# Data Understanding

The data has been retrieved and processed through various sources, and database.

The main source being the data-collisions csv file

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'.

## SEVERITYCODE

Severity codes are as follows:

1. Little to no probability (Clear conditions)
2. Very Low probability (chance or damage property)
3. Low Probability - Chance of Injury
4. Mild Probability - Chance of Serious Injury
5. High Probability - Chance of Fatality

## WEATHER

* Overcast: Overcast sky conditions occur when clouds cover all or most of the sky and cause low visibility conditions
* Rains: Heavy or moderate rainfall, which causes roads to be slippery
* Clear: Clear weather conditions

## ROADCOND

This attribute refers to the road condition for a day

* Wet: Usually refers to wet and snowy conditions on a day
* Dry: Normal road conditions

## LIGHTCOND

This attribute gives information of light conditions when the accident took place and will be useful in predicting in what conditions an accident is probable.

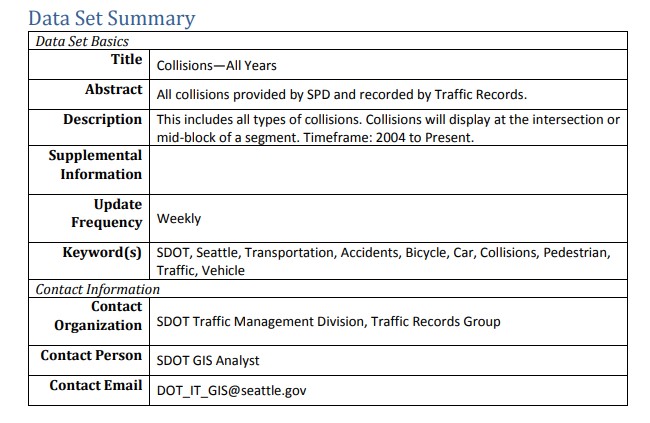
* Daylight: Indicates daylight conditions
* Dark with Street Lights on: Dark conditions but streetlight source was present
* Dark without Street Lights: Pitch dark conditions, only head lights to guide the way
* Dawn: Early Morning, day starts getting more light
* Dusk: late evening, light from the day reduces

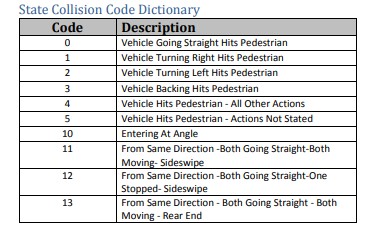
## Data Balancing and Cleaning

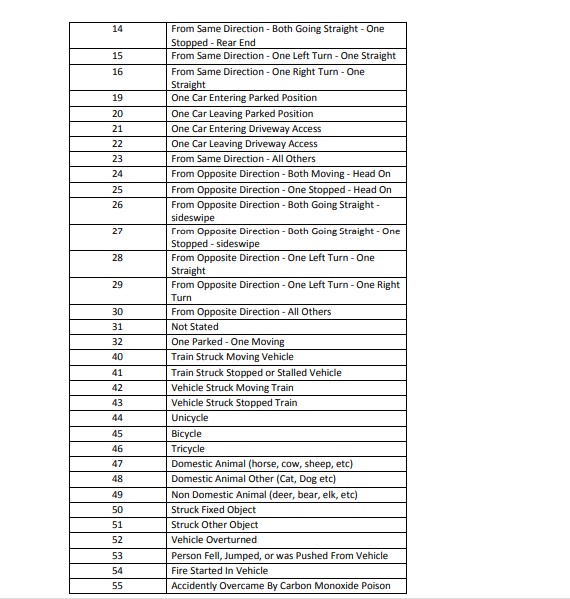
Data is unbalanced and cannot be directly used for analysis

In its 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.







|  |
| --- |
|  |