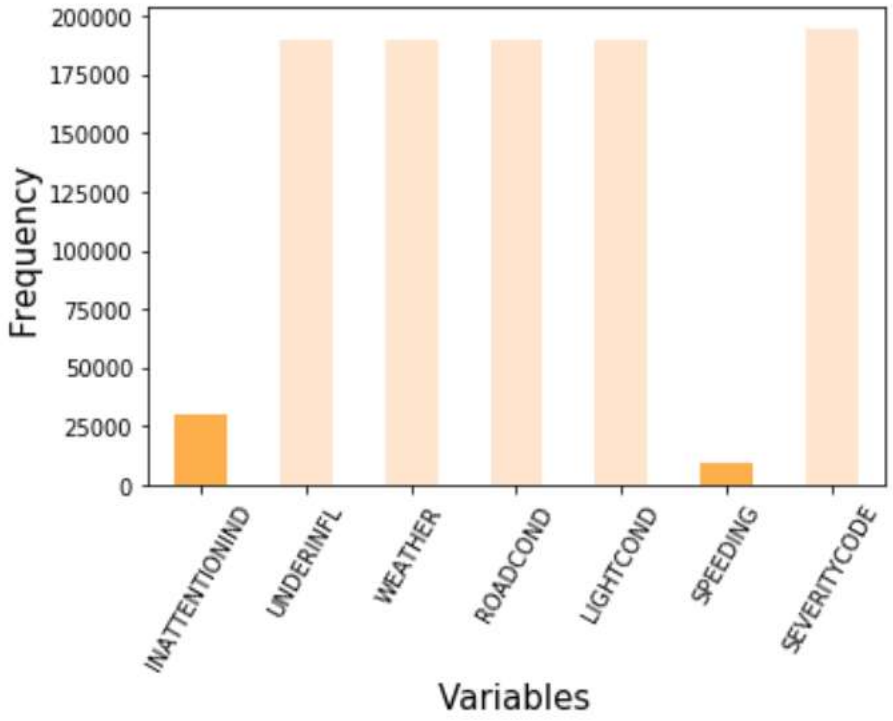
Data

The dataset used for this project is based on car accidents which have taken place within the city of *Seattle, Washington* from the year *2004* to *2020*. This data is regarding the *severity of each car accidents* along with the time and conditions under which each accident occurred. The data set used for this project can be found **here!**. The model aims to predict the severity of an accident, considering that, the variable of Severity Code was in the form of *1* (Property Damage Only) and *2* (Physical Injury) which were encoded to the form of *0* (Property Damage Only) and *1* (Physical Injury). Following that, *0* was assigned to the element of each variable which can be the least probable cause of severe accident whereas a high number represented adverse condition which can lead to a higher accident severity. Whereas, there were unique values for every variable which were either *Other* or *Unknown*, deleting those rows entirely would have led to a lot of loss of data which is not preferred.



In order to deal with the issue of columns having a variation in frequency, arrays were made for each column which were encoded according to the original column and had equal proportion of elements as the original column. Then the arrays were imposed on the original columns in the positions which had *Other* and *Unknown* in them. This entire process of cleaning data led to a loss of almost 5000 rows which had redundant data, whereas other rows with unknown values were filled earlier.