Data

**Data Source**

It is obtained directly from the Coursera page, which is directly from the Seattle Department of Transportation (SDOT). The file originally contains 194,673 rows and 38 columns.

**Feature Identification**

At first glance, it is difficult to identify the relevant features for model prediction. However, as the objective is to *predict* the severity of an accident, any post-accident data is irrelevant.

This is despite the clear correlation between the severity of an accident and the post-accident data column, such as the number of injuries, serious injuries, fatalities and vehicle count.

Of course, any column with SDOT codes are also irrelevant. Thus, with much thought of the relevant data, the features used for the machine learning model will include data columns that fulfil two characteristics:

1. It must be information available before an incident happens, enabling prediction.
2. It has correlation with the severity of an incident.

As such, there are only a few features that fulfil the criteria above. Namely, they are the columns on weather, road condition, light condition and address type.

**Data Cleaning**

All features are categorical variables and contain many unique values, making it difficult to replace null values. Thus, I decided to remove rows that contain null values for all four columns. After this operation, the dataset is left with 187,525 rows.

Next, I noticed there are rows with “Unknown” values for three columns. I decided to further remove these rows due to these variables being categorical ones. After removal of unknown values, 169,781 rows remain, which is still a considerable amount of data.