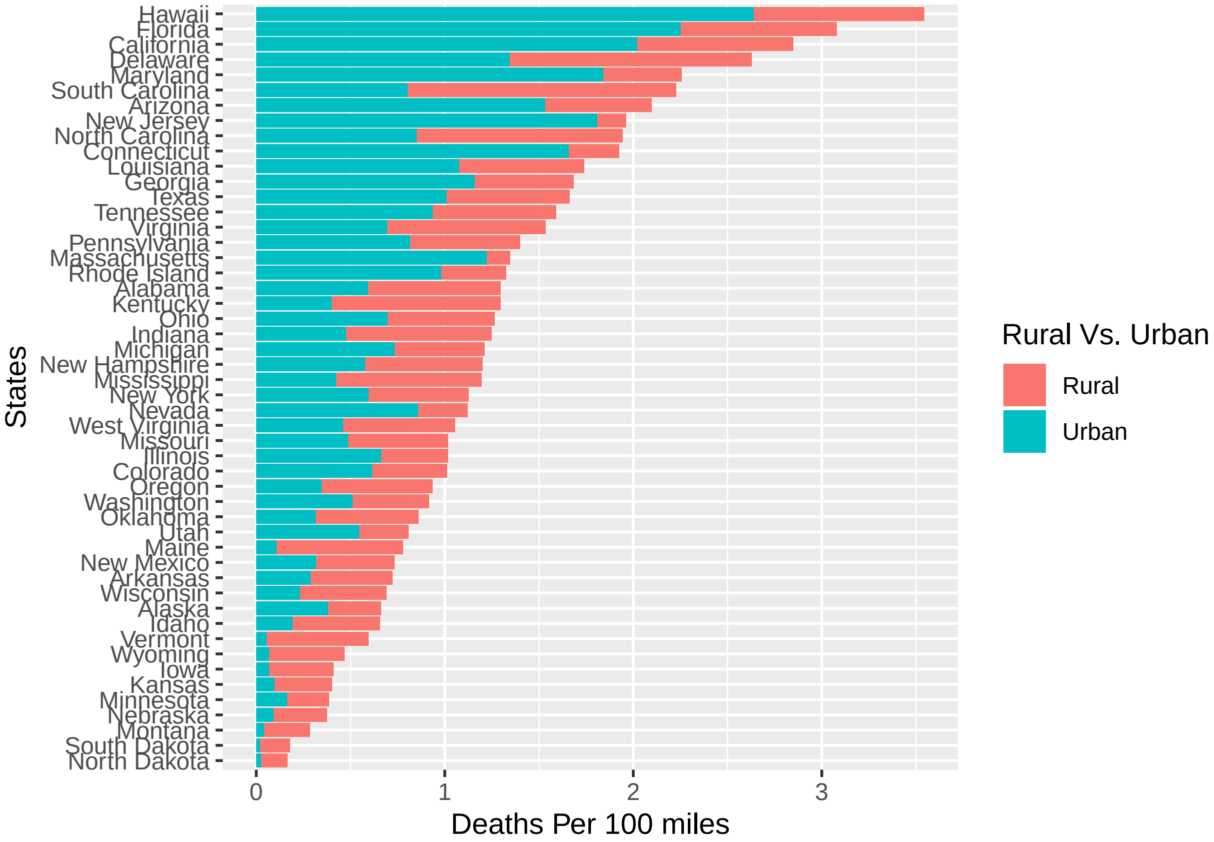
Initial Thoughts

- States with more big cities have more fatal accidents
- States on the east coast have a higher amount of fatal accidents
- Urban locations are more dangerous than rural locations
- States with larger populations are more dangerous

```
# Percent deaths in accidents
dpercent <- VEH_df %>% count(DEATHS) %>% mutate(x = 24790/sum(n))
```

```
state_mapping = c(Alabama = "1", Alaska = "2", Arizona = "4",
    Arkansas = "5", California = "6", Colorado = "8", Connecticut = "9",
    Delaware = "10", `District of Columbia` = "11", Florida = "12",
    Georgia = "13", Hawaii = "15", Idaho = "16", Illinois = "17",
    Indiana = "18", Iowa = "19", Kansas = "20", Kentucky = "21",
    Louisiana = "22", Maine = "23", Maryland = "24", Massachusetts = "25",
    Michigan = "26", Minnesota = "27", Mississippi = "28", Missouri = "29",
    Montana = "30", Nebraska = "31", Nevada = "32", `New Hampshire` = "33",
    `New Jersey` = "34", `New Mexico` = "35", `New York` = "36",
    `North Carolina` = "37", `North Dakota` = "38", Ohio = "39",
    Oklahoma = "40", Oregon = "41", Pennsylvania = "42", `Puerto Rico` = "43",
    `Rhode Island` = "44", `South Carolina` = "45", `South Dakota` = "46",
    Tennessee = "47", Texas = "48", Utah = "49", Vermont = "50",
    Virginia = "51", `Virgin Islands` = "52", Washington = "53",
    `West Virginia` = "54", Wisconsin = "55", Wyoming = "56")
```

```
#Recode the state number to name of the state.
ACC_per_state2<-mutate(VEH_df, STATE= as.character(STATE),
                       STATE = fct_recode(STATE, !!!state_mapping))

#Highest and lowest fatalities
most_least_fatalities <- ACC_df%>%
  mutate( STATE= as.character(STATE) ,
         STATE = fct_recode(STATE, !!!state_mapping)) %>%
  count(STATE)%>%
  arrange(n)%>%
  left_join( miles_of_road, by = c("STATE"= "State"))%>%
  mutate(acc_per_mile= (n/Total) * 100)%>%
  arrange(desc(n))

#Accidents vehicles with and without fatal accidents
ACC_per_state<- ACC_per_state2%>%
  mutate(DEATHS = (DEATHS>0), RUR_URB= as.character(RUR_URB),
         DR_DRINK=as.character(DR_DRINK))

#count the number of accident in each state that happen in either rural or urban locations
ACC_State_RUR_URB<-ACC_per_state%>%
  count(STATE, RUR_URB)%>%
  filter(RUR_URB<3)

#left_join the miles of road dataset for each state.
# Calculate the number of accidents per 100 miles of road
ACC_RUR_URB<-left_join(ACC_State_RUR_URB, miles_of_road, by = c("STATE"= "State"))%>%
  mutate(acc_per_mile= (n/Total) * 100)%>%
  arrange(desc(n))
# ACC_RUR_URB
# plot a bar graph showing what states have the highest number of accident per miles of
# road and whether the majority of the aciddents happened in rural or urban locations
ACC_RUR_URB%>%
  filter(STATE!="District of Columbia")%>%
  ggplot(aes(y=  acc_per_mile, x= reorder(STATE, acc_per_mile),fill= RUR_URB))+
  geom_bar(stat="identity")+
  coord_flip()+
  xlab("States")+
  ylab("Deaths Per 100 miles")+
  scale_fill_discrete(name= "Rural Vs. Urban",  labels= c("Rural", "Urban"))
```



```
#56 percent of the accidents happen in urban locations
ggsave("ACC_RUR_URB_BAR.png",
       width = 30, height = 20, units = "cm")
#56% of accidents happen in urban locations
```

Appendix 2: Question 2 Code (Contributing Factors)

Several categories of factors contributing in the accidents are recorded in FARS dataset. Speeding, alcohol, drugs, driver distractions, drivers' vision obstruction, and vehicle level contributing circumstances are some examples. Perform a comprehensive review of the contributing factors, analyze the related tables in the dataset, and identify the most common contributing factors in each category.

Initial Thoughts

- Probably a relationship between speeding and fatal crashes.

Code Information:

Table	CSV Name	FARS Data Dictionary Location
Driver Distractions	DISTRACT	545
Driver Impaired	DRIMPAIR	465
Vehicle Defects Contributing to Crash	FACTOR	510
Driver Manuevered to Avoid prior to Crash	MANEUVER	542
Non-Motorist Contributing Circumstances	NMCRASH	734
Non-Motorist Impaired	NMIMPAIR	743
Non-Motorist Action	NMPRIOR	730
Motorist Violations	VIOLATN	459
Driver Vision Obscured	VISION	539
Damaged Areas	DAMAGE	371

Case 1: Speeding

Important columns are in VEHICLE.csv. SPEEDREL denotes if the crash was speeding related. TRAV_SP is the speed of the driver and VSPD_LIM is speed limit of road crash occured on.

SPEEDREL Codes	Attributes
0	No
2	Yes, Racing
3	Yes, Exceeded Speed Limit
4	Yes, Too Fast for Conditions
5	Yes, Specifics Unknown
8	No Driver Present
9	Unknown

- Note: It may be interesting to analyze the relationship between speed limit and number of crashes.

```r
vehicle_df = read.csv("../FARS_Data/FARS2018NationalCSV/VEHICLE.csv")
```

```r
frequency_speed = as.data.frame(table(vehicle_df$SPEEDREL))
colnames(frequency_speed) = c("Code", "Freq")
f_s = frequency_speed[order(frequency_speed$Freq, decreasing = TRUE),
    ]
kable(f_s, "latex", booktabs = T, row.names = FALSE)
```

Code	Freq
0	41190
4	3850
3	3432
9	1735
5	1245
8	342
2	78

First, I wanted to investigate the total number of fatal accidents that are speeding related

```r
# Percentage of accidents that were speed related
# If two vehicles are in the same crash, only the ones
# which were speeding will have a speed_rel > 0.
speed_rel <- vehicle_df$SPEEDREL
speed_rel <- mapvalues(speed_rel, from = c(0, 2, 3, 4, 5, 8, 9),
                       to = c(0, 1, 1, 1, 1, 2, 2))
# Some missing values can be collected by comparing the
# travel speed to the speed limit (2077 -> 481)
for (i in 1:length(speed_rel)) {
    if (speed_rel[i] == 2) {
        if (vehicle_df$TRAV_SP[i] < 151) {
            speed_rel[i] = ifelse((vehicle_df$TRAV_SP[i] - vehicle_df$VSPD_LIM[i]) >
                                  0, 1, 0)
        } else {
            speed_rel[i] = 0
        }
    }
}

# Now identify which accidents had at least one accident where speed was related
speed_df = data.frame(cbind("Speed_Related" = speed_rel,
                            "State_Case" = vehicle_df$ST_CASE))
accident_speed = aggregate(x = speed_df[c("Speed_Related")],
                           by = speed_df[c("State_Case")], FUN = max)
accident_speed_percentage = sum(accident_speed$Speed_Related)/
  length(accident_speed$State_Case)
print(sprintf("In Fatal Crashes, Speed is related %s%% of the time",
              round(accident_speed_percentage, digits = 4)*100))
```

```
[1] "In Fatal Crashes, Speed is related 25.29% of the time"
```

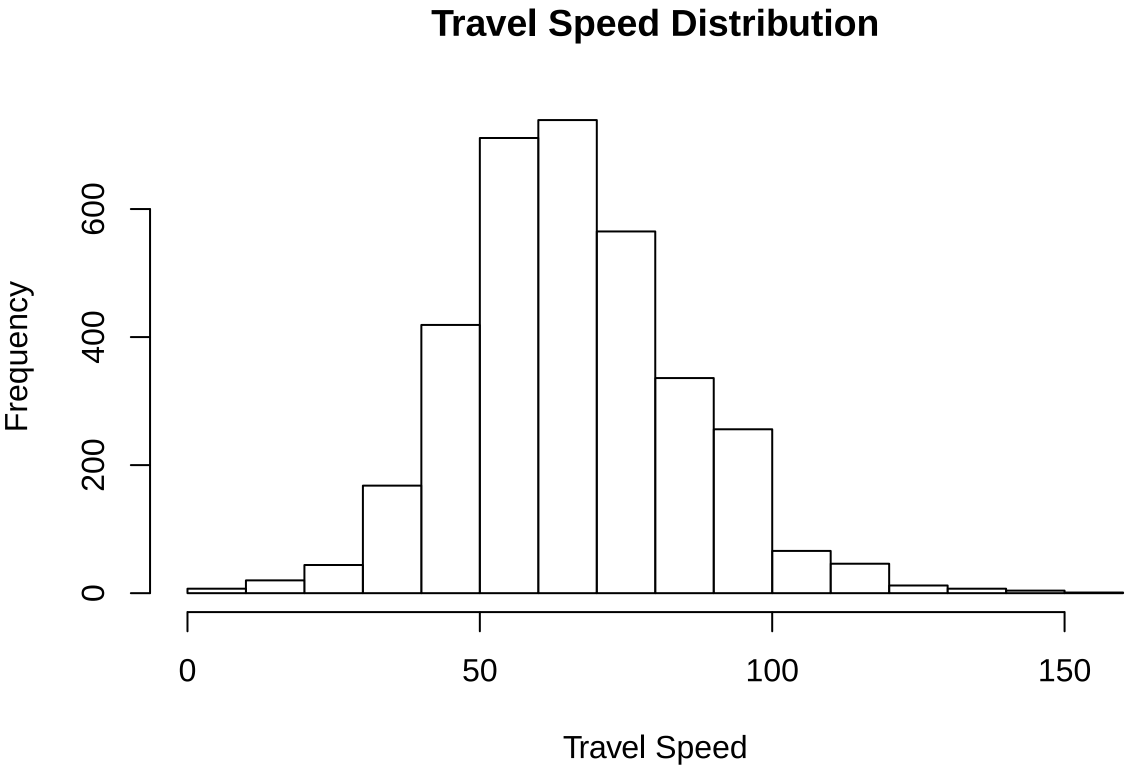
Next, of these accidents how are they distributed among the four types of speeding classified by FARS. One note is that there are accidents with multiple types, so the total percentage of the accidents caused by each type will be over 100.

```r
# Want to check how many accidents are a result of 1) racing 2) Exceeding Speed Limit
# 3) Speeding in bad conditions 4) Other
# First, check if there are multiple wihtin the same accident
speed_rel <- vehicle_df$SPEEDREL
speed_rel <- mapvalues(speed_rel, from = c(0, 2, 3, 4, 5, 8, 9),
                       to = c(0, 1, 2, 3, 4, 0, 0))
speed_df = data.frame(cbind("Speed_Related" = speed_rel,
                            "State_Case" = vehicle_df$ST_CASE))
types_of_speeding = aggregate(x = speed_df[c("Speed_Related")],
                              by = speed_df[c("State_Case")], FUN = sum)

# Since there are multiple speeding related for each,
# we can get the count of state cases joined by the speed related value
speed_df = data.frame(cbind("Speed_Related" = speed_rel,
                            "State_Case" = vehicle_df$ST_CASE))
types_of_speeding = aggregate(x = speed_df[c("State_Case")],
                              by = speed_df[c("Speed_Related")], FUN = length)

# Calculate percentages
total_accidents = length(unique(vehicle_df$ST_CASE))
total_speeding_related_accidents = sum(accident_speed$Speed_Related)
racing_accidents_percentage = types_of_speeding$State_Case[2] /
  total_speeding_related_accidents * 100
exceed_limit_accidents_percentage = types_of_speeding$State_Case[3] /
  total_speeding_related_accidents * 100
bad_cond_accidents_percentage = types_of_speeding$State_Case[4] /
  total_speeding_related_accidents * 100
other_accidents_percentage = types_of_speeding$State_Case[5] /
  total_speeding_related_accidents * 100

# Add the percentages to the total_table
speeding_related_accidents = data.frame(matrix(ncol = 3, nrow = 4))
colnames(speeding_related_accidents) = c("Type of Speeding", "Number of Accidents",
                                         "Percentage of all Accidents")
speeding_related_accidents[,1] = c("Racing", "Exceeded Speed Limit",
                                   "Speeding in Bad Conditions", "Other")
speeding_related_accidents[,2] = types_of_speeding$State_Case[2:5]
speeding_related_accidents[,3] = c(round(racing_accidents_percentage, digits = 4),
                                   round(exceed_limit_accidents_percentage, digits = 4),
                                   round(bad_cond_accidents_percentage, digits = 4),
                                   round(other_accidents_percentage, digits = 4))

# This is slightly over 100 because some accidents had more than one.
total_percentage = racing_accidents_percentage + exceed_limit_accidents_percentage +
  bad_cond_accidents_percentage + other_accidents_percentage
```

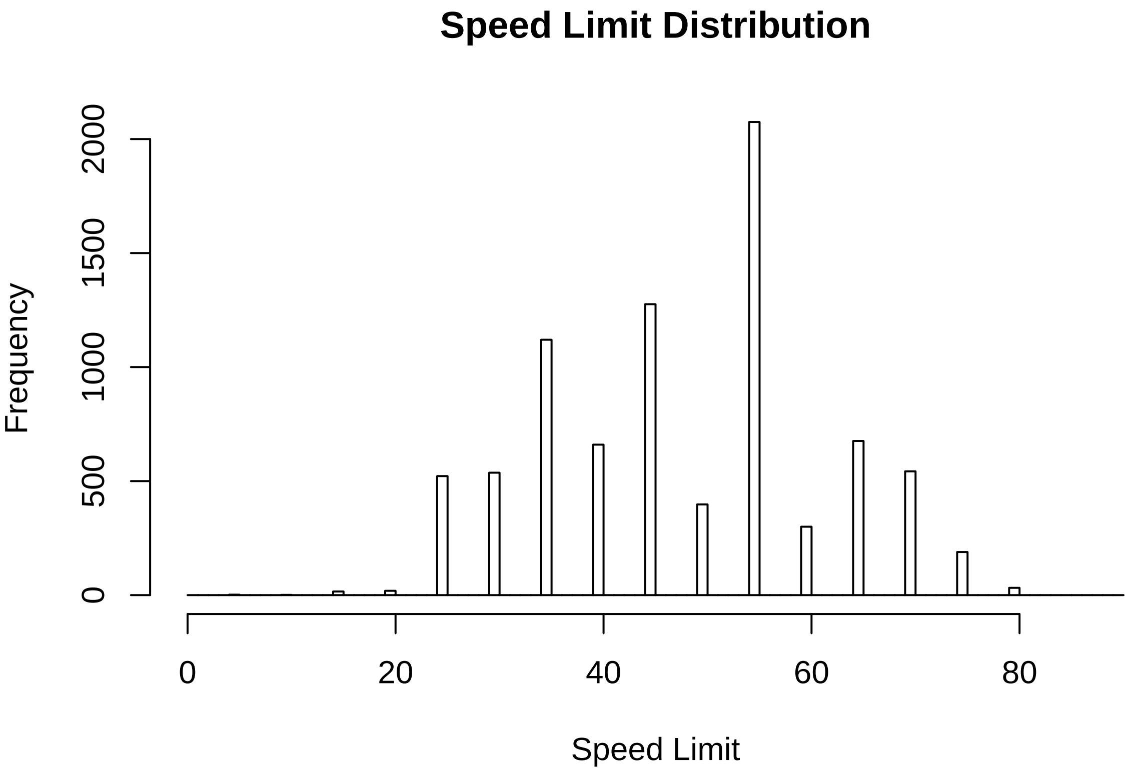
A few interesting ways to slice the speeding data. First, let's look at the distribution of travel speed for vehicles in a fatal crash. This should have tails on both ends, because cars that were going very slow may have been hit by cars going very fast. Second, by looking at the speed limit of the roads the crashes occur on, we can identify the types of roads that most commonly have fatal crashes. My initial assumption is that there will be large clumpings around 55 and 70, since these are common speed limits for the highway and interstate, respectively. Finally, by comparing the travel speed to the speed limit, we can identify, on average, how fast were drivers going over the speed limit before being involved in a fatal crash.

Data coders round all speeds to the nearest integer, so we don't have to worry about decimals here.

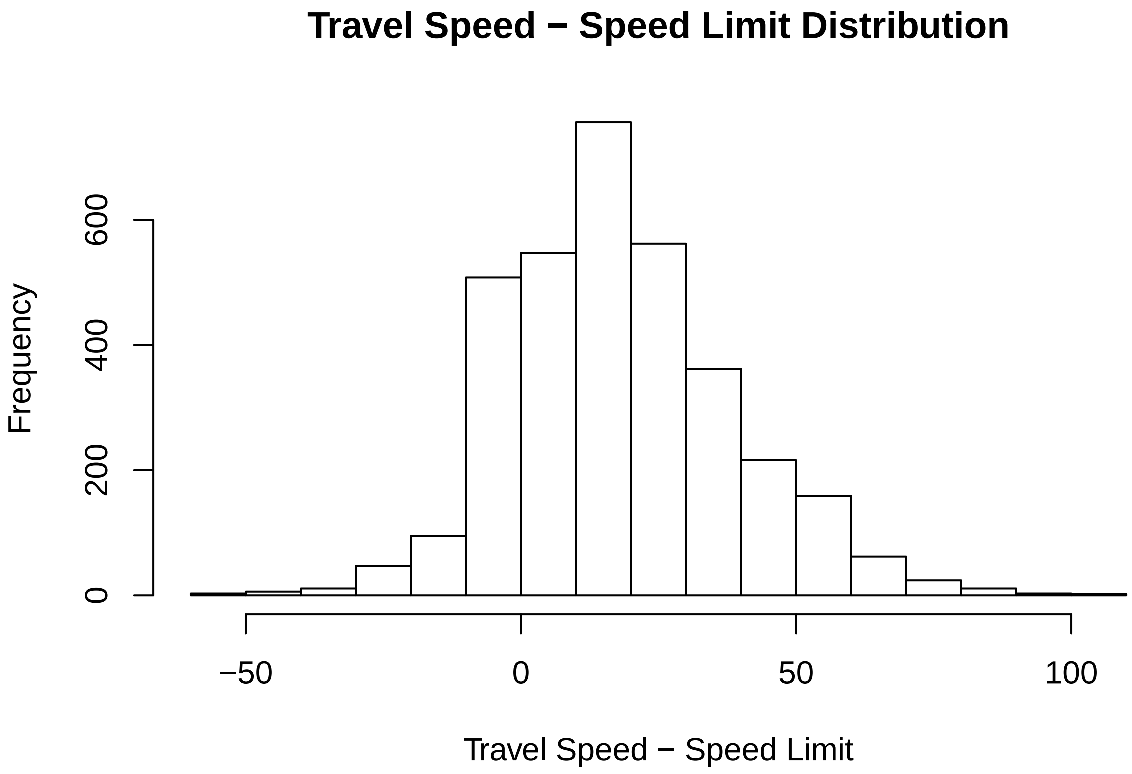
```
speed_rel <- mapvalues(speed_rel, from = c(0, 1, 2, 3, 4),
                       to = c(0, 1, 1,1,1))

# Distribution of how fast drivers were going
travel_speed_distribution = vehicle_df$TRAV_SP[speed_rel == 1]
travel_speed_distribution = travel_speed_distribution[travel_speed_distribution < 152]
travel_speed_distribution = travel_speed_distribution[travel_speed_distribution > 0]
hist(travel_speed_distribution,
     main = "Travel Speed Distribution", xlab = "Travel Speed")
```



```
# Distribution of the speed limit on roads
speed_limit_distribution = vehicle_df$VSPD_LIM[speed_rel == 1]
speed_limit_distribution = speed_limit_distribution[speed_limit_distribution < 96]
speed_limit_distribution = speed_limit_distribution[speed_limit_distribution > 4]
hist(speed_limit_distribution, breaks = seq(0,90,1),
     main = "Speed Limit Distribution", xlab = "Speed Limit")
```



```r
# Distribution of the difference between the driver speed and speed limit.
speed_amt_dist = data.frame(cbind(vehicle_df$TRAV_SP, vehicle_df$VSPD_LIM))
speed_amt_dist = speed_amt_dist[speed_rel == 1,]
speed_amt_dist = speed_amt_dist[speed_amt_dist$X1 < 152,]
speed_amt_dist = speed_amt_dist[speed_amt_dist$X1 > 0,]
speed_amt_dist = speed_amt_dist[speed_amt_dist$X2 < 96,]
speed_amt_dist = speed_amt_dist[speed_amt_dist$X2 > 4,]
hist(speed_amt_dist$X1 - speed_amt_dist$X2,
     main = "Travel Speed - Speed Limit Distribution",
     xlab = "Travel Speed - Speed Limit")
```



Case 2: Alcohol

Important columns are in PERSON.csv. DRINKING denotes whether the officer believed alcohol was involved (0: No, 1: Yes, 8: Not Rep, 9: Unknown). Other important columns are: ALC_DET (method by which police made determination), ALC_STATUS (Whether test was administered 0:No, 2:Yes), ATST_TYP (Type of Test), ALC_RES (Test Result)

ALC_DET Codes	Attributes
1	Evidential Test (breath, blood, urine)
2	Preliminary Breath Test (PBT)
3	Behavioral
4	Passive Alcohol Sensor (PAS)
5	Observed
8	Other (e.g., Saliva test)
9	Not Reported

ALC_RES Codes	Attributes
000-939	Actual Value
940	.94 or Greater
996	Test Not Given
997	AC Test Performed, Results Unknown
998	Positive Reading with No Actual Value
995	Not Reported
999	Reported as Unknown if Tested

```
person_df = read.csv("../FARS_Data/FARS2018NationalCSV/PERSON.csv")
accident_df = read.csv("../FARS_Data/FARS2018NationalCSV/Accident.csv")
# Only get important columns
alc_res = person_df$ALC_RES
drinking = person_df$DRINKING
accident_drinking = accident_df$DRUNK_DR
```

A couple interesting questions for alcohol consumption and fatalities. First, how often are drunk people involved in fatal accidents. A side note to this question direclty involves whether the driver was drunk. Second, is there a relationship between how drunk (i.e. BAC level) and fatalities?

There are a few ways to get drunk driver statistics. First, is using the accident table column "DR_DRINK". This column provides the number of drunk drivers in each accident. Another way is to use the person table column "ALC_RES" and test if the BAC of the driver (i.e. seat pos == 11) is greater than 0.8 (the federal limit). The limit in Utah is slighlty lower (0.5), but only 59 out of 8644 drunk driving cases are in Utah, so using the 0.8 benchmark should be sufficient. Let's compare the number of cases using each method.

Look at who is more likely to die in accidents involved in drinking.

```
# Using Accident table. Need to do > 0, because some accidents have more
# than one drunk driver
accident_drinking = accident_df$DRUNK_DR
num_accident_drinking = accident_drinking[accident_drinking > 0]

# Total number of drunk drivers
sum(num_accident_drinking)
```

```
[1] 8644
```

```
# Total accidents involving at least one drunk driver
length(num_accident_drinking)
```

```
[1] 8379
```

```
# Number of drunk drivers using Person BAC > 0.8 guideline
num_person_drinking = alc_res[person_df$SEAT_POS == 11]
num_person_drinking = num_person_drinking[num_person_drinking < 941]
num_person_drinking = num_person_drinking[num_person_drinking > 80]

# Total number of drunk drivers
length(num_person_drinking)
```

```
[1] 5752
```

```
# Number of accidents drunk drivers were involved using Person BAC > 0.8 guideline
num_pers_acc_drink = data.frame(cbind(alc_res, person_df$ST_CASE))
num_pers_acc_drink = num_pers_acc_drink[person_df$SEAT_POS == 11,]
num_pers_acc_drink = num_pers_acc_drink[num_pers_acc_drink$alc_res > 80,]
num_pers_acc_drink = num_pers_acc_drink[num_pers_acc_drink$alc_res < 941,]
num_pers_acc_drink = aggregate(x = num_pers_acc_drink[c("alc_res")],
                               by =num_pers_acc_drink[c("V2")],
                               FUN = length)

# Total accidents involving at least one drunk driver
nrow(num_pers_acc_drink)
```

```
[1] 5666
```

So based on the accident table, there were 8644 drunk drivers, which were spread across 8379 accidents. Based on the person table, using BAC as a guideline, in total there were 5752 drunk drivers, which were spread across 5666 accidents.

Another interesting line of analysis is to compare how often the police expected drinking was involved, compared to the actual number of drivers that were legally intoxicated

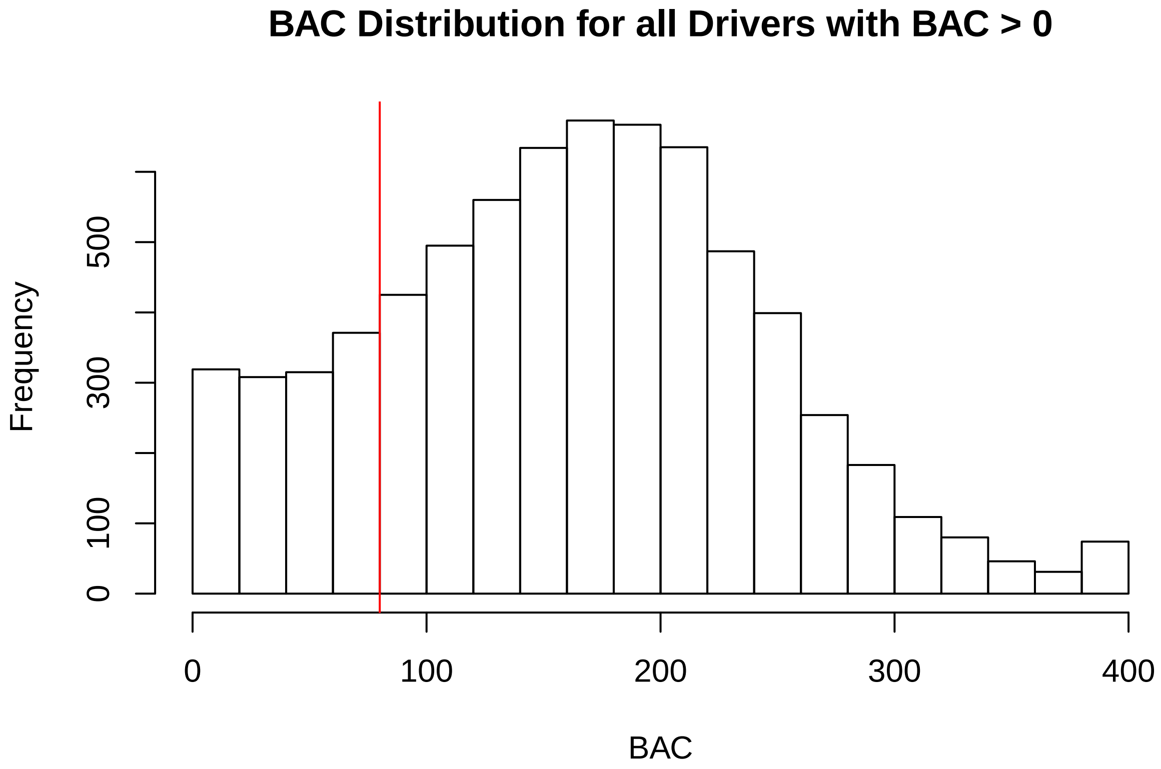
```r
# Number of accidents drunk drivers were involved using Person BAC > 0.8 guideline
num_pers_acc_drink = data.frame(cbind(drinking, person_df$ST_CASE))
num_pers_acc_drink = num_pers_acc_drink[person_df$SEAT_POS == 11,]
num_pers_acc_drink = num_pers_acc_drink[num_pers_acc_drink$drinking == 1,]
num_pers_acc_drink = aggregate(x = num_pers_acc_drink[c("drinking")],
                               by = num_pers_acc_drink[c("V2")],
                               FUN = length)

# Total accidents involving at least one drunk driver
nrow(num_pers_acc_drink)
```

```
[1] 6112
```

Surprisingly, there were roughly 500 accidents where police believed alcohol was involved, but the driver did not have a BAC over 0.8. It is possible that other occupants in the vehicle were drunk (there are 1350 of these cases by modifying SEAT_POS != 11 above) or the driver was distracted by alcohol, but not drunk. However, we are not given further information about why the police believed alcohol was involved, so this is merely speculation.

Examining the BAC of drivers who were involved in fatal accidents may lead to interesting relationships.

```r
# For all people in the person set. Generally, most people
# are not drinking.
alc_res_all = alc_res[person_df$SEAT_POS == 11]
alc_res_all = alc_res_all[alc_res_all < 941]
# Only 45 cases are greater than 400, so these are moved to
# 400 for ease of plotting the histogram
alc_res_all[alc_res_all > 400] = 400
# There are 14268 cases where the BAC is 0, so these are
# removed to show the histogram more clearly
alc_res_non_zero = alc_res_all[alc_res_all > 0]
hist(alc_res_non_zero, main = "BAC Distribution for all Drivers with BAC > 0",
    xlab = "BAC", breaks = seq(0, 400, 20))
abline(v = 80, col = "red")
```



```
# Total percentage of drivers over a BAC of 0.8
drunk_percentage = length(alc_res_all[alc_res_all > 80])/length(alc_res_all)
print(sprintf("Number of drivers in fatal crashes who had BAC over legal limit 0.8 is %s%%",
    round(drunk_percentage * 100, digits = 2)))
```

```
[1] "Number of drivers in fatal crashes who had BAC over legal limit 0.8 is 26.96%"
```

When there are fatalities from drunk driving, there are more often drunk people than sober people. This may be related to drunk people riding home together. The BAC follows pretty close to a normal distribution, which is somewhat surprising to me. I expected a large area below 200, and much less area to the right of 200.

Case 3: Drugs

Data is in both DRUGS.csv and PERSON.csv, need to check that similar data is in both. Using DRUGS.csv right now. For DRUGS.csv make sure to group by person number. One person can have multiple drugs. From PERSON.csv, use DRUGS (0: No, 1:Yes, 8: Not Reported, 9:Unknown) DRUG_DET (Type of Test 1:Evidential, 2:Expert, 3:Behavioral, 7:Other, 8:NR) DSTATUS (Test Given 0:No, 2:Yes, 8:NR, 9:Unknown). DRUGS.csv use DRUGSPEC (Type of Test pg 657) DRUGRES (Drug Test Result)

DRUGRES Codes	Attributes
000	Test Not Given
001	Tested, No Drugs Found/Negative
100-295	Narcotic*
300-399	Depressant*
400-495	Stimulant*
500-595	Hallucinogen*
600-695	Cannabinoid*
700-795	Phencyclidine (PCP)*
800-895	Anabolic Steroid*

DRUGRES Codes	Attributes
900-995	Inhalant*
996	Other Drug
997	Tested for Drugs, Results Unknown
998	Tested for Drugs, Drugs Found, Type Unknown/Positive
095	Not Reported
999	Reported as Unknown If Tested for Drugs

```r
drug_df = read.csv("../FARS_Data/FARS2018NationalCSV/DRUGS.csv")
person_df = read.csv("../FARS_Data/FARS2018NationalCSV/PERSON.csv")
# Need to join these based on state case and state number.
# Group DRUGS_DF by person number first. I think it is an
# outer group to make multiple columns for each drug
```

Need to join drug df and person df based on st_case, veh_no, per_no to get the seat pos of each person. That way the drugged drivers can be extracted.

```r
reduced_person_df = data.frame(cbind(ST_CASE = person_df$ST_CASE,
    VEH_NO = person_df$VEH_NO, PER_NO = person_df$PER_NO, SEAT_POS = person_df$SEAT_POS))
person_drugs_df = merge(x = drug_df, y = reduced_person_df, by = c("VEH_NO",
    "PER_NO", "ST_CASE"), all = TRUE)
```

Now, filter based on only drivers and map all drug values to the categories provided by FARS NHTSA

```r
person_drugs_df = person_drugs_df[person_drugs_df$SEAT_POS == 11,]
person_drugs_df$DRUGRES = mapvalues(person_drugs_df$DRUGRES,
                                    from = 1, to = 2)
person_drugs_df$DRUGRES = mapvalues(person_drugs_df$DRUGRES,
                                    from = c(0, 997, 95, 999), to = rep(1, 4))
person_drugs_df$DRUGRES = mapvalues(person_drugs_df$DRUGRES,
                                    from = seq(100, 295, by = 1), to = rep(3, 196))
person_drugs_df$DRUGRES = mapvalues(person_drugs_df$DRUGRES,
                                    from = seq(300, 399, by = 1), to = rep(4, 100))
person_drugs_df$DRUGRES = mapvalues(person_drugs_df$DRUGRES,
                                    from = seq(400, 495, by = 1), to = rep(5, 96))
person_drugs_df$DRUGRES = mapvalues(person_drugs_df$DRUGRES,
                                    from = seq(500, 595, by = 1), to = rep(6, 96))
person_drugs_df$DRUGRES = mapvalues(person_drugs_df$DRUGRES,
                                    from = seq(600, 695, by = 1), to = rep(7, 96))
person_drugs_df$DRUGRES = mapvalues(person_drugs_df$DRUGRES,
                                    from = seq(700, 795, by = 1), to = rep(8, 96))
person_drugs_df$DRUGRES = mapvalues(person_drugs_df$DRUGRES,
                                    from = seq(800, 895, by = 1), to = rep(9, 96))
person_drugs_df$DRUGRES = mapvalues(person_drugs_df$DRUGRES,
                                    from = seq(900, 995, by = 1), to = rep(10, 96))
person_drugs_df$DRUGRES = mapvalues(person_drugs_df$DRUGRES,
                                    from = c(996, 998), to = rep(11, 2))
```

```r
drug_summary = data.frame(matrix(nrow = 11, ncol = 3))
colnames(drug_summary) = c("Drug Group", "Number of Drugged Drivers",
                           "Percentage of Drugged Drivers")

drug_summary[,1] = c("Unknown", "No Drugs Found", "Narcotic","Depressant","Stimulant",
                     "Hallucinogen", "Cannabinoid","Phencyclidine (PCP)",
                     "Anabolic Steroid","Inhalant", "Other")

for (i in seq(1, 11, by = 1)){
  subtotal = person_drugs_df$DRUGRES[person_drugs_df$DRUGRES == i]
  drug_summary[i, 2] = length(subtotal)
  drug_summary[i, 3] = round(length(subtotal) * 100 / 51872, digits = 2)
}
```

Case 4: Driver Distractions

Data is in DISTRACT.csv. Codes are below.

Distractions

Codes	Attributes
00	Not Distracted
03	By Other Occupant(s)
04	By a Moving Object in Vehicle
05	While Talking or Listening to Cellular Phone
06	While Manipulating Cellular Phone
07	Adjusting Audio or Climate Controls
09	While Using Other Component/Controls Integral to Vehicle
10	While Using or Reaching For Device/Object Brought Into Vehicle
12	Distracted by Outside Person, Object, or Event
13	Eating or Drinking
14	Smoking Related
15	Other Cellular Phone Related
16	No Driver Present/Unknown if Driver Present
17	Distraction/Inattention
18	Distraction/Careless
19	Careless/Inattentive
92	Distraction (Distracted), Details Unknown
93	Inattention (Inattentive), Details Unknown
96	Not Reported
97	Lost in Thought/Day Dreaming
98	Other Distraction
99	Reported as Unknown if Distracted

```
distraction_df = read.csv("../FARS_Data/FARS2018NationalCSV/DISTRACT.csv")
driver_distraction = distraction_df$MDRDSTRD
# Replace 5,6, 15 with 2; 3,4,7,9,10 with 3 etc.
```

```
frequency_distraction = as.data.frame(table(driver_distraction))
colnames(frequency_distraction) = c("Code", "Freq")
f_d = frequency_distraction[order(frequency_distraction$Freq, decreasing = TRUE), ]
kable(f_d, "latex", booktabs = T, row.names = FALSE)
```

Code	Freq
96	24259
0	17533
99	7050
93	1012
16	342
17	308
92	263
12	187
98	153
3	126
15	125
6	116
5	114
10	102
19	48
13	45
9	33
7	29
97	17
4	12
14	9
18	6

I think these can be grouped into the following groups.

Codes	Attributes
00	Not Distracted
05, 06, 15	Cell Phone Distraction
03, 04, 07, 09, 10	In-Car Distraction not cell phone
13, 14	Food / Smoking Related
12, 17, 18, 19, 92, 93, 97, 98	Other Distraction
16, 96, 99	Unknown if Distracted

There are a few different ways to group the data, but some of the commmon distractions I hear about from the media are cell-phone, food/drink, and other in-car distractions.

Some drivers had multiple distractions. Specifically, there were 51872 vehicles involved in crashes and 51889 distractions. Since there are more distractions than vehicles, the percentage of each type will sum to slightly more than 100

```r
driver_distraction = distraction_df$MDRDSTRD
driver_distraction = mapvalues(driver_distraction,
                              from = c(0), to = c(1))
driver_distraction = mapvalues(driver_distraction,
                              from = c(5, 6, 15), to = rep(2, 3))
driver_distraction = mapvalues(driver_distraction,
                              from = c(3, 4, 7, 9, 10), to = rep(3, 5))
driver_distraction = mapvalues(driver_distraction,
                              from = c(13, 14), to = rep(4, 2))
driver_distraction = mapvalues(driver_distraction,
                              from = c(12, 17, 18, 19, 92, 93, 97, 98), to = rep(5, 8))
```

```
driver_distraction = mapvalues(driver_distraction,
                               from = c(16, 96, 99), to = rep(6, 3))

distraction_summary = data.frame(matrix(nrow = 6, ncol = 3))
colnames(distraction_summary) = c("Distraction Group",
                                  "Total Number of Vehicles Distracted",
                                  "Percentage of Vehicles Distracted")

distraction_summary[,1] = c("Not Distracted", "Cell Phone Distraction",
                            "Other In-Car Distraction", "Food/Smoking Distraction",
                            "Other Distraction", "Unknown if Distracted")

for (i in seq(1, 6, by = 1)){
  subtotal = driver_distraction[driver_distraction == i]
  distraction_summary[i, 2] = length(subtotal)
  distraction_summary[i, 3] = round(length(subtotal) * 100 / 51872, digits = 2)
}
```

Case 5: Driver's Vision Obstruction

Codes	Attributes
00	No Obstruction Noted
01	Rain, Snow, Fog, Smoke, Sand, Dust
02	Reflected Glare, Bright Sunlight, Headlights
03	Curve, Hill, or Other Roadway Design Feature
04	Building, Billboard, Other Structure
05	Trees, Crops, Vegetation
06	In-Transport Motor Vehicle (including load)
07	Not In-Transport Motor Vehicle (parked/working)
08	Splash or Spray of Passing Vehicle
09	Inadequate Defrost or Defog System
10	Inadequate Vehicle Lighting System
11	Obstruction Interior to the Vehicle
12	External Mirrors
13	Broken or Improperly Cleaned Windshield
14	Obstructing Angles on Vehicle
95	No Driver Present/Unknown if Driver Present
97	Vision Obscured – No Details
98	Other Visual Obstruction
99	Reported as Unknown

```
vision_df = read.csv("../FARS_Data/FARS2018NationalCSV/VISION.csv")
# Think about grouping htese together better
```

```
frequency_vision = as.data.frame(table(vision_df$MVISOBSC))
colnames(frequency_vision) = c("Code", "Freq")
f_v = frequency_vision[order(frequency_vision$Freq, decreasing = TRUE), ]
kable(f_v, "latex", booktabs = T, row.names = FALSE)
```

Code	Freq
0	48264
99	1726
1	457
95	342
2	256
98	227
6	214
3	137
97	104
7	86
5	68
4	15
14	10
13	9
11	8
9	3
8	2
10	2
12	2

From the frequency table, we see that the most common values are no obstruction and unknown data. For the codes that correspond to obstructions, weather related obstructions, light-based obstructions, and other motor vehicle obstructions are the most common.

First, let's look at how many vehicles involved in fatal crashes were distracted. Some vehicles had multiple distractions so these need to be merged together.

Note: Easiest way to look at data is to create one row for each vision obstruction and divide the frequency by 51872, which is the total number of drivers. This does lead to the percentages summing to over 100%, but does identify how often each type of obstruction is

```
reduced_vehicle_df = data.frame(cbind(ST_CASE = vehicle_df$ST_CASE,
    VEH_NO = vehicle_df$VEH_NO))
vision_vehicle_df = merge(x = vision_df, y = reduced_vehicle_df,
    by = c("VEH_NO", "ST_CASE"), all = TRUE)
```

```
vision_vehicle_df = vision_vehicle_df[vision_vehicle_df$MVISOBSC > 0,]
vision_vehicle_df = vision_vehicle_df[vision_vehicle_df$MVISOBSC != 95,]
vision_vehicle_df = vision_vehicle_df[vision_vehicle_df$MVISOBSC != 99,]

# Total number of vision obstructions after removing
# 1) no obstruction (0) 2) Unknown Driver (95) and 3) Unknown (99)
nrow(vision_vehicle_df)
```

```
[1] 1600
```

There are quite a few vision obstruction types, but relatively few vehicles (i.e. 3%) have any vision obstruction at all. To help identify the major causes of vision obstruction, we group the vision obstruction types as follows:

Codes	Attributes
00	No Obstruction Noted
01	Rain, Snow, Fog, Smoke, Sand, Dust
02	Reflected Glare, Bright Sunlight, Headlights

Codes	Attributes
03	Curve, Hill, or Other Roadway Design Feature
04	Building, Billboard, Other Structure
05	Trees, Crops, Vegetation
06	In-Transport Motor Vehicle (including load)
07	Not In-Transport Motor Vehicle (parked/working)
08	Splash or Spray of Passing Vehicle
09	Inadequate Defrost or Defog System
10	Inadequate Vehicle Lighting System
11	Obstruction Interior to the Vehicle
12	External Mirrors
13	Broken or Improperly Cleaned Windshield
14	Obstructing Angles on Vehicle
95	No Driver Present/Unknown if Driver Present
97	Vision Obscured – No Details
98	Other Visual Obstruction
99	Reported as Unknown

Category	Group
1	No obstruction (00)
2	Weather Related Obstructions (01, 02)
3	Other Exterior Obstructions (03, 04, 05, 06, 07, 08)
4	Vehicle Related Obstructions (09, 10, 11, 12, 13, 14)
5	Other/Unknwon Obstructions (95, 97, 98, 99)

```r
# Weird order to make sure values don't get overwritten
vision_vehicle_df = merge(x = vision_df, y = reduced_vehicle_df,
                          by = c("VEH_NO", "ST_CASE"), all = TRUE)
vision_vehicle_df$MVISOBSC = mapvalues(vision_vehicle_df$MVISOBSC,
                                       from = c(seq(3,8,by=1)), to = rep(3,6))
vision_vehicle_df$MVISOBSC = mapvalues(vision_vehicle_df$MVISOBSC,
                                       from = c(95,97,98,99), to = rep(5,4))
vision_vehicle_df$MVISOBSC = mapvalues(vision_vehicle_df$MVISOBSC,
                                       from = c(seq(9,14,by=1)), to = rep(4,6))
vision_vehicle_df$MVISOBSC = mapvalues(vision_vehicle_df$MVISOBSC,
                                       from = c(1,2), to = rep(2,2))
vision_vehicle_df$MVISOBSC = mapvalues(vision_vehicle_df$MVISOBSC,
                                       from = c(0), to = c(1))
```

```r
vision_summary = data.frame(matrix(nrow = 5, ncol = 3))
colnames(vision_summary) = c("Vision Obstruction",
                             "Number of Vision Obstructions",
                             "Percentage of Vision Obstructions")

vision_summary[,1] = c("No Obstruction", "Weather Related Obstruction",
                       "Other Exterior Obstruction", "Vehicle Related Obstrcution",
                       "Other/Unknown Obstruction")

for (i in seq(1, 5, by = 1)){
  subtotal = vision_vehicle_df$MVISOBSC[vision_vehicle_df$MVISOBSC == i]
  vision_summary[i, 2] = length(subtotal)
  vision_summary[i, 3] = round(length(subtotal) * 100 / 51872, digits = 2)
}
```

```
# First find the rows with
reduced_vision_df = data.frame(cbind(ST_CASE = vision_df$ST_CASE,
    STATE = vision_df$STATE, VEH_NO = vision_df$VEH_NO))
```

```
obs_vehicle = vision_df[vision_df$MVISOBSC > 0, ]
obs_vehicle = obs_vehicle[obs_vehicle$MVISOBSC != 95, ]
obs_vehicle = obs_vehicle[obs_vehicle$MVISOBSC != 99, ]
# Number of distracted vehicles
nrow(obs_vehicle)
```

```
[1] 1600
```

```
# Percentage of distracted vehicles
round(nrow(obs_vehicle)/nrow(vision_df) * 100, digits = 4)
```

```
[1] 3.081
```

Case 6: Summarizing Figures

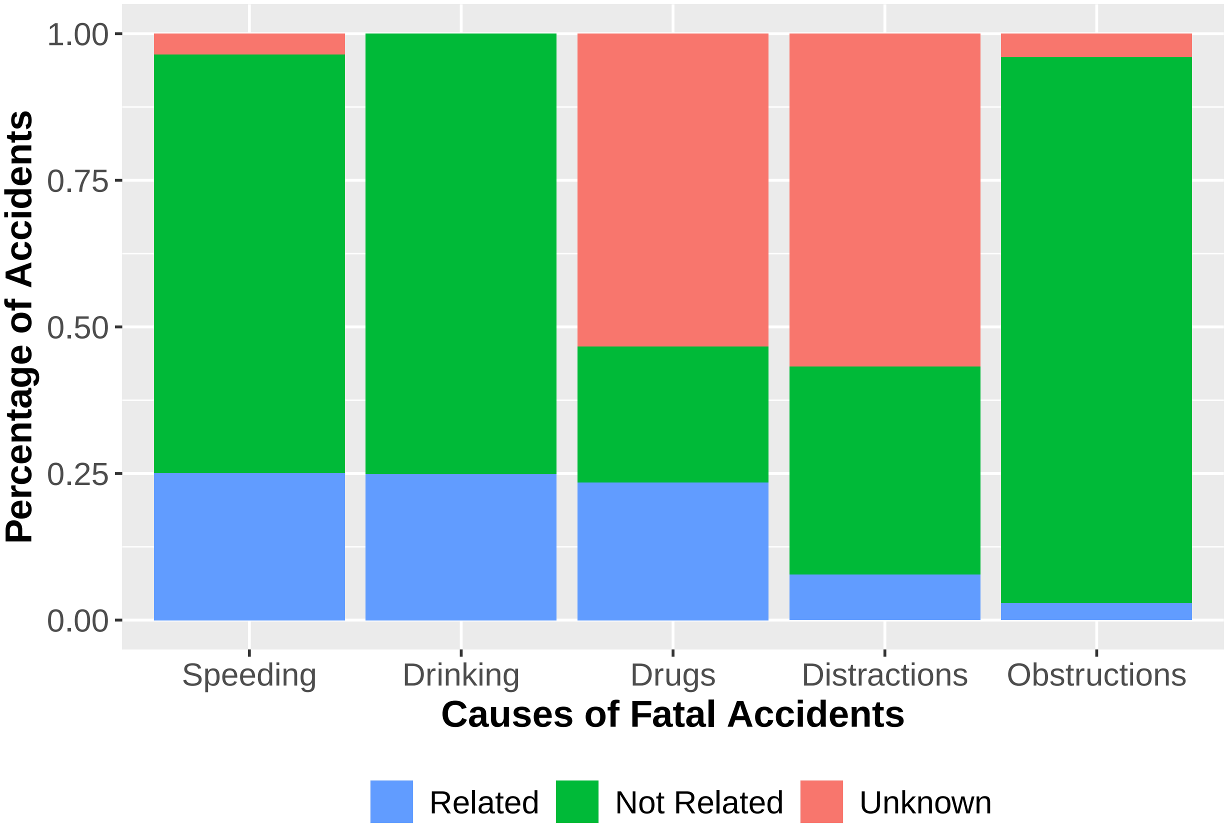
```
# Get percentage of accidents where each factor was involved

# Speeding
speed_rel <- vehicle_df$SPEEDREL
speed_rel <- mapvalues(speed_rel, from = c(0, 2, 3, 4, 5, 8, 9),
                       to = c(1, 2, 2, 2, 2, 0, 0))
speed_df = data.frame(cbind("Speed_Related" = speed_rel,
                            "State_Case" = vehicle_df$ST_CASE))
types_of_speeding = aggregate(x = speed_df[c("Speed_Related")],
                              by = speed_df[c("State_Case")], FUN = max)

# Speeding Results
unknown_speeding = length(types_of_speeding[types_of_speeding == 0])
known_no_speeding = length(types_of_speeding[types_of_speeding == 1])
known_speeding = length(types_of_speeding[types_of_speeding == 2])

# Alcohol
known_drinking = length(accident_df$DRUNK_DR[accident_df$DRUNK_DR > 0])
known_no_drinking = length(accident_df$DRUNK_DR[accident_df$DRUNK_DR == 0])

# Drugged Driving
drug_use = as.data.frame(person_drugs_df$DRUGRES)
colnames(drug_use) = c("Drugs_Found")
drug_use$Drugs_Found <- mapvalues(drug_use$Drugs_Found,
                                  from = c(seq(1,11,1)),
                                  to = c(22,23,rep(24,9)))
drug_df = data.frame(cbind("Drug_Related" = drug_use$Drugs_Found,
                           "State_Case" = person_drugs_df$ST_CASE))
types_of_drugs = aggregate(x = drug_df[c("Drug_Related")],
                           by = drug_df[c("State_Case")], FUN = max)
```

```
# Drug Results
unknown_drugs = length(types_of_drugs$Drug_Related[types_of_drugs$Drug_Related == 22])
known_no_drugs = length(types_of_drugs$Drug_Related[types_of_drugs$Drug_Related == 23])
known_drugs = length(types_of_drugs$Drug_Related[types_of_drugs$Drug_Related == 24])
# There are 86 cases where drugs are not reported these are added to the unknown
unknown_drugs = unknown_drugs + 86

# Distraction
distraction_use = as.data.frame(driver_distraction)
colnames(distraction_use) = c("Distractions")
distraction_use$Distractions = mapvalues(distraction_use$Distractions,
                                        from = c(seq(1,6,1)),
                                        to = c(21, rep(22, 4), 20))
distraction_summary_df = data.frame(cbind("Distraction" = distraction_use$Distractions,
                                         "State_Case" = distraction_df$ST_CASE))
types_of_dist = aggregate(x = distraction_summary_df[c("Distraction")],
                                 by = distraction_summary_df[c("State_Case")],
                                 FUN = max)

# Distraction Results
unknown_distraction = length(types_of_dist$Distraction[types_of_dist$Distraction == 20])
known_no_distraction = length(types_of_dist$Distraction[types_of_dist$Distraction == 21])
known_distraction = length(types_of_dist$Distraction[types_of_dist$Distraction == 22])


# Obstruction Results
obstruction_use = as.data.frame(vision_vehicle_df$MVISOBSC)
colnames(obstruction_use) = c("Obstruction")
obstruction_use$Obstruction = mapvalues(obstruction_use$Obstruction,
                                       from = c(seq(1,5,1)),
                                       to = c(21, rep(22, 3), 20))
obstruction_summary_df = data.frame(cbind("Obstruction" = obstruction_use$Obstruction,
                                         "State_Case" = vision_vehicle_df$ST_CASE))
types_of_obst = aggregate(x = obstruction_summary_df[c("Obstruction")],
                                 by = obstruction_summary_df[c("State_Case")],
                                 FUN = max)

# Distraction Results
unknown_obstruction = length(types_of_obst$Obstruction[types_of_obst$Obstruction == 20])
known_no_obstruction = length(types_of_obst$Obstruction[types_of_obst$Obstruction == 21])
known_obstruction = length(types_of_obst$Obstruction[types_of_obst$Obstruction == 22])
```

```
# Now turn the above into a table and then a stacked barplot
speeding_data = c(known_speeding, known_no_speeding, unknown_speeding)
drinking_data = c(known_drinking, known_no_drinking, 0)
drug_data = c(known_drugs, known_no_drugs, unknown_drugs)
distraction_data = c(known_distraction, known_no_distraction,
    unknown_distraction)
obstruction_data = c(known_obstruction, known_no_obstruction,
    unknown_obstruction)

test_four = data.frame(rbind(speeding_data, drinking_data, drug_data,
    distraction_data, obstruction_data))
```

```r
summary_info_crashes = c(0, 0, 0)
named_columns = c("Speeding", "Drinking", "Drugs", "Distraction",
    "Obstruction")
named_columns = c(seq(1, 5, 1))
named_rows = c("Related", "Not Related", "Unknown")

for (i in 1:nrow(test_four)) {
    for (j in 1:ncol(test_four)) {
        summary_info_crashes = rbind(summary_info_crashes, c(named_columns[i],
            named_rows[j], test_four[i, j]/33654))
    }
}

summary_info_crashes = data.frame(summary_info_crashes)
summary_info_crashes$X3 = as.numeric(as.matrix(summary_info_crashes$X3))
summary_info_crashes = summary_info_crashes[-1, ]
summary_info_crashes = summary_info_crashes[-6, ]
row.names(summary_info_crashes) <- 1:nrow(summary_info_crashes)
```

```r
summary_plot = ggplot(summary_info_crashes,
                      aes(fill=factor(X2,
                                      levels = c("Unknown", "Not Related", "Related")),
                          y=as.numeric(X3), x=X1)) +
  geom_bar(position="stack", stat="identity") +
  xlab("Causes of Fatal Accidents") +
  ylab("Percentage of Accidents") +
  scale_x_discrete(breaks=1:5, labels=c("Speeding", "Drinking",
                                        "Drugs", "Distractions", "Obstructions"))+
  theme(legend.position="bottom",
        legend.direction="horizontal",
        legend.title = element_blank()) +
  guides(fill = guide_legend(reverse = TRUE)) +
  theme(plot.title = element_text(size = 16, hjust = 0.5),
        axis.text = element_text(size = 12),
        axis.title = element_text(size = 14, face="bold"),
        legend.text=element_text(size=12))

summary_plot
```


1.00
0.75
0.50
0.25
0.00
Percentage of Accidents
Speeding
Drinking
Drugs
Distractions
Obstructions
Causes of Fatal Accidents
Related
Not Related
Unknown

Appendix 3: Question 3 Code (Crash Patterns)

The identification of crash patterns is important for policy makers to assist drivers in avoiding dangerous behavior and driving times. Crash patterns can include the time of the crash (i.e. month, week, day, hour), weather conditions, type of road, and type of vehicle. We first investigate individual factors and then hypothesize some potential combinations of factors that may be related in identifying crash patterns

Initial Thoughts

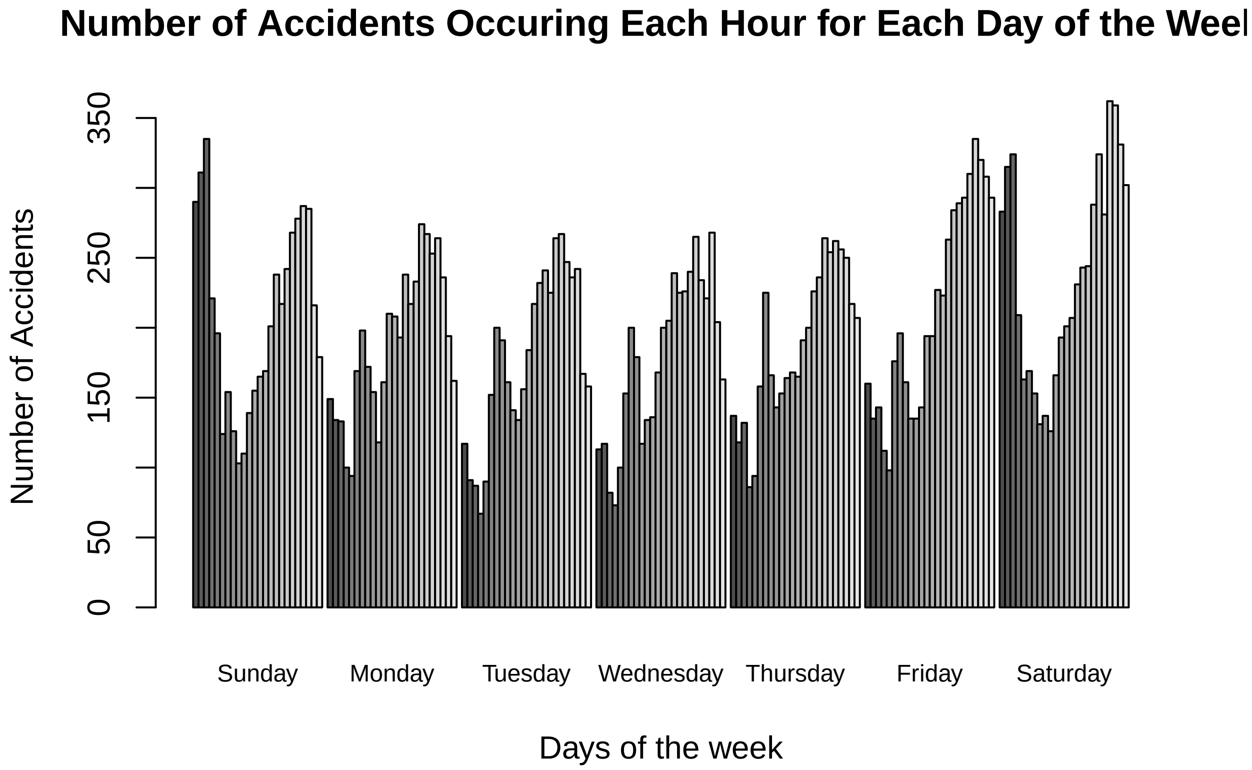
- There is probably a strong correlation between the time of day and crashes. Most miles are driven around rush hour during the week, while the most dangerous miles are driven at night on the Weekends, due to drunk/drugged driving.
- In general, it will be interesting to investigate the most likely times for crashes to occur during the day and year
- We examine the effect of body type, but only in relation to other factors, since the direct effect is asked about in Question 5
- Does the type of road have an impact on fatal crashes? Initial guess is that more lanes would lead to more fatal crashes, because these roads generally have higher speeds
- How does weather affect the US? Generally we expect more crashes in snow/rain/fog, but would be interesting to look at how this effect varies by state (i.e. are Southern drivers significantly worse at driving in snow)
- Curious if there is a relationship between the number of accidents on a given type of road and the time of day
- Expect to see spikes in accidents near holidays

First, read in the data including our external data sources giving information about the total miles of road and state population/area. Next, take a few columns from the Accident table and join with the vehicle table.

```
ACC_df <- read_csv("../FARS_Data/FARS2018NationalCSV/ACCIDENT.csv")
VEH_df <- read_csv("../FARS_Data/FARS2018NationalCSV/VEHICLE.csv")
PER_df <- read_csv("../FARS_Data/FARS2018NationalCSV/PERSON.csv")
miles_of_road <-
  read_excel("../Background_Information/2013 miles of road per state_abr.xlsx")

state_population <-
  read_excel("../Background_Information/2014 state population and total area.xlsx")
a <- select(ACC_df, ST_CASE, DAY_WEEK, RUR_URB, FUNC_SYS)
# Get the day of the week for each accident and road information
VEH_df <- left_join(VEH_df, a, by = c(ST_CASE = "ST_CASE"))
VEH_df <- VEH_df %>% mutate(time_of_day = if_else(HOUR < 12, "Morning", "Afternoon"))
ACC_df <- ACC_df %>% mutate(time_of_day = if_else(HOUR < 12, "Morning", "Afternoon"))
```

Lets look at when during the day and week an accident is most likely to happen.

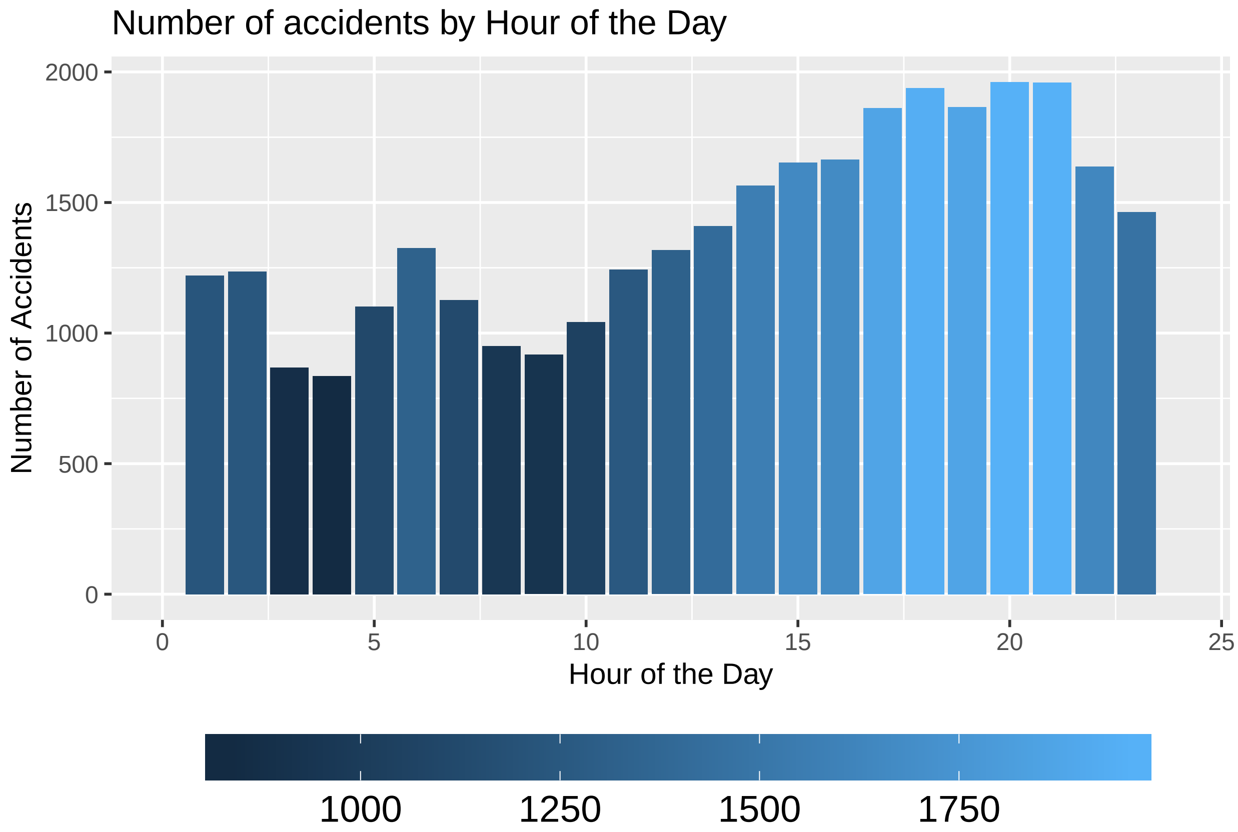
```
ACC_df$DAY_WEEK <- as.character(ACC_df$DAY_WEEK)
Acc_per_hour_day <- ACC_df %>%
  filter(HOUR < 24) %>%
  mutate(DAY_WEEK = fct_recode(DAY_WEEK, Sunday = "1", Monday = "2", Tuesday = "3",
                               Wednesday = "4", Thursday = "5",
                               Friday = "6", Saturday = "7"))
h <- table(Acc_per_hour_day$HOUR, Acc_per_hour_day$DAY_WEEK)
barplot(h, beside = T, cex.names = 0.75,
        ylab = "Number of Accidents",
        xlab = "Days of the week",
        main = "Number of Accidents Occuring Each Hour for Each Day of the Week")
```



For weekdays, there is a spike at 6 AM, followed by a decrease until about 10 AM. After 10 AM, there is a slow increase until rush hour (5 - 7 PM), then a reduction in crashes until the next day at 6 AM. There are large peaks on weekend nights (Friday and Saturday Night) from roughly 8 PM to 2 AM. Our initial hypothesis is that this is strongly correlated with drunk drivers and will be tested shortly.

Next, let's combine all days together and see how accidents change by hour of the day

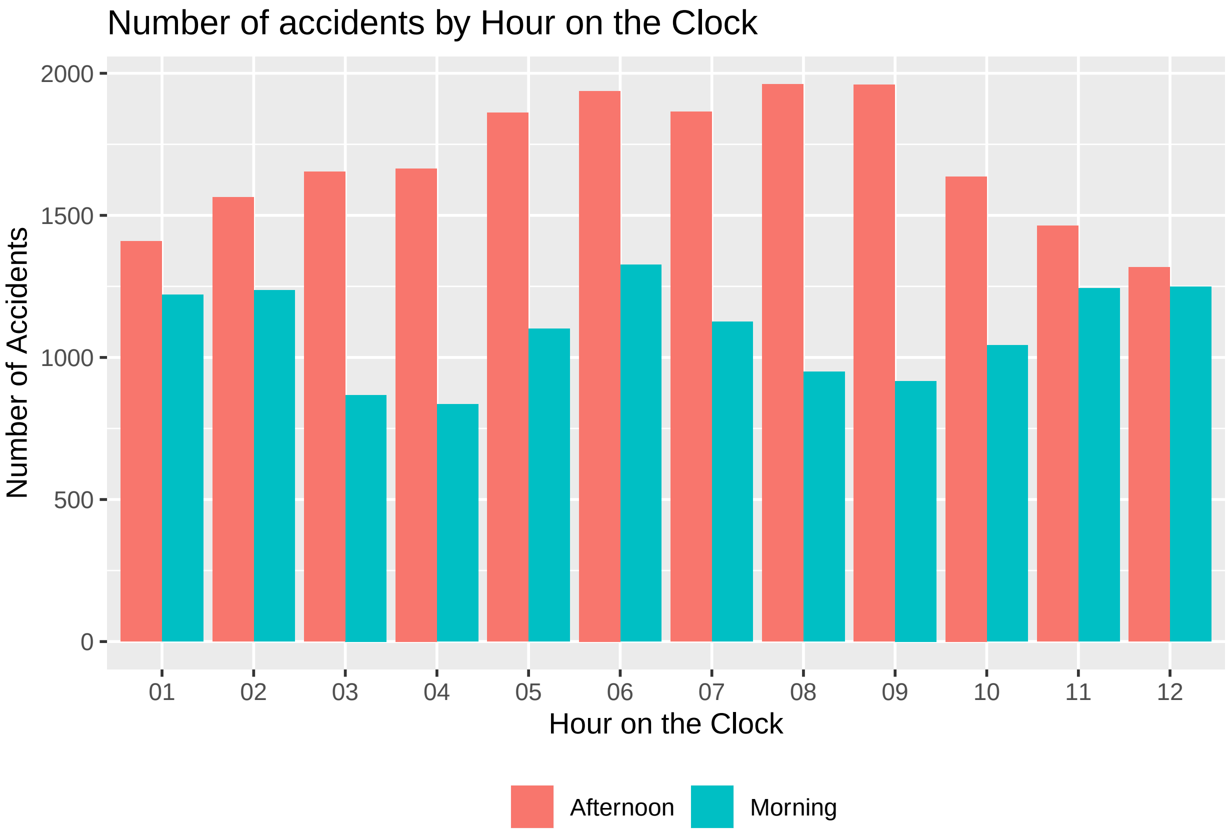
```
ggplot(data = Acc_per_hour_day, aes(x = HOUR, y = ..count.., fill = ..count..)) +
  geom_bar() +
  xlim(c(0, 24)) +
  xlab("Hour of the Day") +
  ylab("Number of Accidents") +
  ggtitle("Number of accidents by Hour of the Day") +
  theme(legend.position="bottom",
        legend.direction="horizontal",
        legend.text = element_text(size=14),
        legend.key.width = unit(2.5, "cm"),
        legend.title = element_blank())
```



Similar to the previous plot, there is a spike around 6 AM, a decrease until 10 AM and then a slow increase the remainder of the day. This plot shows that most accidents occur between 4 and 9 PM. Thus, in general it is more dangerous to drive in rush hour traffic than late night weekend traffic on average.

Another visualization is to see how the times of AM and PM compare to each other.

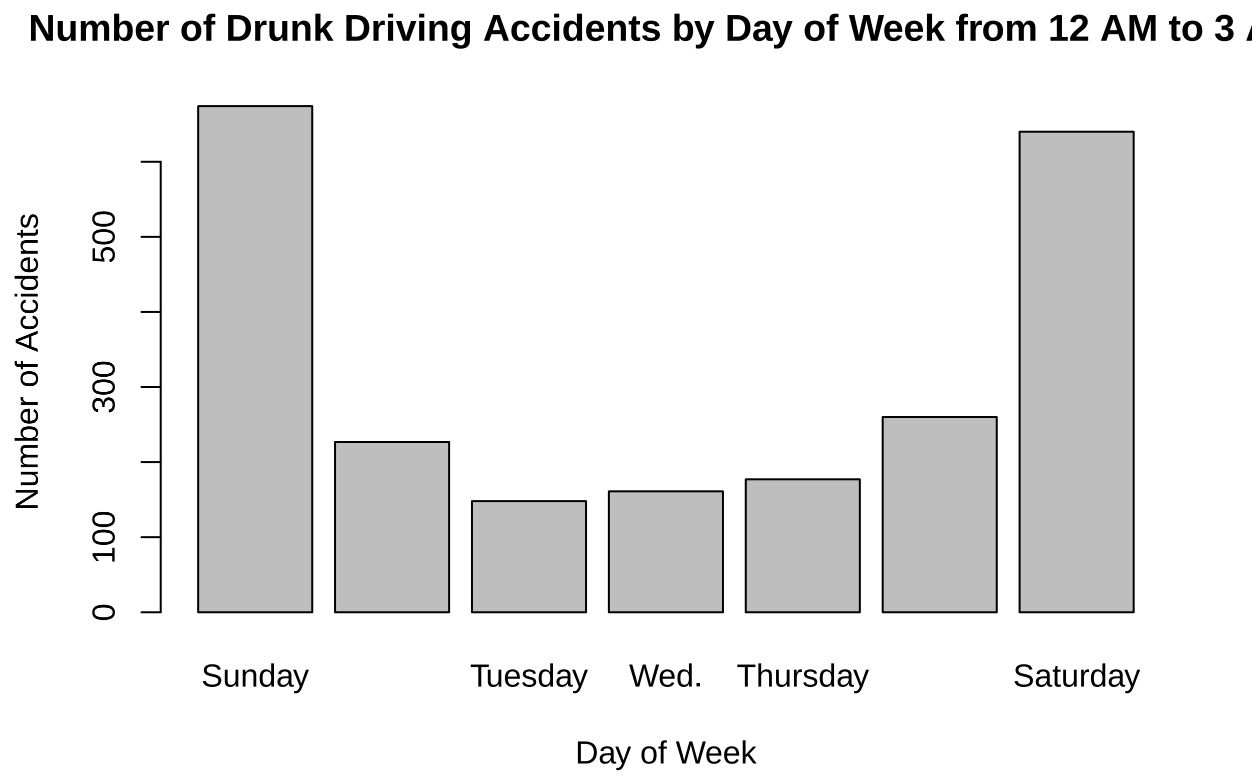
```r
ACC_df$HOUR2 <- format(strptime(ACC_df$HOUR, "%H"), "%I")
hourly_accident_plot <- ACC_df %>% filter(HOUR2 < 13) %>%
  ggplot(aes(x = HOUR2, y = ..count.., shade = time_of_day, fill = ..count..)) +
  geom_bar(mapping = aes(x = HOUR2, y = ..count.., fill = time_of_day),
           position = "dodge") +
  xlab("Hour on the Clock") +
  ylab("Number of Accidents") +
  ggtitle("Number of accidents by Hour on the Clock") +
  theme(legend.position="bottom",
        legend.direction="horizontal",
        legend.title = element_blank())
hourly_accident_plot
```



We find that the afternoon hours (i.e. 12 PM to 12 AM) are more dangerous. Interestingly, 11 AM and 11 PM have roughly the same amount of accidents.

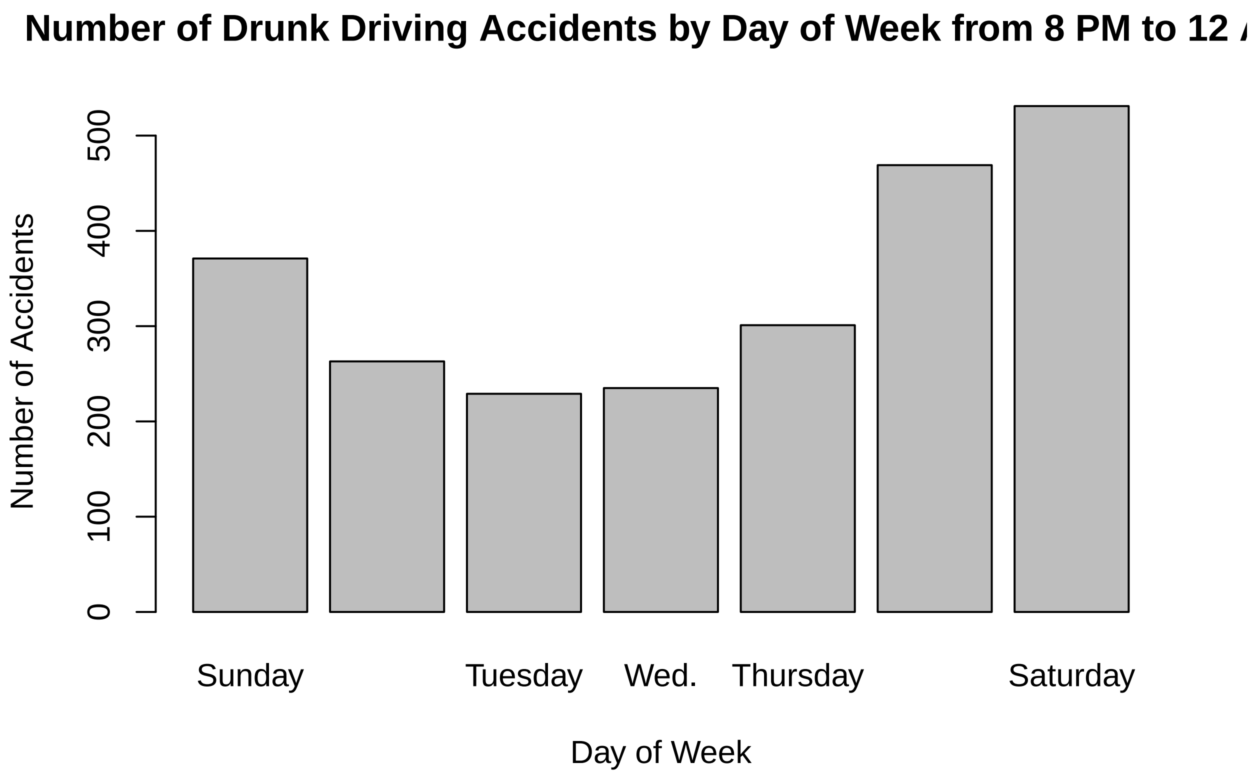
There are large spikes of accidents from 8 PM to 3 AM for Friday Night and Saturday Night. First, let's do a preliminary check and make sure drinking is correlated by only examining accidents from 12 AM to 3 AM.

```
drunk_early_morning<-ACC_df%>%
  select(ST_CASE, HOUR, DAY_WEEK, DRUNK_DR)%>%
  count(DAY_WEEK, HOUR, DRUNK_DR)%>%
  filter(HOUR<= 3, DRUNK_DR>=1)
drunk_0_to_3 <- drunk_early_morning%>%
  group_by(DAY_WEEK)%>%
  mutate(x= sum(n))%>%sample_n(1)%>%
  group_by()%>%
  mutate(Pecent_accident_from_0_to_3 = x/sum(x))
Late_drunk_acc<-aggregate(drunk_early_morning$n,
                          by= list(Category=drunk_early_morning$DAY_WEEK),FUN= sum)
barplot(Late_drunk_acc$x,
        xlab = "Day of Week",
        names = c("Sunday", "Monday", "Tuesday", "Wed.", "Thursday", "Friday", "Saturday"),
        ylab = "Number of Accidents",
        main = "Number of Drunk Driving Accidents by Day of Week from 12 AM to 3 AM")
```



Next, lets look at drinking accidents from 8 PM to 12 AM.

```
drunk_late_night<-ACC_df%>%
  select(ST_CASE, HOUR, DAY_WEEK, DRUNK_DR)%>%
  count(DAY_WEEK, HOUR, DRUNK_DR)%>%
  filter(HOUR>= 20, HOUR < 24, DRUNK_DR>=1)

drunk_20_to_24 <- drunk_late_night%>%
  group_by(DAY_WEEK)%>%
  mutate(x= sum(n))%>%sample_n(1)%>%
  group_by()%>%
  mutate(Pecent_accident_from_20_to_24 = x/sum(x))

Late_drunk_acc_night<-aggregate(drunk_late_night$n,
                                by= list(Category=drunk_late_night$DAY_WEEK),FUN= sum)
barplot(Late_drunk_acc_night$x,
        xlab = "Day of Week",
        names = c("Sunday", "Monday", "Tuesday", "Wed.", "Thursday", "Friday", "Saturday"),
        ylab = "Number of Accidents",
        main = "Number of Drunk Driving Accidents by Day of Week from 8 PM to 12 AM")
```



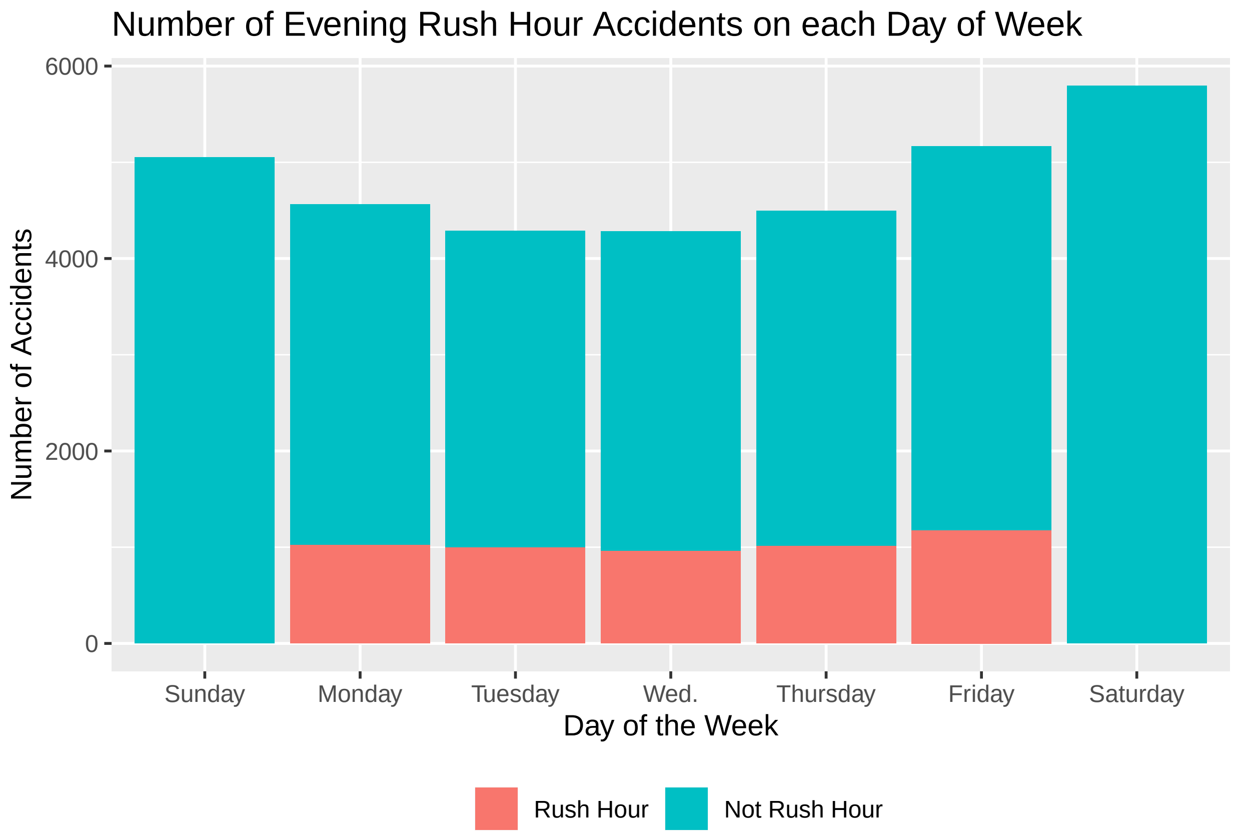
There is an increase in drunk driving accidents from 8 PM to 12 AM on Friday and Saturday night, but it seems that the majority of drunk driving accidents occur after 12 AM as shown by the previous figure.

Next, we examine the effect of the evening rush hour on fatal accidents. This is a time when many people are driving, so it is likely that a large number of accidents occur in this time window. It may be interesting to look at the 6-8 AM rush hour combined with this. Number of accidents during rush hour compared to the number of accidents not during rush hour

```r
Rush_hour_acc<-ACC_df%>%
  select(ST_CASE, HOUR, DAY_WEEK)%>%
  count(DAY_WEEK, HOUR)%>%
  filter(16<=HOUR, HOUR< 20, DAY_WEEK != 1, DAY_WEEK != 7)

rush_hour_accidents <- aggregate(Rush_hour_acc$n,
                                 by= list(Category=Rush_hour_acc$DAY_WEEK),FUN= sum)

Not_rush_hour_acc<-ACC_df%>%
  select(ST_CASE, HOUR, DAY_WEEK)%>%
  count(DAY_WEEK, HOUR)%>%
  subset(HOUR >=20 | HOUR<16 | DAY_WEEK == 1 | DAY_WEEK == 7)

not_rush_hour_accidents <- aggregate(Not_rush_hour_acc$n,
                                     by= list(Category=Not_rush_hour_acc$DAY_WEEK),
                                     FUN= sum)

rush_hour_acc_percent <- sum(rush_hour_accidents[,2]) /
  (sum(rush_hour_accidents[,2]) + sum(not_rush_hour_accidents[,2]))

print(sprintf("Percentage of Accidents that Occur during Rush Hour is %s%%",
              round(rush_hour_acc_percent*100, digits = 3)))
```

```
[1] "Percentage of Accidents that Occur during Rush Hour is 15.413%"
```

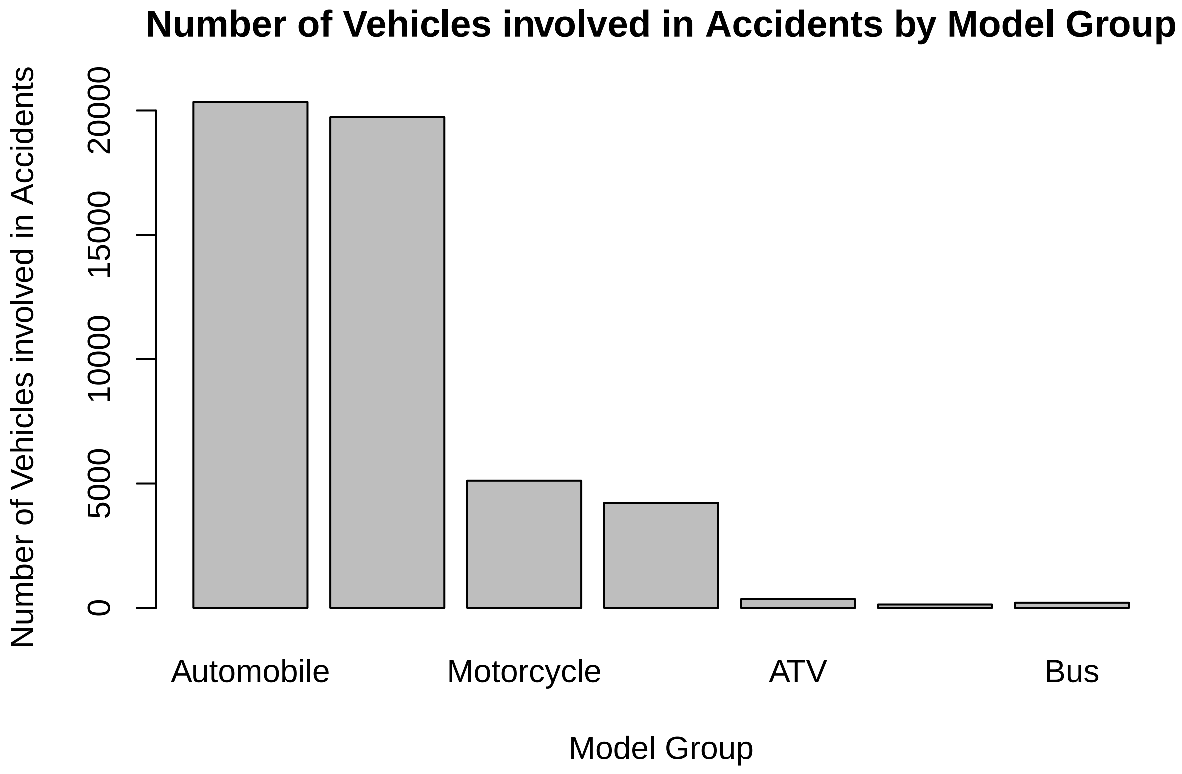
```
rush_hour_accidents = cbind(rush_hour_accidents, "Rush Hour")
colnames(rush_hour_accidents) <- c("Day", "Number", "Category")
not_rush_hour_accidents = cbind(not_rush_hour_accidents, "Not Rush Hour")
colnames(not_rush_hour_accidents) <- c("Day", "Number", "Category")

all_accidents = rbind(rush_hour_accidents, not_rush_hour_accidents)

ggplot(all_accidents, aes(fill=all_accidents$Category,
                          y=all_accidents$Number,
                          x=all_accidents$Day)) +
  geom_bar(position=position_stack(reverse = TRUE), stat="identity") +
  xlab("Day of the Week") +
  ylab("Number of Accidents") +
  ggtitle("Number of Evening Rush Hour Accidents on each Day of Week") +
  scale_x_discrete(breaks=1:7,
                   labels=c("Sunday", "Monday", "Tuesday", "Wed.",
                            "Thursday", "Friday", "Saturday")) +
  theme(legend.position="bottom",
        legend.direction="horizontal",
        legend.title = element_blank())
```



Overall, evening rush hour accidents comprise 15% of accidents, while only being 12% of total time during the week.

Next, let's look at how the type of vehicle is related to crash patterns.

```r
Truck <- VEH_df %>% count(MODEL) %>% filter(MODEL > 400, MODEL < 500)
Truck <- sum(Truck$n)

Automobile <- VEH_df %>% count(MODEL) %>% filter(MODEL < 400)
Automobile <- sum(Automobile$n)

Motorcycles <- VEH_df %>% count(MODEL) %>% filter(MODEL > 700, MODEL < 710)
Motorcycles <- sum(Motorcycles$n)

Heavy_Truck <- VEH_df %>% count(MODEL) %>% filter(MODEL > 880, MODEL < 900)
Heavy_Truck <- sum(Heavy_Truck$n)

ATV <- VEH_df %>% count(MODEL) %>% filter(MODEL > 730, MODEL < 740)
ATV <- sum(ATV$n)

MotorHome_Van <- VEH_df %>% count(MODEL) %>% filter(MODEL > 849, MODEL < 871)
MotorHome_Van <- sum(MotorHome_Van$n)

Bus <- VEH_df %>% count(MODEL) %>% filter(MODEL > 900, MODEL < 990)
Bus <- sum(Bus$n)

k <- rbind(Automobile, Truck, Motorcycles, Heavy_Truck, ATV, MotorHome_Van, Bus)
k <- as.data.frame(k)
barplot(k$V1,
        xlab = "Model Group",
        names = c("Automobile", "Truck", "Motorcycle",
                  "Heavy Truck", "ATV", "MotorHome/Van", "Bus"),
        ylab = "Number of Vehicles involved in Accidents",
        main = "Number of Vehicles involved in Accidents by Model Group")
```



Next, lets look at the road type and how that is relatead to crashes. Need to bring in the FUNC SYS and RURURB tables to identify what these are. First, we examine the frequency of each attribute in the FARS dictionary for both FUNCSYS and RURURB

FUNC SYS Codes	Attributes
1	Interstate
2	Principal Arterial - Other Freeways and Expressways
3	Prinicipal Arterial - Other
4	Minor Arterial
5	Major Collector
6	Minor Collector
7	Local
96	Trafficway not in State Inventory
98	Not Reported
99	Unknown

```
kable(ACC_df %>% count(FUNC_SYS) %>% rename(count = n), "latex",
    booktabs = T)
```

FUNC_SYS	count
1	4324
2	1418
3	10144
4	7069
5	4694
6	1061
7	4314
96	61
98	551
99	18

From the data, we see that most accidents occur on arterial roadways. For future analysis, we can group these into 1) interstate 2) arterial 3) collector 4) local 5) Unknown

RUR_URB Codes	Attributes
1	Rural
2	Urban
6	Trafficway not in State Inventory
8	Not Reported
9	Unknown

```
kable(ACC_df %>% count(RUR_URB) %>% rename(count = n), "latex",
    booktabs = T)
```

RUR_URB	count
1	14760
2	18285
6	61
8	526
9	22

Surprisingly, there is only a slight increase in the number of accidents on urban roads. Our intuition was that more accidents would happen on urban roads at higher speeds. Finally, we can look at the type of road by signage. This is given in the Route column of the Accident Table

ROUTE Codes	Attributes
1	Interstate
2	US Highway
3	State Highway
4	County Road
5	Local Street - Township
6	Local Street - Municipality
7	Local Street - Frontage Road
8	Other
9	Unknown

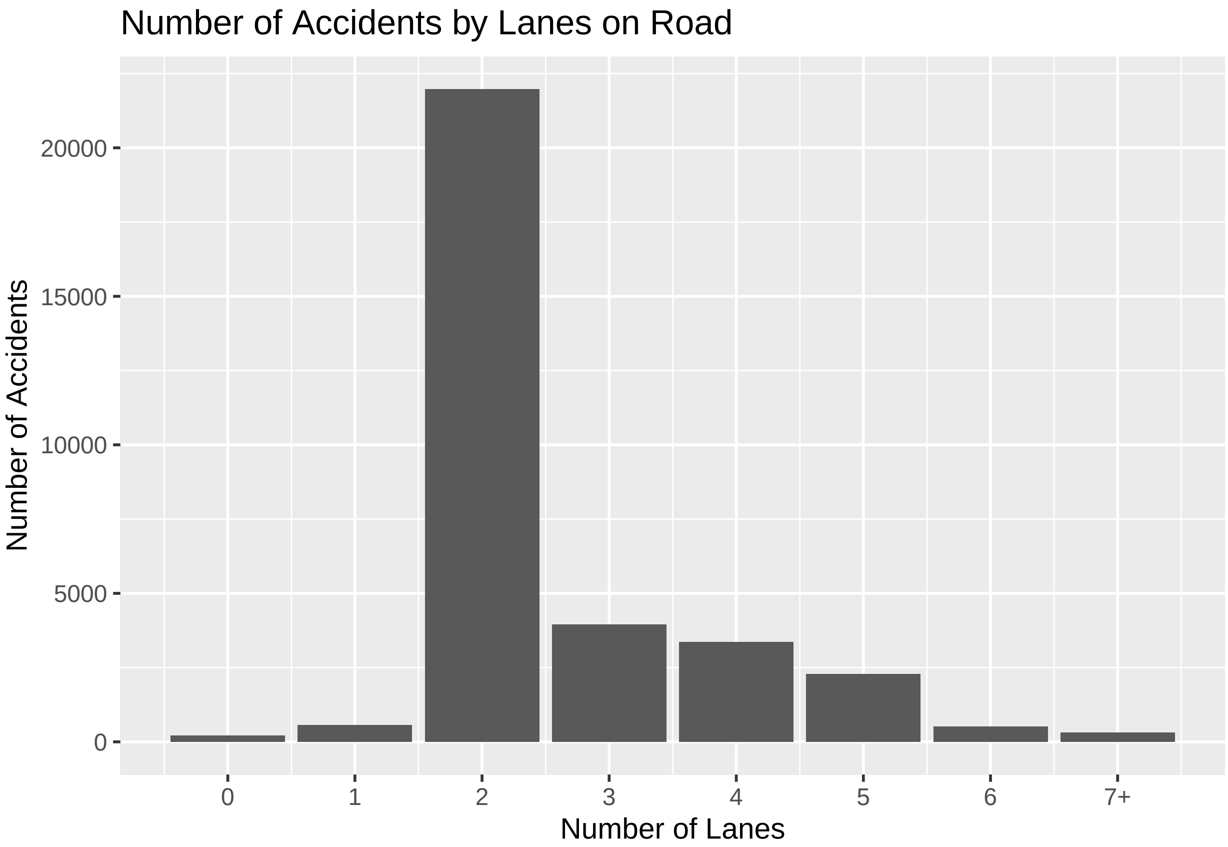
```r
kable(ACC_df %>% count(ROUTE) %>% rename(count = n), "latex",
    booktabs = T)
```

ROUTE	count
1	4217
2	5555
3	9890
4	4325
5	804
6	5090
7	297
8	1473
9	2003

Interestingly, most accidents happen on state highways by far. The next most common are US highways and local streets

Next, lets look at how crashes are related to the number of lanes on a road. From the FARS Data, 0 is non-trafficway/driveway, 8 is not reported and 9 is unknown. All others are the number of lanes for the road

```r
accident_from_veh_df <- VEH_df[!duplicated(VEH_df$ST_CASE),]

accident_from_veh_df %>% select(VNUM_LAN, VPROFILE) %>% filter(VNUM_LAN < 8) %>%
  ggplot(aes(x=VNUM_LAN, y=..count.. ))+ geom_bar() +
  xlab("Number of Lanes") +
  ylab("Number of Accidents") +
  ggtitle("Number of Accidents by Lanes on Road") +
  scale_x_continuous(breaks=0:7, labels=c("0", "1", "2", "3", "4", "5", "6", "7+")) +
  theme(legend.position="bottom",
        legend.direction="horizontal",
        legend.title = element_blank())
```



```
# Majority of roads are flat in the united states but it
# is interesting to note many accidents occur on a slope,
```

From this frequency plot, we can see that fatal accidents happen most frequently on 2-lane roads. One callout here is that in the FARS dictionary, a 2 lane road with a dedicated turn lane at the time of the accident is classified as a 3-lane road. So it is very likely that many of the accidents on 3 lane roads are on 2-lane roads with a turn lane. Similarly, some of the 2-lane accidents could be on 1-lane roads with a dedicated turn lane at the time of the crash.

Next steps

Next, lets examine the relationship between weather and crashes. First lets examine the frequency of each weather type occuring. The relevant codes are below:

WEATHER Codes	Attributes
00	No Additional Atmospheric Conditions
01	Clear
02	Rain
03	Sleet or Hail
04	Snow
05	Fog, Smog, Smoke
06	Severe Crosswinds
07	Blowing Sand, Soil, Dirt
08	Other
10	Cloudy
11	Blowing Snow
12	Freezing Rain or Drizzle
98	Not Reported

WEATHER Codes	Attributes
99	Reported as Unknown

```
kable(ACC_df %>% count(WEATHER) %>% rename(count = n), "latex",
    booktabs = T)
```

WEATHER	count
1	22263
2	2755
3	64
4	391
5	364
6	53
7	10
8	40
10	4896
11	32
12	25
98	2525
99	236

Most weather conditions for fatal crashes were either clear or cloudy. Note, the Data dictionary explicitly says not to assume any additional weather pattern when cloudy is recorded unless mentioned. Out of the weather conditions that generally are thought to negatively affect driver performance, rain was the most common, followed by snow and fog/smog/smoke. Lets exmaine the percentage of accidents where weather was involved.

```
weather_related_accidents <-
  ACC_df %>%
  count(WEATHER) %>%
  filter(WEATHER > 1, WEATHER != 10, WEATHER < 98)
print(sprintf("Percentage of fatal crashes where inclement weather was involved was %s%%",
              round( sum(weather_related_accidents$n)*100/nrow(ACC_df), digits = 3)))
```

```
[1] "Percentage of fatal crashes where inclement weather was involved was 11.095%"
```

The weather and amount of light outside may be related. We can use the LGT_COND column which gives information on if the road was lit. For this, we assume that in both dawn and dusk the road is not adequately lit, since the data dictionary denotes these times as before the sun rises and after the sun sets respectively.

```
light_and_weather_accidents <- ACC_df %>% filter(WEATHER > 1,
    WEATHER != 10, WEATHER < 98, LGT_COND != 2, LGT_COND < 4)

dark_and_weather_accidents <- ACC_df %>% filter(WEATHER > 1,
    WEATHER != 10, WEATHER < 98) %>% filter(LGT_COND != 1, LGT_COND !=
    3, LGT_COND < 6)

print(sprintf("Percentage of fatal crashes with light and inclement weather was involved: %s%%",
    round(nrow(light_and_weather_accidents) * 100/nrow(ACC_df),
        digits = 4)))
```

```
[1] "Percentage of fatal crashes with light and inclement weather was involved: 6.9561%"
```

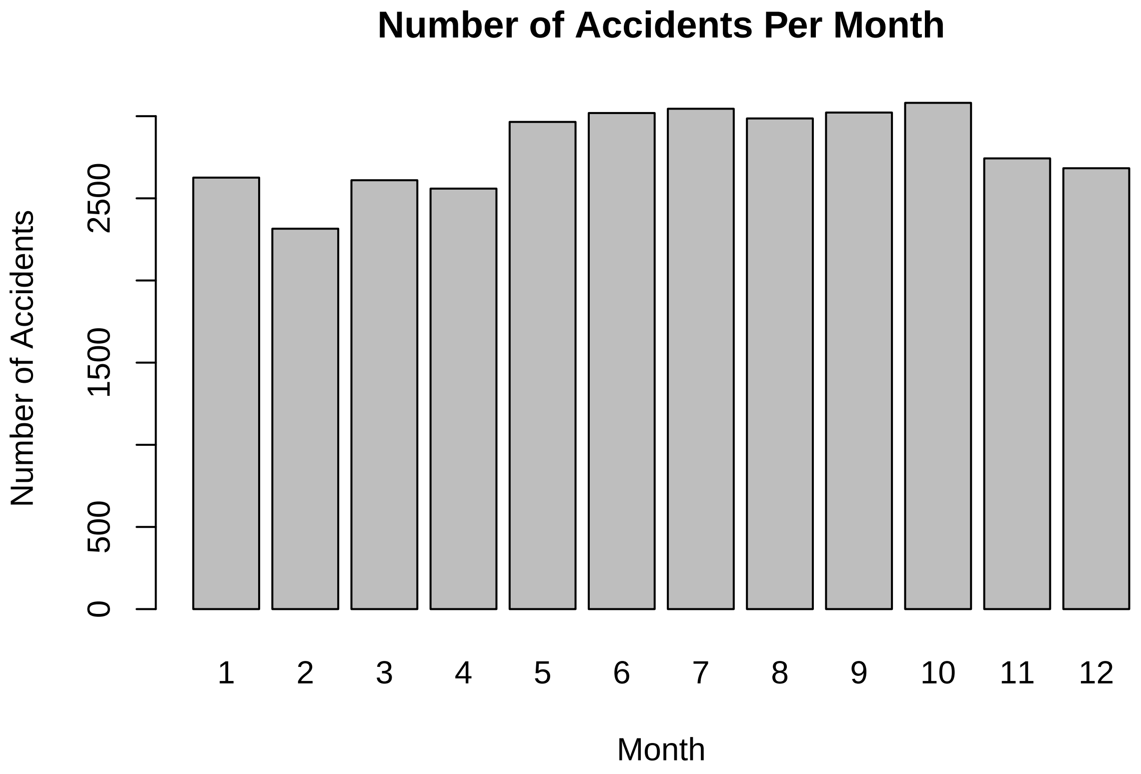
```
print(sprintf("Percentage of fatal crashes with dark and inclement weather was involved: %s%%",
    round(nrow(dark_and_weather_accidents) * 100/nrow(ACC_df),
        digits = 4)))
```

```
[1] "Percentage of fatal crashes with dark and inclement weather was involved: 3.9995%"
```

We find that weather related accidents are almost twice as frequent when there is sufficient lighting outside compared to when it is dark outside.

Next we investigate yearly trends in fatal accident data. First we examine if there are certain months which are more or less likely to have fatal accidents.

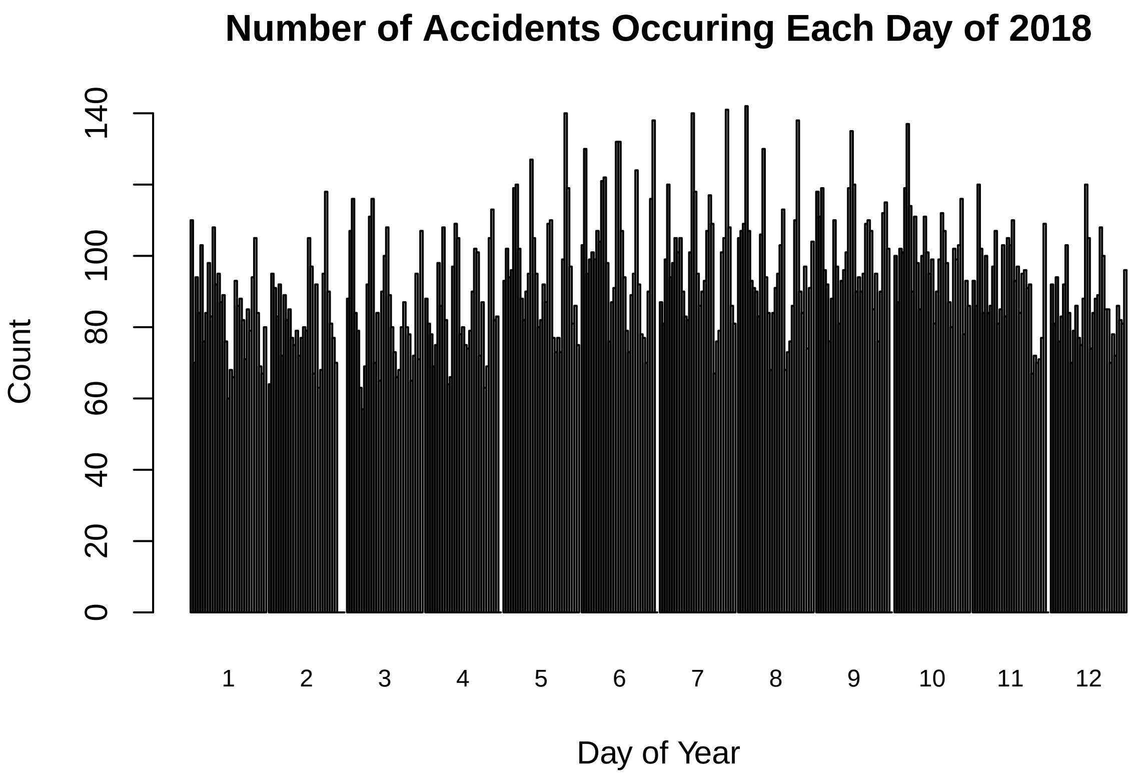
```
acc_per_month = ACC_df %>% count(MONTH)
barplot(acc_per_month$n,
        xlab = "Month",
        names = seq(1,12,1),
        ylab = "Number of Accidents",
        main = "Number of Accidents Per Month")
```



In general, there are more accidents in the summer compared to the winter.

Next lets look and see if there are any daily trends throughout the year

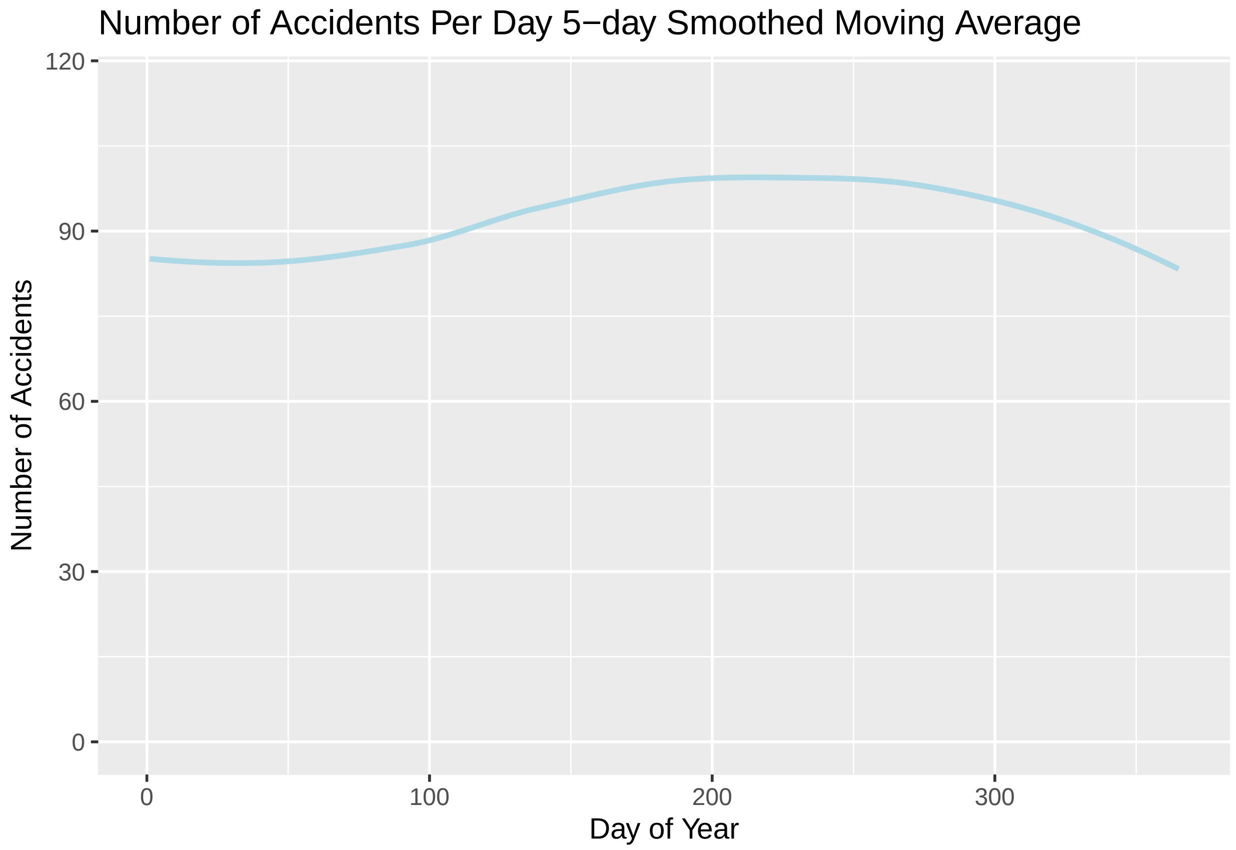
```
month_day_accidents <- table(ACC_df$DAY, ACC_df$MONTH)
barplot(month_day_accidents, beside = T, cex.names = 0.75,
        ylab = "Count",
        xlab = "Day of Year",
        main = "Number of Accidents Occuring Each Day of 2018")
```



The above plot is quite noisy, so lets try plotting a 5-day moving average for the year. The main takeaway is that there are not many days that fall outside of the weekly average for that day of the week. There are consistent spikes 7 days apart, showing the large number of accidents that occur on Saturday more generally. Interestingly, there were was a spike for New Years Day, but no spike for traveling at Thanksgiving or Christmas, which is a common time for many people to travel.

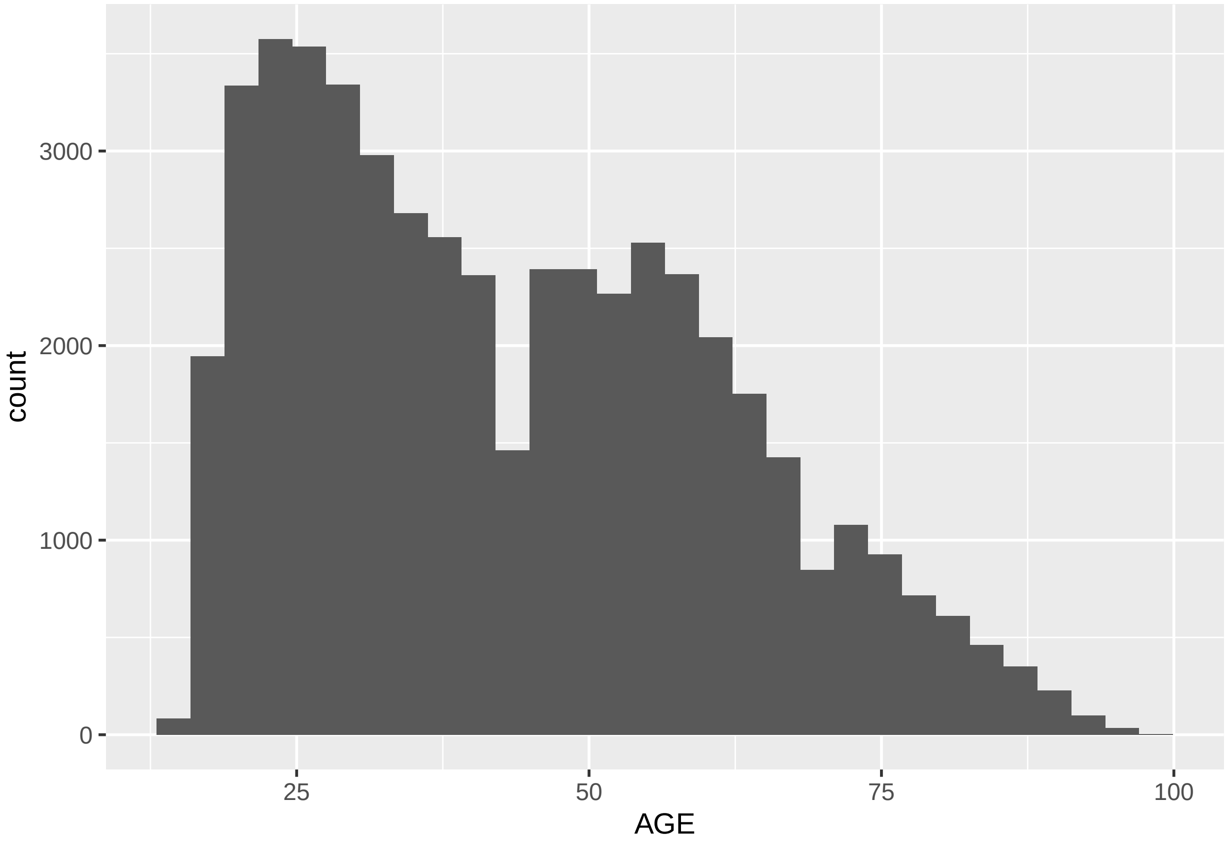
```
month_day_acc_long <- as.data.frame(month_day_accidents) %>% filter(Freq > 0)
month_day_mov_aver = as.data.frame(movavg(month_day_acc_long$Freq, 5, 's'))
month_day_mov_aver = cbind(seq(1,365,1), month_day_mov_aver)
colnames(month_day_mov_aver) = c("Day", "Mov_Aver")
ggplot(data = month_day_mov_aver,aes(x=Day, y=Mov_Aver))+
  geom_smooth(color="lightblue",se=F)+
  xlab("Day of Year")+
  ylab("Number of Accidents") +
  ggtitle("Number of Accidents Per Day 5-day Smoothed Moving Average") +
  ylim(c(0,115))
```

```
`geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



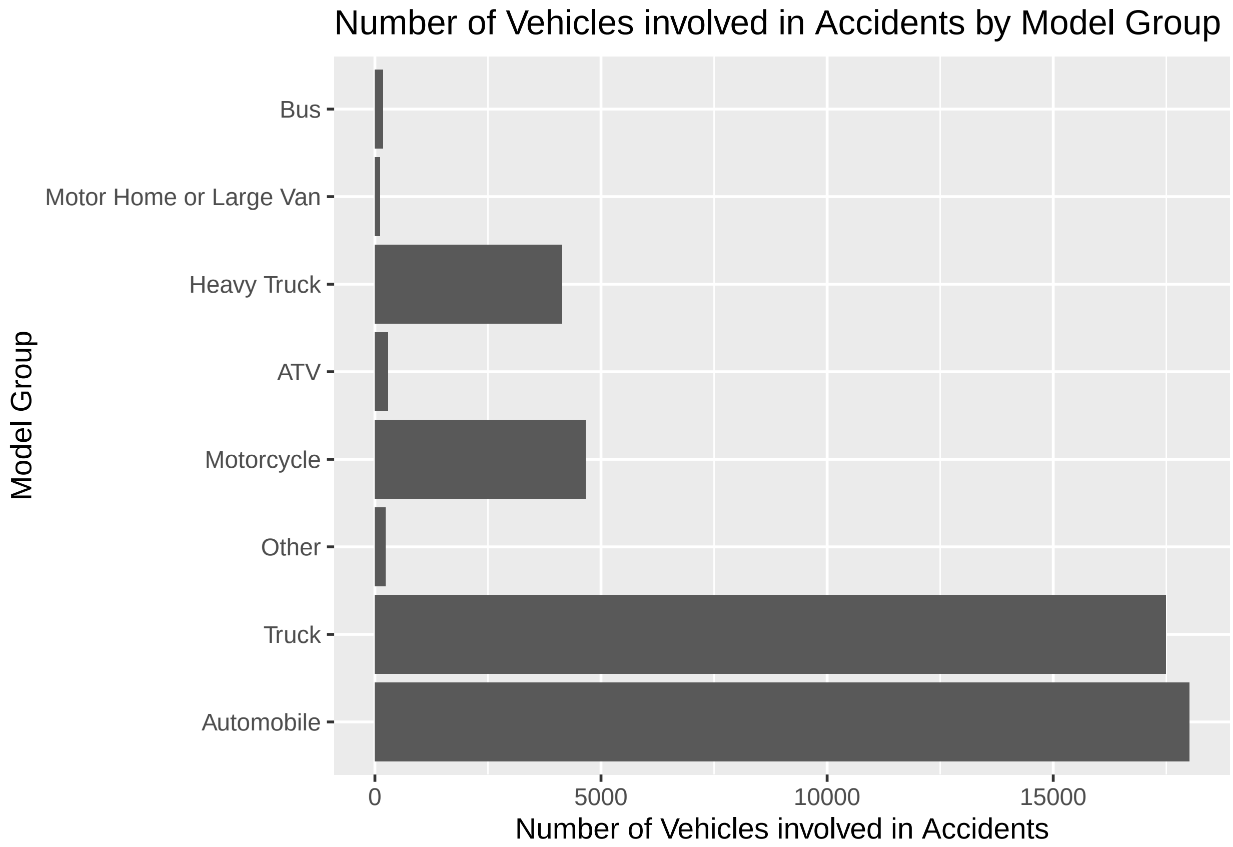
Overall, this plot reinforces that the monthly trends are stronger than individual dates where people tend to travel. My intuition was that holidays would have large spikes, but the overall seasonal and day of week trends clearly dominate the pattern of fatal crashes.

```
Driver_age<-PER_df%>%
  filter(SEAT_POS==11)%>%
  select(ST_CASE,AGE,SEX)%>%
  filter(AGE<100, AGE>14)
Driver_age%>%
  ggplot(aes(x=AGE))+geom_histogram()
```

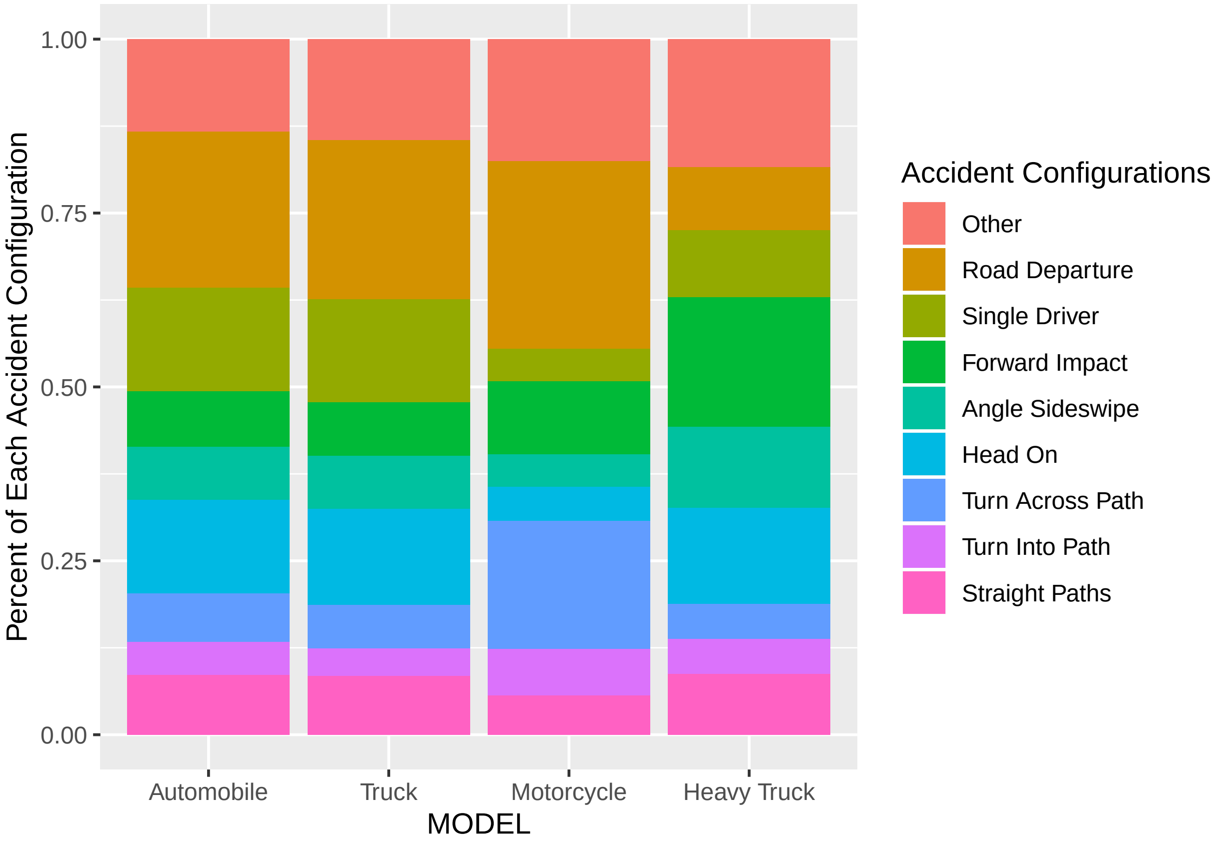


```
#The figure above show at what ages there are the highest number of fatal accidents.

Driver_age_violation2<-left_join(VEH_df,Driver_age, by=c("ST_CASE"="ST_CASE"))%>%
  filter(AGE<100, AGE>14)%>%
  mutate(ST_CASE=paste(ST_CASE,".", VEH_NO))%>%
  group_by(ST_CASE)%>%
  filter(row_number(ST_CASE)<=1)%>%
  filter(PREV_ACC<9)%>%
  mutate(PREV_Violation= PREV_DWI+PREV_ACC+ PREV_DWI+PREV_OTH+
           PREV_SPD+PREV_SUS1+ PREV_SUS2+PREV_SUS3)%>%
  mutate(Date= as.Date( paste(2018, MONTH,DAY,sep="-"), "%Y-%m-%d"),
         LastDate= as.Date( paste(LAST_YR, LAST_MO, 1,sep="-"), "%Y-%m-%d"),
         time_since_last= Date- LastDate,
         time_since_last= as.character(time_since_last),
         time_since_last= parse_double(time_since_last))%>%
  mutate(FirstDate= as.Date(paste(FIRST_YR, FIRST_MO, 1, sep = "-"), "%Y-%m-%d"),
         time_btw_First_last= LastDate- FirstDate,
         num_days_perAlt_before= time_btw_First_last/PREV_Violation,
         num_days_perAlt_before= round(num_days_perAlt_before),
         num_days_perAlt_before= as.character(num_days_perAlt_before),
         num_days_perAlt_before= parse_number(num_days_perAlt_before))
```

```
Driver_Configuration<-Driver_age_violation2%>%
  group_by()%>%
  mutate(ACC_TYPE=as.character(ACC_TYPE),
         ACC_Configurations = fct_collapse(ACC_TYPE,
              "Other"="0",
              "Road Departure"= c("1",2:10),
              "Single Driver"= c(11:16),
              "Forward Impact"= c(20:33),
              "Forward Impact"=c(38:43),
              "Angle Sideswipe"= c(42:49),
              "Head On"=c(50:53),
              "Forward Impact"=c(54:62),
              "Angle Sideswipe"= c(64:67),
              "Turn Across Path"=c(68:75),
              "Turn Into Path"=c(76:85),
              "Straight Paths"= c(86:91),
              "Other"= c(92:93),
              "Other"= c(98:99)))%>%
  mutate(AGE= as.character(AGE),
         age= fct_collapse(AGE,
              "15-25" = c("15","16","17","18","19","20","21","22","23","24","25"),
              "26-35"= c(26:35),
              "36-45"= c(36:45),
              "46-55"= c(46:55),
              "56-65"= c(56:65),
              "66-100"= c(66:100)
              ))

Driver_Configurations_Model<-Driver_Configuration%>%
  mutate(MODEL= as.character(MODEL),
         MODEL= fct_collapse(MODEL,
                "Automobile"= c("1",2:400),
                "Truck"= c(401:500),
                "Motorcycle"= c(700:710),
                "Heavy Truck"=c(800:809,880:900),
                "ATV"=c(730:740),
                "Motor Home or Large Van"=c(849:871),
                "Bus"=c(900:990),
                "Other"=c(598,799,997,998,999)))

Driver_Configurations_Model%>%
  count(MODEL)%>%
  ggplot(aes(x= MODEL, y= n))+
  geom_col()+coord_flip()+
  xlab("Model Group")+
  ylab("Number of Vehicles involved in Accidents")+
  ggtitle("Number of Vehicles involved in Accidents by Model Group")
```



```
Driver_Configurations_Model%>%
group_by(MODEL)%>%
  count(MODEL, ACC_Configurations)%>%
  mutate(x=sum(n),
         y= n/x,
         z=sum(y))%>%
  filter(x>500)%>% #filter out vehichles involved in less than 1 percent of accidents.
  ggplot(aes(x= MODEL, y= y, fill= ACC_Configurations))+
  geom_col(position = "fill")+
  ylab("Percent of Each Accident Configuration")+
  theme(legend.position = "right")+
  guides(fill= guide_legend(title = "Accident Configurations"))
```



Trucks and Automobile have similar values for each crash type with little variation. Motorcycles have significantly larger number of accidentsmnmnm that happen when turning across path likely due to not looking for motorcyles. Motorcycles also have a higher number of turn into path so people are not looking for them when they are merging. Large trucks have a higher number of rear ends likely to due to more blind spots and not being able to see every obstacle.

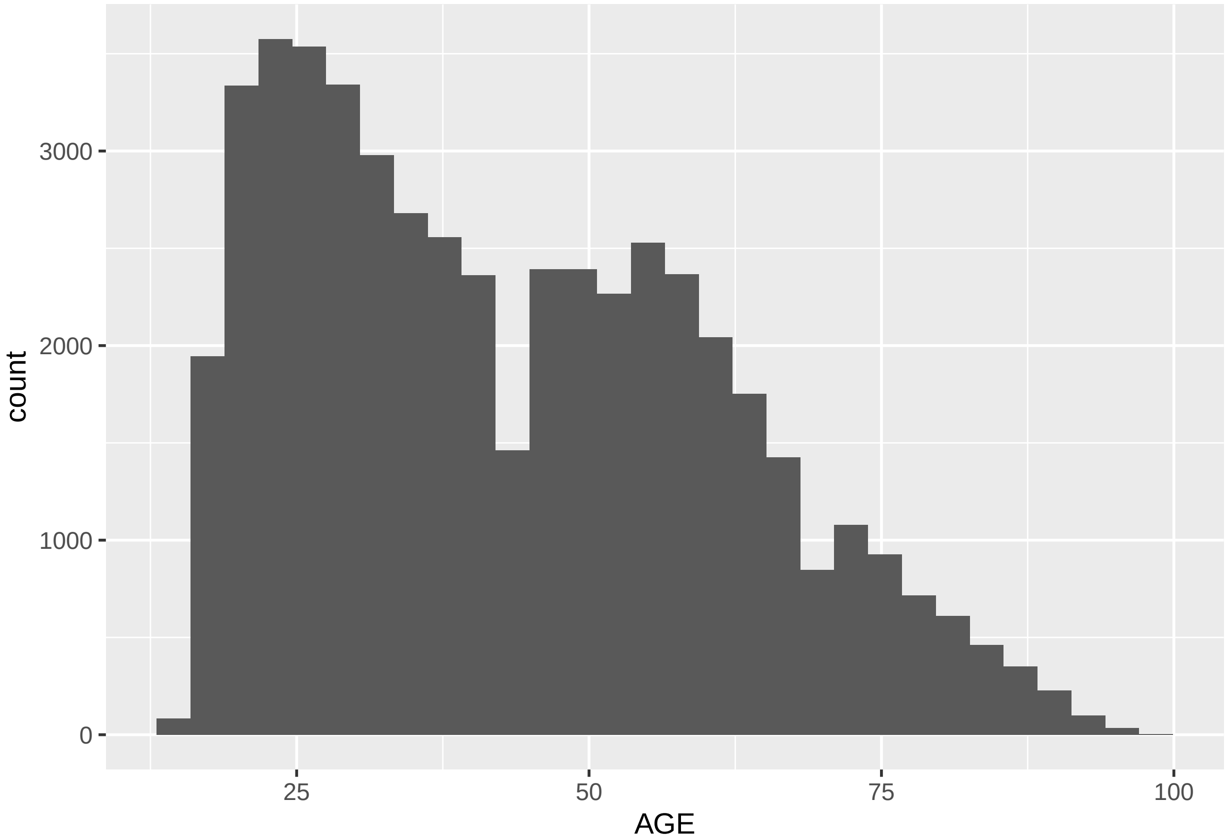
Appendix 4: Question 4 (Driver Profiling)

```
ACC_df <- read_csv("../FARS_Data/FARS2018NationalCSV/ACCIDENT.csv")
VEH_df <- read_csv("../FARS_Data/FARS2018NationalCSV/VEHICLE.csv")
PER_df <- read.csv("../FARS_Data/FARS2018NationalCSV/PERSON.csv")
miles_of_road <-
  read_excel("../Background_Information/2013 miles of road per state.xlsx")
state_population <-
  read_excel("../Background_Information/2014 state population and total area.xlsx")
a<-select(ACC_df, ST_CASE, DAY_WEEK, RUR_URB, FUNC_SYS)

VEH_df<-left_join(VEH_df,a, by= c("ST_CASE"="ST_CASE"))
VEH_df<-VEH_df%>%
  mutate(time_of_day= if_else(HOUR<12, "Morning", "Afternoon"))
```

Initial Thougths

- The time between first and last offense divided by the amount of previous altercations could show be people that are prone to getting in accidents
- The time between last and time of the accident could show who is higher risk
- Drivers that driver is drinking and driver is speeding is considered risky behavior
- Obese drivers could be in more accidents due to health conditions
- Men are risky drivers than women

```
Driver_age<-PER_df%>%
  filter(SEAT_POS==11)%>%# seat 11 is identified as the driver in each accident
  select(ST_CASE,AGE,SEX)%>%
  filter(AGE<100, AGE>14)
Driver_age%>%
  ggplot(aes(x=AGE))+geom_histogram()
```

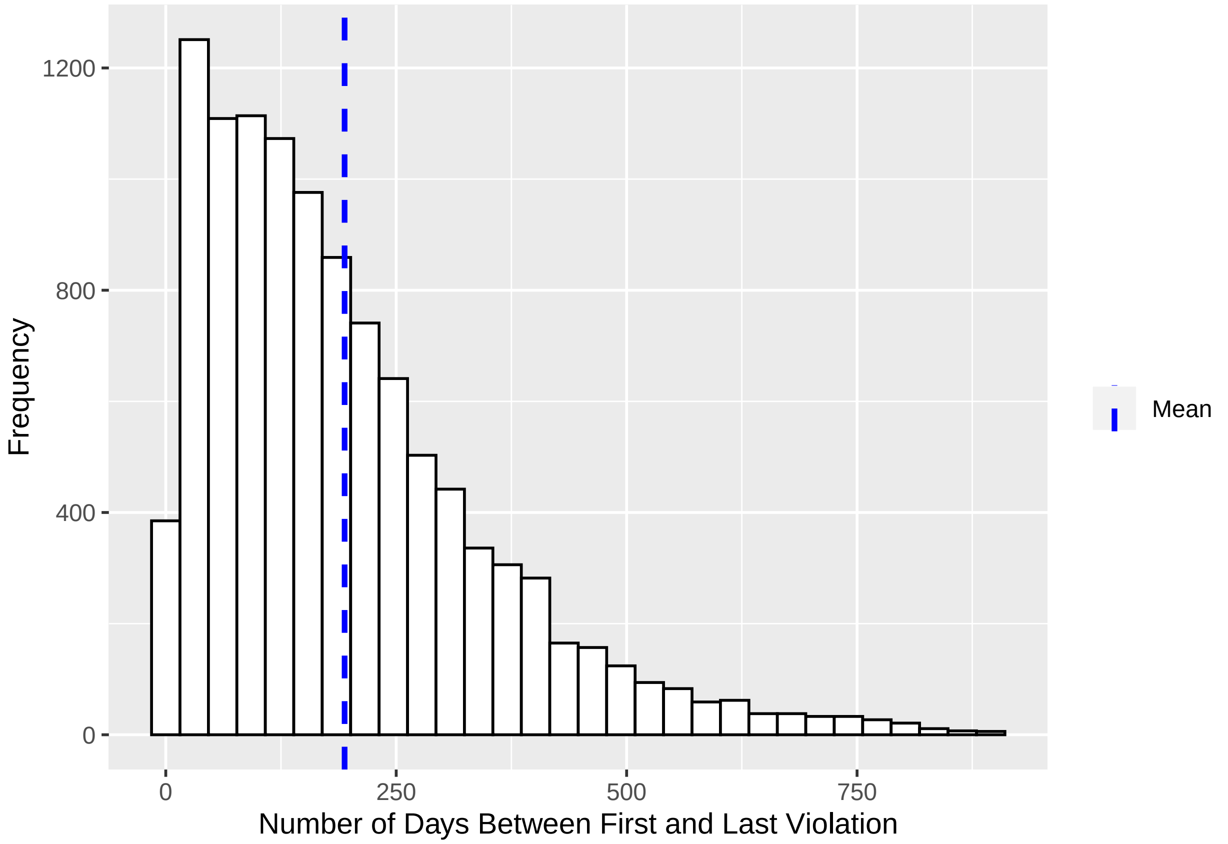


The figure above show at what ages there are the highest number of fatal accidents.

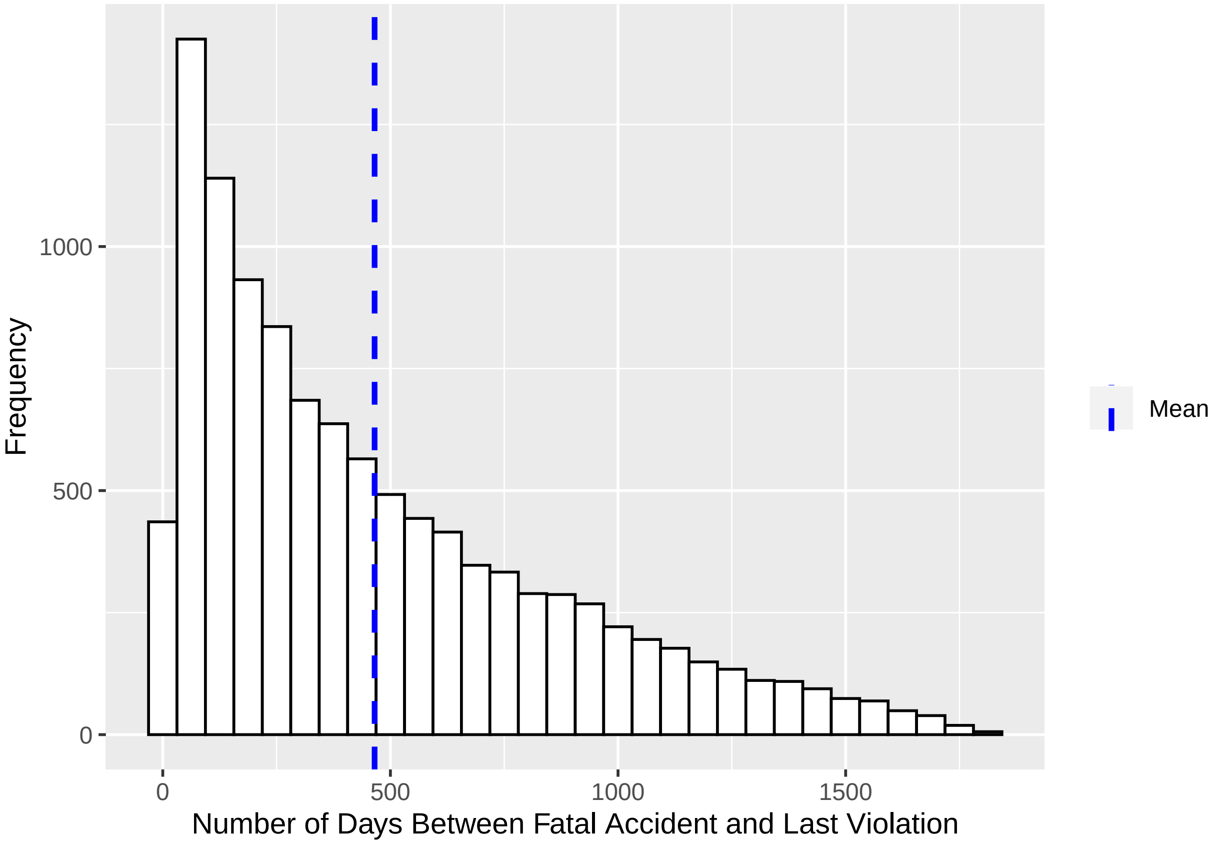
```
Driver_age_violation2<-left_join(VEH_df,Driver_age, by=c("ST_CASE"="ST_CASE"))%>%
  filter(AGE<100, AGE>14)%>%
  mutate(ST_CASE=paste(ST_CASE,".", VEH_NO))%>%
  group_by(ST_CASE)%>%
  filter(row_number(ST_CASE)<=1)%>%
  filter(PREV_ACC<9)%>%
  mutate(PREV_Violation= PREV_DWI+PREV_ACC+ PREV_DWI+PREV_OTH+
           PREV_SPD+PREV_SUS1+ PREV_SUS2+PREV_SUS3)%>%
  mutate(Date= as.Date( paste(2018, MONTH,DAY,sep="-"), "%Y-%m-%d"),
         LastDate= as.Date( paste(LAST_YR, LAST_MO, 1,sep="-"), "%Y-%m-%d"),
         time_since_last= Date- LastDate,
         time_since_last= as.character(time_since_last),
         time_since_last= parse_double(time_since_last))%>%
  mutate(FirstDate= as.Date(paste(FIRST_YR, FIRST_MO, 1, sep = "-"), "%Y-%m-%d"),
         time_btw_First_last= LastDate- FirstDate,
         num_days_perAlt_before= time_btw_First_last/PREV_Violation,
         num_days_perAlt_before= round(num_days_perAlt_before),
         num_days_perAlt_before= as.character(num_days_perAlt_before),
         num_days_perAlt_before= parse_number(num_days_perAlt_before))
Driver_age_violation<- Driver_age_violation2%>%
  select(AGE, DEATHS, PREV_Violation, time_since_last,
         num_days_perAlt_before, PREV_ACC)%>%
  filter(PREV_Violation>0)%>%
  filter(num_days_perAlt_before>0)
# Driver_age_violation
# Removed drivers with not previous record because we cannot
# determine if they are at high risk when evaluating their driving record
```

Days between first and last (time_btw_First_last) and Days since last and current accident (time_since_last) is calculated. Than, previous accidents, DWI, Speeding, Suspension, other are summed. This creates a previous violation variable that we used to calculate time per violation (num_days_perAlt_before). Showing how often a person is getting in violation.

```
Driver_age_violation%>%
  filter(num_days_perAlt_before>1)%>%
  ggplot(aes(x= num_days_perAlt_before))+
  geom_histogram(color="black",fill="white")+
  geom_vline(aes(xintercept=mean(num_days_perAlt_before),
                 color="blue"), linetype="dashed", size=1)+
  theme(legend.box = "right")+
  scale_color_identity(name = " ", breaks = c("blue"),
                       labels = c("Mean"), guide = "legend")+
  xlab("Number of Days Between First and Last Violation")+
  ylab("Frequency")
```



```
Driver_age_violation%>%
  ggplot(aes(x= time_since_last))+
  geom_histogram(color="black", fill="white")+
  geom_vline(aes(xintercept=mean(time_since_last),
                 color="blue"), linetype="dashed", size=1)+
  theme(legend.box = "right")+
  scale_color_identity(name = " ", breaks = c("blue"),
                       labels = c("Mean"), guide = "legend")+
  xlab("Number of Days Between Fatal Accident and Last Violation")+
  ylab("Frequency")
```



```
mean(Driver_age_violation$num_days_perAlt_before)
```

```
[1] 194.0356
```

```
mean(Driver_age_violation$time_since_last)
```

```
[1] 465.3042
```

Number of Days per altercation

The mean number of altercantions is around 200 and majority of the people in fatal accidents are commiting an altercation less than 200 days

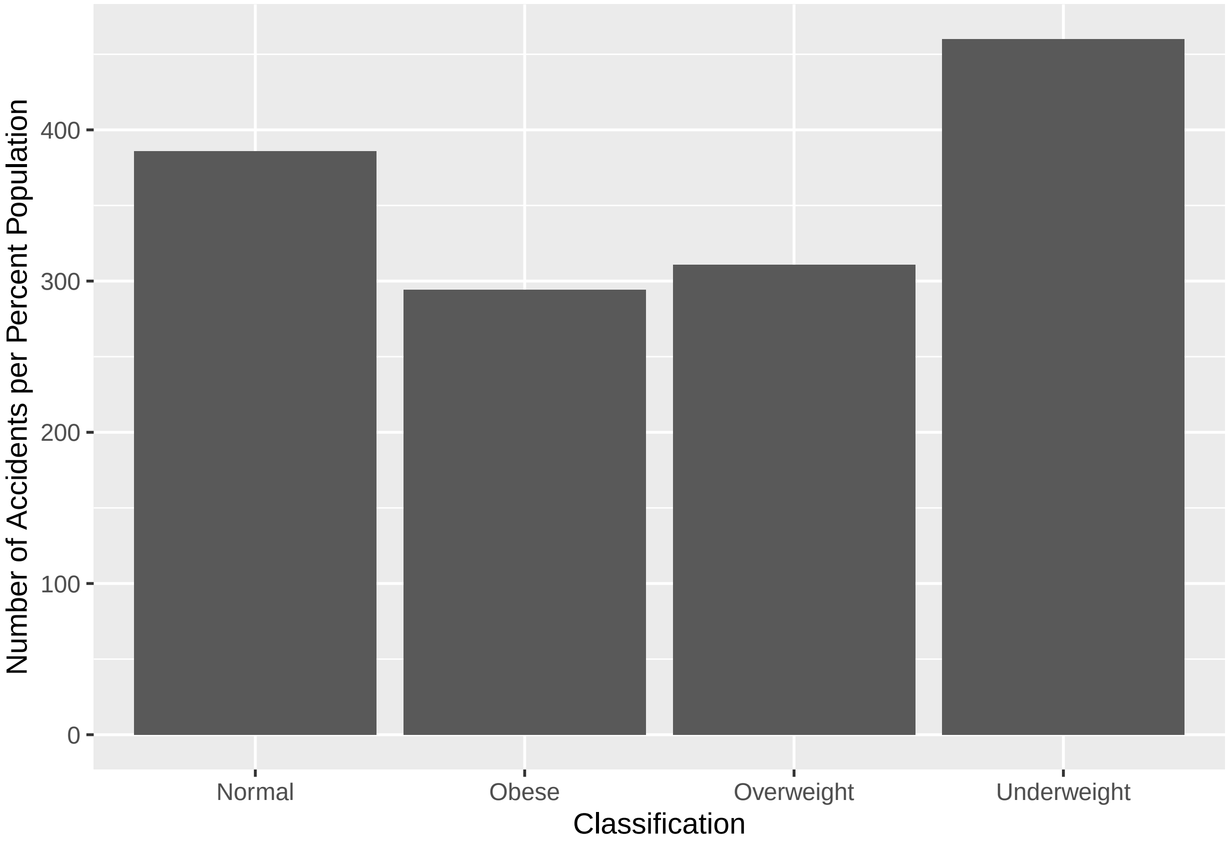
Number of Days Since Only Altercation

There is a lot people with good driving records and 1 accient. The mean number of altercations moves to around 450 days for people with 1 previous accident

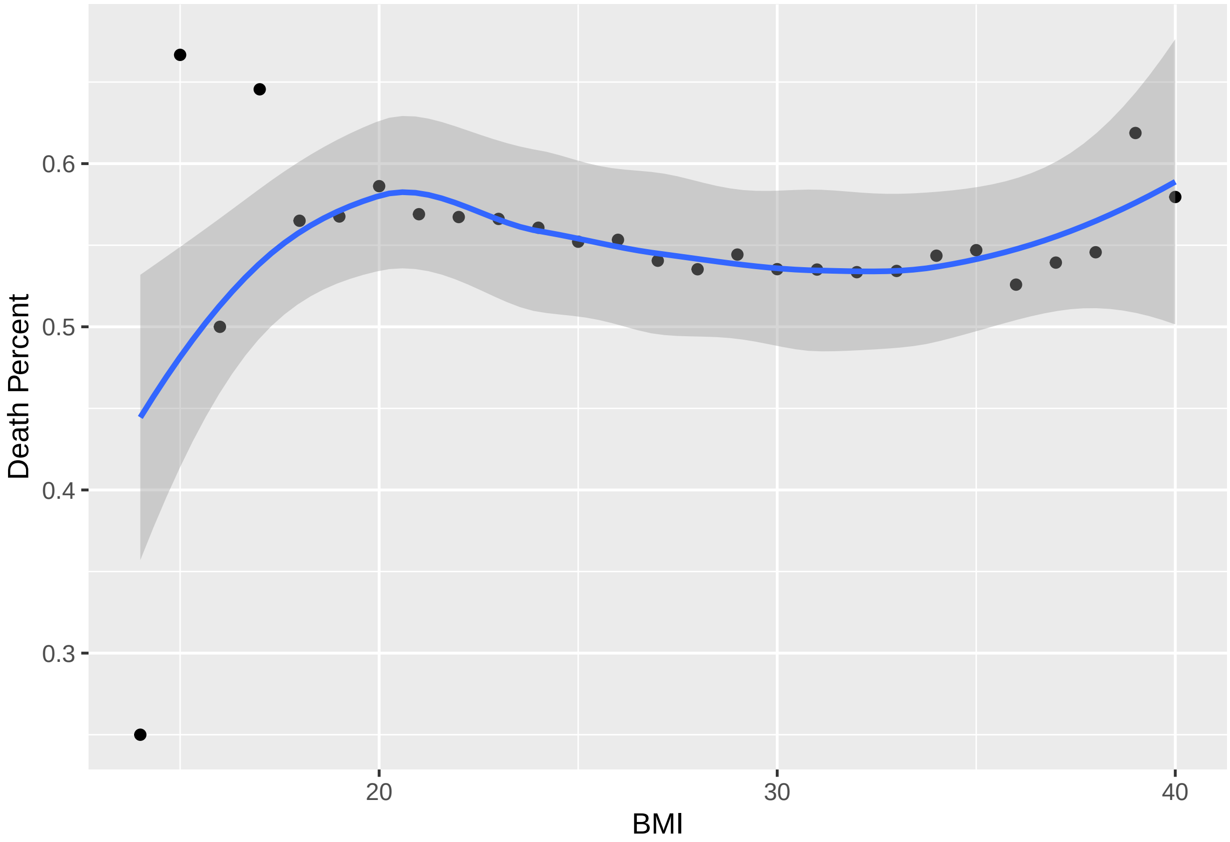
Number of Days Since Last Altercation

Removed occurance with 1 altercation because mistake happen but people having altercations frequently (more than 1 once) are higher risk drivers. Before people were averaging an altercation every 200 days but the days since last altercation is averaging 400 days. There is high spike of getting in a fatal accident less than 100 days since the last altercation.

```
# 25-29.9 = overweight, 30< = obese, 16> severely underweight,
# 16-18.5 =underweight, normal = 18.5-25
Pop_BMI_percent<- c(1.40,33,36,29.6)
Classification<- c("Underweight", "Normal", "Overweight", "Obese")
Pop_BMI<- data.frame(Pop_BMI_percent, Classification)

BMI_dist<-VEH_df%>%
  filter(DR_WGT<571, DR_HGT<88)%>%
  mutate( BMI= (703*DR_WGT)/((DR_HGT)^2))%>%
  mutate( BMI= round(BMI))%>%
  count(BMI, DEATHS=DEATHS>0)%>%
  filter(BMI >13, BMI <75)%>%
  pivot_wider(names_from= "DEATHS", values_from = "n")%>%
  rename(Deaths= "TRUE", Survive ="FALSE")
# BMI_dist
BMI_dist %>%
  mutate( Classification = cut(BMI, breaks = c(-Inf, 18.5,25,30, Inf),
           labels= c("Underweight", "Normal", "Overweight", "Obese")),
          num_accidents = Survive +Deaths,
          Deaths= replace_na(Deaths,0),
          num_accidents = replace_na(num_accidents,0))%>%
  group_by(Classification)%>%
  mutate(x= sum(Deaths),
         y= sum(num_accidents))%>%
  sample_n(1)%>%
  left_join(Pop_BMI, by=c("Classification"= "Classification"))%>%
  mutate(z= y/ Pop_BMI_percent)%>%
  ggplot(aes(x=Classification, y= z))+geom_col()+
  ylab("Number of Accidents per Percent Population")
```



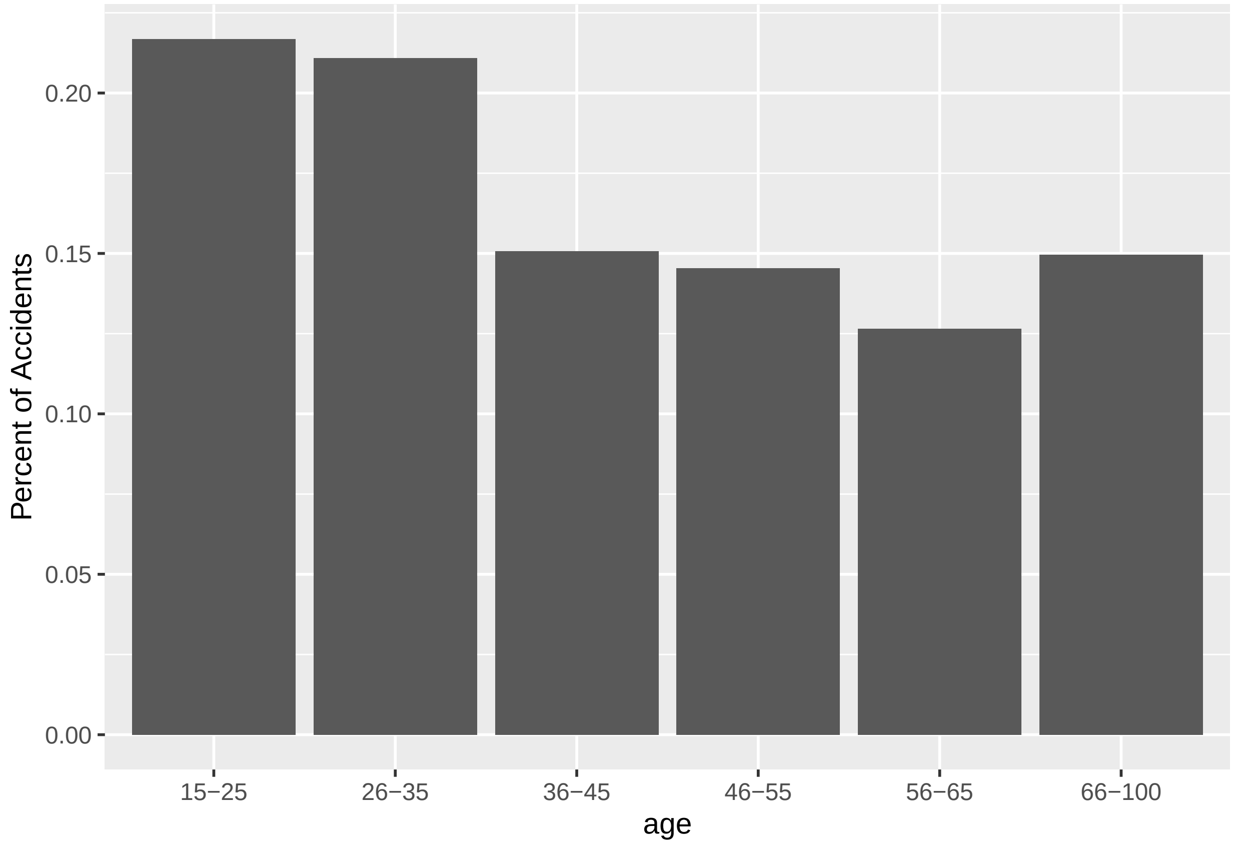
```
BMI_dist_40<-BMI_dist%>%
  filter(BMI>39)%>%
  mutate( Survive = sum(Survive, na.rm=T),
          Deaths= sum(Deaths, na.rm = T))%>%
  filter(BMI==40)


BMI_dist2<-BMI_dist%>%
  filter(BMI<40)%>%
  rbind(BMI_dist_40)%>%
  mutate(Death_percent= Deaths/ (Survive+ Deaths))

BMI_dist2%>%
  ggplot(aes(x= BMI, y= Death_percent))+
  geom_point()+
  geom_smooth()+
  ylab("Death Percent")
```



BMI

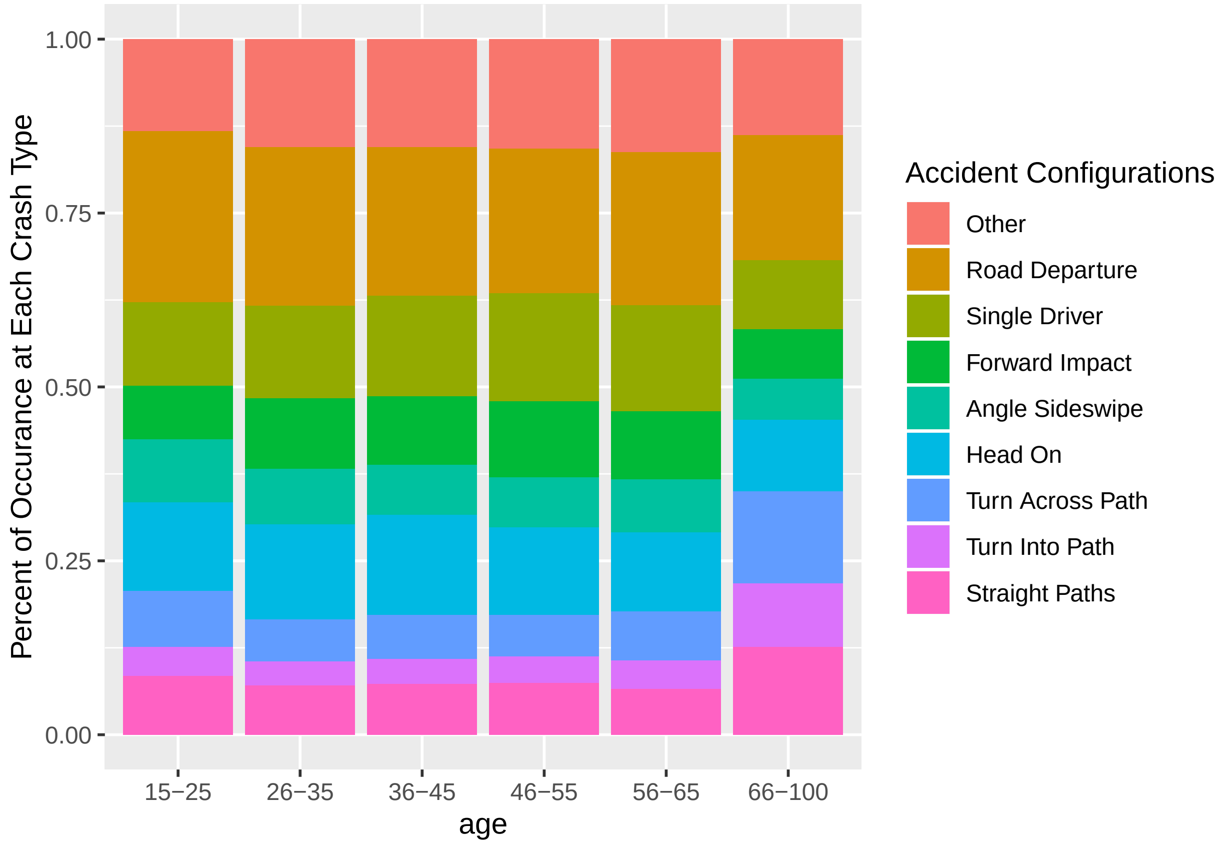
When the number of accident is divided by the percent of the US population in each category the results show that underweight drivers tend to be in more fatal accident The death percentage for drivers at different BMI is relatively the same with a little curve downward for people at BMI from 20 to 32. BMI of 15 stands as a lower death rate likely due to a lot less drivers and parental support while driving.

```
Driver_Configuration <- Driver_age_violation2 %>% group_by() %>%
    mutate(ACC_TYPE = as.character(ACC_TYPE), ACC_Configurations = fct_collapse(ACC_TYPE,
        Other = "0", `Road Departure` = c("1", 2:10), `Single Driver` = c(11:16),
        `Forward Impact` = c(20:33), `Forward Impact` = c(38:43),
        `Angle Sideswipe` = c(42:49), `Head On` = c(50:53), `Forward Impact` = c(54:62),
        `Angle Sideswipe` = c(64:67), `Turn Across Path` = c(68:75),
        `Turn Into Path` = c(76:85), `Straight Paths` = c(86:91),
        Other = c(92:93), Other = c(98:99))) %>% mutate(AGE = as.character(AGE),
    age = fct_collapse(AGE, `15-25` = c("15", "16", "17", "18",
        "19", "20", "21", "22", "23", "24", "25"), `26-35` = c(26:35),
        `36-45` = c(36:45), `46-55` = c(46:55), `56-65` = c(56:65),
        `66-100` = c(66:100)))
Driver_Configuration %>% count(age) %>% mutate(x = n/sum(n)) %>%
    ggplot(aes(x = age, y = x)) + geom_col() + ylab("Percent of Accidents")
```

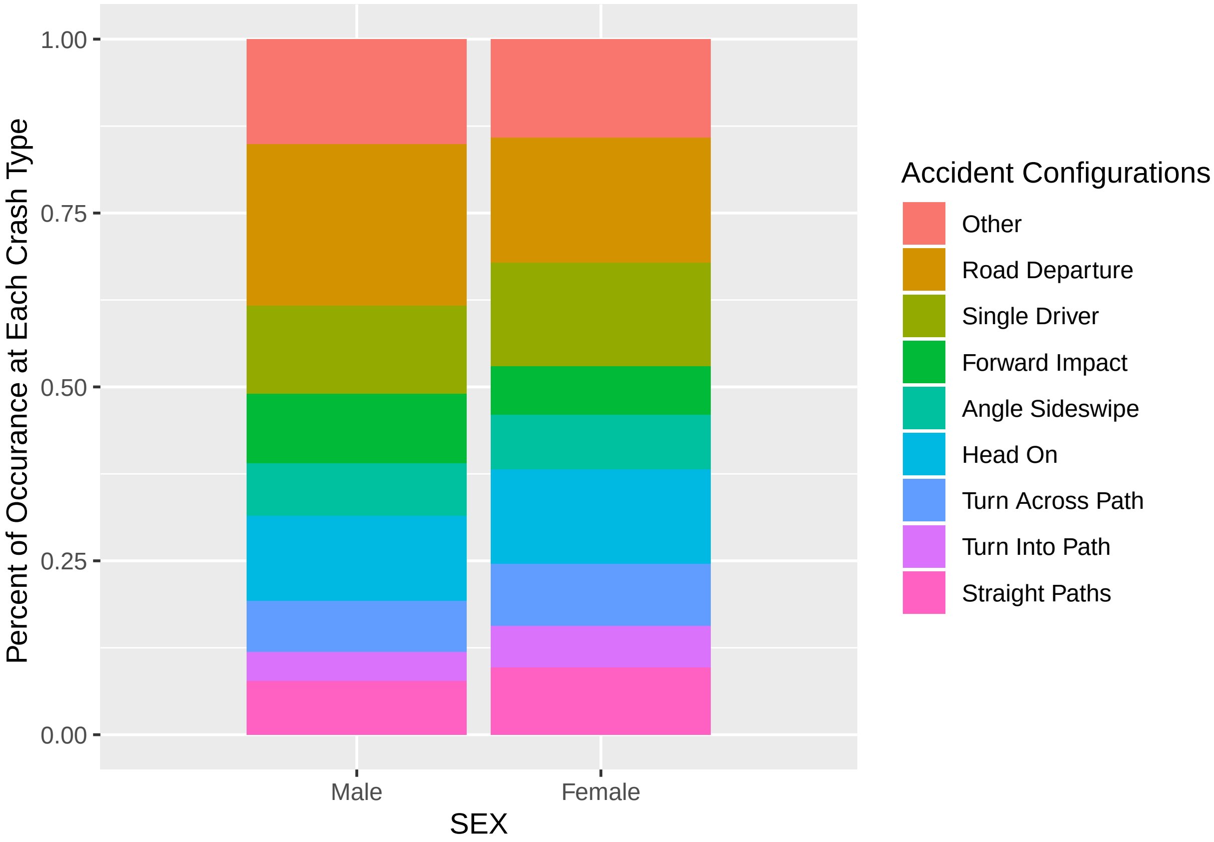


Crash type is divided into crash configurations. Show different configurations and what type of crash is defined. Age is divided into categories of 10. Starting at the earliest age of 15

```
Driver_Configuration%>%
  group_by(age)%>%
  count(age, ACC_Configurations)%>%
  mutate(x=sum(n),
         y= n/x,
         z=sum(y))%>%
  group_by(ACC_Configurations)%>%
  mutate(u= sum(y))%>%
  filter(u>0.01)%>%

  ggplot(aes(x= age, y= y, fill= ACC_Configurations))+
  geom_col()+
  ylab("Percent of Occurance at Each Crash Type")+
  guides(fill= guide_legend(title = "Accident Configurations"))
```



```
#straight path, turn into path, turn across path increases significantly over the age of 66.
# could be due to lack of awareness or color blindness.
# Not stopping at signs or lights. single driver drop significantly
Driver_Configuration%>%
  group_by(SEX)%>%filter(SEX<3)%>%
  count(SEX, ACC_Configurations)%>%
  mutate(x=sum(n),
         y= n/x,
         z=sum(y))%>%#view()%>%
  group_by(ACC_Configurations)%>%
  mutate(u= sum(y))%>%
  filter(u>0.01)%>%
  ggplot(aes(x=SEX, y=y, fill= ACC_Configurations))+
  geom_col()+
  ylab("Percent of Occurance at Each Crash Type ")+
  xlim("Male", "Female")+
  guides(fill= guide_legend(title = "Accident Configurations"))
```



```
#removed accident configures that occur less than 1 percent of the time

# Males are involved in fatal accidents almost 3 times as much as females
kable(Driver_Configuration%>%
        group_by()%>%
        count(SEX)%>%
        filter(SEX<3)%>%
        mutate(x= n/sum(n)) %>%
        rename(count = n, percentage = x),
      "latex", booktabs = T, row.names = FALSE)
```

SEX	count	percentage
1	33634	0.7453848
2	11489	0.2546152

Age

Each Crash type is configured to the percent of fatal accidents that happen at each age group. straight path, turn into path, turn across path increases significantly over the age of 66. could be due to lack of awareness or color blindness. Not stopping at signs or lights. single driver drop significantly. Angle sideswip, opposite side, staight paths and turn across path slighly increase. I would expect these to be true because I would think younger drivers are less aware of the rightaway. Thus, they turn at the wrong times. Rear Ends tend to decrease. This is opposite of what I would think to happen because I would suspect a younger driver to pay less attention follow someonw closer than what is needed to prevent an accident. All other crash type are consistent between groups.

Sex

Rear ends and single drivers are more common crash types among men. Straight path and turn across path are more common crash types among women

Appendix 5: Question 5 Code (Vehicle Vulnerability)

Analyzing patterns related to vehicle attributes such as type, make, model, body type, etc. combined with type of damage, rollover, fire/explosion, and frequency of incidents can reveal vehicle vulnerabilities. Employ an exploratory approach similar to what is discussed in the previous question to hypothesize and validate vehicle vulnerabilities.

Initial thoughts:

- Jeep/ SUV are probably more likely to rollover compared to other vehicle types, this may be an interesting relationship to investigate
- Frequency of incidents should be interesting, but care will need to be taken that the general observations are not correlated with demand
- Fire/explosion seem exceedinly rare, so investigating the types of cars that fall in this category would be interesting. It would also be interesting to see if there is a common sequence of events that leads to fires, or if the car itself is the important covariate.
- An interesting point to consider is that this data is survivorship biased heavily. Finding the safest car based on the data is not the same as finding the safest car overall, because the safest car may have never been in an accident and thus not in this dataset.
- Are certain types/ make/model more likely to be in speeding related accidents?
- Get Reliability Data for each make / model
- Get Safety Data for various body types

Code Information:

Table/Column	CSV Name	FARS Data Dictionary Location	Other
Damaged Areas	DAMAGE	371	NA
Vehicle/MAKE	VEHICLE	189	NA
Vehicle/MODEL	VEHICLE	192	NA
Vehicle/BODY_TYP	VEHICLE	298	NA
Vehicle/MOD_YEAR	VEHICLE	316	4 digit model year
Vehicle/ROLLOVER	VEHICLE	367	Rollover Type (0: None, 1:Tripped, 2:Untripped, 9:Unknown)
Vehicle/ROLINLOC	VEHICLE	369	Rollover Location
Vehicle/IMPACT1	VEHICLE	371	Initial Impact
Vehicle/DEFORMED	VEHICLE	380	Extent of Damage
Vehicle/FIRE_EXP	VEHICLE	420	Fire/Explosion Occurence (0: No, 1: Yes)
Vehicle/SPEEDREL	VEHICLE	464	Speeding Related

ROLINLOC Codes	Attributes
0	None
1	Roadway
2	Shoulder
3	Median/Separator
4	In Gore
5	On Roadside
6	Outside of Trafficway
7	In Parking Lane/Zone

ROLINLOC Codes	Attributes
9	Unknown

IMPACT 1 Codes	Attributes
00	Non-Collision
01-12	Clock Points
13	Top
14	Undercarriage
61	Left
62	Left-Front Side
63	Left-Back Side
81	Right
82	Right-Front Side
83	Right-Back Side
18	Cargo/Vehicle Parts Set-In-Motion
19	Other Objects Set-In-Motion
20	Object Set in Motion, Unknown if Cargo/Vehicle Parts or Other
98	Not Reported
99	Reported as Unknown

MDAREAS Codes	Attributes
01-12	Clock Values
13	Top
14	Undercarriage
15	No Damage
99	Damage Areas Unknown

DEFORMED Codes	Attributes
0	None
2	Minor Damage
4	Functional Damage
6	Disabling Damage
8	Not Reported
9	Unknown

SPEEDREL Codes	Attributes
0	No
2	Yes, Racing
3	Yes, Exceeded Speed Limit
4	Yes, Too Fast for Conditions
5	Yes, Specifics Unknown
9	Unknown

```
vehicle_df = read.csv("../FARS_Data/FARS2018NationalCSV/VEHICLE.csv")
damages_df = read.csv("../FARS_Data/FARS2018NationalCSV/DAMAGE.csv")
accident_df = read.csv("../FARS_Data/FARS2018NationalCSV/ACCIDENT.csv")
```

The first item to investigate is if the type of car is an important factor in crashes/fatalities. Let's first get all of the types of cars in the dataset.

```
vehicle_body_types = vehicle_df$BODY_TYP
```

Lots of vehicle types. These are grouped into 9 categories (plus an other category) in the Data Dictionary. I am grouping them based on these predefined groupings.

Category	Group
1	Automobile
2	Automobile Derivative
3	Utility Vehicle
4	Van
5	Light Truck
6	Bus
7	Heavy Truck
8	Motor Home
9	Motorcycle/Moped
10	Other

```
vehicle_body_types <- mapvalues(vehicle_body_types,
                                from = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 17),
                                to = rep(1, 10))
vehicle_body_types <- mapvalues(vehicle_body_types,
                                from = c(10, 11, 12, 13),
                                to = rep(2, 4))
vehicle_body_types <- mapvalues(vehicle_body_types,
                                from = c(14, 15, 16, 19),
                                to = rep(3, 4))
vehicle_body_types <- mapvalues(vehicle_body_types,
                                from = c(20, 21, 22, 28, 29),
                                to = rep(4, 5))
vehicle_body_types <- mapvalues(vehicle_body_types,
                                from = c(33, 34, 39, 40, 41, 45, 48, 49),
                                to = rep(5, 8))
vehicle_body_types <- mapvalues(vehicle_body_types,
                                from = c(50, 51, 52, 55, 58, 59),
                                to = rep(6, 6))
vehicle_body_types <- mapvalues(vehicle_body_types,
                                from = c(60, 61, 62, 63, 64, 66, 67, 71, 72, 78, 79),
                                to = rep(7, 11))
vehicle_body_types <- mapvalues(vehicle_body_types,
                                from = c(42, 65, 73),
                                to = rep(8, 3))
vehicle_body_types <- mapvalues(vehicle_body_types,
                                from = c(80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90),
                                to = rep(9, 11))
vehicle_body_types <- mapvalues(vehicle_body_types,
                                from = c(91, 92, 93, 94, 95, 96, 97, 98, 99),
                                to = rep(10, 9))
```

Next, lets see how the various groups compare on number of incidents, extent of damage, rollover percentage, fire/explosion percentage, and speed related percentage. Before this, some data preparation is needed.

```
# Extent of Damage, assuming that not reported and unknown
# are same category to be removed
damage_extent <- vehicle_df$DEFORMED
damage_extent <- mapvalues(damage_extent, from = c(0, 2, 4, 6,
    8, 9), to = c(1, 2, 3, 4, 5, 5))

# Number of Missing/Not Reported Values (4455)
length(damage_extent[damage_extent == 5])
```

```
[1] 4455
```

```
# Rollover Percentage
rollover <- vehicle_df$ROLLOVER
rollover <- mapvalues(rollover, from = c(0, 1, 2, 9), to = c(0,
    1, 1, 2))

# Some missing values can be collected by examining the
# ROLINLOC column. If this value is not 0 or 9, then the
# vehicle did rollover. (430 -> 29)
for (i in 1:length(rollover)) {
    if (rollover[i] == 2) {
        rollover[i] = ifelse(vehicle_df$ROLINLOC[i] != 9 & vehicle_df$ROLINLOC[i] !=
            0, 1, 2)
    }
}

# Number of Missing Values (29)
length(rollover[rollover == 2])
```

```
[1] 29
```

```
# Fire/Explosion
fire_exp <- vehicle_df$FIRE_EXP

# Speed Related Incidents
speed_rel <- vehicle_df$SPEEDREL
speed_rel <- mapvalues(speed_rel, from = c(0, 2, 3, 4, 5, 8,
    9), to = c(0, 1, 1, 1, 1, 2, 2))

# Some missing values can be collected by comparing the
# travel speed to the speed limit (2077 -> 1596)
for (i in 1:length(speed_rel)) {
    if (speed_rel[i] == 2) {
        if (vehicle_df$TRAV_SP[i] < 151) {
            speed_rel[i] = ifelse((vehicle_df$TRAV_SP[i] - vehicle_df$VSPD_LIM[i]) >
                0, 1, 0)
        } else {
            speed_rel[i] == 2
        }
    }
}
```

```
# Number of Missing Values (1596)
length(speed_rel[speed_rel == 2])
```

```
[1] 1596
```

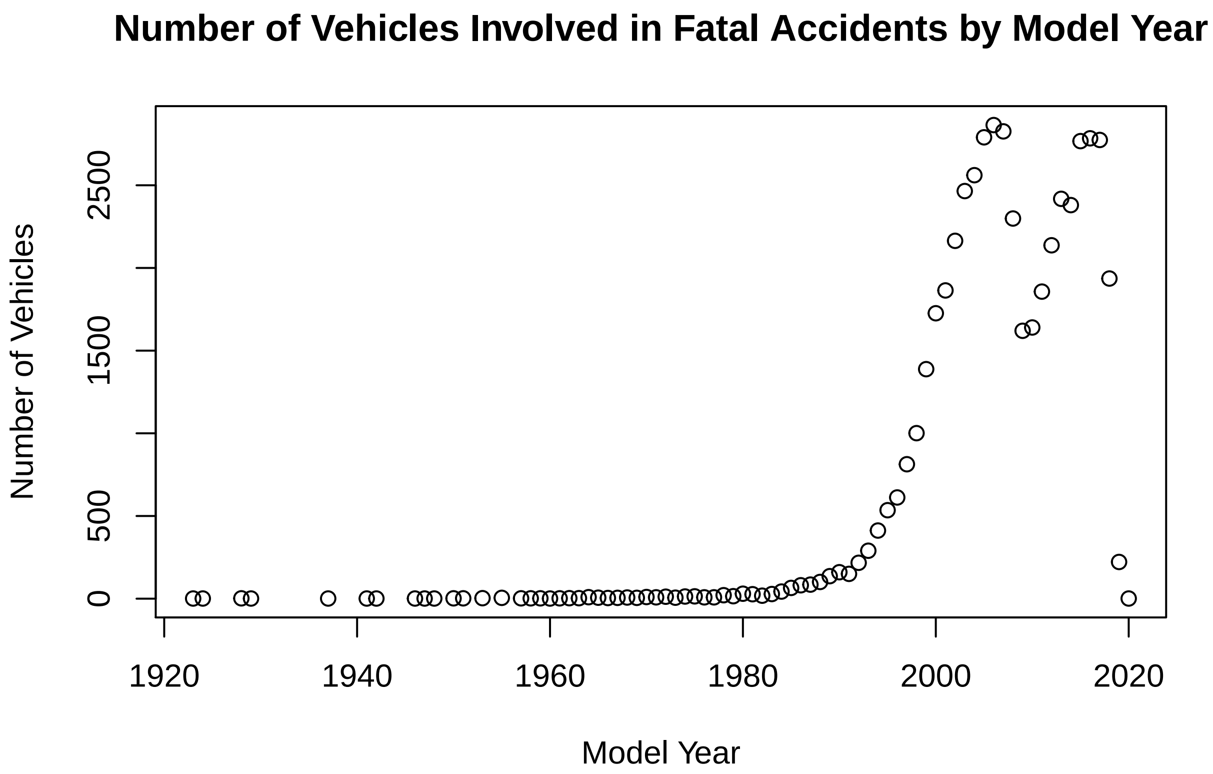
Let's get the number of occurences for each vehicle type and the percentage of rollover, fire/exp, speed rel, and the 4 types of damage

```
body_type_vulnerability <- data.frame(matrix(ncol = 9, nrow = 10))
colnames(body_type_vulnerability) <- c("Group", "Accidents", "Rollover Percentage",
                                       "Fire/Exp Percentage", "Speed Rel Percentage",
                                       "No Damage Percentage", "Minor Damage Percentage",
                                       "Functional Damage Percentage",
                                       "Disabling Damage Percentage")
vehicle_combined = data.frame(vehicle_body_types, rollover,
                              fire_exp, speed_rel, damage_extent)
for (i in 1:10) {
    vehicle_rollover = vehicle_combined[vehicle_combined$vehicle_body_types == i &
                                          vehicle_combined$rollover != 2, ]
    vehicle_fire_exp = vehicle_combined[vehicle_combined$vehicle_body_types == i,]
    vehicle_speedrel = vehicle_combined[vehicle_combined$vehicle_body_types == i &
                                          vehicle_combined$speed_rel != 2, ]
    vehicle_damage = vehicle_combined[vehicle_combined$vehicle_body_types == i &
                                        vehicle_combined$damage_extent != 5,]
    body_type_vulnerability[i, 1] = i
    body_type_vulnerability[i, 2] = length(vehicle_body_types[vehicle_body_types == i])
    body_type_vulnerability[i, 3] = round(sum(vehicle_rollover$rollover)
                                          /nrow(vehicle_rollover) * 100, digits = 3)
    body_type_vulnerability[i, 4] = round(sum(vehicle_fire_exp$fire_exp)
                                          /nrow(vehicle_fire_exp) * 100, digits = 3)
    body_type_vulnerability[i, 5] = round(sum(vehicle_speedrel$speed_rel)
                                          /nrow(vehicle_speedrel) * 100, digits = 3)
    body_type_vulnerability[i, 6] = round(sum(vehicle_damage$damage_extent == 1)
                                          /nrow(vehicle_damage) * 100, digits = 3)
    body_type_vulnerability[i, 7] = round(sum(vehicle_damage$damage_extent == 2)
                                          /nrow(vehicle_damage) * 100, digits = 3)
    body_type_vulnerability[i, 8] = round(sum(vehicle_damage$damage_extent == 3)
                                          /nrow(vehicle_damage) * 100, digits = 3)
    body_type_vulnerability[i, 9] = round(sum(vehicle_damage$damage_extent == 4)
                                          /nrow(vehicle_damage) * 100, digits = 3)
}
```

Can turn the vulnerability calculation into a function for body_type, make, model etc.

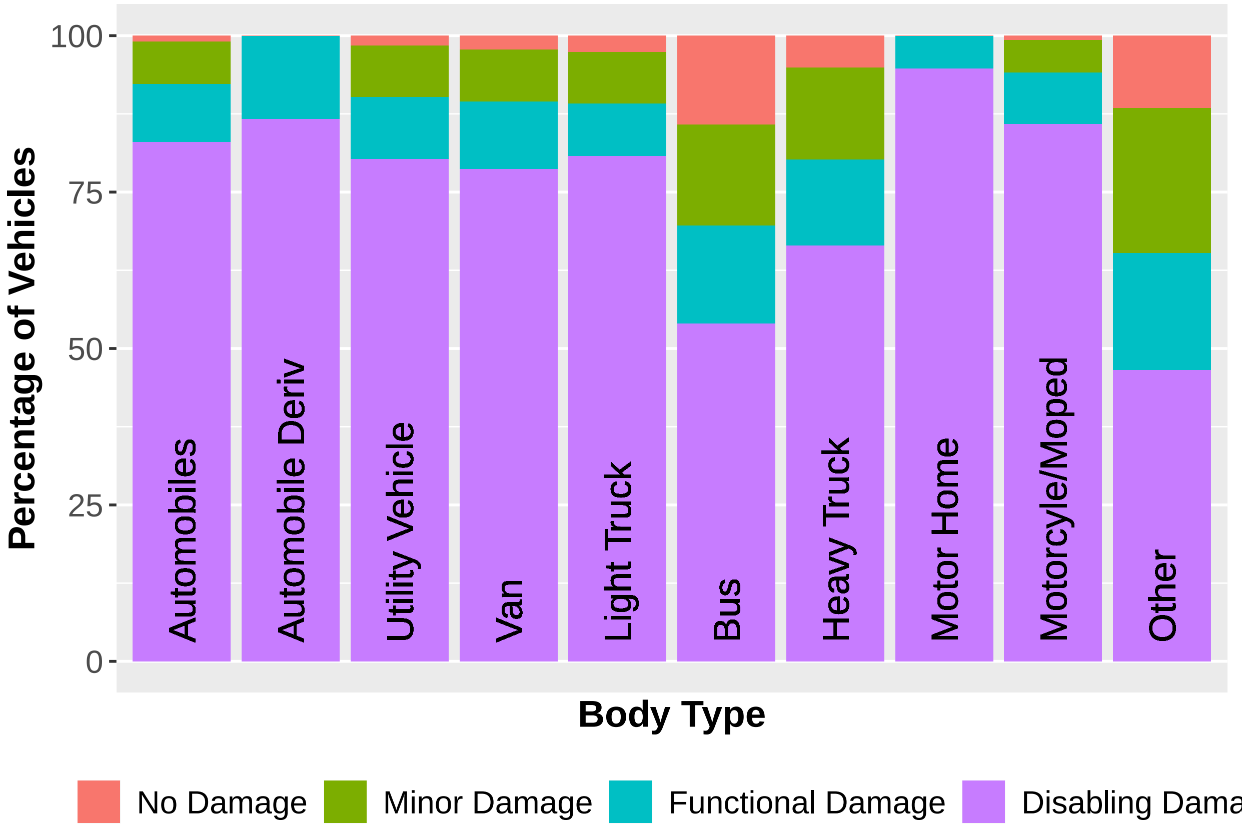
```
get_vulnerability <- function(x) {
    vulnerability <- data.frame(matrix(ncol = 9, nrow = length(sort(unique(x)))))
    colnames(vulnerability) <- c("Group", "Accidents", "Rollover Percentage",
                                 "Fire/Exp Percentage", "Speed Rel Percentage",
                                 "No Damage Percentage", "Minor Damage Percentage",
                                 "Functional Damage Percentage",
                                 "Disabling Damage Percentage")
    vehicle_combined = data.frame(x, rollover, fire_exp, speed_rel, damage_extent)
    # Not all categories before preprocessing start from 1
    counter = 1
    for (i in sort(unique(x))) {
        vehicle_rollover = vehicle_combined[vehicle_combined$x == i &
                                              vehicle_combined$rollover != 2, ]
        vehicle_fire_exp = vehicle_combined[vehicle_combined$x == i, ]
        vehicle_speedrel = vehicle_combined[vehicle_combined$x == i &
                                              vehicle_combined$speed_rel != 2, ]
        vehicle_damage = vehicle_combined[vehicle_combined$x == i &
                                            vehicle_combined$damage_extent != 5, ]
        vulnerability[counter, 1] = i
        vulnerability[counter, 2] = length(x[x == i])
        vulnerability[counter, 3] = round(sum(vehicle_rollover$rollover)
                                          /nrow(vehicle_rollover) * 100, digits = 3)
        vulnerability[counter, 4] = round(sum(vehicle_fire_exp$fire_exp)
                                          /nrow(vehicle_fire_exp) * 100, digits = 3)
        vulnerability[counter, 5] = round(sum(vehicle_speedrel$speed_rel)
                                          /nrow(vehicle_speedrel) * 100, digits = 3)
        vulnerability[counter, 6] = round(sum(vehicle_damage$damage_extent == 1)
                                          /nrow(vehicle_damage) * 100, digits = 3)
        vulnerability[counter, 7] = round(sum(vehicle_damage$damage_extent == 2)
                                          /nrow(vehicle_damage) * 100, digits = 3)
        vulnerability[counter, 8] = round(sum(vehicle_damage$damage_extent == 3)
                                          /nrow(vehicle_damage) * 100, digits = 3)
        vulnerability[counter, 9] = round(sum(vehicle_damage$damage_extent == 4)
                                          /nrow(vehicle_damage) * 100, digits = 3)
        counter = counter + 1
    }
  return(vulnerability)
}
```

```
# Body Type
body_type_vulnerability = get_vulnerability(vehicle_body_types)

# Make There are a lot of makes, so need a better way to
# preprocess this data.
make_vulnerability = get_vulnerability(vehicle_df$MAKE)

# Model Sorting by model leads to the same observation as
# body type.
model_vulnerability = get_vulnerability(vehicle_df$MODEL)

# Model Year
model_yr_vulnerability = get_vulnerability(vehicle_df$MOD_YEAR)
plot(model_yr_vulnerability[1:78, 1], model_yr_vulnerability[1:78,
    2], main = "Number of Vehicles Involved in Fatal Accidents by Model Year",
    xlab = "Model Year", ylab = "Number of Vehicles")
```

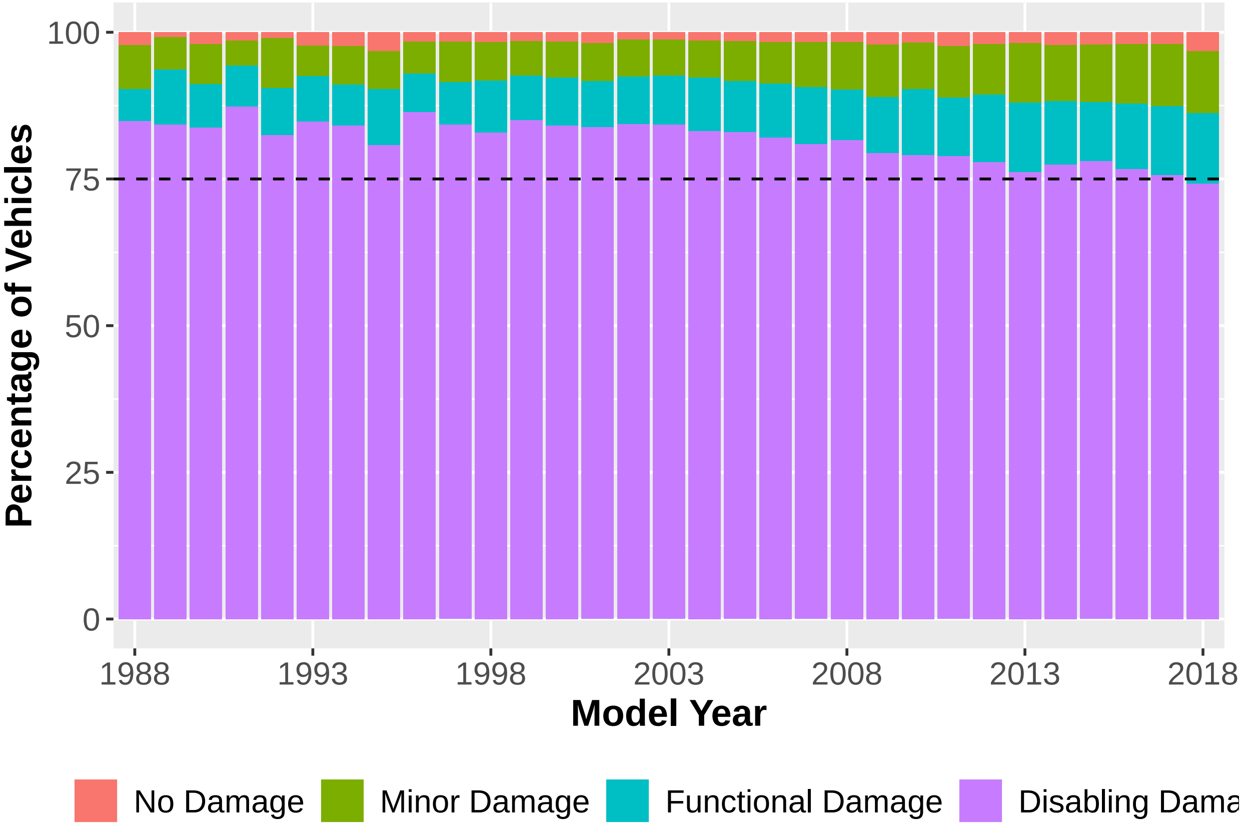


Interestingly, there is a sharp downturn in the number of vehicles in the mid 2000s. This is due to the economic recession. Additionally, there were multiple vehicles from the 1920's involed in fatal accidents.

```
# Now turn the above into a table and then a stacked barplot
summary_info_body_type = c(0,0,0)
named_columns = c("Automobile", "Automobile Derivative","Utility Vehicle",
                  "Van","Light Truck","Bus","Heavy Truck","Motor Home",
                  "Motorcycle/Moped","Other")
named_rows = c(0,0,0,0,0,"No Damage", "Minor Damage",
               "Functional Damage", "Disabling Damage")

for (i in 1:nrow(body_type_vulnerability)){
  for (j in 6:ncol(body_type_vulnerability)){
    summary_info_body_type = rbind(summary_info_body_type,
                                   c(named_columns[i], named_rows[j],
                                     body_type_vulnerability[i,j]))
  }
}

summary_info_body_type = data.frame(summary_info_body_type)
summary_info_body_type$X3 = as.numeric(as.matrix(summary_info_body_type$X3))
summary_info_body_type = summary_info_body_type[-1,]
row.names(summary_info_body_type) <- 1:nrow(summary_info_body_type)

summary_plot = ggplot(summary_info_body_type,
                      aes(fill=factor(X2,
                                      levels = c("No Damage", "Minor Damage",
                                                 "Functional Damage", "Disabling Damage")),
                          y=as.numeric(X3),
                          x=factor(X1,
                                   levels = c("Automobile", "Automobile Derivative",
                                              "Utility Vehicle","Van","Light Truck",
                                              "Bus","Heavy Truck","Motor Home",
                                              "Motorcycle/Moped","Other")))) +
  geom_bar(position="stack", stat="identity") +
  xlab("Body Type") +
  ylab("Percentage of Vehicles") +
  scale_x_discrete(breaks=1:10, labels=c("Automobile", "Automobile Derivative",
                                         "Utility Vehicle","Van","Light Truck",
                                         "Bus","Heavy Truck","Motor Home",
                                         "Motorcycle/Moped","Other")) +
  theme(legend.position="bottom",
        legend.direction="horizontal",
        legend.title = element_blank(),
        plot.title = element_text(size = 16, hjust = 0.5),
        axis.text = element_text(size = 12),
        axis.title = element_text(size = 14, face="bold"),
        legend.text=element_text(size=12),
        axis.text.x = element_text(angle = 45, hjust = 1)) +
  geom_text(x = 1, y = 3, label = "Automobiles",      angle = 90, hjust = 0, size = 5) +
  geom_text(x = 2, y = 3, label = "Automobile Deriv", angle = 90, hjust = 0, size = 5) +
  geom_text(x = 3, y = 3, label = "Utility Vehicle",  angle = 90, hjust = 0, size = 5) +
  geom_text(x = 4, y = 3, label = "Van",              angle = 90, hjust = 0, size = 5) +
  geom_text(x = 5, y = 3, label = "Light Truck",      angle = 90, hjust = 0, size = 5) +
  geom_text(x = 6, y = 3, label = "Bus",              angle = 90, hjust = 0, size = 5) +
  geom_text(x = 7, y = 3, label = "Heavy Truck",      angle = 90, hjust = 0, size = 5) +
  geom_text(x = 8, y = 3, label = "Motor Home",       angle = 90, hjust = 0, size = 5) +
  geom_text(x = 9, y = 3, label = "Motorcyle/Moped",  angle = 90, hjust = 0, size = 5) +
  geom_text(x = 10, y = 3, label = "Other",           angle = 90, hjust = 0, size = 5)

summary_plot
```



```
# Now turn the above into a table and then a stacked barplot
model_yr_summary = model_yr_vulnerability[model_yr_vulnerability$Group > 1987,]
model_yr_summary = model_yr_summary[model_yr_summary$Group < 2019,]

summary_info_model_year = c(0,0,0)
named_columns = c(seq(1988,2018,1))
named_rows = c(0,0,0,0,0,"No Damage", "Minor Damage",
               "Functional Damage", "Disabling Damage")

for (i in 1:nrow(model_yr_summary)){
  for (j in 6:ncol(model_yr_summary)){
    summary_info_model_year = rbind(summary_info_model_year,
                                    c(named_columns[i], named_rows[j],
                                      model_yr_summary[i,j]))
  }
}

summary_info_model_year = data.frame(summary_info_model_year)
summary_info_model_year$X3 = as.numeric(as.matrix(summary_info_model_year$X3))
summary_info_model_year = summary_info_model_year[-1,]
row.names(summary_info_model_year) <- 1:nrow(summary_info_model_year)

every_nth = function(n) {
  return(function(x) {x[c(TRUE, rep(FALSE, n - 1))]})
}

summary_plot = ggplot(summary_info_model_year,
                      aes(fill=factor(X2, levels = c("No Damage", "Minor Damage",
                                                     "Functional Damage",
                                                     "Disabling Damage")),
                          y=as.numeric(X3),
                          x=factor(X1, levels = c(seq(1988,2018,1))))) +
  geom_bar(position="stack", stat="identity") +
  xlab("Model Year") +
  ylab("Percentage of Vehicles") +
  geom_hline(yintercept = 75, linetype = "dashed") +
  scale_x_discrete(breaks=every_nth(n = 5)) +
  theme(legend.position="bottom",
        legend.direction="horizontal",
        legend.title = element_blank(),
        plot.title = element_text(size = 16, hjust = 0.5),
        axis.text = element_text(size = 12),
        axis.title = element_text(size = 14, face="bold"),
        legend.text=element_text(size=12),
        axis.text.x = element_text(angle = 0, hjust = 0.5))

summary_plot
```


Percentage of Vehicles
Model Year
No Damage
Minor Damage
Functional Damage
Disabling Dama