**Introduction/ Business Problem**

In this project, we are given a dataset about accidents happened in a particular area. Information such as number of vehicles involved, number of people involved, and severity of the accident are provided. From this dataset, we are interested to know how weather, light condition, road condition and other criteria affect the severity of accident. In the end of this project, we will build a classification model to predict the severity of accident based on the information.

**Data**

This dataset contains 38 column and 194673 rows. We only select 11 columns which we think contribute more to determine the severity of accidents. The 11 columns are SEVERITYCODE, LOCATION, SEVERITYDESC, COLLISIONTYPE, PERSONCOUNT, VEHCOUNT, JUNCTIONTYPE, WEATHER, ROADCOND, LIGHTCOND and SPEEDING.

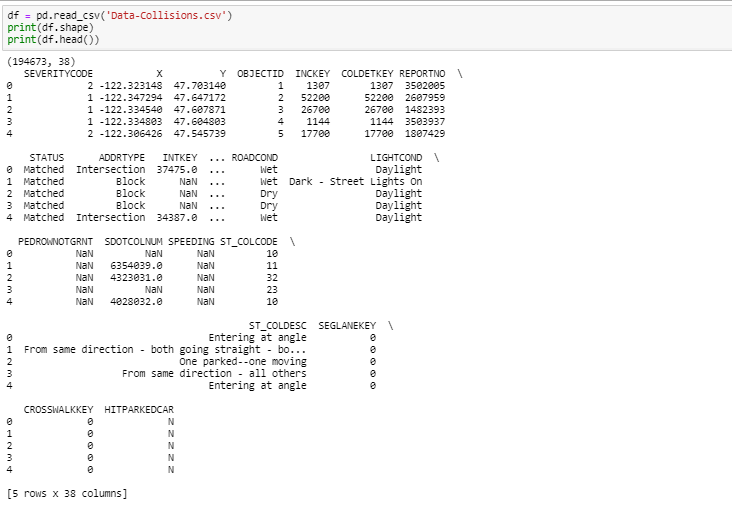


Figure 1. Full Dataset

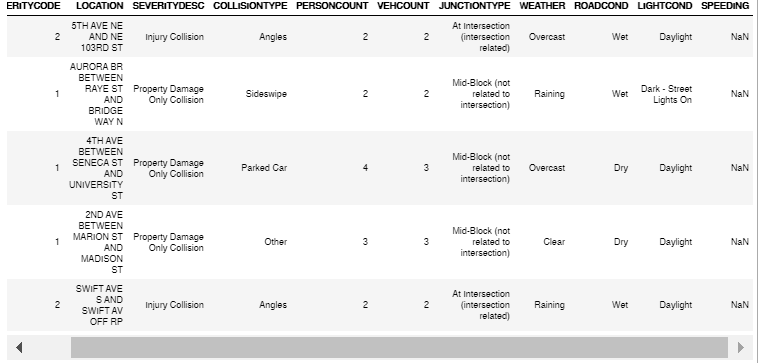
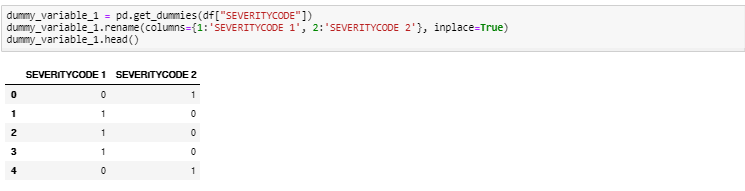


Figure 2. Dataset with selected columns

We split the SEVERITYCODE column into 2 by getting dummy variables for future usage.



In the next section, we will discuss the top 20 locations of most accidents happened, and how weather, light condition and road condition affect severity of accident by generating pie charts respectively.