Background

According to WHO, the number of road traffic deaths rising steadily up to 1.35 million in 2016. It is the 8th leading cause of death, less likely to survive than AIDs.

Problem

Prevention is always better than cure. This project is to examine the factors of car accidents with the dataset

Data sources

The data I will us is obtained from this class's, provided by SDOT Traffic Management Division, Traffic Records Group, from 2004 to Present.

Data cleaning and Feature selection

There were 194673 samples and 38 features in the data. Looking into the features, some id-type features don't help the analysis and thus dropped out. Duplicated and details of the crashes will also be removed.

The remaining features and data types are shown as follow:

['SEVERITYCODE', 'X', 'Y', 'LOCATION', 'INCDATE', 'INATTENTIONIND', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND', 'SPEEDING', 'HITPARKEDCAR', 'PEDCYLCOUNT']

```
SEVERITYCODE       int64
X                float64
Y                float64
LOCATION          object
INCDATE           object
INATTENTIONIND    object
UNDERINFL         object
WEATHER           object
ROADCOND          object
LIGHTCOND         object
SPEEDING          object
HITPARKEDCAR      object
PEDCYLCOUNT        int64
dtype: object
```

There is one more problem regarding the prediction group, the samples are so imbalanced and will be resampling to equal amount

```
1    136485
2     58188
Name: SEVERITYCODE, dtype: int64
```