1. Data Understanding

The data provided by SPD and traffic Records of Seattle consists of 38 columns and around 2 lakh accidents (represented by each row). Different factors and measures for each record are put in the table format.

Now the data has several information for each accident. What kind of accident? What was the severity, Environment condition, traffic conditions, direction of cars, whether or not a pedestrian was involved etc.

Some are related to the driver, some to environmental condition and some to manual error or design of the road.

Our goal here is to find out how much role a situation played for an accident. For example, we may want to see how much role inattention plays for an accident or is it the road condition that plays important role here.

Important fields:

Some important fields from the data table are:

1. SEVERITYCODE (Text):

Severity of the accident is identified by severity code as:

• 3—fatality • 2b—serious injury • 2—injury • 1—prop damage • 0—unknown

This will also be our target variable in modelling.

2. ROADCOND (Text)

Condition of the road. (example : Wet, Dry)

3. LIGHTCOND (text)

Light conditions during the collision

4. WEATHER (text)

Description of weather during collision

5. SPEEDING (Text : Y/N)

Weather or not speeding was a factor in the collision.

6. COLLISIONTYPE (Text)

7. LOCATION (Text)

8. JUNCTIONTYPE (Text)
   Category of junction at which accident took place.

9. INATTENTIONIND (text)

Whether or not collision was due to inattention.

10. UNDERINFL (Text)

Was the driver under the influence of alcohol or drugs or not?

Data preparation and cleaning:

For gathering the data, we will use Pandas library in Jupyter notebook. For preparing the data for analysis, we need to drop irrelevant data columns. Fix data types and formats if required.

Modelling and evaluation:

For modelling and evaluation, we will K-nearest neighbour and Linear regression. We will find out K and other evaluation KPIs to support our model.