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

DATA UNDERSTANDING

I decided to directly download the dataset from the official Seattle GeoData website (https://data-seattlecitygis.opendata.arcgis.com/). The advantage of this is that this dataset is updated weekly, so the model can be directly updated with the new entries each week. and continuously updated) reports all accidents that have been happening in Seattle from 2004 to present days.

The dataset is a csv file, consisting of 221738 rows and 40 columns at the time of writing (oct 7th, 2020). Each column, or attribute, is helpful in the overall description of the accident. For instance, they describe the address of the accident and the type of location (intersections or segments), whether the accident involved cars, bikers or pedestrians, and the total number of people involved in each accident. A number of attributes describes the external conditions surrounding the accident, such as the weather, the visibility, the road conditions, or the presence of drivers under influence.

An important column is the description of the severity of the accident, which has 5 possible outcomes: fatality, serious injury, inkury, or property damage only, plus a category for unknown severity. This will become the dependent variable (y) in our analysis, with other conditions being the independent variables (x): type of location, weather, light condition, road condition, type of collision. These are the attributes that I will use to train the machine learning models.

To help in this decision, during this phase I will use some simple plots to visualize the data and explore correlations: for example, plotting the number of people involved in accidents as a function of weather conditions.