Understanding Traffic Accidents in the United States

Christian Steinhofer, Jimi Abbott, Jerry Cai, Jessica Duan

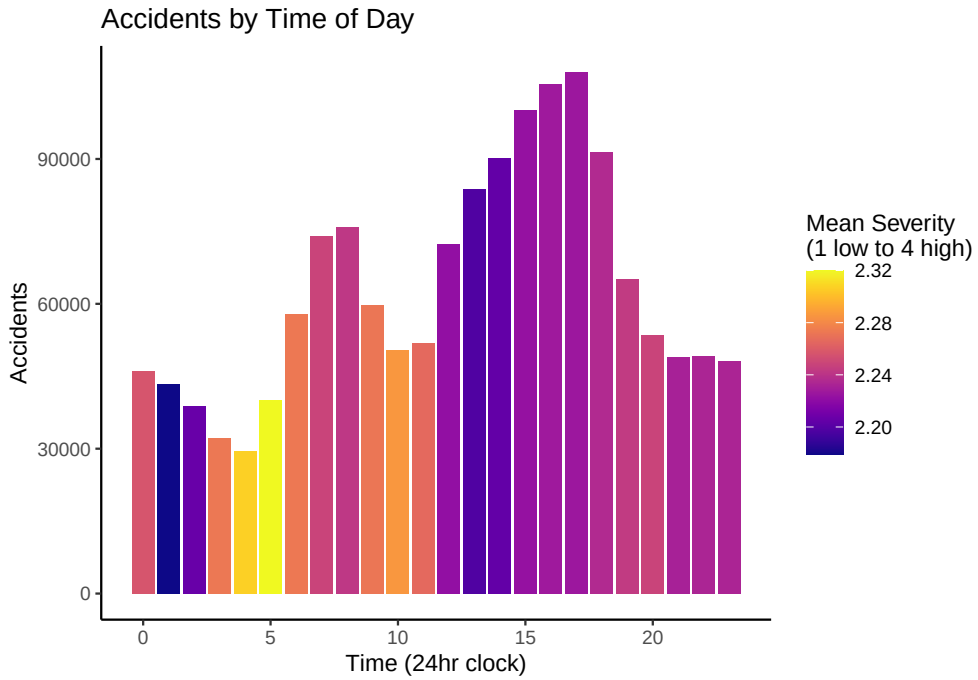
Dateset Introduction

The US Bureau of Transportation Statistics estimates that there were roughly 276 million registered highway vehicles in the US in 2019. In these vehicles Americans traveled over 3 trillion miles that same year. However, one effect of high automobile usage is a large number of traffic accidents and crashes: about 7 million in 2019. With that, we did research on US accidents and found a nationwide traffic accident dataset which contains details about car accidents in 49 states of the United States. The dataset contains information that was collected from February 2016 to December 2020. The data are collected through multiple APIs which are captured by institutions such as US and state departments of transportation. Our goal is to understand reasons that cause the accidents and variables that correlate with the severity of the accidents, which is key to further application such as predicting car accidents and the effect that it may have.

Visualization

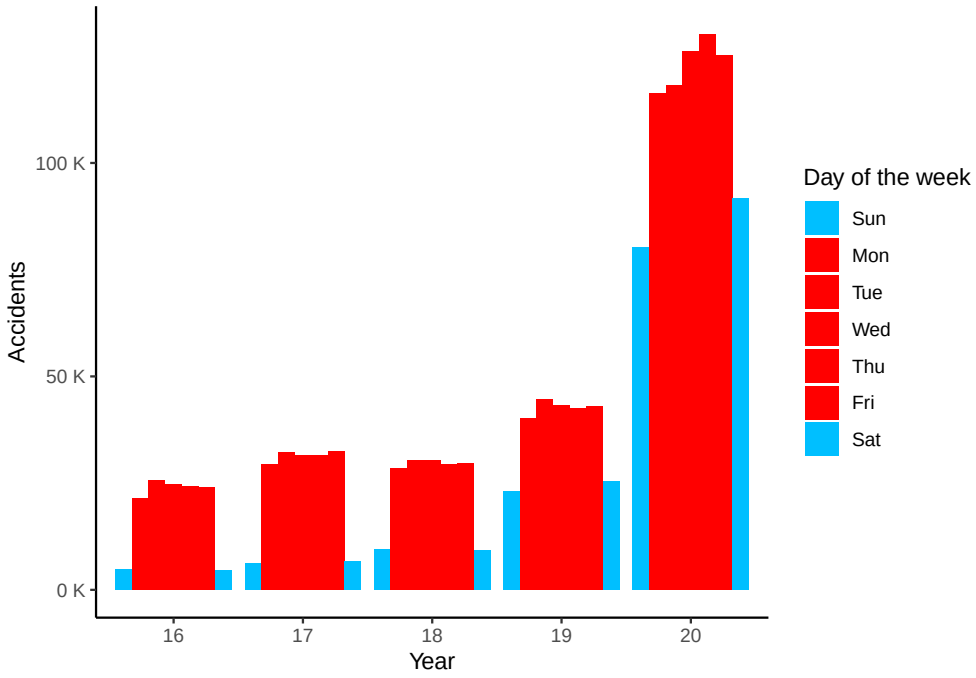
The dataset contains 47 variables. The first thing we decided to do is to drop values that do not have much relation with severity and number of accidents. There are variables that contains True and False values, and all of them are more than 90% False such as Junction, No_exit. Variables like Civil_Twilight, Nautical_Twilight,Astronomical_Twilight define day and night in different ways but we have the exact time of the accidents and we are able to classify them into more detailed time intervals. After cleaning the dataset, we are left with the time the accidents happened, the severity of the accidents, the location of the accidents including states,high ways,and coordinates, and different weather conditions. We will show visualization of those variables to see if there are any patterns.

Accidents by Hour



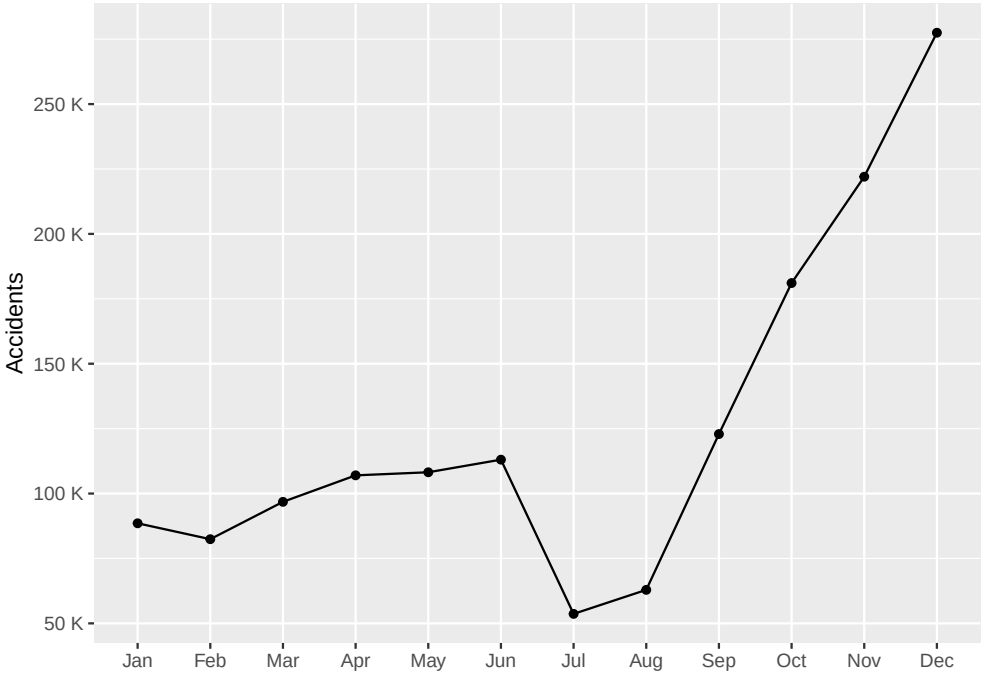
This is a plot showing the accidents that have occurred in each hour of the day, with the fill color showing the average severity of these accidents with each hour. The very first thing we notice is that most accidents happen during the work day. More specifically, we see that at the start of the work day or when people start driving to work, the number of accidents spike. Then, the next couple of hours they die down, but during lunch break, around 12-2 they spike back up again and keep increasing until the end of the average workday, around 5-6 pm when they start to trend down again. In terms of severity, most time intervals are pretty close together, average being centered around severity level 2 out of 4. Interestingly, the most severe accidents happen very early in the morning, which could be attributed to when driving is most reckless due to either tiredness or other physical factors. Thus, we decide the exact time of the day may be a factor.

Accidents by Day of the Week



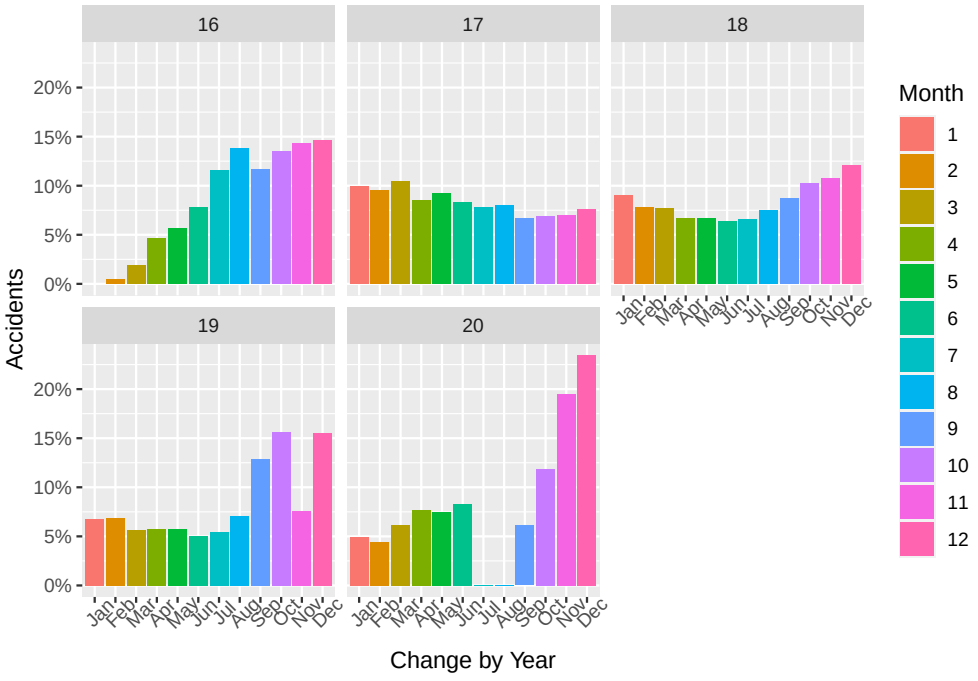
Next, we have a plot that shows the total number of accidents for each day by each year.The main thing we're looking at here is the difference between weekends and weekdays, and more importantly the difference between weekends and workdays. Here, we see that the majority of accidents occur on workdays, during higher stress times because in workdays people need to drive which means that there will be more vehicles on the road and higher probability of accidents.

Change by Month



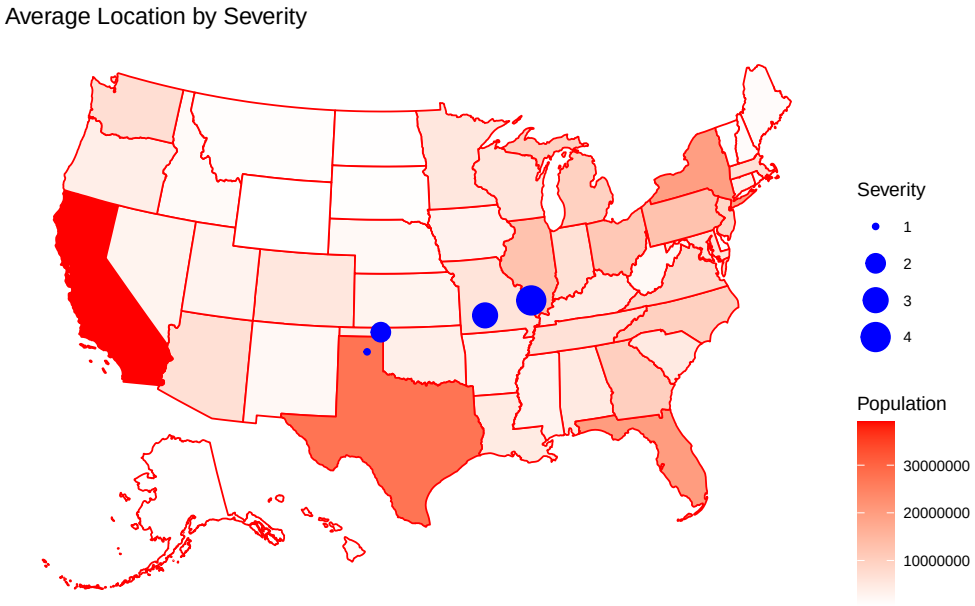
Now, we are inspecting the accidents that are classified into each month. From the plot, we learn that there is a sudden increase from July and August to September, reaching a high spot in November and December,and a sudden decrease in January. We believe that the main reason is because of big holidays including thanksgiving and Christmas along with reopening of schools.

Change by Year



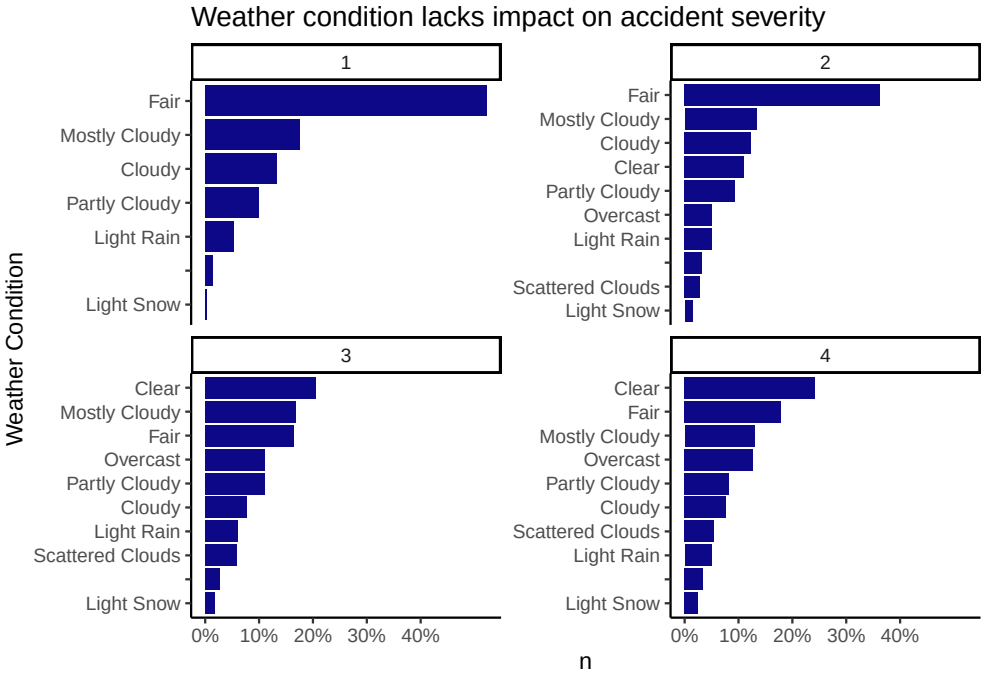
This plot further classified the accidents by month into each year, with the y-axis showing the percentage of accidents that each month takes in each year. For each year, December usually takes about 12 to 15 percent. One uncommon thing we find in this plot is that there are almost no data collected for 2020 July and August. We tend to believe that this is because the COVID hit since the pandemic starts in march 2020 and reaches its peak in July. We also believe that COVID definitely affected the way that the data for accidents was recorded because of how little data there is between July and August. Further, the increase of number of accidents in 2020 November and December are because people have quarantined for a long period of time and finally begin to go out, especially for Thanksgiving and Christmas.

Severity Map



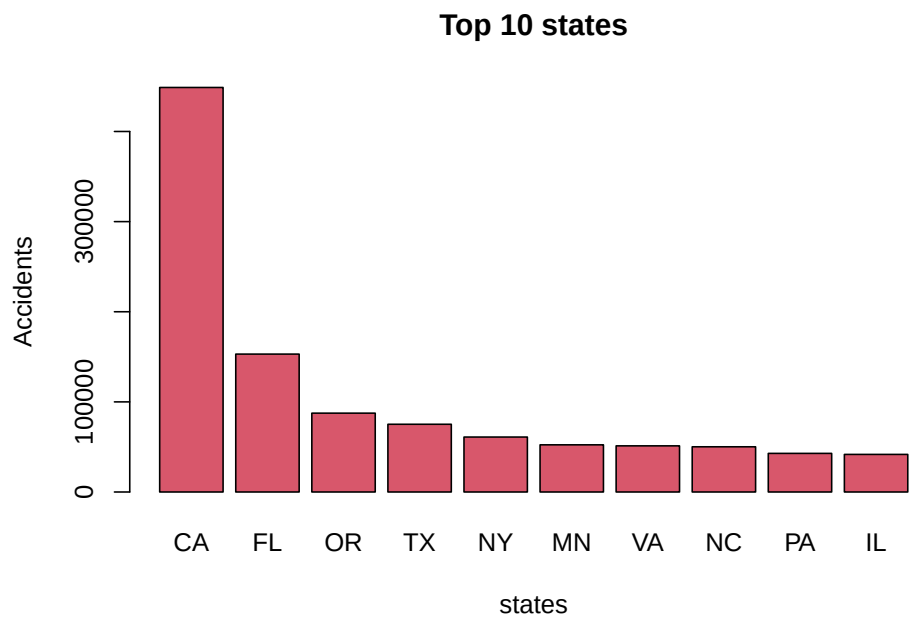
This is the average location of accidents by severity. What stands out about this map is that as the accidents become more severe, their average location moves further northeast. This leads us to believe that this is because colder weather conditions are much more likely to cause worse accidents. But, as we'll see in the next visualization, that might not be the case.

Relationship with Weather Conditions



It's normal for one to think that the severity of the accidents will be closely related to the weather conditions because bad weather such as snowy will cause high severity accidents. The plot is showing the distribution of different weather conditions under different severity levels. By inspecting the plot, the distribution of those weather variables seems similar, which may indicate that weather conditions don't relate that much in terms of the severity level. After running tests on these variables, the difference in means of severity between accidents with wind speed greater than 30 mph and wind speed less than 30 mph is statistically significant via the t-test: the t-test shows that accidents with wind speed greater than 30 mph have greater severity than those less than 30 mph.

Top 10 states in Accidents

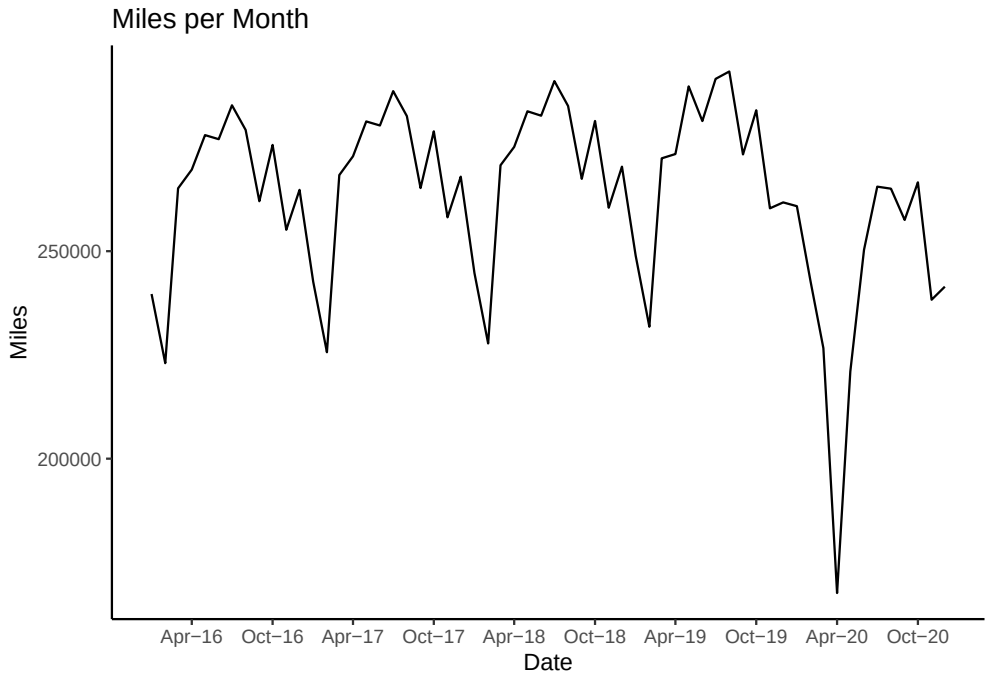


This is a barplot showing the top 10 states that have the most number of accidents. Since these states take almost 80% of the data, we will focus on these states when doing analysis.

Cars and COVID

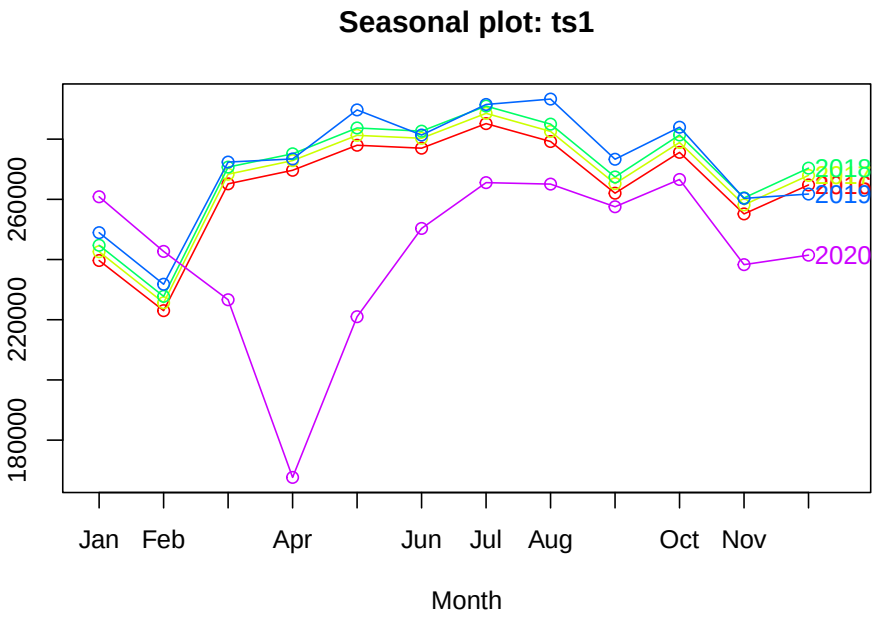
Secondary Dataset: Miles Driven

With our secondary data set, we wanted to provide some context for the main data set of car accidents by examining the number of miles driven every month. This secondary data set is collected by the U.S. Federal Highway Administration, which includes monthly data for the number of miles driven in the entire U.S from January 2016 to December 2020. We believe that this information will provide better insights into patterns in car accidents. We knew that driving habits decreased in early 2020 due to COVID and lockdowns across the country, and we wanted to be able to better quantify these changes from previous years to understand patterns in the main dataset of accidents.



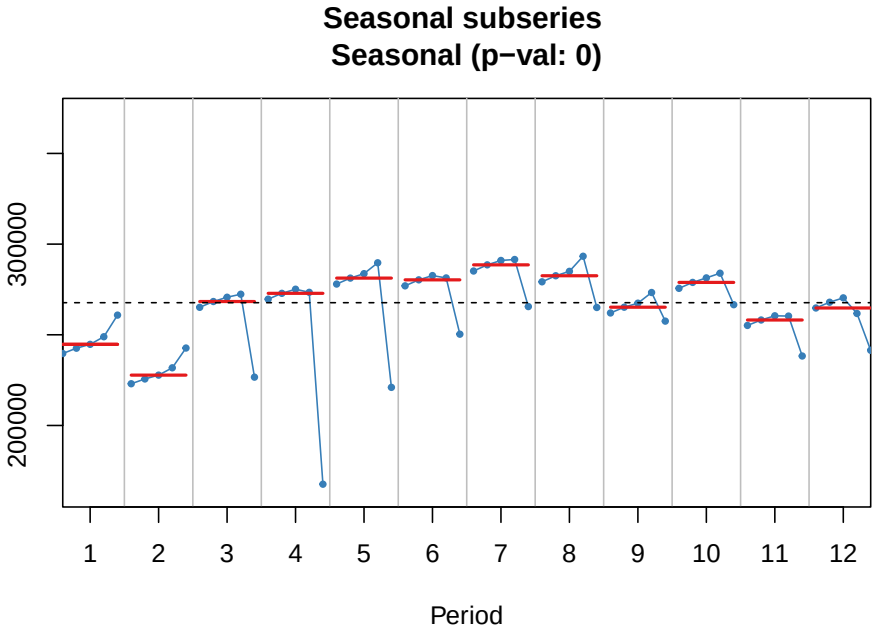
Seasonal Trends

Here we see a seasonal trend where the number of miles driven peaks in the summer months of July and August and reaches a low in February. In 2020, things are clearly different, as we notice a large decrease in April of 2020 brought by lockdowns. Delving into yearly trends, we see that the number of miles driven each year has been increasing; the red line is 2016, and blue line is 2019. Something interesting about this is that in the first 2 months of 2020, as shown by this purple line, the number of miles driven was much higher than usual. Then, in March and April, there was a decline in the number of miles, and this slowly recovered into the summer of 2020 where it came close to previous years, before declining again in late winter of 2020.



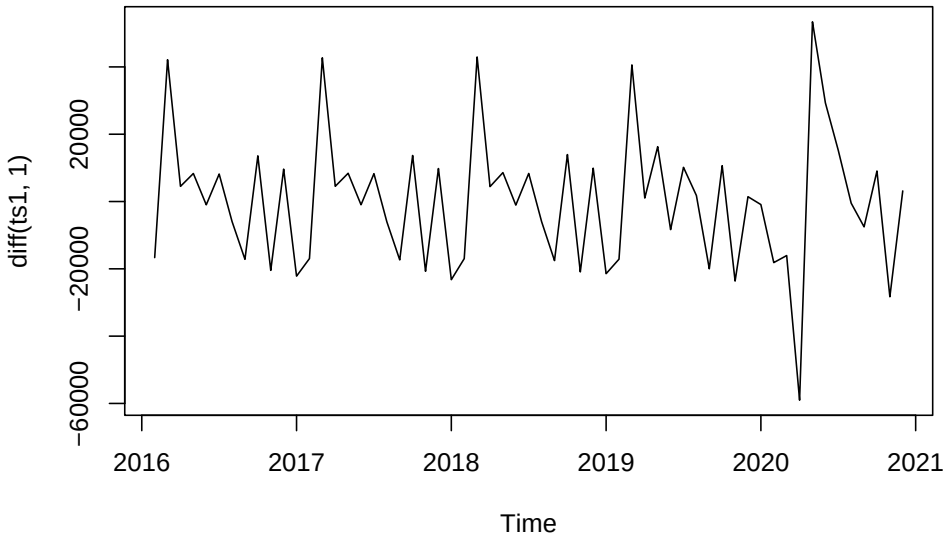
Changes in 2020

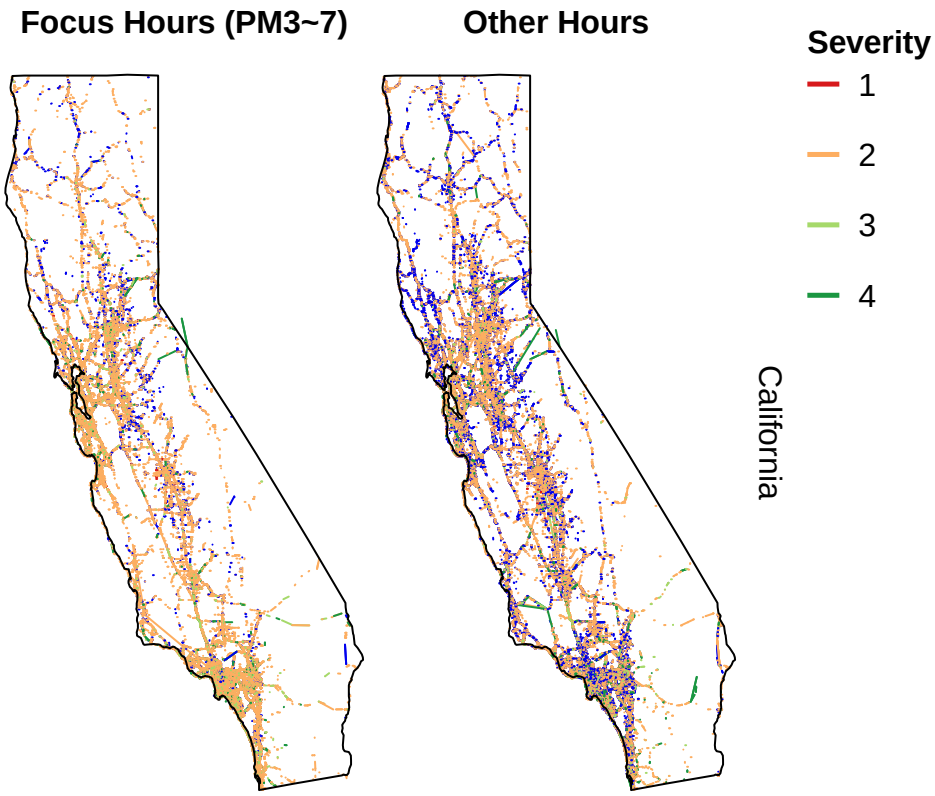
This is another way of visualizing changes by looking at the changes from year to year for each month. Here we see that there was a slow growth from 2016 to 2019 for each month. In 2020, these changes were all reversed, with the most drastic decrease in April 2020. The results of statistical testing confirms evidence of seasonality during the yearly period.

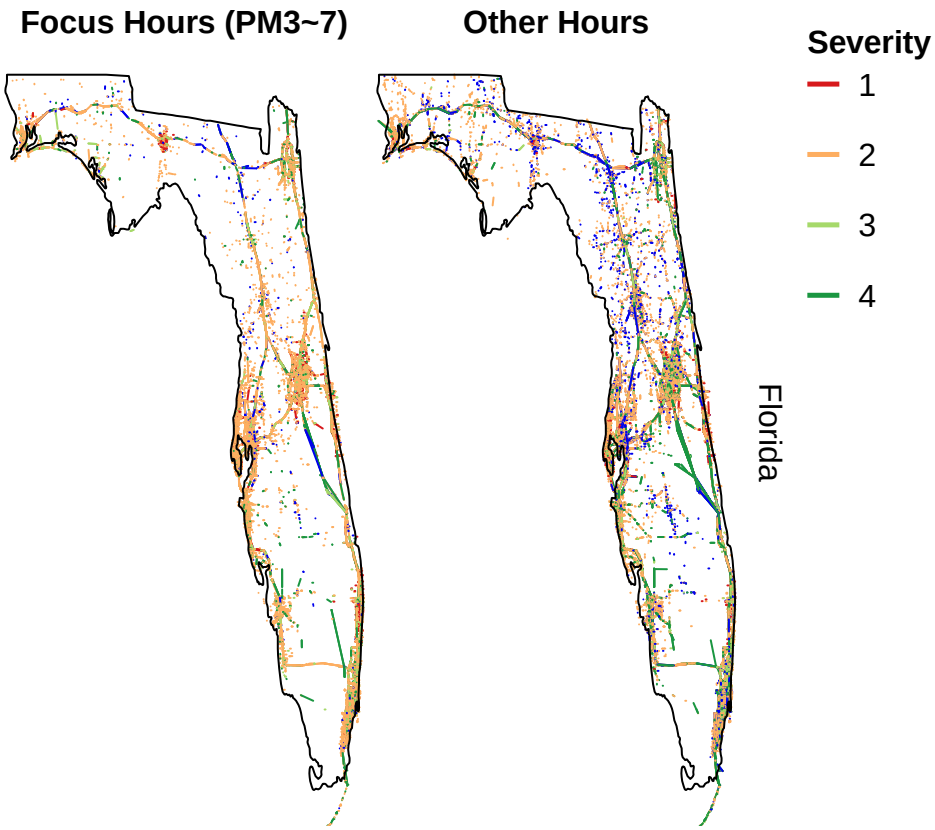


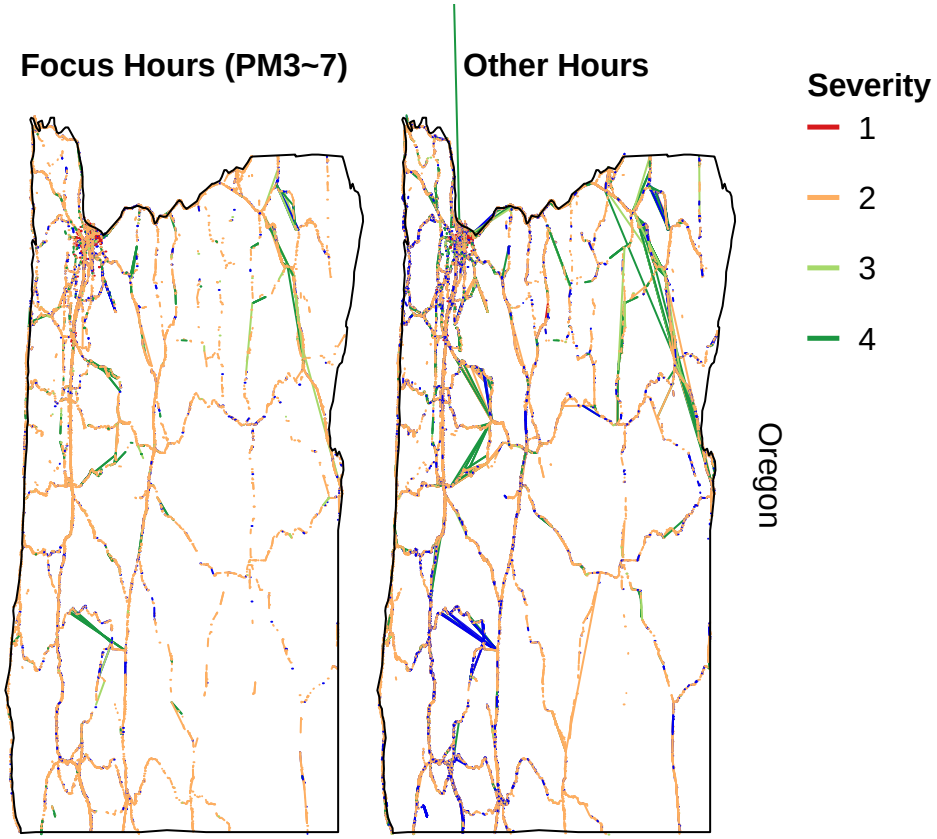
Changes in 2020

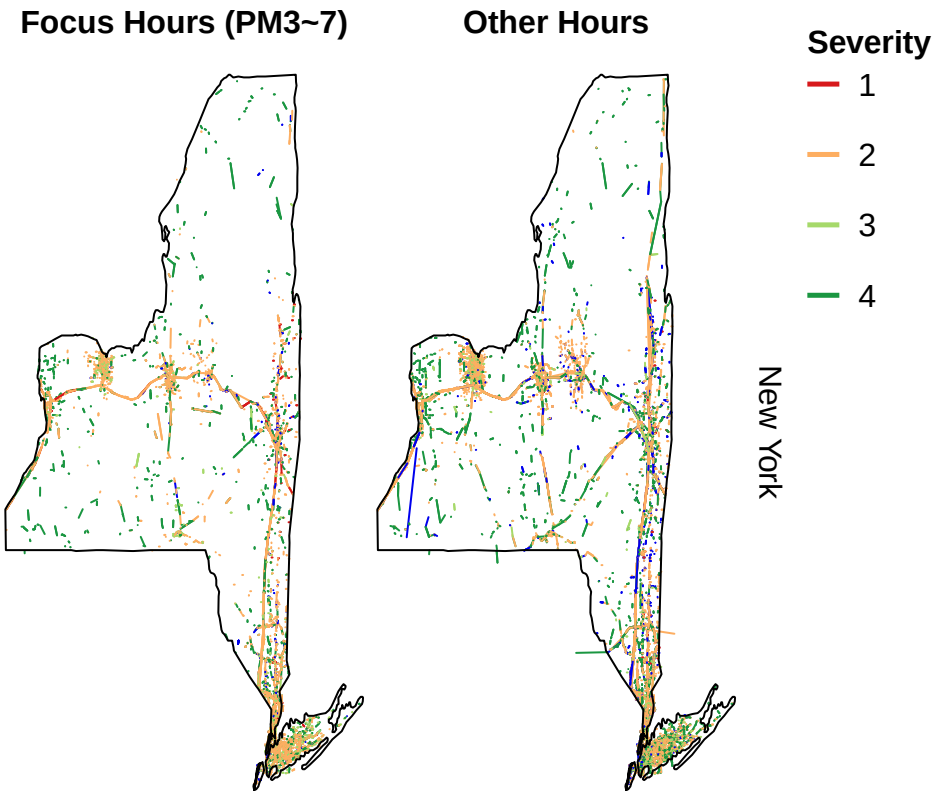
Finally, by examining the differences from month to month, we see that the seasonal trend pattern still remains while accounting for the numbers in 2020.

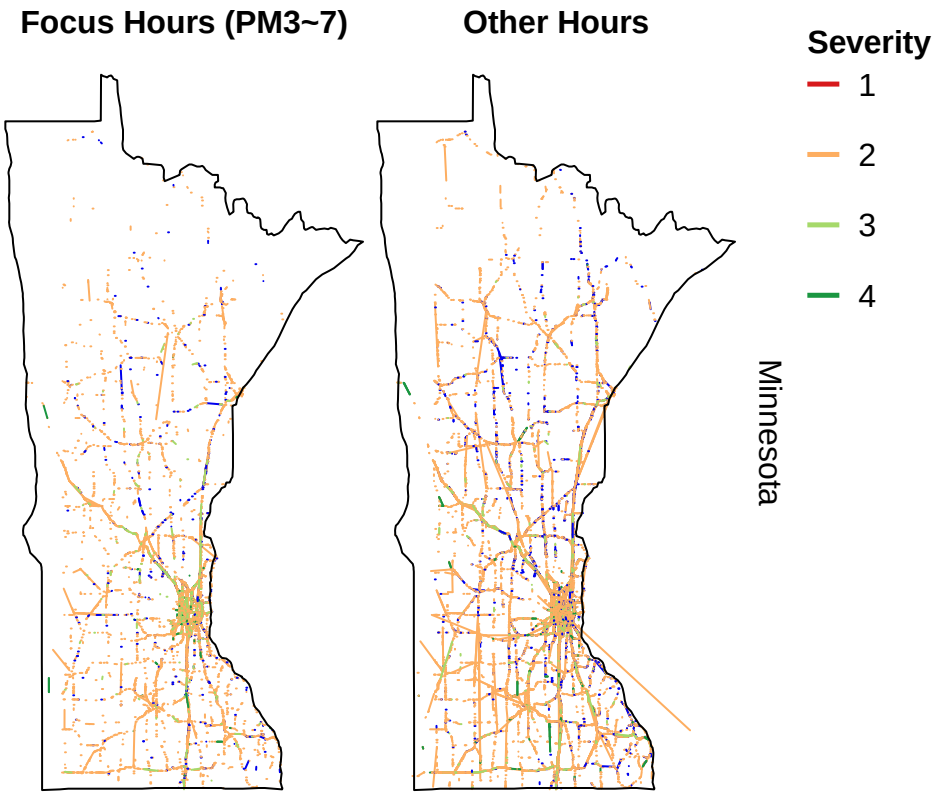


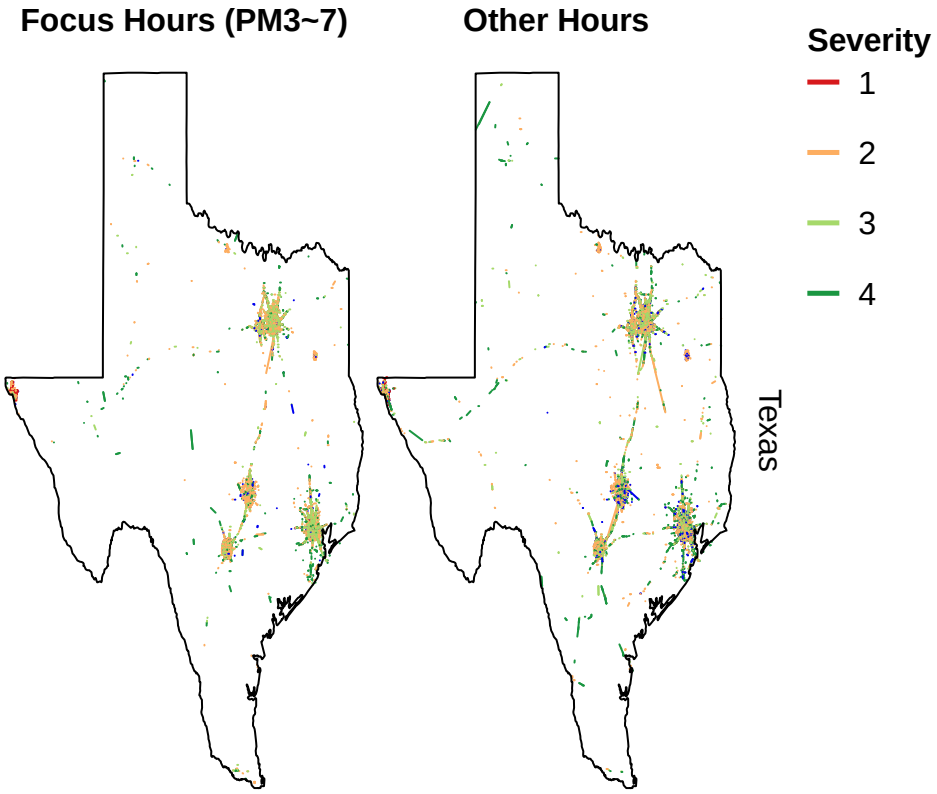

Focus Hours (PM3~7)
Other Hours
Severity
1
2
3
4
California


Focus Hours (PM3~7)
Other Hours
Severity
1
2
3
4
Florida


Focus Hours (PM3~7)
Other Hours
Severity
1
2
3
4
Oregon


Focus Hours (PM3~7)
Other Hours
Severity
1
2
3
4
New York


Focus Hours (PM3~7)
Other Hours
Severity
1
2
3
4
Minnesota



As we have suggests previously, the severity, the exact time in a day, wind speed, and the top number of accidents of state. Our killer plot is a heat map of the top 6 states where the points on the map are the accidents with color showing the severity(1-4), we have divided the plot into the focus hour (3-7 pm) and non-focus hours for the different state to better understand how has time influence the accidents. The points with blue are those accidents that experience wind speed that's more than 30 mph.

1. From the plot, we can see that most accidents are around level 2 severity and it's obvious that lots of the accidents are concentrated in big cities, and the points assemble the highways and roads in the state, We can see that high severity accidents tend to happen on the roads that are going out of the city, which correspond our thought that more accidents happened on holidays because people are going out of the cities to go back to home or for traveling.
2. For wind speed, California, Oregon, and Minnesota have a relatively high number of blue points which means that the number of accidents with wind speed higher than 30 mph is more common there, thus location is a factor that affects the occurrence and severity of accidents.
3. Since we have divided our plot into two different time intervals, we notice that most of the high severity accidents tend to happen outside the focus hour, and we think it's probably because the focus hour did not include the nighttime, focus hour usually would not happen high severity accidents because there are many vehicles and most likely accidents would be scratched or rear-end which is considered low severity.

Conclusion

We have learned interesting things by inspecting the interesting variables that seem to have a great relationship with the number of accidents and the severity of the accidents. Our killer plot shows that the location is

a factor that affects the occurrence of the accidents with different states having different weather conditions, One variable we investigate is the wind speed. Accidents usually happen from September to December because of school's reopening, holidays including Thanksgiving and Christmas. We also investigate the COVID influence on the number of car accidents by introducing a secondary dataset of vehicle miles by month with almost the same time interval. We find that with a decrease in vehicle miles there is also a decrease in car accidents in 2020 starting from March to August. During the analysis of the dataset, we have also found some limitations in the primary dataset. The severity level is not clear enough as it can not clearly account for the financial loss and the effect that the accidents may have on the roads. The data for July and August 2020 is missing for a large amount as well.

Sources

US car accidents dataset:https://arxiv.org/abs/1906.05409 \

Vehicle Miles Traveled(US): https://fred.stlouisfed.org/series/TRFVOLUSM227NFWA#0