The purpose of this project is to, broadly, determine which locations in the United States are the most dangerous in regards to motor vehicle collisions. Which locations have the most collisions, for instance. In addition, an attempt will be made to determine which characteristics of that locale makes them so dangerous. Is it possible that weather could be causing accidents? Perhaps the visibility is bad during a time of year leading to a spike in collisions. In addition, the time it takes to clear the accident from the scene is of particular interest.

The final report of this project will be of interest to many groups including local governments and insurance companies. Local governments, for example, will be interested in seeing a pattern where accidents appear to be occurring frequently and to then make the roadways safer there. This could be accomplished by signage changes or better local education about driver safety in whichever conditions prevail in that region. For instance, if accidents on bridges increase during the winter time, the local government might be interested in placing signage which informs drivers that bridges will begin to ice over long before the air temperature drops to freezing. On the other hand, insurance companies could use the findings to make better cost analyses concerning charge rates in the area. In an area where accidents are frequent, the company would prefer to charge a higher monthly rate to compensate for increased claims. In addition, local road authorities could use the report in order to redesign the roadways in a way to minimize the time it takes to clear accidents. This could be accomplished by building emergency lanes or possibly by increasing the lanes available to commuters so flux of traffic in an affected area is not diminished to a high extent.

Furthermore, the findings could be used by local authorities to determine the impact construction has in an area. If a location has lanes closed for construction, the same load of traffic is being forced through a smaller area. For instance, a four lane highway typically sees 100,000 commuters each morning. When two lanes are closed for construction, the remaining two lanes now have 100,000 commuters to accommodate. Therefore, since the traffic on the two remaining lanes has doubled, it could be assumed that accidents increased due to the higher congestion. If given a date of change, it could be studied how the change affected the local area and the length of travel. However, using statistics, the date the change was implemented could also be inferred due to a substantial difference in the rate of accidents in the location.

The dataset used in this study was obtained from kaggle.com (link: https://www.kaggle.com/sobhanmoosavi/us-accidents). It consists of over 3.5 million records obtained from Mapquest and Bing. The data has 49 columns in total. The most useful of these include the city and state of the accident, the longitude and latitude coordinates of the accident, the severity, the start and end time, temperature, weather conditions, visibility, and time of day.

The first task was to explore the relationship between accidents and the temperature. It is commonly said that colder weather leads to more accidents. However, this did not seem to be the case. According to the following plot, accidents during freezing weather were a minority.

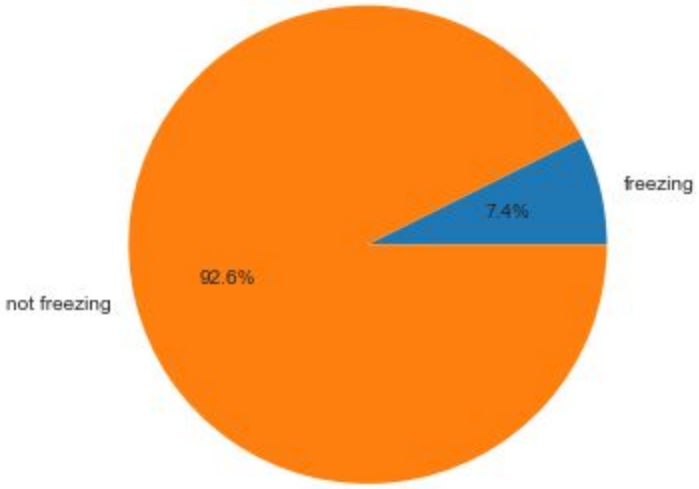


Fig 1. Proportion of accidents during freezing and non freezing temperatures

As can be seen by Figure 1, accidents during freezing temperatures is only 7.4% of all collisions. This is an interesting finding which seems to fly contrary to normal, everyday knowledge. Perhaps freezing temperatures lead to more severe accidents rather than just more accidents as a whole. Figure 2 shows the relationship between freezing temperatures and the severity of accidents.

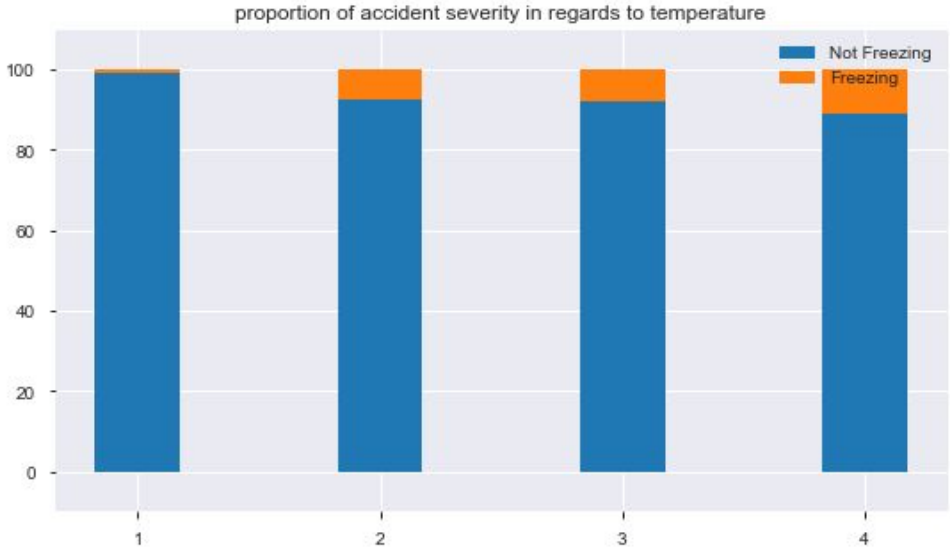


Fig 2. Proportion of accident severity in regards to temperature

As can be seen in Figure 2, the proportion of severity four accidents in freezing temperatures is significantly higher than that of severity one accidents. In fact there are almost twice as many severity four accidents in freezing temperatures than there are severity three accidents in freezing temperatures. This finding does seem to suggest that, while freezing temperatures may not increase the *number* of accidents, they do increase the occurrence of higher severity accidents.

The next angle of investigation was the different time measurements: hours, days, years, etc. A quite interesting pattern emerged when accidents per hour were plotted in Figure 3.

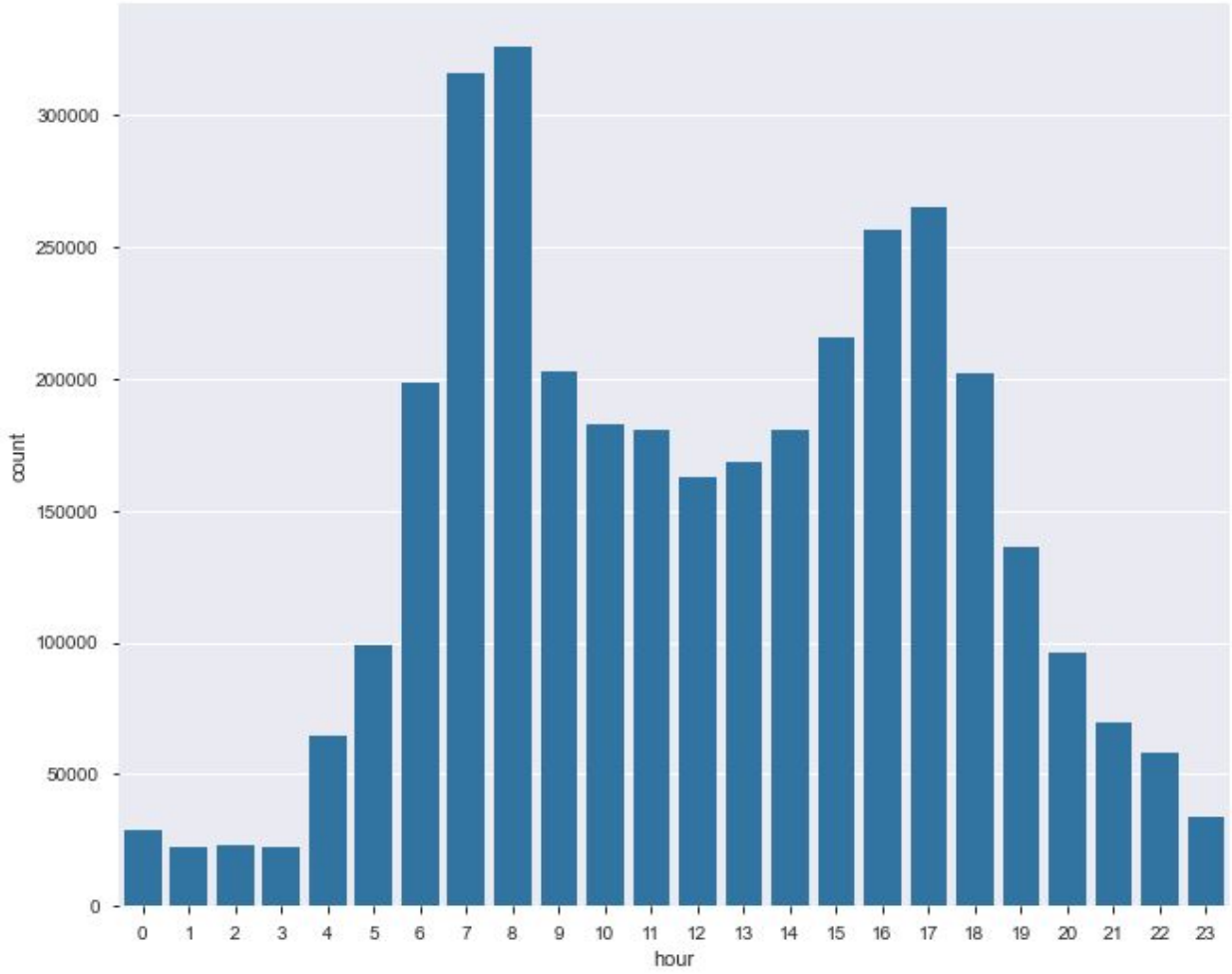


Fig 3. Accidents in each hour of the day

There appear to be spikes in hours seven and eight and again at hours sixteen and seventeen. Why would this be? The increases are substantial compared to the rest of the day. The most likely explanation is that those are the times people are travelling the most. At hours 7 and 8, people are traveling to work; at hours 16 and 17 people are going home. This increase in traffic concentration leads to more collisions.

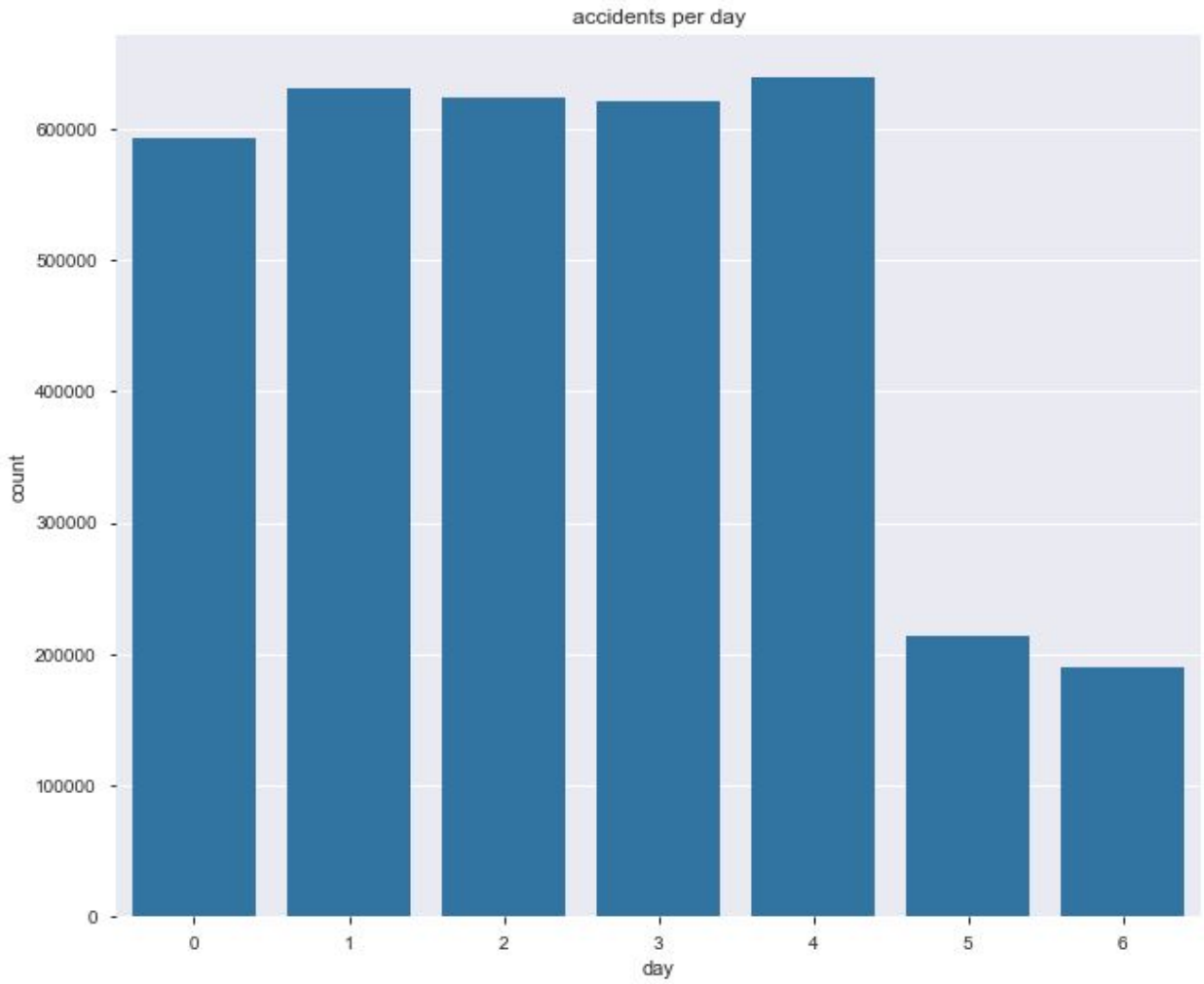


Fig 4. accidents per day

In Figure 4, day zero is a Monday. Given this context, the chart appears to show behaviour one would expect. Most accidents occur Monday to Friday, the standard American work week. The weekend constitutes a minority of accidents since most people prefer to stay home on the weekend. Again, this chart verifies behaviour one would expect.

The next investigation involves the first appearance of two variables which are of particular interest: duration and severity. These two variables appear repeatedly in many other studies of this data. In order to get duration, we simply subtracted the start time and the end time columns from each record. The code is seen below.

```
df['duration'] = ((df['End_Time'] - df['Start_Time']).dt.total_seconds()/60)
```

Snip 1. Creating duration column in the dataframe

The code in Snip 1 simply creates a new column in the data frame, *'duration'*, by subtracting the start time from the end time. This is then divided by 60 in order to convert it to minutes from seconds. The duration is simply the time it takes to clear an accident from the scene to allow unimpeded flow of traffic again.

On the other hand, severity measures how badly traffic was interrupted by the accident, not how bad the accident itself was. While it is not directly measuring the severity of the accident, it can be extrapolated that accidents which result in increased traffic interruption tend to be more severe. Therefore, it can safely be said for the case of this project that the severity of the accident is both the measure of traffic interruption as well as the severity of the accident itself.

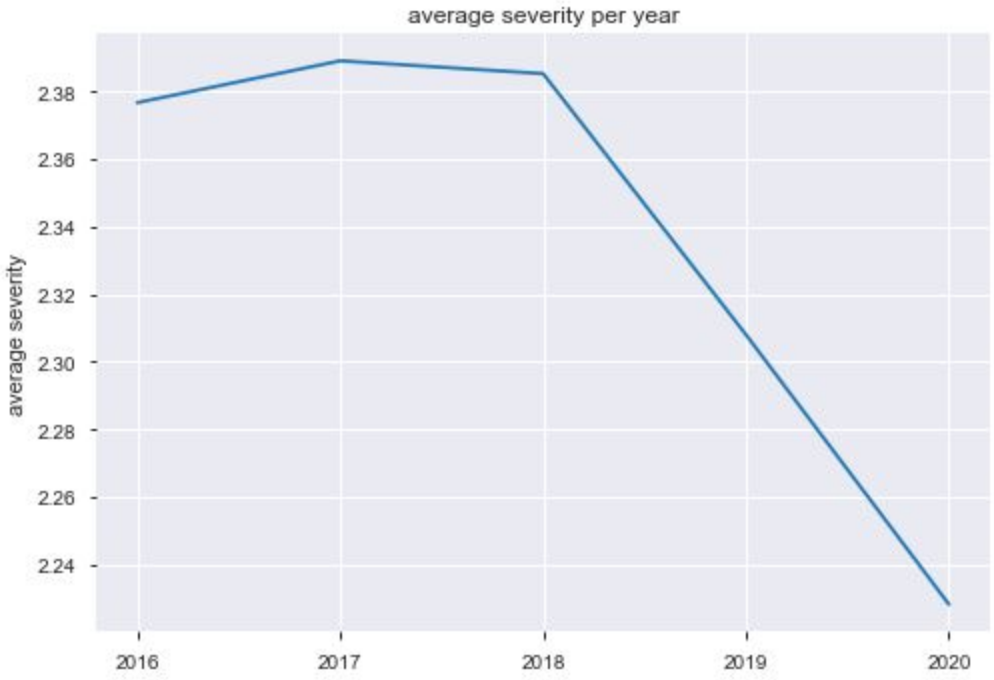


Fig 5. Average severity per year

Figure 5 shows the average severity of accidents from 2016 to 2020. However, it is best to not use 2016 as part of the analysis since this a comparison of whole years, whereas, 2016 is not represented in their entirety. In the case of 2016, data collection did not begin until well into the year, so it is not a good representation for the year. Therefore, we will focus our analysis from 2017 onward.

As can be seen in Figure 5, the severity drops drastically from 2017 to 2020. This is interesting as no obvious reason can be deduced from the dataset itself. The only explanation is for the year 2020 when a pandemic struck the United States, less people were traveling which may have contributed. However, there is no explanation for the decrease in 2019 in the dataset itself.

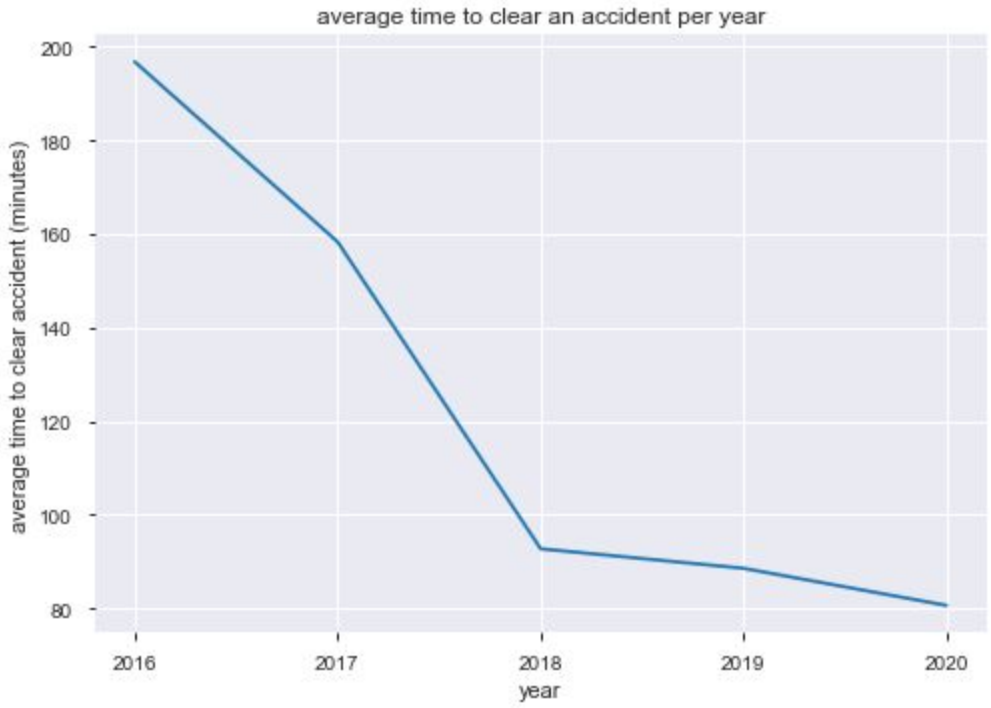


Fig 6. Average time to clear an accident per year

Above, figure 6 depicts the average time it takes crews to clear an accident from the scene. As can be seen, that time has decreased drastically. This makes sense since the severity has decreased as well. Less severe accidents are cleared much more quickly than severe accidents.

```
def str_type(text):
    if '-' in text or 'Fwy'in text or 'Expy' in text or 'Highway'in
text or 'Hwy'in text :
        result = 'Highway'
    else:
        result = 'others'
    return result

df['street_type'] = df['Street'].apply(str_type)
```

Snip 2. Separating records by street type

Snippet 2 depicts the method used to differentiate the street type where each collision occurred. The records were divided into two large categories: *'Highway'* and *'other'*. This was originally to only be used to count how many accidents occurred on each type of road. However,

further down, an interesting correlation was seen.

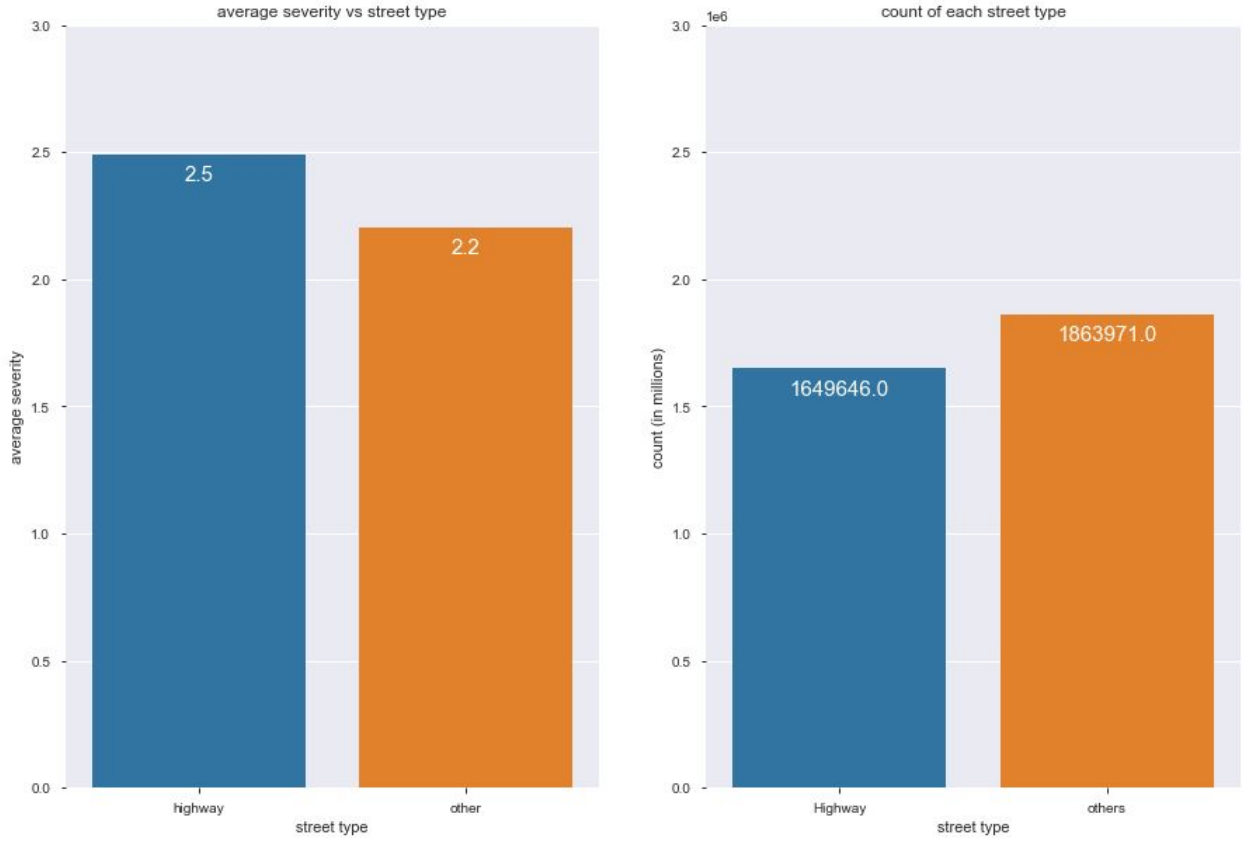


Fig 7
7.1 Average Severity vs Street type
7.2 Count of each street type

As seen in Figure 7.1, highways depicted a severity approximately 7.5% higher than surface streets. The mean of severity is 2.5 and 2.2 respectively. In order to put this into context, it is shown in Figure 7.2 that there are approximately 200,000 more accidents on surface streets which comprise most of the *'other'* category. Despite more chances of a higher severity accident, the surface streets show a lower severity than highways. This is very interesting and will be explored further in this project.

With these findings in mind, the data set was reduced to the ten cities which demonstrated the most accidents. This was done via the following code snippet.

```
df_city = df.groupby('City').size().to_frame('count_city')
df_city = df_city.reset_index().sort_values('count_city', ascending =
False)[:11]
```

Snip 3. Finding the top 10 cities

This code returns 11 cities because of a technical issue encountered further in the code. However, it was narrowed down to 10 by eliminating the last one. Given these cities, the following plot was generated.

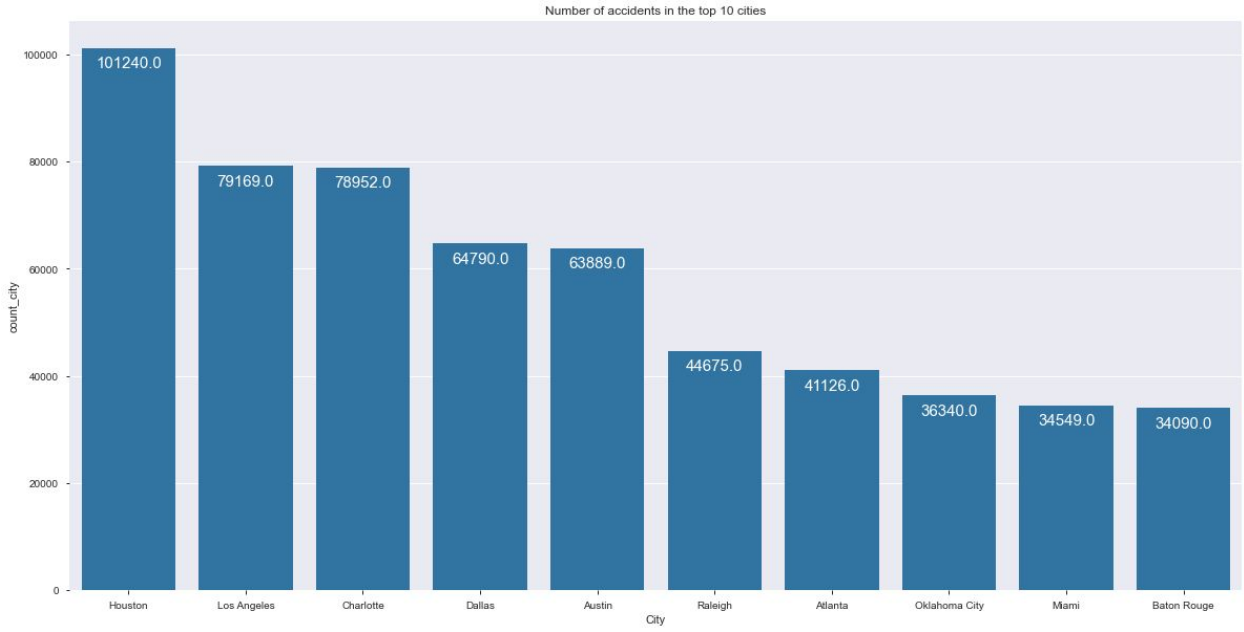


Fig 8. Number of accidents in the top 10 cities

Figure 8 shows the top 10 cities as well as the total number of accidents recorded in that city. Houston has recorded the most accidents by a fair margin of around 20,000 more accidents than the next highest city, Los Angeles.

This is very interesting. All of these very large cities have one specific characteristic in common. They all have high mileage counts of highways. Perhaps the average severity shown in Figure 9 will help clarify this. As can be seen in Figure 9, the average severity of most of these areas is closer to the highway average severity seen above in Figure 7.1. Perhaps there is, indeed a relationship taking form here.

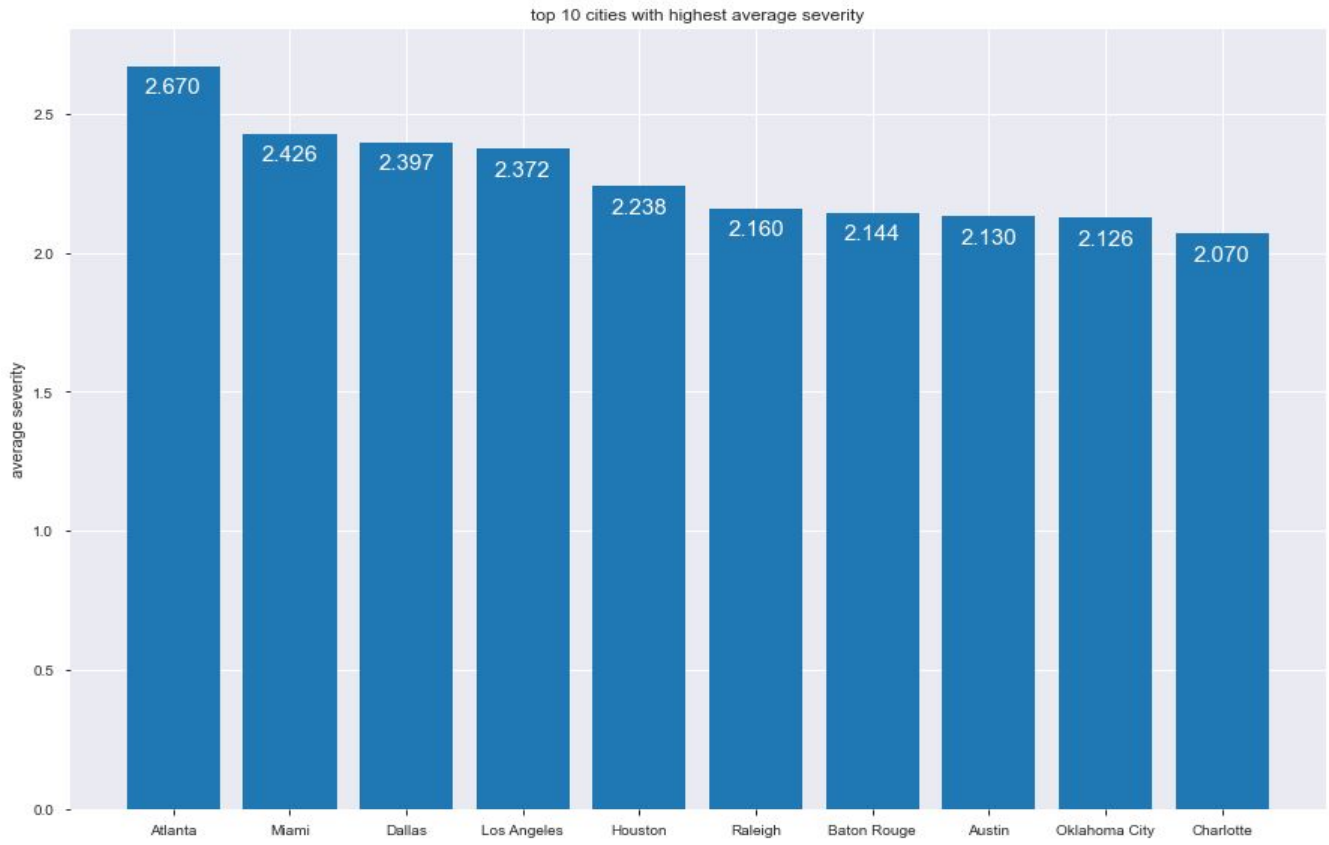
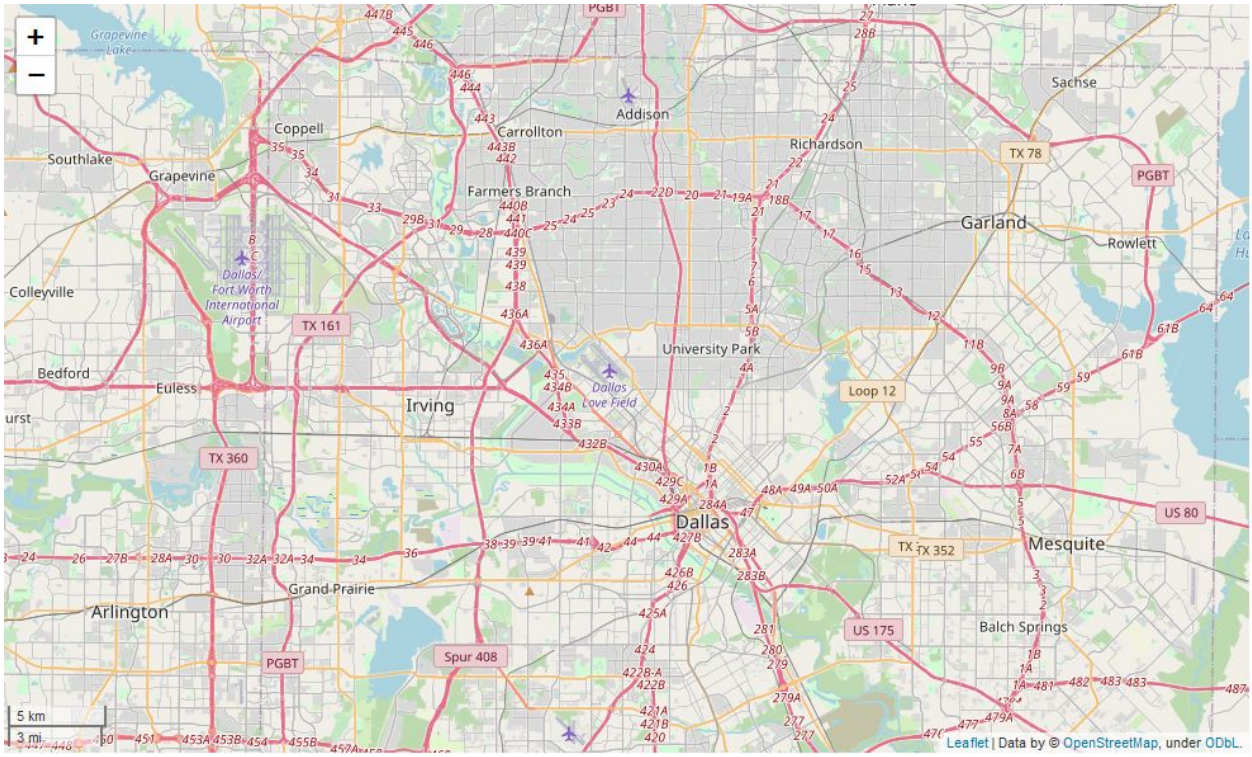
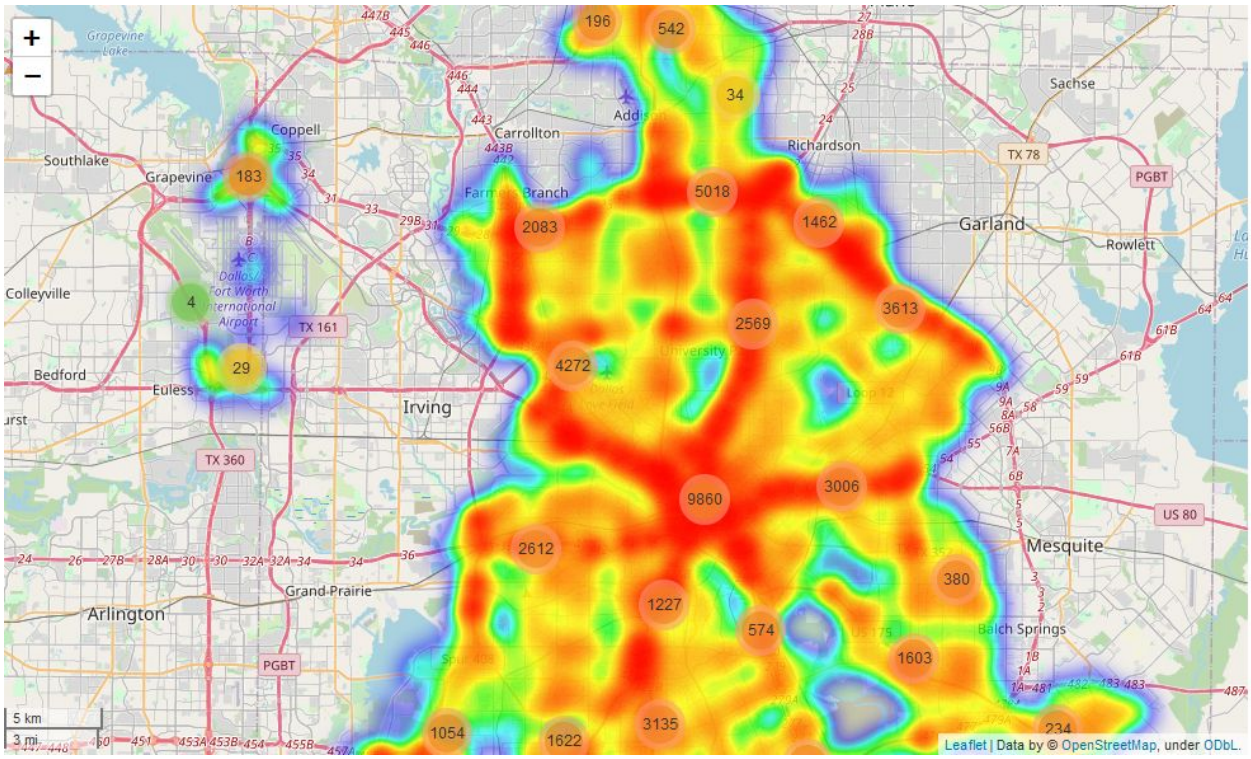


Fig 9. Average severity in the top 10 cities

The data was then narrowed down to each city in turn and the coordinates were plotted on a map. The relationship taking shape above was clearly demonstrated. For example, Dallas, TX was isolated and plotted.



Map 1. Dallas, TX with no mapping



Map 2. Dallas, TX with heatmap applied

Map 1 above depicts a view of Dallas, TX. Notice the locations of the highways. As one can see, they constitute a large part of the travel infrastructure for local residents. Map 2 shows the exact same area, except a heat map has been applied. The relationship between accidents and highways has become amazingly clear. The red of the heatmap is following the highways exactly. The surface streets have very few red spots, whereas the highways are traced almost perfectly by the red hot spots.

Given the data, a relationship has emerged which suggests that highways make an area more dangerous. At the very least, highways seem to demonstrate hot spots for collisions to occur. This will be further investigated for the final report.