Final Project - US Accidents

Travis Ritter and Logan Cole

Due May, 2, 2022

Introduction

Dataset We chose a data set that includes information about over 2.8 million car accident in the United States from the February 2016 through December 2021. It can be found here: https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

What data does this data set contain? It contains many variables and, as mentioned, over 2.8 millions observations. Some of the variables we plan to use include things like weather conditions at time of accident, location of the accident, and at which type of roadway feature did it happen at (i.e. stop sign, traffic lights, roundabout, etc.). It contains much more than we will likely use (47 variables), but those are just a few that we will be using.

Why did we choose this data set? We chose it because it seemed interesting, it was sufficiently large and messy to deem worthy of our final project. Also, the questions that could be answered from it are relevant to most and interesting to propose and try to answer

Why should people care? For our data set this is an easy question, it will hopefully, visually explain to readers how accidents happen. Also, what affects accident's likelihood and/or severity and what they could potentially do to avoid such accidents in the future.

Load in Data

```
data <- read.csv("US_Accidents_Dec21_updated.csv",
                 na = c("", " "))
```

Note: The order of our problems here, may not reflect their order on our power point presentation.

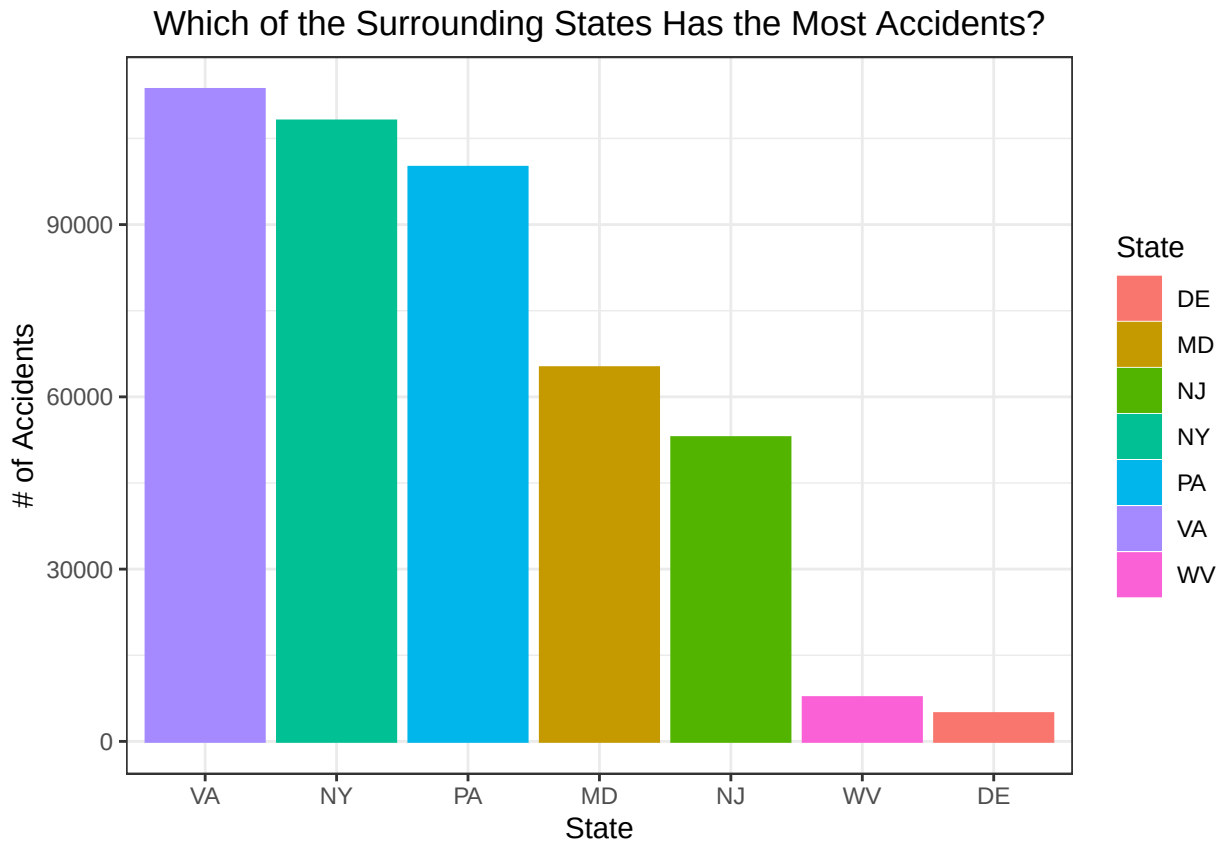
Question #1: Which Neighboring States of PA Have the Most Accidents?

```
stateDrivers <- data %>%
  select(State) %>%
  drop_na(State) %>%
  filter(State == "VA" |
         State == "NY" |
```

```
          State == "PA" |
          State == "MD" |
          State == "NJ" |
          State == "WV" |
          State == "DE")

p1 <- ggplot(stateDrivers, aes(x = State, color = State, fill = State)) +
  geom_bar() +
  scale_x_discrete(limits = c("VA", "NY", "PA", "MD", "NJ", "WV", "DE")) +
  xlab("State") +
  ylab("# of Accidents") +
  ggtitle("Which of the Surrounding States Has the Most Accidents?") +
  theme(plot.title = element_text(hjust = 0.5))

(p1)
```



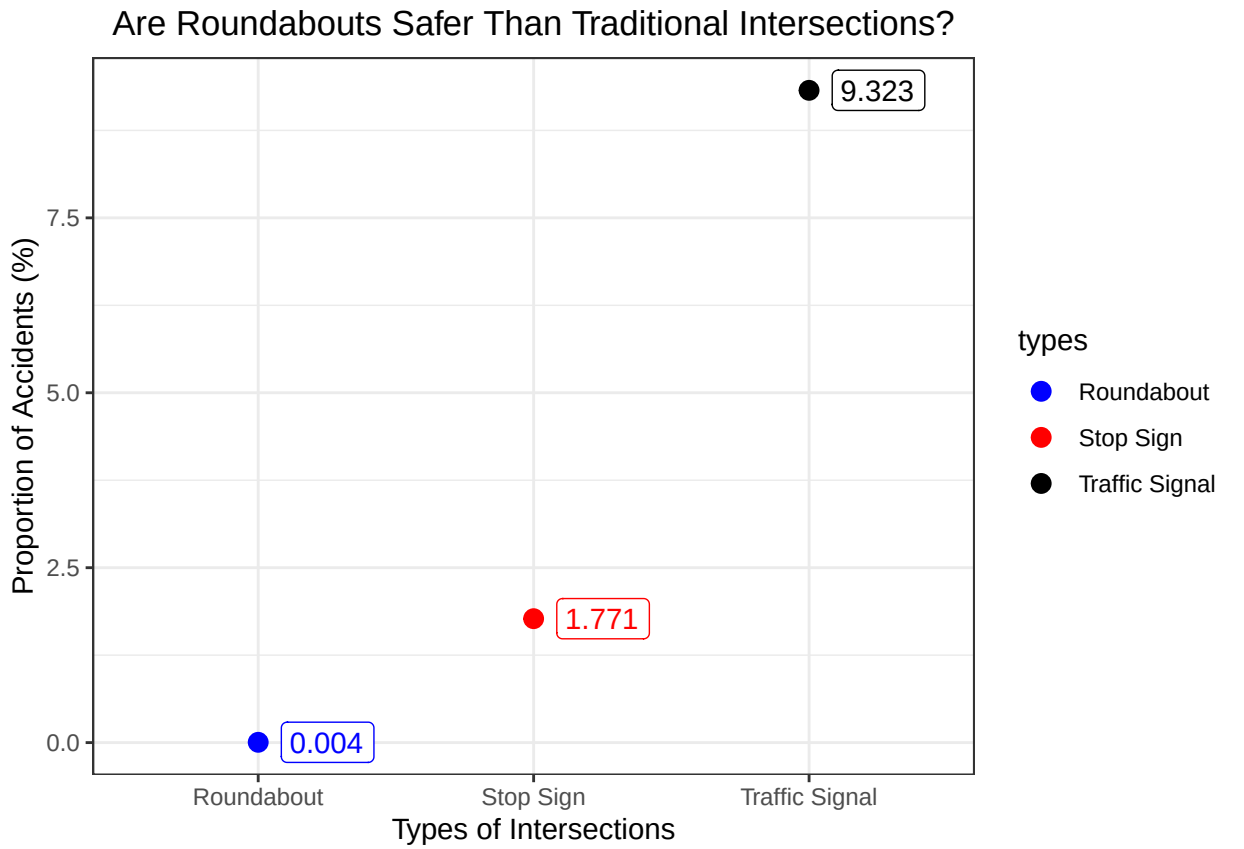
Comment: For this graph we decided to limit the number of states we looked at to just a few of those close to PA, to simply make it easier to digest and compute. What we see after we tidy the data and graph it is that Virginia has the most accidents, followed by New York and PA. Obviously, these are highly populated states, so this should be no suprise, but visually representing the actual difference is interesting and we think worthwhile.

Question #2: Are Roundabouts Safer Than Traditional Forms of Intersection Management?

```
intersections <- data %>%
  select(Roundabout, Stop, Traffic_Signal) %>%
  drop_na(Roundabout, Stop, Traffic_Signal) %>%
  transmute(Roundabout = Roundabout == "True",
            Signal = Traffic_Signal == "True",
            Stop = Stop == "True") %>%
  summarize(Roundabout = mean(Roundabout == T) * 100,
            "Stop Sign" = mean(Stop == T) * 100,
            "Traffic Signal" = mean(Signal == T) * 100)   %>%
  pivot_longer(cols = c(Roundabout, "Stop Sign", "Traffic Signal")) %>%
  transmute(types = name, proportion = value)

p1 <- ggplot(intersections, aes(x = types, y = proportion,
                                color = types, label = round(proportion, 3))) +
  geom_point(size = 3) +
  geom_label(show.legend = F, hjust = -.25) +
  scale_color_discrete(type = c("blue", "red", "black")) +
  xlab("Types of Intersections") +
  ylab("Proportion of Accidents (%)") +
  ggtitle("Are Roundabouts Safer Than Traditional Intersections?") +
  theme(plot.title = element_text(hjust = 0.5))

(p1)
```

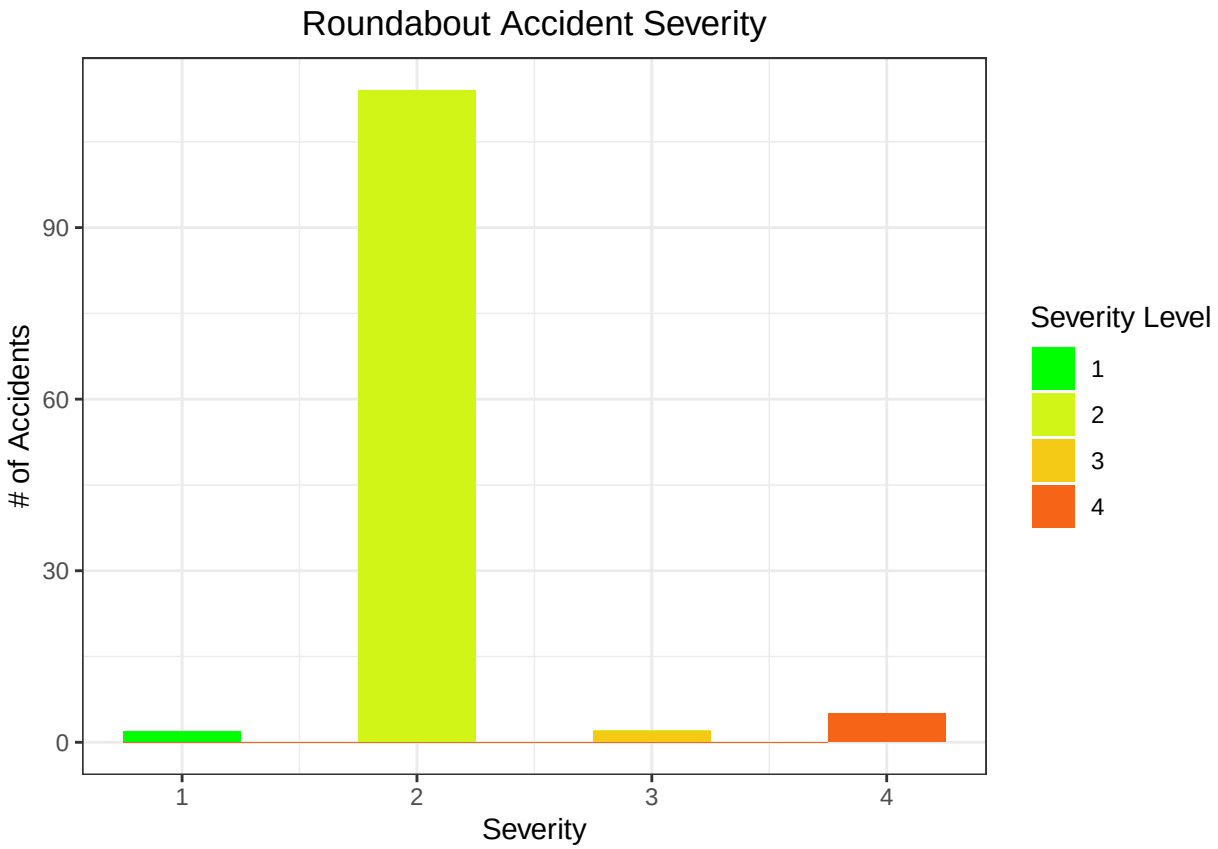


Comment: Here we filtered the data and summarized it to get the calculated proportion of accidents that happened at them. We chose stop signs and traffic signals as they are the most common form of intersection management, and compared the proportion of all the accidents in our data set that happened at each one. What we found was interesting, roundabouts only represent four thousandths of a percent of all accidents. If this is due to their safety or the fact that they are not widely used in America, or both, is hard to pinpoint, but it does make a good case for their safety. Stop signs are higher at around 2%, and out of all accidents almost 10% happen at or around traffic lights. So be careful next time you want to speed through a yellow light!

Question #3: We Know Few Accidents Happen at Roundabouts, but are They Especially Severe?

```
severity <- data %>%
  select(Severity, Roundabout) %>%
  filter(Roundabout == "True") %>%
  drop_na(Roundabout, Severity)

ggplot(severity, aes(x = Severity, fill = as.factor(Severity))) +
  geom_histogram(binwidth = .5) +
  ylab("# of Accidents") +
  ggtitle("Roundabout Accident Severity") +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_fill_discrete("Severity Level", type = c("green", "#d1f517", "#f5ca17", "#f56417"))
```

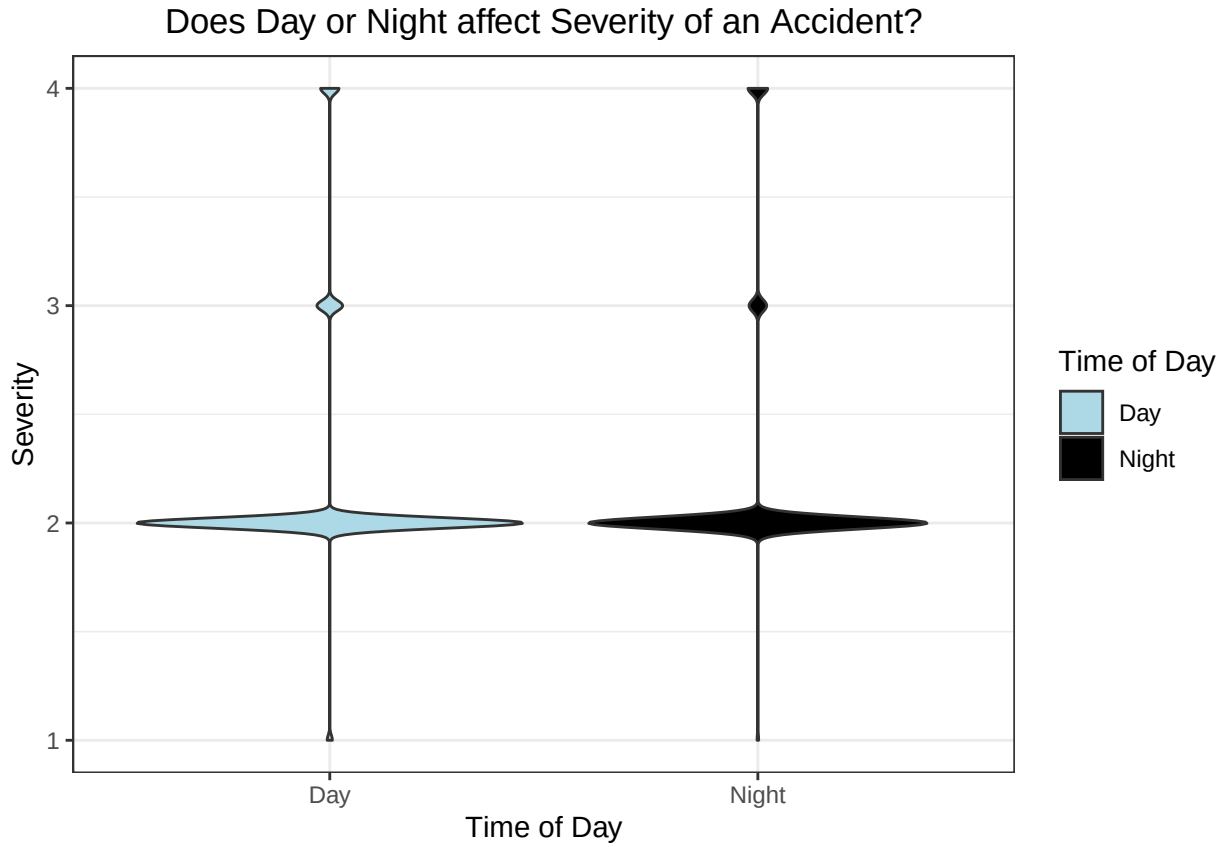


Comment: What this graph is telling us is no, of the very few accidents that happen at roundabouts, they are not particularly severe. It is worth noting severe in the context of this data set means how much traffic was impeded due to said accident. So not only do few accidents happen, but they are generally low scale accidents. This is evidence of the usefulness of roundabouts that America should try to utilize.

Question #4: Day vs. Night, Which One is More Dangerous?

```
timeOfDay <- data %>%
  select(Sunrise_Sunset, Severity) %>%
  drop_na(Sunrise_Sunset, Severity)


ggplot(timeOfDay, aes(x = Sunrise_Sunset, y = Severity, fill = Sunrise_Sunset)) +
  geom_violin() +
  scale_fill_discrete("Time of Day", type = c("light blue", "black")) +
  xlab("Time of Day") +
  ggtitle("Does Day or Night affect Severity of an Accident?") +
  theme(plot.title = element_text(hjust = 0.5))
```



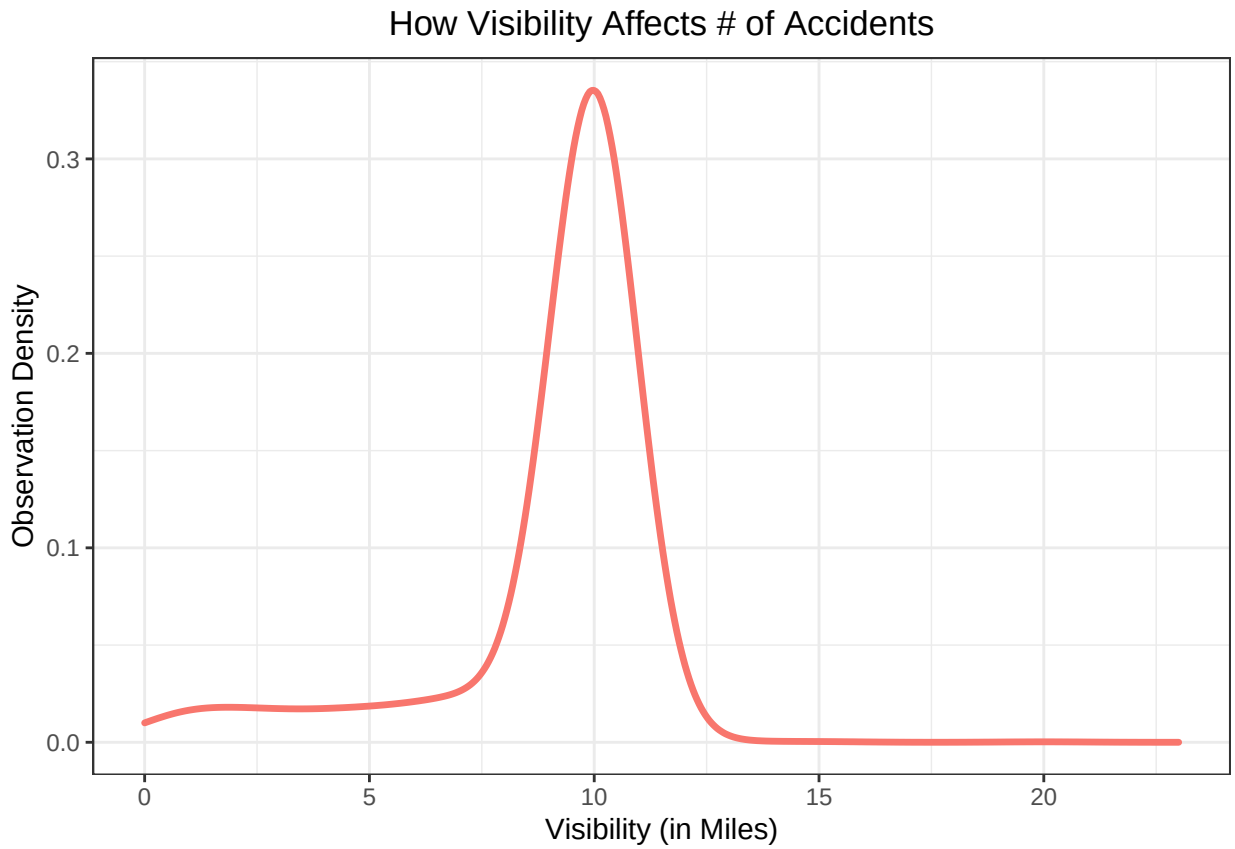
Comment: In this section we compare the severity of crashes that happen during daytime and night time, we filter the data and graphed it. Common sense would tell us that night should be more dangerous, because there is less of the road visible. However, this graph shows us that the severity of the of crashes is not really better or worse in day or night time. However it would stand to reason that less people driver during the night, yet the amount of accidents between the two is very similar. So it would take further investigation to see the proportion of drivers on the road to accidents, during day or night.

Question #5: How Does Visibility Affect the # of Accidents?

```
visibility <- data %>%
  select(Visibility.mi.) %>%
  drop_na(Visibility.mi.) %>%
  filter(Visibility.mi. < 25)

p1 <- ggplot(visibility, aes(x = Visibility.mi., color = "red")) +
  geom_density(adjust = 9, size = 1.2, show.legend = F) +
  xlab("Visibility (in Miles)") +
  ylab("Observation Density") +
  ggtitle("How Visibility Affects # of Accidents") +
  theme(plot.title = element_text(hjust = 0.5))

(p1)
```



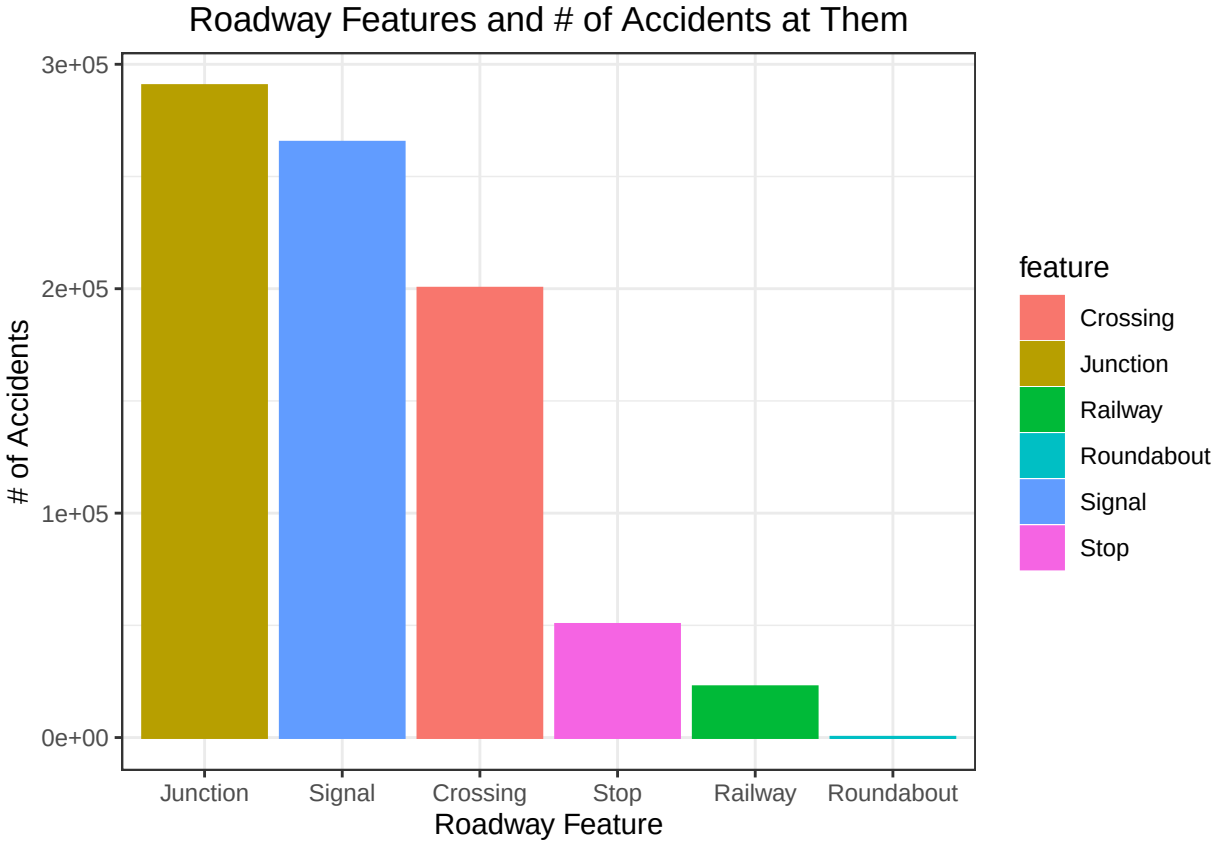
Comment: For this section we filtered out just the visibility information, and this gave us a long list of visibility values. The numbers correspond to the average visibility in the location in miles. So, the higher the value the less obscured the road in front of you is. Surprisingly, the majority of crashes happen when visibility is relatively high, at 10 miles. One would expect that more crashes would happen when visibility is dampened, which is sort of the case because there are very few accidents after a visibility of around 10; so more visibility may mean less accidents. This is likely due to the fact that on any day that is not unusually foggy or rainy, visibility stays fairly static.

Question #6: Which Roadway Features Contributes to the Most Accidents?

```
features <- data %>%
  select(Junction, Traffic_Signal,
         Crossing, Stop,
         Railway, Roundabout) %>%
  transmute(Junction = Junction == "True",
            Signal = Traffic_Signal == "True",
            Crossing = Crossing == "True",
            Stop = Stop == "True",
            Railway = Railway == "True",
            Roundabout = Roundabout == "True") %>%
  pivot_longer(cols = c(Junction, Signal,
                        Crossing, Stop,
                        Railway, Roundabout)) %>%
  filter(value) %>%
  transmute(feature = name)
```

```
p1 <- ggplot(features, aes(x = feature, fill = feature, color = feature)) +
  geom_bar() +
  scale_x_discrete(limits = c("Junction", "Signal", "Crossing",
                              "Stop", "Railway", "Roundabout")) +
  xlab("Roadway Feature") +
  ylab("# of Accidents") +
  ggtitle("Roadway Features and # of Accidents at Them") +
  theme(plot.title = element_text(hjust = 0.5))

(p1)
```



Comment: For this section, we selected a few prominent roadway features (Junctions, Traffic Signals, Crosswalks, Stop Signs, Railway Crossings, and Roundabouts) and compared the amount of crashes that happened at each one. Junctions and Signals are both close contenders for most dangerous, with nearly 300,000 incidents of crashes each! Stop signs are less dangerous, and due to their prominence, the fact that they are not the most is proof positive in favor of the idea that stop signs are a decent way to control traffic flow. Again, roundabouts are proven to be effective, as we can barely see the line that roundabouts created as its so few.

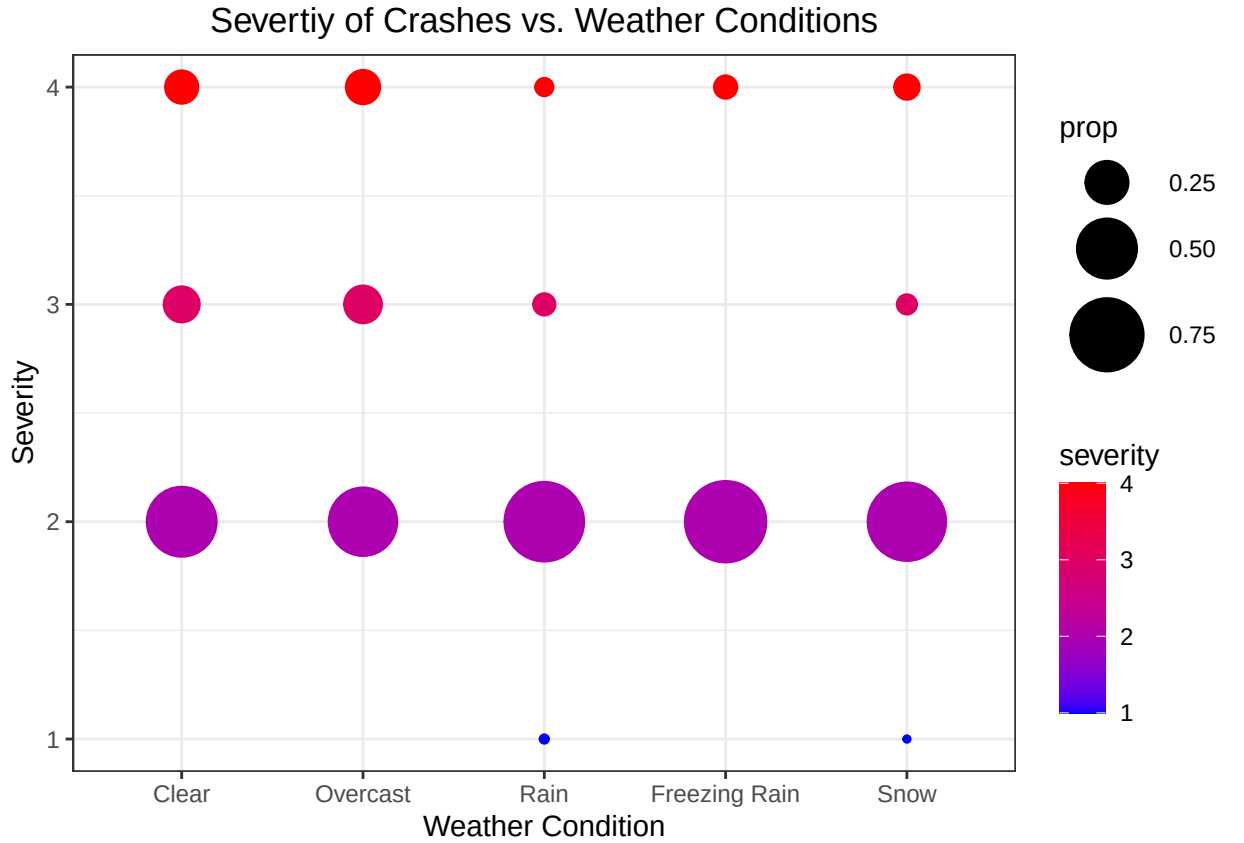
Question #7: How do Weather Conditions Affect the # of Accidents

```
weather <- data %>%
  select(Weather_Condition, Severity) %>%
  drop_na(Weather_Condition, Severity) %>%
```

```
  filter(Weather_Condition == "Snow" |
         Weather_Condition == "Overcast" |
         Weather_Condition == "Freezing Rain" |
         Weather_Condition == "Rain" |
         Weather_Condition == "Clear")

p1 <- ggplot(weather, aes(x = as.factor(Weather_Condition),
                          y = Severity, color = Severity)) +
  geom_count(aes(size = after_stat(prop))) +
  scale_size_area(max_size = 14) +
  scale_color_gradient(name = "severity", low = "blue", high = "red") +
  scale_x_discrete(limits = c("Clear", "Overcast", "Rain", "Freezing Rain", "Snow")) +
  xlab("Weather Condition") +
  ylab("Severity") +
  ggtitle("Severtiy of Crashes vs. Weather Conditions") +
   theme(plot.title = element_text(hjust = 0.5))

(p1)
```



Comment: Here we selected just five out of dozens of different potential weather conditions available in our data set, because more than that and the plot looks cluttered. What we found was interesting, as it seems that no matter the weather, most accidents tend to be rated a two out of at most four on the severity scale. What we hypothesized was the inclement weather like rain, snow, etc. would have a higher proportion of more sever accidents, but actually we found the exact opposite! Days that were clear or cloudy actually had a higher proportion of level four severity accidents. This may be due to people being less attentive because they feel falsely secure with the conditions of the roads. Either way it is interesting to proved wrong by

data.

Overall Findings

It is hard to come up with a one-for-all conclusion, and conclusion is not even the right word because there is always some degree of doubt on our findings. However we can make a few assumptions/speculations about what our visualizations demonstrate:

Speculation #1: Roundabouts appear to be safer than other traditional methods of intersection management. However, our proof for this is that there are few accidents at roundabouts and those accidents tend to be not severe, and the fact is that America does not have many roundabouts, which diminishes the significances of these findings.

Speculation #2: Visibility and Weather Conditions do not affect accident proneness or severity as much as one would reasonably think. We can see from our plots regarding visibility and weather, that accidents happen even when visibility is high and the weather is clear. In fact, especially regarding weather, it is actually the opposite case, as seen by the fact that there were a higher proportion of highly severe crashes when it was clear, than when it was snowing or raining! The reason for this may be due simply to the volume of vehicular and pedestrian traffic on nice days vs. inclement days, but that cannot be said for certain. Another thing we thought, was that perhaps people are lured into a sense of false security when the roads are in favorable condtion, leading to nonchalance and more accidents (again, cannot be proven).

Speculation #3: Time of day does not affect accident severity. Based on our plot about it, the severity of crashes was extremely similar between day and night. The difference is not big enough to prove one way or the other that day or night is more dangerous than the other. However, since there were about the same number of observations for both night and day, and it stands to reason that there are less drivers at night. Perhaps night time driving actually is more dangerous than daytime, as would seem appropriate. This cannot be concluded with more certainty without supplemental research.