The project's goal is to identify critical characteristics that influence accident severity and create a model that can reliably forecast the severity of an accident. To be more specific, this model is expected to forecast the likelihood of a severe accident without any detailed information about the accident, such as driver qualities or vehicle type. The catastrophe could have occurred recently, and details are still unavailable, or it could be predicted by other models.

**The outcome of EDA:**

We can certainly state that our model can forecast the severity of a vehicle accident with high accuracy. All of the other independent variables in this dataset, such as time, date, distance, latitude, temperature, humidity, and so on, influence the severity of an accident. More significant severities, which account for incapacitating or deadly accidents, can usually be attributed to a longer duration, more traffic, or the type of weather or season that was present at the time of the accident.

As a result of this exploratory data analysis, we can conclude that traffic accidents in the United States have been getting significantly worse. Not only has the severity of accidents grown in recent years, but there has also been a surge of accidents, particularly during a time when a pandemic was just getting started, meaning that there were more incidents with fewer drivers on the road. In this dataset, 75% of all traffic incidents had a severity score of 2, indicating that the people involved were disabled (required assistance getting out of the accident) and had mild injuries. Solving the problem of minimizing traffic accidents will ease the burden on drivers and everyone.

Some of the findings from this analysis are:

* The top 5 states by accidents include populous ones like Los Angeles, Houston, and Florida.
* Less than 5% of cities have more than 1000 accidents between February 2016 and December 2020.
* New York is a populous city with less than 7500 accidents in the accounting period.
* The hypothesis that weekdays see more accidents during morning and evening rush hours is corroborated by the data. Weekends have the peak during the afternoon.
* The weather conditions seem to be "Fair,” “Clear," and "Mostly Cloudy" in most cases. The assumption of a 10mile visibility also doesn't conclusively explain the accidents due to these factors.

**What do you feel was missed during the analysis?**

The year-by-year analysis appears to reveal missing data for several months. As a result, there's not much that can be said decisively about the overall upward trend in the second half of the year.

**Were there any variables you felt could have helped in the analysis?**

Other variables, such as driver operational conditions during the accident, would help uncover more insights from this dataset.

**Were there any assumptions made you felt were incorrect?**

One assumption most people think of by population in the city, a greater number of accidents occur. But after the analysis of the data, New York City is a big exception which reported only 7500 accidents.

**What challenges did you face, and what did you not fully understand?**

Since the dataset is huge, with 2.8 million records, I felt it was challenging to prepare the data set by cleaning and identifying the correct variables from the data set that would be a good fit for the predictive analysis. I started by considering some variables like location and other weather-related variables but later realized they were not relevant and had to discard them. I did not fully understand the various statistical predictive models, and I hope to learn more in my other data science courses.