Data source & description

The data used in this study was provided by Coursera, and had records on all types of collisions from 2004 through to present day.
It has 194.673 rows and initially 37 different independent variables. I will use the SEVERITYCODE as the dependent variable Y, with different independent variables to identify the cause and severity of road accidents.

SEVERITYCODE	Corresponds to the severity of the collision: 3 – fatality 2b – serious injury 2 – injury 1 – prop damage 0 – unknown

Other important variables to include are:

ADDTYPE	Collision address, i.e.: - Alley - Block - Intersection
LOCATION	Description of the general location of the collision
PERSONCOUNT	Total number of people involved in the collision
PEDCOUNT	Total number of pedestrians involved in the collision
PEDCYLCOUNT	Number of cyclists involved in the collision
VEHCOUNT	Number of vehicles involved in the collision
INCCDTTM	Date and time of accident
JUNCTIONTYPE	Category of junction where the accident happened
WEATHER	Description of weather conditions at time of collision
ROADCOND	Condition of road at time of collision
LIGHTCOND	Light conditions at time of collision
SPEEDING	Whether or not this was a factor in the collision (Y/N)

As there are null values in some records, the data needs to be pre-processed before any further processing.

The data will be analysed to predict the likelihood of a collision and it's severity given a number of variables such as weather conditions, road conditions, time of day, period of year, etc.
These predictions will help the Department of Transport of Seattle to implement preventive measures as required.