**Data Understanding**

The dataset provided by SDOT consists of 194674 instances, with 36 attributes in total. However, a significant portion of the attributes contain a large amount of Missing values and do not contribute to the prediction of severity.

Attributes which do not contribute to the prediction of severity are:

1. INCKEY (Key corresponding to the incident)
2. COLDETKEY (A secondary Key for the incident)
3. REPORTNO (A key for the report)
4. OBJECTID (ESRI unique identifier)
5. STATUS
6. INTKEY (Key corresponding to the intersection)
7. SDOTCOLNUM

The above attributes do not contribute to the severity prediction; hence they are not taken into consideration in building the predictive model.

Some of the attributes in the data contain over 95% of missing data. Hence rather than speculating and filling the missing data, we do not take these into consideration. These attributes are removed from consideration. 2 Attributes, **EXCEPTRSDESC** and **EXCEPTRSNCODE,** only consist of Not Enough Information. These are removed from consideration too. Such attributes are:

1. PEDROWNOTGRNT
2. EXCEPTRSDESC
3. EXCEPTRSNCODE
4. INATTENTIONIND

Collision codes are worth investigating into to see if certain collisions have effect on the severity. Hence while we drop the description column of each collision, we keep the collision code column. We keep **ST\_COLCODE and SDOT\_COLCODE** while dropping their corresponding description columns, **ST\_COLDESC and SDOT\_COLDESC.** Using the collision codes, we can investigate if certain collision types have higher impact on severity, hence in the final model, such collision codes can be given a higher weight.

Furthermore, the incident time and data is split into Year, Month, Date and time of day to create more meaningful features that are easier to work with and incorporate into a prediction model. This step will result in 4 more attributes. Namely:

1. YEAR
2. MONTH
3. DAY (Day of week with 0 being Monday and 6 being Sunday)
4. TIMEOFDAY

Afterwards we drop INCDATE and INCDTTM from the dataset.

The final set of attributes will be:

>>dfcombined.columns.tolist()

['X',

'Y',

'ADDRTYPE',

'SEVERITYDESC',

'COLLISIONTYPE',

'PERSONCOUNT',

'PEDCOUNT',

'PEDCYLCOUNT',

'VEHCOUNT',

'JUNCTIONTYPE',

'SDOT\_COLCODE',

'UNDERINFL',

'WEATHER',

'ROADCOND',

'LIGHTCOND',

'SPEEDING',

'ST\_COLCODE',

'SEGLANEKEY',

'CROSSWALKKEY',

'HITPARKEDCAR',

'TIMEOFDAY',

'YEAR',

'MONTH',

'DAY',

'SEVERITYCODE']

These are the attributes along with the SEVERITYCODE that will be used for data visualization and

building the prediction model.