Introduction/Business Understanding

The problem is Seattle has car accidents that can be prevented. Accidents happen at all times, but if the main causes of accidents are determined, advance warning or mitigating methods can be performed. For example, certain intersections may be more susceptible to accidents due to heavy usage or the way they are constructed. As a result, better streetlights can be added (only protected left and right turns) or traffic personnel can be used to direct the cars. If it is determined that accidents occur most of a time a driver is speeding, has a high blood alcohol level, or was not paying attention, the data can be used as evidence for enacting harsher laws and regulations. In addition, the data can be advertised to the public to show them the consequences of driving under these conditions. This will hopefully dissuade people in the future. Finally, there are also uncontrollable factors such as weather and road conditions. If certain patterns are discovered to cause many accidents, local government can know when to send alerts to the public to drive more cautiously or even avoid the roads entirely.

The target audience of this analysis is the Seattle government and transportation department. It should identify key causes of accidents and allow them to identify trends for when accidents can be prevented. This will reduce the number of accidents and injuries for the city.

Data

The data comes from collision and accident reports in Seattle during the years 2004-present. It was collected by the Seattle Police Department and Traffic Records department. The data will be used to identify the key variables that cause accidents. For example, the "WEATHER" column can be used to show the types and number of accidents that occur for different categories. In addition, the "INTKEY" column can be grouped and the sum of the accidents in that intersection can be calculated. This list can be sorted descending to identify the more dangerous intersections that need improvements or closer monitoring. Finally, a supervised learning model will be used to come up with a formula that can predict the severity of an accident based on the inputs.

The data has 37 independent variables and 194,673 records. The dependent variable, "SEVERITYCODE", has numbers that correspond to different levels of severity caused by the accident. Many of the columns are object types. In addition, other columns that appear to be integer types are also actually objects, because the numbers correspond to different categories. Finally, some columns and rows have null values, which will be dealt with during the data pre-processing phase.