Stats 199 Research

Lisa Kaunitz

03/14/2020

Contents

1 Abstract **2**

2 Introduction **3**
2.1 Variables of the Study 3

3 Methods **5**
3.1 Exploratory Data Analysis 5
3.1.1 States 6
3.1.2 Severity 6
3.1.3 Accident Signals 7
3.1.4 Weather Conditions 8
3.1.5 Time 9
3.1.6 Humidity 9

4 Results and Applications **10**

5 Discussion **11**
5.1 Challenges of the study 11
5.2 Reccomendations for Future Work 11

6 References **12**

7 Appendix **12**

1 Abstract

In this study, I examined the factors that cause vehicle accidents. Using the U.S. Accidents dataset collected by Sobhan Moosavi with 3 million records, I explored the significance of variables on accidents with a focus on extracting that information and applying it to optimizing technology found in autonomous vehicles. First I cleaned the original dataset and transformed some of the variables. Then I did in-dept exploratory data analysis on numeric variables that deemed to be significant in the overall count of accidents. Notably, humidity was one of the most important factors when determining accidents, and upon further investigation, the humidity was one of the only variables that would play a role in irritability which can be a cause for human error. Consequently, I concluded that the greatest risk factors presented in autonomous vehicle crashes are those that go along with human error. If we are able to implement full autonomy of infrastructure and vehicles, then our roads will be much safer.

2 Introduction

It is likely that if you have been driving for the past 18 years you will have experienced the ramifications of being in an automobile crash multiple times (Belsky). In fact, every time a driver gets on the road, they have a 1 in 103 chance of dying in a car crash. While these numbers are staggering, the automobile industry has slowly added more safety features to cars: first starting off with the introduction of backup cameras, then self-parking capabilities, and now a level of autonomous driving. All of these features have been leading up to full autonomy on the road.

In this study, I developed two research questions concentrating on the significance of factors that cause accidents, and how we can use that information to optimize autonomous machine learning and AI algorithms to create safer vehicles on the road. These two things led me to create one overarching **Research Question:** What are the causes of accidents in the US, and how does that relate to accidents when it comes to implementing preventative technology in autonomous vehicles to minimize the mean squared error?

2.1 Variables of the Study

For this study, I cut down the original 49 variables to 36 and then created a new `Total_Time` variable. I decided that for the purposes of this research I did not need to look at `Source` because the majority of the data was collected from MapQuest, which would not create any significance in my study. Next, I decided to get rid of `Start_Lat`, `Start_Lng`, `End_Lat` and `End_Lng` because the "End" variables all had NA's in them so even if I wanted to know the distance, I would not have been able to. Additionally, because all of the data was collected from the US, I decided the `Country` variable was redundant to have in the data. I also got rid of the `Timezone`, `Street Number` and `Airport Code` variables because I decided they would not have much of a different significance. In regards to location, I am more interested in looking at the state rather than the airport code; if I were to hone in more on a specific region or state, then this variable would be useful to me, but for the purpose of exploratory data analysis I was more interested in analyzing accidents relative to each state. Finally, I decided to delete the `Start_Time` and `End_Time` variables, and create a new variable out of the difference between these two columns called `Total_Time`. I had to clean up this variable a little because I ended up with extreme values that did not have any place in this study, such as negative time observations, and reported accidents that take up to 29,772 hours. These were unrealistic so I used if-else statements to make all values less than 0 and greater than 400 min to be NA's.

Lastly, I decided that `Weather_Conditions` would be significant in my research because a lot of what the driver and vehicles have to adjust to is the conditions of the road beneath them. When I looked at the unique conditions included in the original dataset, I found that there were 62, however, many of them would be able to fit in one of five main categories: rain, snow, low visibility, and clear. As a result, I cleaned the column such that all variables that would fit into each of these groups would be regrouped. My final change to the original dataset was nullifying the `Weather_Timestamp` column because I was able to transform the weather conditions. The final dataset called *accidents.csv* has the following codebook of variables with all of the cleaning and transformations I mentioned above.

Data Codebook		
Number	Attribute	Description
1	ID	This is a unique identifier of the accident record.
2	TMC	A traffic accident may have a Traffic Message Channel (TMC) code which provides more detailed description of the event.
3	Severity	Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay as a result of the accident) and 4 indicates a significant impact on traffic (i.e., long delay).
4	Distance(mi)	The length of the road extent affected by the accident.
5	Description	Shows natural language description of the accident.
6	Street	Shows the street name in address field.
7	Side	Shows the relative side of the street (Right/Left) in address field.
8	City	Shows the city in address field.
9	County	Shows the county in address field.
10	State	Shows the state in address field.
11	Temperature(F)	Shows the temperature (in Fahrenheit).
12	Wind Chill(F)	Shows the wind chill (in Fahrenheit).
13	Humidity	Shows the humidity (in percentage).
14	Pressure(in)	Shows the air pressure (in inches).
15	Visibility(mi)	Shows visibility (in miles).
16	Wind Direction	Shows wind direction.
17	Wind Speed(mph)	Shows wind speed (in miles per hour).
18	Precipitation(in)	Shows precipitation amount in inches, if there is any.
19	Weather Condition	Shows the weather condition (rain, snow, thunderstorm, fog, etc.)
20	Amenity	A POI annotation which indicates presence of amenity in a nearby location.
21	Bump	A POI annotation which indicates presence of speed bump or hump in a nearby location.
22	Crossing	A POI annotation which indicates presence of crossing in a nearby location.
23	Give Way	A POI annotation which indicates presence of give way in a nearby location.
24	Junction	A POI annotation which indicates presence of junction in a nearby location.
25	No Exit	A POI annotation which indicates presence of no exit in a nearby location.
26	Railway	A POI annotation which indicates presence of railway in a nearby location.
27	Roundabout	A POI annotation which indicates presence of roundabout in a nearby location.
28	Station	A POI annotation which indicates presence of station in a nearby location.
29	Stop	A POI annotation which indicates presence of stop in a nearby location.
30	Traffic Calming	A POI annotation which indicates presence of traffic calming in a nearby location.
31	Traffic Signal	A POI annotation which indicates presence of traffic signal in a nearby location.
32	Turning Loop	A POI annotation which indicates presence of turning loop in a nearby location.
33	Sunrise Sunset	Shows the period of day (i.e. day or night) based on sunrise/sunset.
34	Civil Twilight	Shows the period of day (i.e. day or night) based on civil twilight.
35	Nautical Twilight	Shows the period of day (i.e. day or night) based on nautical twilight.
36	Astronomical Twilight	Shows the period of day (i.e. day or night) based on astronomical twilight.
37	Total Time	Shows the total time of the accident from Start to End (in minutes).

3 Methods

3.1 Exploratory Data Analysis

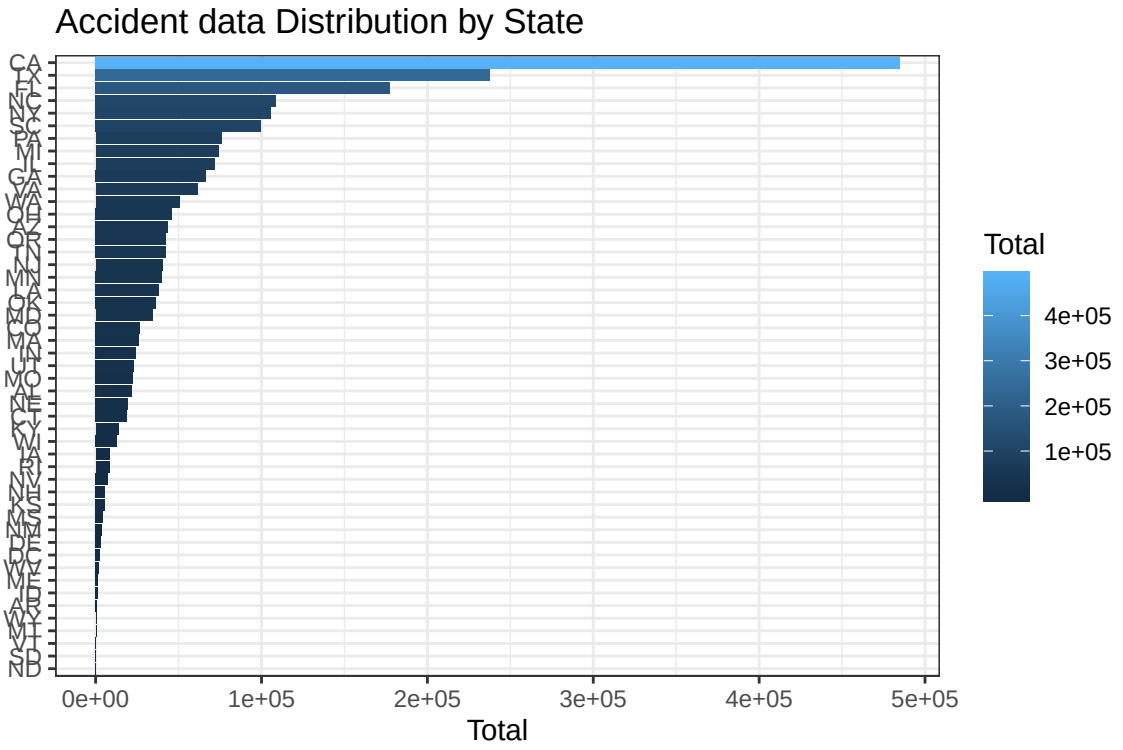
The exploratory data analysis is the most important part of my research in regards to finding solutions for autonomous vehicle safety. First, I decided to take a preliminary look at the summary statistics for the numeric variables. Then I looked at the data by *States* and used a function found in the data notebook by Manuel T to observe the relationships between the *Severity* of accidents by state and a graphical representation of *Accident Signals* that were most likely to occur, a critical factor in creating new insights for autonomous vehicles. Second, I wanted to investigate the significance of *Weather Conditions* on accidents. This is where I believe most improvements and recommendations can be made. Next, I took a look at how *Time* would play a role in accidents, here I was expecting to see most accidents take place during rush hours. Finally, I thought it was worth looking into factors that would have an impact on human error such as *Humidity*, which can influence irritability and may lead to an increase in accidents.

	Distance(mi)	Temperature(F)	Wind_Chill(F)	Humidity(%)	Pressure(in)
X	Min. : 0.0000	Min. :-77.80	Min. :-65.9	Min. : 4.00	Min. : 0.00
X.1	1st Qu.: 0.0000	1st Qu.: 48.90	1st Qu.: 19.2	1st Qu.: 50.00	1st Qu.:29.92
X.2	Median : 0.0000	Median : 63.00	Median : 28.7	Median : 68.00	Median :30.03
X.3	Mean : 0.2879	Mean : 61.23	Mean : 26.0	Mean : 65.93	Mean :30.04
X.4	3rd Qu.: 0.0100	3rd Qu.: 75.90	3rd Qu.: 36.4	3rd Qu.: 85.00	3rd Qu.:30.15
X.5	Max. :333.6300	Max. :170.60	Max. : 45.2	Max. :100.00	Max. :33.04
X.6		NA's :62265	NA's :1852370	NA's :64467	NA's :57280

	Visibility(mi)	Wind_Speed(mph)	Precipitation(in)	Total_Time
X	Min. : 0.00	Min. : 1.2	Min. : 0.0	Min. : 1.217
X.1	1st Qu.: 10.00	1st Qu.: 5.8	1st Qu.: 0.0	1st Qu.: 29.683
X.2	Median : 10.00	Median : 8.1	Median : 0.0	Median : 30.833
X.3	Mean : 9.12	Mean : 8.8	Mean : 0.1	Mean : 94.759
X.4	3rd Qu.: 10.00	3rd Qu.: 11.5	3rd Qu.: 0.0	3rd Qu.: 59.717
X.5	Max. :140.00	Max. :822.8	Max. :10.8	Max. :400.000
X.6	NA's :71360	NA's :442954	NA's :1979466	NA's :4745

From the results above we can see that `Distance(mi)`, `Pressure(in)`, `Visibility(mi)` and `Precipitation(in)` do not look significant at first glance. It goes against my intuition that precipitation, distance affected and visibility would not have much statistical significance because they seem like they would be the most influential factors in someone's driving capabilities. One would assume that the more it rained (higher precipitation), the driver would have a harder time viewing the road or they would have a higher likelihood of hydroplaning and causing a crash. Similarly, I was surprised to see that visibility does not seem like a significant factor because I would expect that if the driver has low visibility, then they would end up in more accidents. Nonetheless, one of the reasons that these variables do not have enough variation for significance is that the majority of the data collected comes from California, Florida, and Texas - states which are not indicative of the precipitation in the whole country and do not have much precipitation yearly as it is. It is important to keep in mind that the data was not collected at an equal distribution between all states.

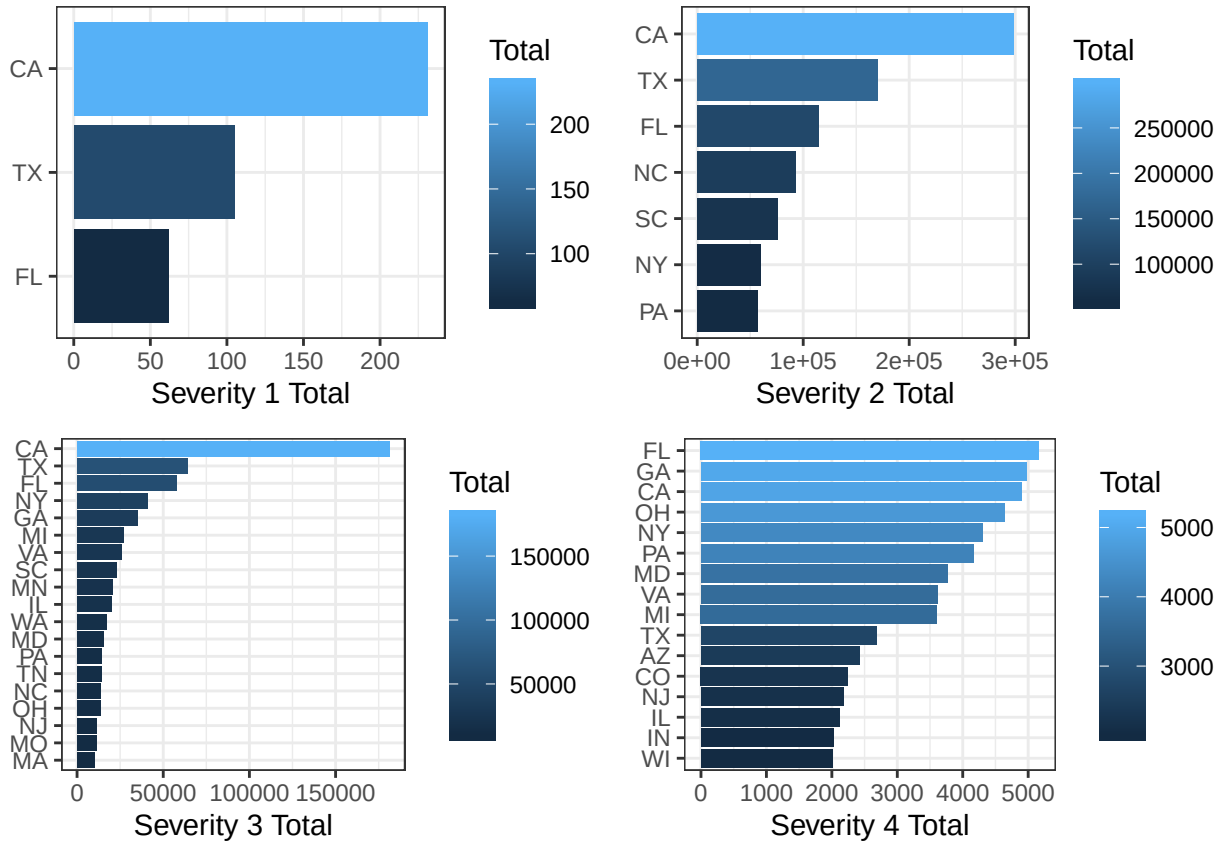
3.1.1 States



At first glance, we can see that California has the most total accidents with Texas and Florida proceeding, however, this is not simply because those states have the most accidents. We must not only take into consideration the proportion of drivers in each state, but also the overall data collected per state. The driver density and collected data go hand in hand because if a large state like California has more people on the road than Rhode Island, then there is a higher likelihood of accidents just due to numbers. The original dataset on Kaggle clears up this issue and mentions that the majority of data was collected from CA, TX, and FL which follows the graph above. While this skews that data when trying to make generalizations for the country as a whole, it is still valuable because the same types of accidents happen all over the country and an autonomous vehicle would not make extreme adjustments due to its location, rather the terrain it is driving on.

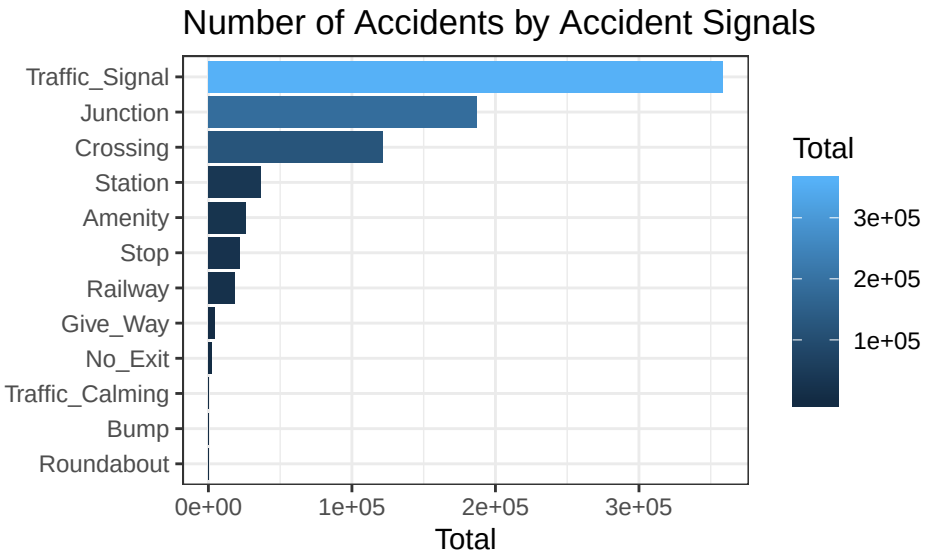
3.1.2 Severity

The graphs below were visualized using the function created by Manual T



The graphs demonstrate that the severity of accidents is correlated with the states where most of the accidents happen. We can see that California is the top state in the first three severity levels, while Florida accounts for most accidents at the highest severity level. While it is interesting to see how the severity varies between each state, it is important to note that because these two variables are highly correlated, we would not be able to make overall assumptions based on the graphics above. Specifically taking a look at the fourth graph, we can see that Florida leads with the most accidents, however, the difference between the other states is not significant when compared to the number of entries we have.

3.1.3 Accident Signals

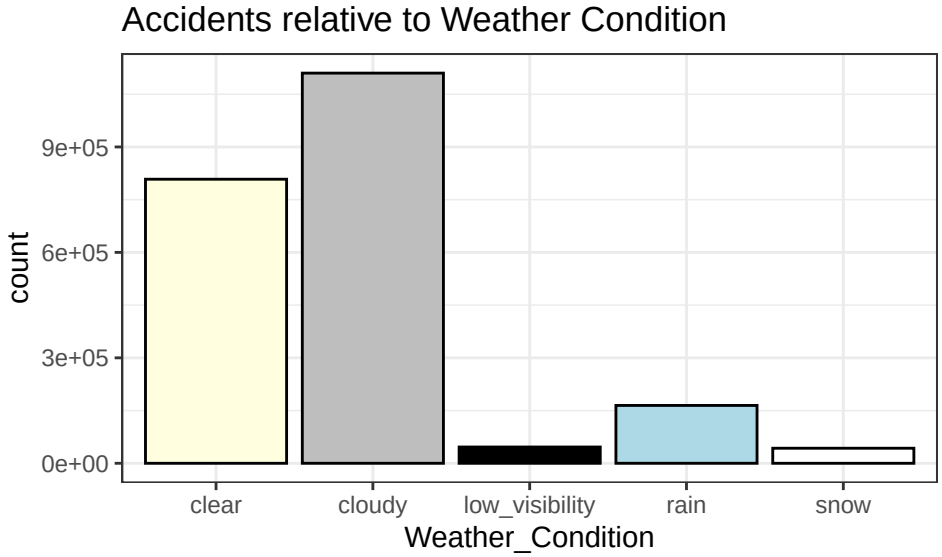


We can see that most accidents tend to happen at Traffic Signals, then Junctions, then Crossings. This information is We can see from the code above produced by Manuel T, most accidents tend to happen at Traffic Signals, Junctions, and Crossings. This information is important when relating it to what would be the best priority when looking at autonomous vehicles. If most accidents happen at traffic signals, then we can try to see all the factors that lead to traffic signals. It is obviously important for image recognition to be able to properly tell the difference between the traffic lights, however, it is also worth thinking about *why* traffic signals have the most accidents. Psychologically speaking, it makes sense that when a driver sees a green light they are more prone to worry less about what they are doing when going through traffic signals because a green light simulates a "go" response. Most of the variation from traffic signal accidents must come from:

- Vehicles coming down the perpendicular street turning right on red and the onward traffic not slowing down because they have the right of way.
- Vehicles making an unprotected left turn getting hit by oncoming traffic that also has a green light indicating they have the right of way.

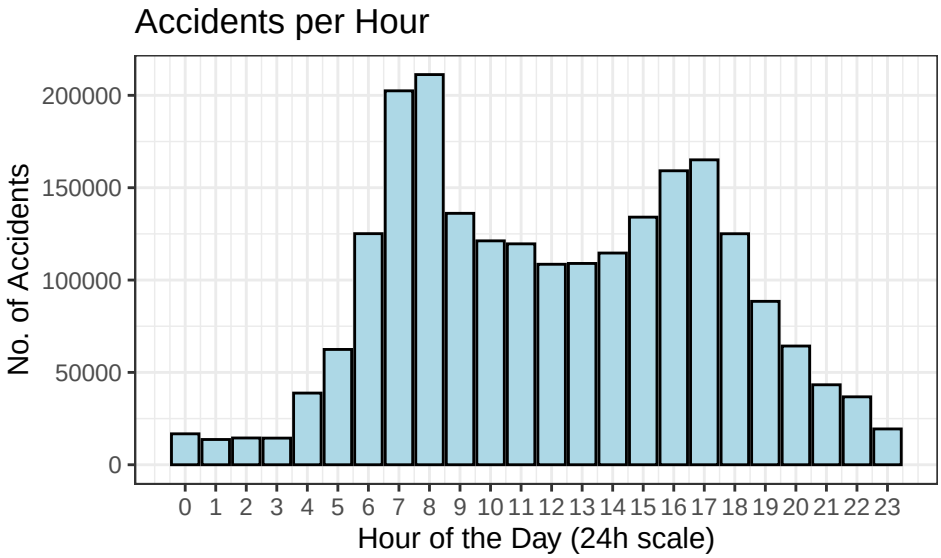
This information checks out with our intuition and current leading statistics. This is the very reason why one might find automated red-light cameras at traffic signals. While they are put in place to prevent accidents and encourage attention, a study in the *Scientific American* found that "when [the cameras] were removed, angle accidents increased by 26 percent. However, all other types of accidents decreased by 18 percent". This was because when people are about to run a red light, instead of making it through the intersection they try to avoid getting fined so they stop as soon as possible, which results in more rear-ended accidents (Gallagher).

3.1.4 Weather Conditions



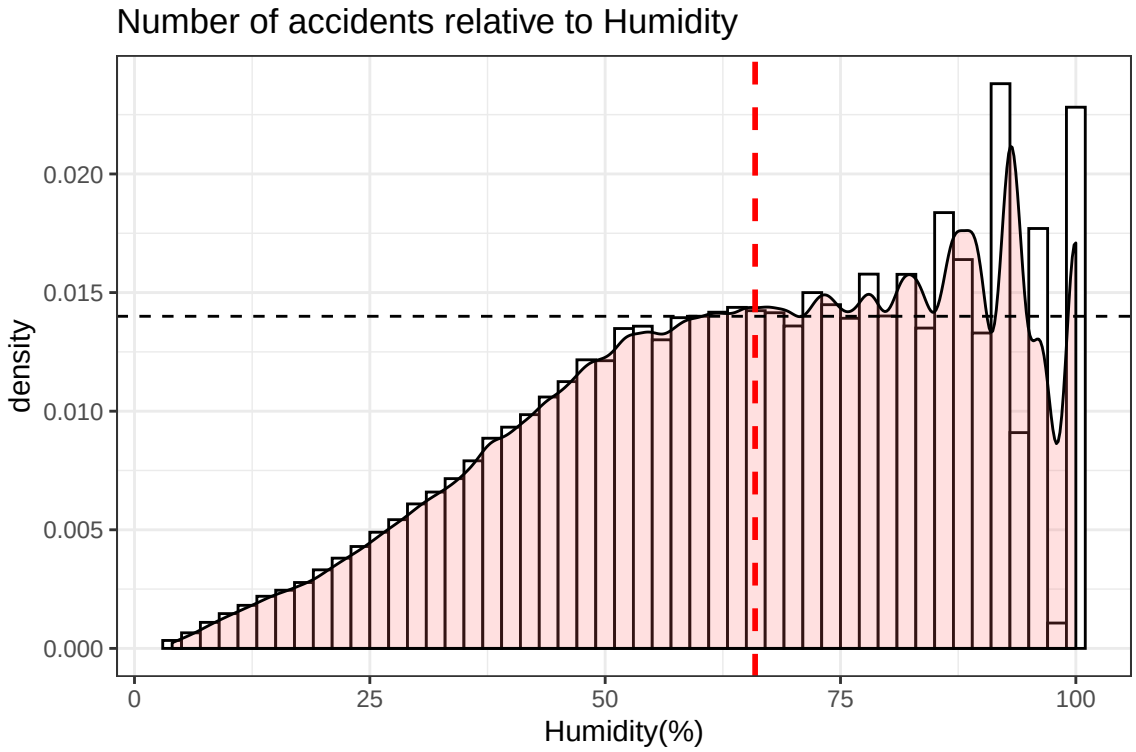
The graph above indicates the number of accidents that occur with each weather condition. Most happen with it is cloudy and clear, and there appear to be very few accidents that occur when there is low visibility, rain or snow, which goes against my initial hypothesis and intuition. It would make sense that most accidents occur when weather elements work against the driver such as low visibility, rain, and snow. While this does not follow my intuition, it *does* follow earlier findings from the initial summary statistics that show that precipitation does not have much significance on accidents.

3.1.5 Time



We can see from the graph above that the accidents distribution tend to be bimodal, with most of them happening around 7-8 am and then 4-5 pm which would follow my intuition because those times are when the majority of people commute to and from work, resulting in rush hour and navigating with more drivers and traffic than the other hours. One thing that stands out to me is that most of the accidents would happen in the morning rush hour times as opposed to the afternoon rush hour times because the NHTSA, along with other studies, has found that most accidents tend to happen on our way's from work, between 3 to 6 pm, because we are more tired and irritable rather than in the morning when we are fresh.

3.1.6 Humidity



One of the biggest significant factors appears to be humidity. Humidity seems to be a factor that would have

equal variance because although most of the data carry weight from three states, the distribution of humidity seems to be average as a whole. Nonetheless, the graph above shows insight that needs to be discovered a little more. We can see that as the humidity starts to rise the number of accidents follow, however, the trend is disrupted at the mean of about 65%. Instead of steadily increasing or leveling out, the distribution begins to become fairly volatile. The humidity variable is unique because it does necessarily directly correlate with the temperature or even the weather condition. High percentages of humidity can be found on rainy days, and with temperatures on the lower side. The volatility of the graph is indicated by the whole area to the right of the benchmark (set at the red vertical line). If we look at the number of accidents that happen in the first and fourth quadrants of the split-up graph, we will find that nearly 54% lie within this range. However, because the graph is so volatile, I am more interested in looking at the first quadrant that is split up by this graph. The total amount of accidents that happened in the first quadrant was 26.7%, which means that if we were able to have fully functioning autonomous vehicles on the road, then we would be able to prevent around 27% of accidents simply from irritability and human error.

4 Results and Applications

An interesting application for traffic signals would be to implement communication between traffic signals and vehicles. This technology is already created and is currently in testing as described by Ivana Kottasova "top automakers including Volkswagen, Honda (HMC), Ford (F) and BMW (BMWYY) are experimenting with technology that allows cars and traffic lights to communicate and work together to ease congestion, cut emissions and increase safety". While this idea is not new, it can be expanded and into the different weights on accidents seen in this study such as traffic junctions and crossings. Moreover, as more research and development is done on fully autonomous vehicles, in addition to the communication between vehicle and infrastructure, we can expect to see vehicle-to-vehicle communication or V2V. There is currently a partnership that is testing V2V communication and they have found that it must not only work wirelessly to share data on location, speed, and direction, but it must also be able to send up to ten messages per second in order to sense what is not immediately in front of them (Krishna).

This study did not find extreme weather conditions to be significant as seen from the graph in section 3.1.4. However, this is not necessarily indicative of the truth because the data collected in this study was heavily weighted by states that do not have much extreme weather such as California and Florida. The results from this study conclude that weather conditions do not play a significant roll in accidents in terms of human error. If we were to only look at the results of the weather conditions we would determine that accidents are prone to happen in any condition and have more significant factors that may go into determining human error. Nevertheless, it is still important that the technology created for autonomous vehicles has the ability to adjust to extreme weather conditions such as icy roads, torrential rains, and low visibility due to thick fog.

While our time variable followed the data during rush hours, the applications from this study would not indicate a need for higher frequency sensors during those times, instead, the focus should be on creating very accurate radar and lidar sensors such that when cars are in constant and close proximity, that they are able to find their way out safely. This goes beyond the scope of this study, however, it is worth noting that if all the vehicles on the road as well as the infrastructure were fully autonomous, then we would theoretically be able to minimize traffic and "rush hour" as a whole.

This study was helpful in concluding the overall causes of accidents in the US. The most significant factors are those that play a role in human error. For the means of this study, humidity was the most significant factor. Due to the fact that this data was taken on for purposes other than simple exploratory analysis, there was not a column of variables that could be a cause of human error such as sleep, diet, happiness, and more on the human condition when getting behind the wheel. However, when it comes to implementing preventative technology in autonomous vehicles, I found that around 27% of accidents can be prevented from human error solely on the humidity.

5 Discussion

5.1 Challenges of the study

The main challenge of this study was working with the dataset. The data cleaning was not as rigorous, however, the size of the data was a factor that held me back from running a random forest as well as creating a model to predict severity. While I was able to write out the code for a random forest, my computer was not able to go through each of the variables for the trees because it would crash. Instead of creating a smaller subset of the data, I chose to leave this out of the study because I did not think that a smaller subset would be useful or significant for my conclusions because it would leave out important factors and I would run the risk miss classifying the data.

The next challenge was working with a dataset that was not collected or meant for the exact purposes of this research. While this dataset was very helpful in the overall initial findings, I am interested in looking at data that has to do with factors of human error. Due to the fact that a great majority of accidents happen due to human error, it would have been compelling to see data that would show exactly how much autonomous vehicles can prevent accidents solely from eliminating all human error.

5.2 Reccomendations for Future Work

Because I was working with accident data that was not specially curated for my research, I would be interested in seeing variables on the human condition to be added to this study. If I am specifically looking at human error vs. what machine learning can correct, I think it would be very helpful to see the main factors that go along with human error that is reported by the NHTSA, such as Alcohol-Related crashes, Speeding, Red Light Running, Fatigue, Distracted Driving, and Cell Phone Use.

In the future, I would want to create a model that predicts the severity, as well as if an accident will occur based on a binary response (0 = no accidents, 1 = accident). In order to do this, it would be best to run a random forest and see which variables have the highest node purity and MSE within the dataset. For a future dataset, instead of the severity of traffic delays, severity should measure the injury of the accident and even include fatalities.

This topic can not be ignored as it is the future of transportation. The role of autonomy is increasing in our society in almost every industry, and it will be the future of how our communities run. Future research should be done on how autonomous vehicles and infrastructure can prevent overall traffic and create a continuous flow. Eventually, a perfect flow of traffic would mean that no one would ever have to drive a car themselves or even own one. All cars would be owned by companies and a person would call their ride on an app and input their destination. While this future has many critics, it is the point to where we are evolving, and in order to ensure the highest possible safety, we must first do detailed and precise research on preventing the randomness that comes along with everyday life.

6 References

"Cost of Auto Crashes & Statistics." RMIIA, www.rmiia.org/auto/traffic_safety/Cost_of_crashes.asp. Gallagher, Justin. "Red Light Cameras May Not Make Streets Safer." Scientific American, Scientific American, 16 Aug. 2018, www.scientificamerican.com/article/red-light-cameras-may-not-make-streets-safer/.

"How Many Car Accidents Does the Average Person Have?" Belsky, Weinberg & Horowitz, LLC, 29 Apr. 2019, www.belsky-weinberg-horowitz.com/how-many-car-accidents-does-the-average-person-have/.

Kottasová, Ivana. "Cars and Traffic Signals Are Talking to Each Other." CNN, Cable News Network, 29 Oct. 2018, www.cnn.com/2018/10/29/business/volkswagen-siemens-smart-traffic-lights/index.html.

Krishna, Swapna. "Self-Driving Cars Are Safer When They Talk to Each Other." Engadget, 19 July 2019, www.engadget.com/2017/06/24/self-driving-cars-mcity-augmented-reality/.

Manuel T. (2020 January). "What causes the accidents (EDA)", Version 17. Retrieved Feb 2020 from https://www.kaggle.com/sobhanmoosavi/us-accidents.

Ray, Aaron. "The Most Dangerous Times on the Road." BACtrack, BACtrack, 30 June 2015, www.bactrack.com/blogs/expert-center/35042821-the-most-dangerous-times-on-the-road.

Teoten. "What Causes the Accidents? (EDA)." Kaggle, Kaggle, 3 Jan. 2020, www.kaggle.com/teoten/what-causes-the-accidents-eda.

7 Appendix

The following code was used to produce this report.

```
# Cleaning the Data
library(readr)
US_Accidents_May19 <- read_csv("~/Desktop/199 Research/US_Accidents_May19.csv")
```

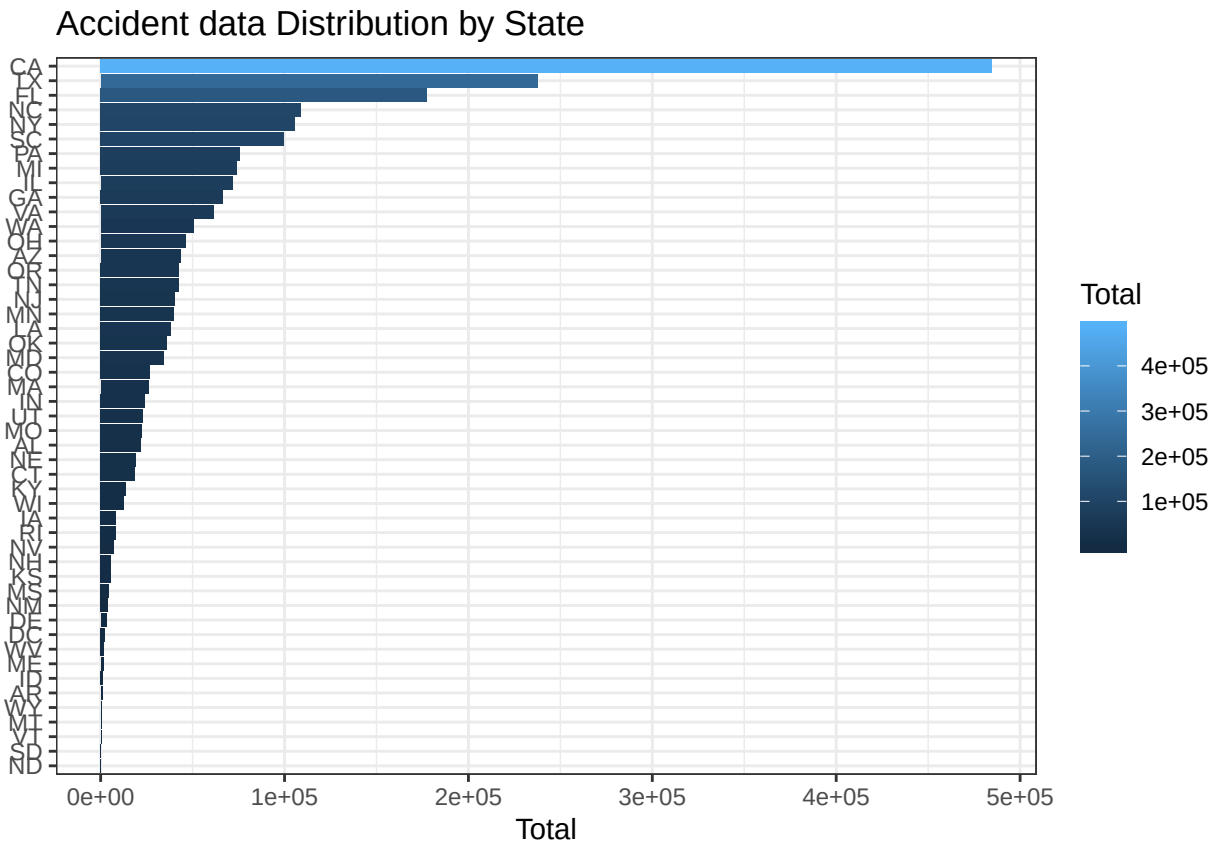
```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   TMC = col_double(),
##   Severity = col_double(),
##   Start_Time = col_datetime(format = ""),
##   End_Time = col_datetime(format = ""),
##   Start_Lat = col_double(),
##   Start_Lng = col_double(),
##   End_Lat = col_logical(),
##   End_Lng = col_logical(),
##   `Distance(mi)` = col_double(),
##   Number = col_double(),
##   Weather_Timestamp = col_datetime(format = ""),
##   `Temperature(F)` = col_double(),
##   `Wind_Chill(F)` = col_double(),
##   `Humidity(%)` = col_double(),
##   `Pressure(in)` = col_double(),
##   `Visibility(mi)` = col_double(),
##   `Wind_Speed(mph)` = col_double(),
##   `Precipitation(in)` = col_double(),
##   Amenity = col_logical(),
##   Bump = col_logical()
##   # ... with 11 more columns
## )
```

```
## See spec(...) for full column specifications.
```

```
# cleaning the data a little bit initially: getting rid of the Source variable becuase they all come fr
accidents <- US_Accidents_May19[,-c(2, 7:10, 19:22)] # Source, Start Lat, Start Lng, End Lat, End Lng,

# NEW TIME VARIABLE

# diff.time is the difference of time in minutes from end to start time of the accident
accidents$Total_Time <- difftime(US_Accidents_May19$End_Time, US_Accidents_May19$Start_Time, units="mins
accidents$Total_Time <- as.numeric(accidents$Total_Time)
summary(accidents$Total_Time) # It is not realistic to have nevative time, nor 1786320 which is 29,772

# Now, we can get rid of the Start_Time  and End_Time variables. We also dont need the Street Number
accidents <- accidents[,-c(4,5,8)]

# Fix Total_Time variable so we dont have negative time values:
accidents$Total_Time = ifelse(accidents$Total_Time > 0, accidents$Total_Time, NA) # got rid of 13 value
# Creating a cap on the amount of time the accident took (getting rid of extreme values)
accidents$Total_Time = ifelse(accidents$Total_Time > 400, NA, accidents$Total_Time) # total of 4,745 NA

# CLEANING WEATHER CONDITIONS:

library(dplyr)
unique(accidents$Weather_Condition)

# With 62 differnt reported weather conditions, we can group them up into 5 main groups: rain, snow, lo

# Rain
accidents <- accidents  %>%
  mutate(Weather_Condition = replace(Weather_Condition,
                                     Weather_Condition == "Light Rain" |
                                     Weather_Condition == "Rain" |
                                       Weather_Condition == "Light Drizzle" |
                                       Weather_Condition == "Heavy Rain" |
                                       Weather_Condition == "Mist" |
                                       Weather_Condition == "Drizzle" |
                                       Weather_Condition == "Rain Showers" |
                                       Weather_Condition ==  "Light Thunderstorms and Rain" |
                                       Weather_Condition == "Light Rain Showers" |
                                       Weather_Condition == "Light Freezing Rain" |
                                       Weather_Condition == "Heavy Drizzle" |
                                       Weather_Condition =="Heavy Thunderstorms and Rain" |
                                       Weather_Condition == "Thunderstorms and Rain" |
                                       Weather_Condition == "Heavy Rain Showers", "rain" ))

# Snow
accidents <- accidents %>% mutate(Weather_Condition = replace(Weather_Condition,
                                                              Weather_Condition == "Light Freezing Drizz
                                                                Weather_Condition == "Light Snow" |
                                                                Weather_Condition == "Snow" |
                                                                Weather_Condition == "Hail" |
                                                                Weather_Condition == "Blowing Snow" |
                                                                Weather_Condition == "Heavy Snow" |
                                                                Weather_Condition == "Ice Pellets" |
```

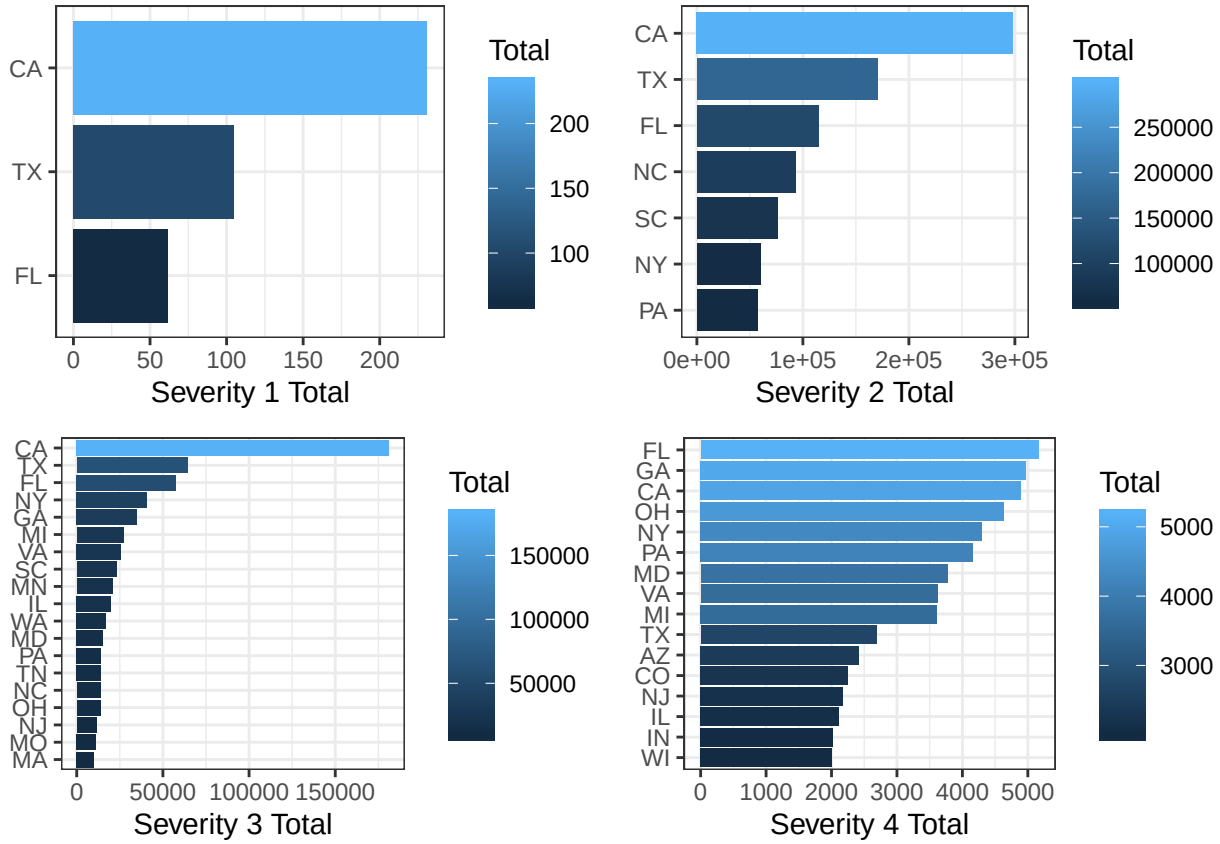
```
                                                    Weather_Condition == "Low Drifting Snow"
                                                    Weather_Condition == "Light Thunderstorm
                                                    Weather_Condition == "Light Ice Pellets"
                                                    Weather_Condition == "Snow Showers" |
                                                    Weather_Condition == "Light Snow Showers
                                                    Weather_Condition == "Heavy Thunderstorm
                                                    Weather_Condition == "Snow Grains" |
                                                    Weather_Condition == "Heavy Blowing Snow
                                                    Weather_Condition ==  "Heavy Freezing Dr
                                                    Weather_Condition == "Light Blowing Snow
                                                    Weather_Condition == "Small Hail" |
                                                    Weather_Condition == "Heavy Thunderstorm
                                                    Weather_Condition == "Light Snow Grains"
                                                    Weather_Condition == "Heavy Ice Pellets"
                                                    Weather_Condition == "Heavy Freezing Rai
                                                    Weather_Condition == "Light Hail" |
                                                    Weather_Condition == "Thunderstorms and

# Low Visibility
accidents <- accidents %>% mutate(Weather_Condition = replace(Weather_Condition,
                                                    Weather_Condition == "Haze" |
                                                    Weather_Condition == "Fog" |
                                                    Weather_Condition == "Shallow Fog" |
                                                    Weather_Condition == "Light Haze" |
                                                    Weather_Condition == "Smoke" |
                                                    Weather_Condition == "Patches of Fog" |
                                                    Weather_Condition ==  "Light Freezing Fo
                                                    Weather_Condition == "Light Fog" |
                                                    Weather_Condition ==  "Dust Whirls" |
                                                    Weather_Condition == "Heavy Smoke" |
                                                    Weather_Condition == "Widespread Dust"
                                                    Weather_Condition ==  "Volcanic Ash" |
                                                    Weather_Condition == "Blowing Sand" |
                                                    Weather_Condition == "Blowing Sand" |
                                                    Weather_Condition == "Funnel Cloud" |
                                                    Weather_Condition == "Sand", "low_visibi

# Cloudy
accidents <- accidents %>% mutate(Weather_Condition = replace(Weather_Condition,
                                                    Weather_Condition == "Overcast" |
                                                    Weather_Condition == "Mostly Cloudy" |
                                                    Weather_Condition == "Partly Cloudy" |
                                                    Weather_Condition == "Scattered Clouds"
                                                    Weather_Condition == "Thunderstorm" |
                                                    Weather_Condition == "Light Thunderstorm
                                                    Weather_Condition == "Squalls" , "cloudy

# Clear
accidents <- accidents %>% mutate(Weather_Condition = replace(Weather_Condition,
                                                    Weather_Condition == "Clear", "clear"))

# Checking for the new variables:
unique(accidents$Weather_Condition) # there are now only 4 groups and then the NA's remain..
```

```r
# We for the visual representation we want to change weather conditions to be a factor with four differ
accidents$Weather_Condition <- as.factor(accidents$Weather_Condition)

# Getting rid of the Weather_Timestamp variable
accidents$Weather_Timestamp <- NULL

# SUMMARY STATISTICS

library(dplyr)
library(xtable)
options(xtable.comment = FALSE)
options(xtable.timestamp = "")
#Summary of Numeric Variables
numsum <- accidents %>%
    select(`Distance(mi)`,
           `Temperature(F)`,
           `Wind_Chill(F)`,
           `Humidity(%)`,
           `Pressure(in)`) %>%
  summary()
numsum1 <- accidents %>%
    select( `Visibility(mi)`,
           `Wind_Speed(mph)`,
           `Precipitation(in)`,
           Total_Time) %>%
  summary()
xtable(numsum)
xtable(numsum1)

# STATE

library(ggplot2)
accidents %>%
    select(State) %>%
    group_by(State) %>%
    summarise(Total = n()) %>%
    ggplot() +
    geom_bar(aes(y = Total,
                 x = reorder(State, Total, FUN = abs),
                 fill = Total),
             stat = 'identity') +
    coord_flip() +
    labs(x = NULL) +
    theme(legend.position="none") +
  theme_bw() +
  ggtitle("Accident data Distribution by State")
```

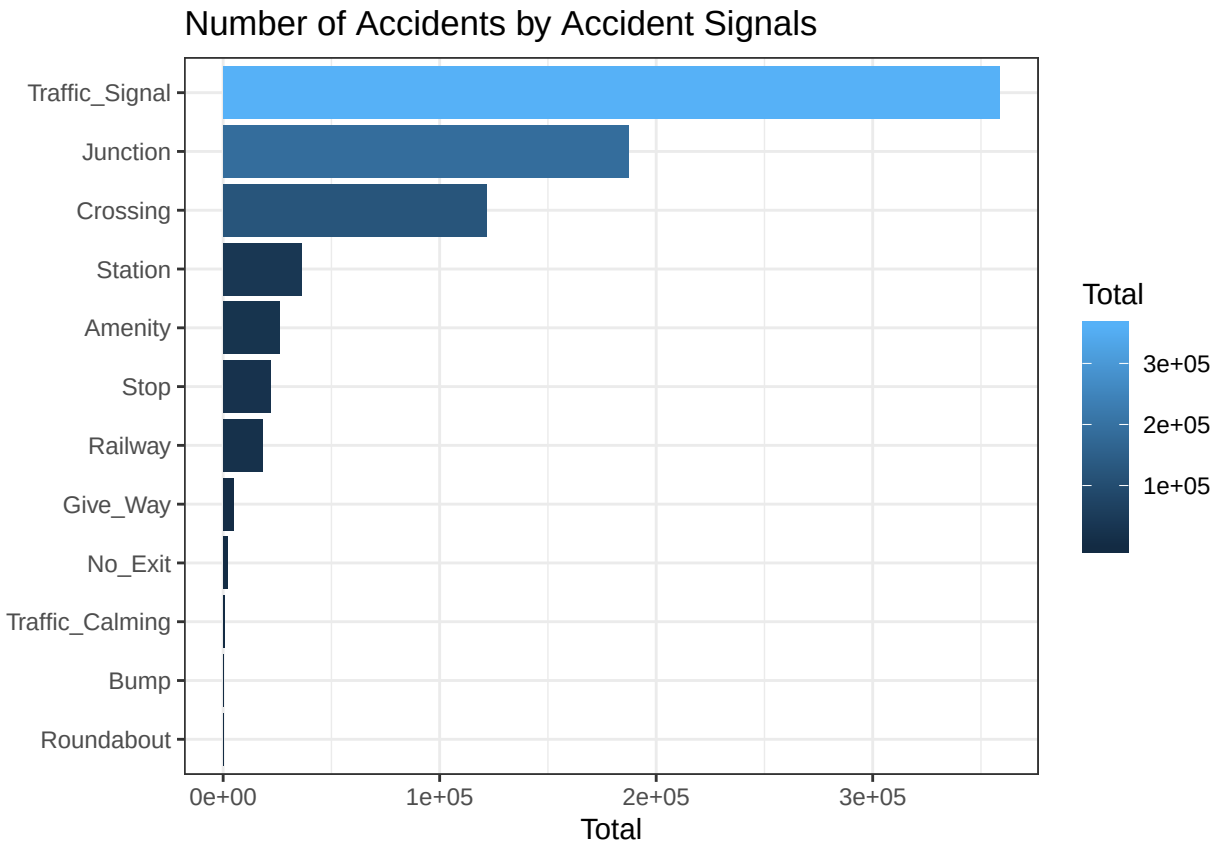



```
# SEVERITY

library(cowplot)
## Function created by Manuel T.
plot_SevState <- function(sev, min = 10000) {
    accidents %>%
        select(State, Severity) %>%
        filter(Severity == sev) %>%
        group_by(State) %>%
        summarise(Total = n()) %>%
        filter(Total > min) %>%
        ggplot() +
        geom_bar(aes(y = Total,
                     x = reorder(State, Total, FUN = abs),
                     fill = Total),
                 stat = 'identity') +
        coord_flip() +
        labs(x = NULL, y = paste("Severity", sev, "Total")) +
        theme(legend.position="none") +
    theme_bw()
}

## Visualization
plot_grid(
    plot_SevState(1, min = 50),
    plot_SevState(2, min = 50000),
    plot_SevState(3),
```

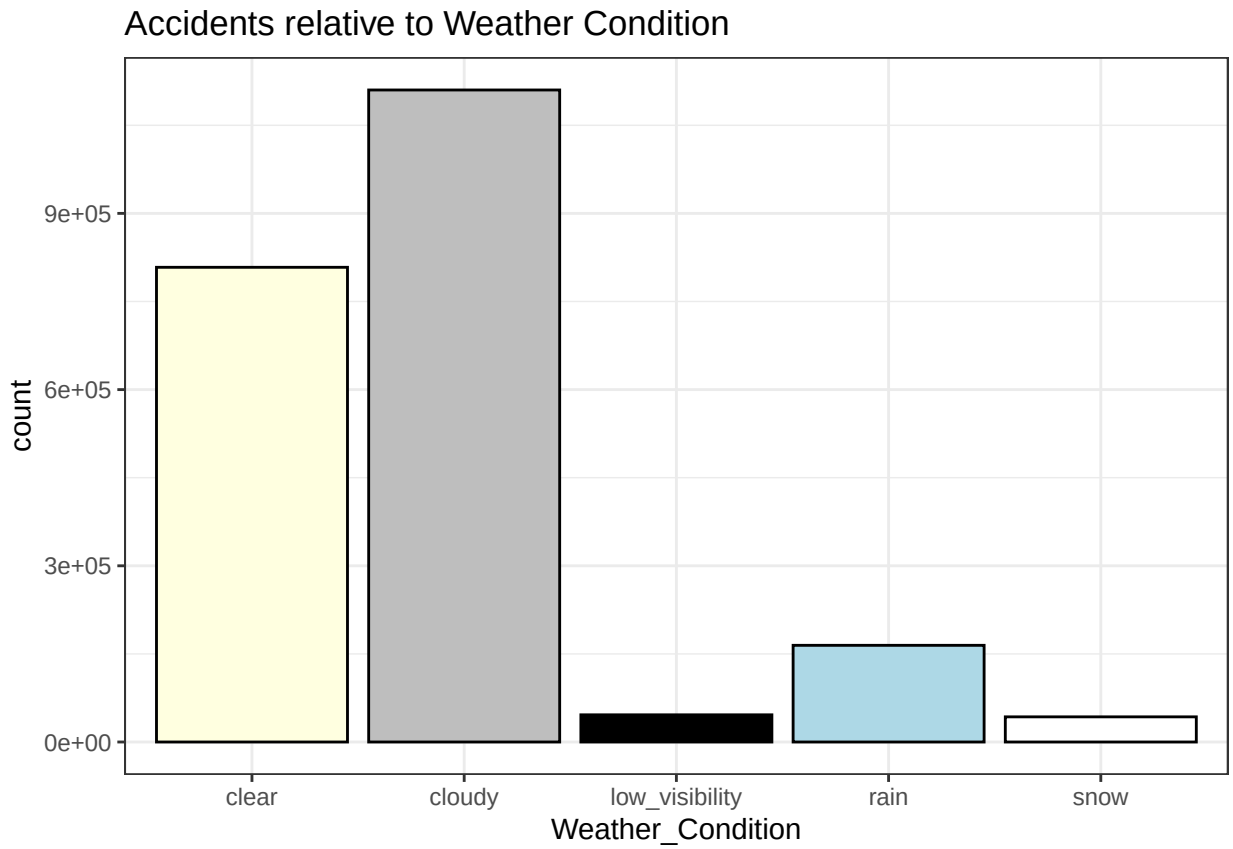
```
    plot_SevState(4, min = 2000))
```



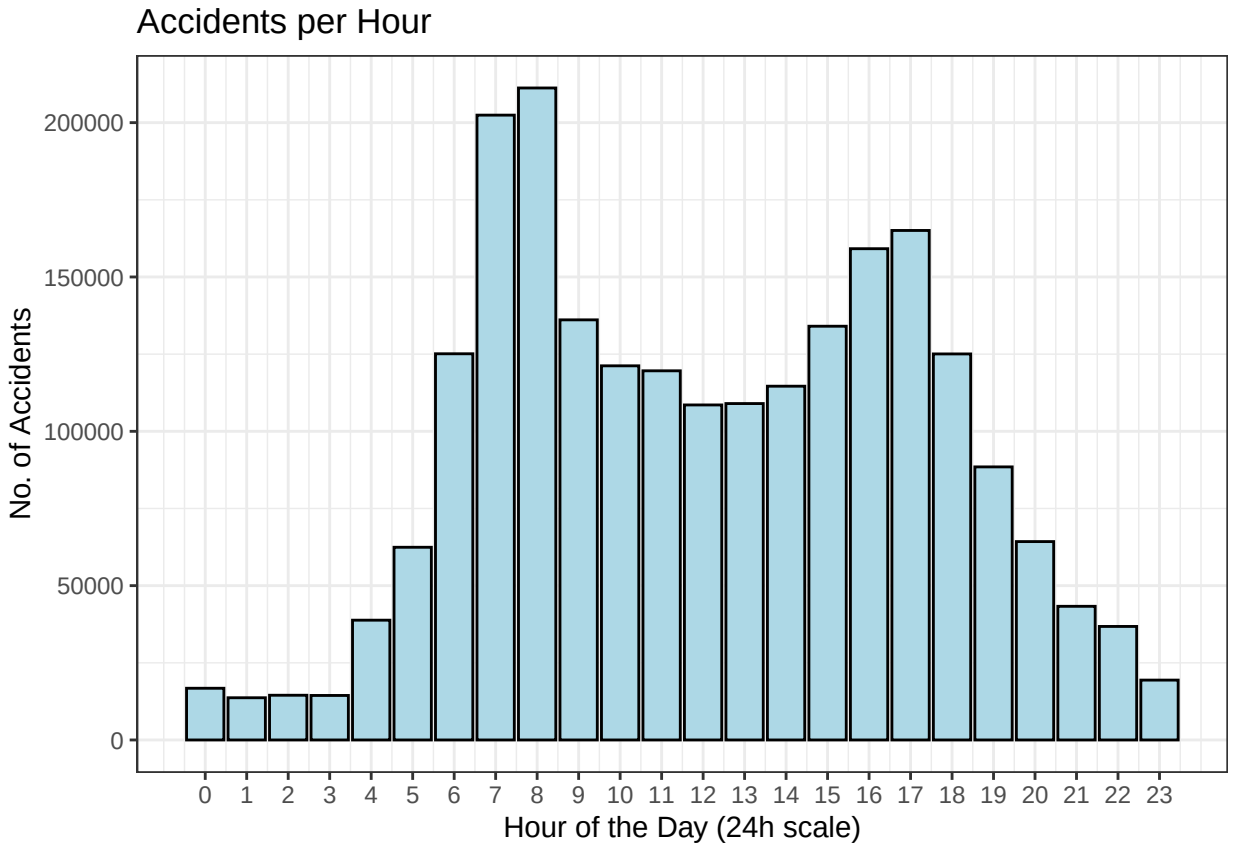

```
# ACCIDENT SIGNALS

library(tidyr)
library(ggplot2)
signals <- accidents %>%
    select(Amenity:Turning_Loop) %>%
    pivot_longer( cols = Amenity:Turning_Loop,
        names_to = 'Annotation',
        values_to = 'Trues') %>%
    filter(Trues == TRUE) %>%
    group_by(Annotation) %>%
    summarise(Total = n())

signals %>%
    ggplot() +
    geom_bar(aes(y = Total,
                 x = reorder(Annotation, Total, FUN = abs),
                 fill = Total),
             stat = 'identity') +
    coord_flip() +
    labs(x = NULL) +
    theme(legend.position="none") +
  theme_bw() +
  ggtitle("Number of Accidents by Accident Signals")
```



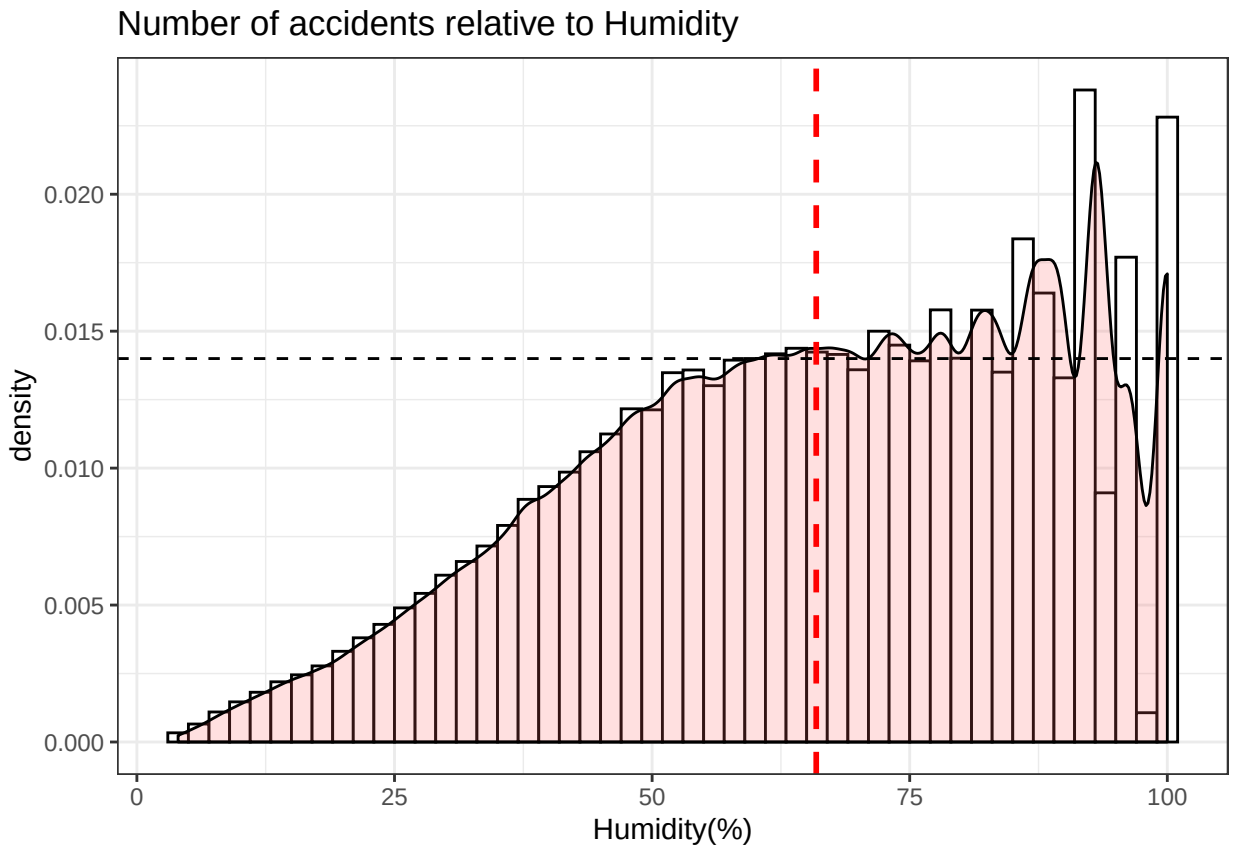

```
# WEATHER CONDITIONS

library(ggplot2)

ggplot(data = subset(accidents, !is.na(Weather_Condition)) , aes(x = Weather_Condition)) +
  geom_bar(color = "black", fill = c("light yellow", "grey", "black", "light blue", "white")) +
  theme_bw() +
  ggtitle("Accidents relative to Weather Condition")
```




```
# TIME

library(lubridate)
ggplot(US_Accidents_May19, aes(x = hour(Start_Time))) +
  geom_histogram(stat = 'count', color = "black", fill = "light blue") +
  theme_bw() +
  labs(x = "Hour of the Day (24h scale)", y = "No. of Accidents") +
  ggtitle("Accidents per Hour") +
  scale_x_continuous(breaks = seq(0, 23, by = 1))
```




```
# HUMIDITY

ggplot(data = subset(accidents, !is.na(`Humidity(%)`)), x = `Humidity(%)`) +
  geom_histogram(aes(x = `Humidity(%)`, y = ..density..), binwidth = 2, color = "black", fill = "white"
  geom_density(aes(x = `Humidity(%)`, y = ..density..), alpha = 0.2, fill = "#FF6666") +
  theme_bw() +
  geom_vline(aes(xintercept = mean(`Humidity(%)`, na.rm = T)), color = "red", linetype = "dashed", size
  geom_hline(yintercept = 0.014, color = "black", linetype = "dashed", size = .5) +
  ggtitle("Number of accidents relative to Humidity")
```


Number of accidents relative to Humidity
density
0.020
0.015
0.010
0.005
0.000
0
25
50
75
100
Humidity(%)