1. **A description of the data and how it will be used to solve the problem. (15 marks)**

The dataset has an initial total of 194673 observations characterized by 38 different features. The feature that needs to be predicted is the “SEVERITYDESC”. It is a categorical one and takes two distinct values “Property Damage Only Collision” and “Injury Collision”.

Other features are both numerical and categorical. We have the environment, accident, location, type of location and accident, people and cars involved, and other accident describing features. The GitHub repository includes the data set in the csv format and a pdf file with a detailed description of the features and values included in the dataset. Therefore, please check the pdf file whenever is necessary to clarify the nature of the features or values.

The dataset contains many categorical features. Some of them take nominal values, thus no natural order to their values is present, instead are used for encoding purposes only, however other features describe the encodings. “SDOT\_COLDESC” and “SDOT\_COLCODE” are an example of such features. Either feature will have to be hot encoded eventually, thus as of now the preferred feature to keep will be the one that comes as a description, while another one will be dropped.

There are features that can be derived from one or more categorical features or values overlaps. Comparing “WEATHER” and “ROADCONDITION”, the “Rain” value relates to “Wet” condition. Thus, there is a strong correlation between some features, however, some of them will be kept and used to extract the most intelligence during the data exploration.

Eventually, those features that present excessive correlation will be removed as they are redundant and are harmful for the model.

The numerical features presented in the dataset are counters that indicate the number of pedestrians, cyclists, or vehicles involved in an accident.