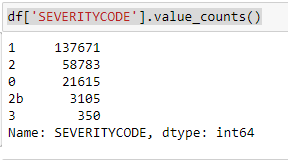
Capstone Project

Car Accident Severity

# About the dataset

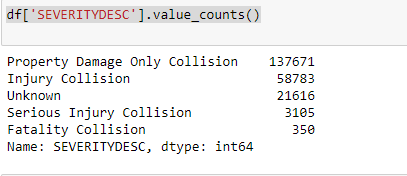
The data for this project is taken from an open source website - [**https://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab\_0/data**](https://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0/data)

The data consists of 39 independent variables and 221525 rows. The dependent variable, “**SEVERITYCODE**”, contains codes that correspond to different levels of severity caused by an accident.



Severity codes are as follows:

* 0 : Unknown
* 1 : Property Damage Only Collision
* 2 : Injury Collision
* 2b : Serious Injury Collision
* 3 : Fatality Collision

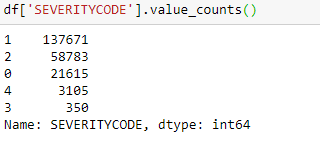


Furthermore, because of the existence of null values in some records, the data needs to be preprocessed before any further processing.

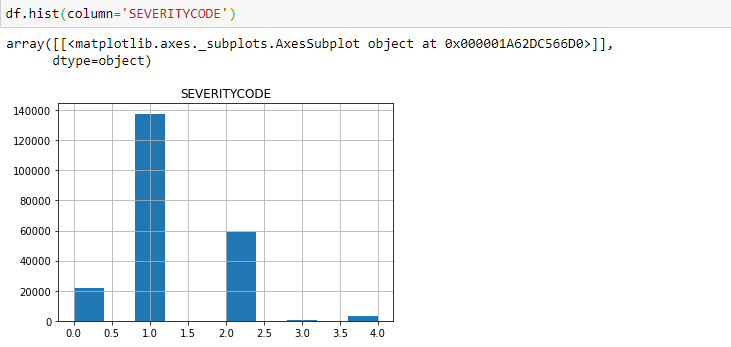
# Data Preprocessing

The dataset in the original form is not ready for data analysis. In order to prepare the data, first, we need to drop the non-relevant columns. In addition, most of the features are of object data types that need to be converted into numerical data types. We have to convert the **SEVERITYCODE** our target variable into numerical data type too.

For this we have updated the value of **SEVERITYCODE** ‘2b’ to ‘4’ and then updated the column data type.



To get a good understanding of the dataset, I have checked different values in the features. The results show, the target feature is imbalance, so we use a simple statistical technique to balance it.



The number of rows in class 1 is almost three times bigger than the number of rows in class 2. It is possible to solve the issue by down sampling the class 1.