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

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, what 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