**Data Understanding**

The dataset has 221525 rows and 40 different variables. We will take

Severity Code to be our dependent variable and various variables will be used as our independent variables. The data is quite large and has a lot of missing values that will have to be treated and variables will have to be dropped.

The dependent variable has 5 values:-

|  |  |
| --- | --- |
| 0 | Unknown |
| 1 | Prop Damage |
| 2 | Injury |
| 2b | Serious Injury |
| 3 | Fatality |

Other important variables are: -

1. X and Y

The longitude and latitude of the accident

1. ADDRTYPE

Collision Address type: -

* Alley
* Block
* Intersection

1. SEVERITYDESC

A detailed description of the severity of the collision

1. COLLISIONTYPE

The type of collision

1. PERSONCOUNT

The total number of people involved in the collision

1. PEDCOUNT

The number of pedestrians involved in the collision

1. PEDCYLCOUNT

The number of bicycles involved in the collision

1. VEHCOUNT

The number of vehicles involved in the collision

1. INJURIES

The number of total injuries in the collision

1. SERIOUSINJURIES

The number of serious injuries in the collision. This is entered by the state

1. FATALITIES

The number of fatalities in the collision

1. INCDATE

The date of the incident

1. INCDTTM

The date and time of the incident

1. JUNCTIONTYPE

Category of junction at which collision took place

1. INATTENTIONIND

Whether or not collision was due to inattention

1. UNDERINFL

Whether or not a driver involved was under the influence of drugs or alcohol

1. ROADCOND

The condition of the road during the collision

1. LIGHTCOND

The light conditions during the collision

1. PEDROWNOTGRNT

Whether or not the pedestrian right of way was not granted

1. SPEEDING

Whether or not speeding was a factor in the collision.

1. HITPARKEDCAR

Whether or not the collision involved hitting a parked car.