Data Understanding

Our predictor or target variable will be 'SEVERITYCODE' because it is used measure the severity of an accident from 0 to 5 within the dataset. Attributes used to weigh the severity of an accident are 'WEATHER', 'ROADCOND' and 'LIGHTCOND'.

Severity codes are as follows:

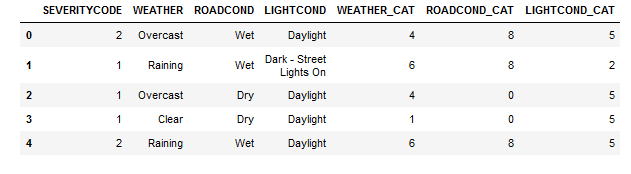
\* 0 : Little to no Probability (Clear Conditions)

* 1 : Very Low Probablility - Chance or Property Damage
* 2 : Low Probability - Chance of Injury
* 3 : Mild Probability - Chance of Serious Injury
* 4 : High Probability - Chance of Fatality

Extract Dataset & Convert

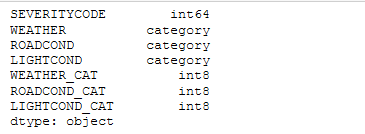
In it's original form, this data is not fit for analysis. For one, there are many columns that we will not use for this model. Also, most of the features are of type object, when they should be numerical type.

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

[](https://res.cloudinary.com/practicaldev/image/fetch/s--chHdQUzJ--/c_limit%2Cf_auto%2Cfl_progressive%2Cq_auto%2Cw_880/https:/dev-to-uploads.s3.amazonaws.com/i/icadxir8nc4agrh2mv49.png)

With the new columns, we can now use this data in our analysis and ML models!

Now let's check the data types of the new columns in our dataframe. Moving forward, we will only use the new columns for our analysis.

[](https://res.cloudinary.com/practicaldev/image/fetch/s--Z-rqpa6U--/c_limit%2Cf_auto%2Cfl_progressive%2Cq_auto%2Cw_880/https:/dev-to-uploads.s3.amazonaws.com/i/ho4kty0dz8gyqkvx170w.png)

**Balancing the Dataset**

Our target variable SEVERITYCODE is only 42% balanced. In fact, severitycode in class 1 is nearly three times the size of class 2.

We can fix this by downsampling the majority class.

[Alt Text](https://res.cloudinary.com/practicaldev/image/fetch/s--ab3RblIA--/c_limit%2Cf_auto%2Cfl_progressive%2Cq_auto%2Cw_880/https:/dev-to-uploads.s3.amazonaws.com/i/owt8a0ptc68nzkntehue.png)

Perfectly balanced.