**Introduction/Business Problem**

In order to reduce the frequency of car collisions in a location , an algorithm must be developed to predict the severity of an accident based on the current weather conditions, visibility conditions etc.

In most of the scenarios, Accidents occurs due to irresponsible drivers who does not pay attention while driving or drive above speed limit or drunk and drive etc. Apart from the above-mentioned scenarios, weather, visibility and road conditions related accidents are the uncontrollable factors. So these factors can be predicted bases on a particular model/ algorithm.

This model will alert drivers to remind them to be more careful based on the current weather conditions, visibility conditions.

**Data Section**

The data consists of 37 independent variables and 194,673 rows. The dependent variable, “SEVERITYCODE”, contains numbers that correspond to different levels of severity caused by an accident from 0 to 4.

Severity codes are as follows:

0: Little to no Probability (Clear Conditions)

1: Very Low Probability — Chance or Property Damage

2: Low Probability — Chance of Injury

3: Mild Probability — Chance of Serious Injury

4: High Probability — Chance of Fatality

Furthermore, because of the existence of null values in some records, the data needs to be pre-processed before any further processing.

The dataset in the original form is not ready for data analysis. In order to prepare the data, first, we need to drop the non-relevant columns. In addition, most of the features are of object data types that need to be converted into numerical data types.

After analysing the data set, I have decided to focus on only four features, severity, weather conditions, road conditions, and light conditions, among others.

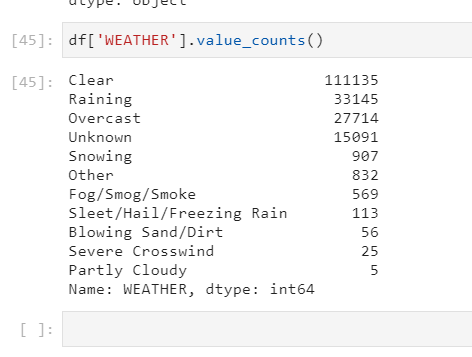
We must use label encoding to covert the features to our desired data type.

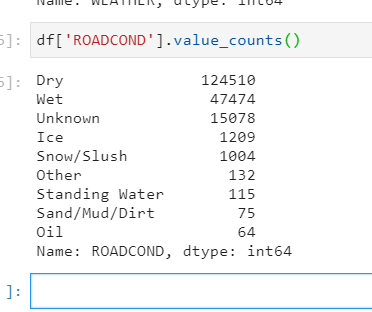


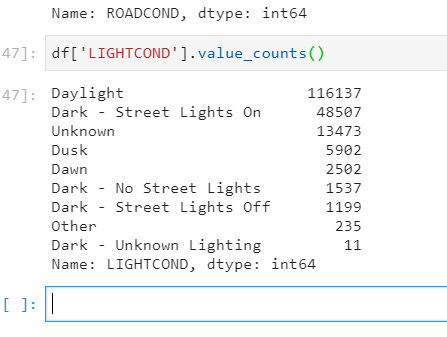
I have selected the following features to predict the severity of accidents. Among all the features, the following features have the most influence in the accuracy of the predictions:

WEATHER, ROADCOND, LIGHTCOND

I have used Value\_counts() for all the 3 to see the breakdowns of accidents occurring during the different conditions.







After balancing “SEVERITYCODE” feature, and standardizing the input feature, the data has been ready for building machine learning models.

# I have used 3 Machine learning models:

* K Nearest Neighbour (KNN)
* Decision Tree
* Linear Regression