**Data Understanding**

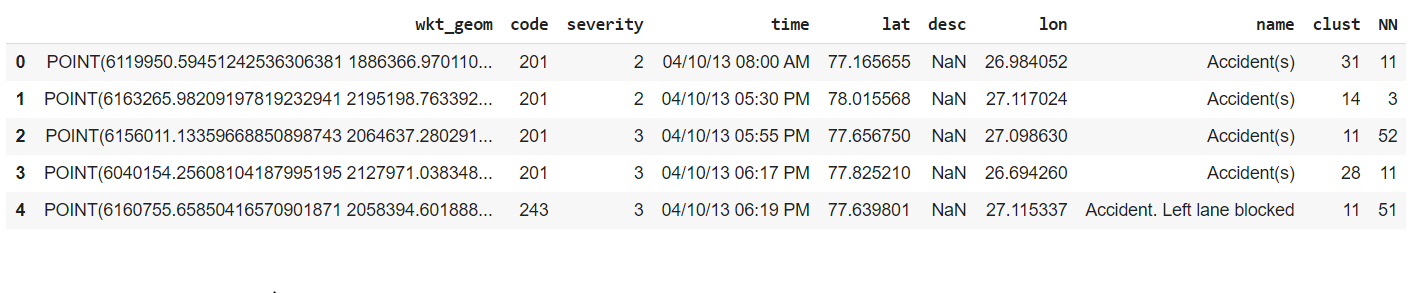
Our predictor or target variable will be 'SEVERITYCODE' because it is used measure the severity of an accident from 0 to 5 within the dataset. Attributes used to weigh the severity of an accident are 'Latitude', 'longitude' and 'time'.

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**

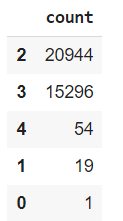
In it's original form, this data is not fit for analysis. For one, there are many columns that we will not use for this model. Also, most of the features are of type object, when they should be numerical type.



#### **Balancing the Dataset**

Our target variable SEVERITYCODE is unbalaced . In fact, severitycode in class 4,1,0 is nearly negligible as compare to 2 and 3.

We can fix this by downsampling the minority class.



**Pre-Processing the time features: creating months, weekday year and hour features because we cant use time directly , its in DD:MM:YY HH:MM:SS format**

Perfectly balanced.