Statistical/Hypothetical Question

According car accidents statistics, there are nearly 6 million car accidents in the United States every year. Approximately 1.25 million people die in road crashed each year, in addition, 20-50 million people are injured, many of them even have long-term disabilities. Due to car accidents, many people lose properties, health, especially, people lose their loved ones. Thus, I determine to analyze a countywide car accident dataset, my goal is to find out the importance of different climatic factors correspond to the accidents and severity.

Outcome of your EDA

I was thinking that climatic factors have an important impact on accidents, especially weather conditions, visibility. There will be more accidents in snow or rainy days, and bad visibility will cause more accidents. But the analysis result gave me a different picture.

When you are seeing the histograms of weather-related variables, most accidents were occurred around 75 degree, not frozen and snow days, 75 degree is a very comfortable temperature. Most accidents were happened in pressure 28 in, 29 in, and wind speed from 2 mph to 6 mph, not strong wind. Most accidents were occurred in daytime, and visibility 10 mile. And I checked the accidents' counts under the different weather condition, most are clear and cloudy day, not snow and rainy day.

Then I made a PMF of visibility for severity 2 and severity 3, I can not see obviously different between them. So did CDF of visibility. All weather-related variables have a little correlation of severity. For hypothesis test, I got the result visibility have a correlation of severity and it is significant. The last, I did a multiply linear regression analysis, all the weather-related variables, temperature, humidity, air pressure, visibility, wind speed, the period of day (day / night) are significant, but R2 is low, this model is not good. The climatic factors do not have a great effect on severity.

What do you feel was missed during the analysis?

I think analyzing climate should be limited in a state or nearby state, different states have different climate even though there are in the same season, the analysis is biased and cannot get a precise result.

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

I think the number of days a place has like clear days, rainy days, cloudy days etc. might help normalize the counts of the accidents for different weather conditions, because in most places, one have more clear days than snow or rainy days.

Were there any assumptions made you felt were incorrect?

Visibility and severity have strong correlation.

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

There were some statistical functions, I was not quite familiar with them yet, and I need to spend much time to understand and use them in my project.