# Data

### DATA SOURCE.

Data provided by Coursera will be used fot this project. It could be downloaded from

<https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv>.

Dataset was downloaded to local drive and opened for exploration and description. A file with metadata associated to the dataset was downloaded also.

### DATA UNDERSTANDING

Dataset consists of a CSV file containing data about car accidents in the city of Seattle in the period from January 2004 to May 2020. There are 194673,00 records of traffic accidents, with 38 columns representing different features of every incident. Among the features can be found the *Severity Code,* which will be the target variable, having the following values and meaning:

| **Value** | **Meaning** |
| --- | --- |
| 3 | fatality |
| 2b | serious injury |
| 2 | injury |
| 1 | prop damage |
| 0 | unknown |

Features that could be used as predictors are:

| **Feature** | **Description** |
| --- | --- |
| ROAD AND WEATHER CONDITIONS |  |
| ADDRTYPE | Collision address type: • Alley / • Block • Intersection |
| WEATHER | A description of the weather conditions during the time of the collision. |
| ROADCOND | The condition of the road during the collision. |
| LIGHTCOND | The light conditions during the collision. |
| DRIVER RELATED FACTORS |  |
| SPEEDING | Whether or not speeding was a factor in the collision. (Y/N) |
| INATTENTION | Whether or not, driver's innatettion was the cause of the accident |
| UNDERINFL | Whether or not, driver was under influence of alcohol or drugs |
| OTHERS |  |
| LOCATION | Description of the general location of the collision |
| SEVERITYDESC | A detailed description of the severity of the collision |
| INCDATE | Date of the incident |
| X | X coordinate of accident location |
| Y | Y coordinate of accident location |

### Feature selection.

After checking Metadata and basic statistic of the dataset, most of features were considered useless, with insufficient records or redundant for analysis. Features and reason to keep them are shown in the following table:

| **FEATURE** | REASON FOR KEEPING IT |
| --- | --- |
| X / Y COORDINATES | To see if any location have higher ocurrency of accidents than other |
| ADDRTYPE | To find relatioship between |
| INCDATE / INCDTTM | To see how is fluctuates accident ocurrence during the year and during the day |
| WEATHER / ROADCOND / LIGHT | These are expected to be the features with the major impact on the prediction of target value |

Features like INATENTTION , SPEEDING or UNDERINFL, related to driver conditions, were also kept so far because they might be considered important factors for predicting accident ocurrence and severity. Nevertheless, some issues were found on these records.

| **FEATURE** | **ISSUE** |
| --- | --- |
| SPEEDING | There are only 9333 records. The rest are empty values |
| UNDERINFL | There are categorical ( Y/N ) and numerical (1/0) values |

Further analysis will help to decide if these features should be considered for the assesment.

It will be estimated the correlation between the target variable and every feature.  
Finally, various Machine Learning algorithm will be evaluated to quantify their performance in the prediction of car collision ocurrence and severity of these accidents.