1. **Introduction/Business Problem**

The problem at hand is how to predict the severity of a car accident given the dataset. Car accidents can happen in a multitude of ways and not all are alike. They can happen anywhere, but do they happen more frequently in certain places? Different vehicles have different accident rates and our dataset includes many variables outside of just these two. The end goal of this project is to help predict the severity based on all these variables, but also have it become interpretable by a large audience. Our audience would be neighborhoods where we see the common predictors are more likely than the norm. Value resides in how well you can identify the predictors, which can hopefully be used to prevent accidents or at a minimum reduce severity in the long run. I’m very curious about the dataset and trying multiple approaches to see what methods yield the most valuable insights.

1. **Data**

The data is structured well and consists of 194673 observations of 37 variables, numeric and categorical. The first column is the severity code which is either 1 or 2 which indicates that we will use some logistic regression to solve if an accident was or wasn’t severe rather than the specific severity on a continuous numeric scale. There is also locational data so we can see if there is a higher density of severe accidents in a specific location. Most of the variables that we will use as predictors are categorical which are better used to solve the problem because communities can take action if they fall into the vulnerable categories, rather than adjust a numerical measurement accordingly.