Coursera Capstone / Data understanding

Data Extract Summary

The dataset for this project is collision data from the City of Seattle, was downloaded from: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

Detailed information and metadata regarding this dataset is here: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf

There are 38 columns, three of which represent the collision severity (code, code.1, and description). This leaves 35 columns of potential predictive attributes. There are 194,673 rows of data. The data ranges in date from Jan 1 2004 to May 2020.

When looking only at data on collisions at intersections, the dataset reduces in size to 65070 rows.

The dataset has collision severity listed as one of: 3—fatality; 2b—serious injury; 2—injury; 1—property damage; 0—unknown. However, in this particular dataset, only severity levels of 1 and 2 are to be found.

Attribute Determination

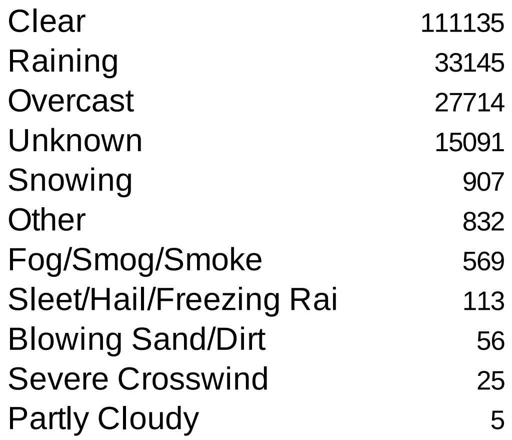
The data has the following columns: SEVERITYCODE X Y OBJECTID INCKEY COLDETKEY REPORTNO STATUS ADDRTYPE INTKEY LOCATION EXCEPTRSNCODE EXCEPTRSNDESC SEVERITYCODE.1 SEVERITYDESC COLLISIONTYPE PERSONCOUNT PEDCOUNT PEDCYLCOUNT VEHCOUNT INCDATE INCDTTM JUNCTIONTYPE SDOT_COLCODE SDOT_COLDESC INATTENTIONIND UNDERINFL WEATHER ROADCOND LIGHTCOND PEDROWNOTGRNT SDOTCOLNUM SPEEDING ST_COLCODE ST_COLDESC SEGLANEKEY CROSSWALKKEY HITPARKEDCAR

Of this list of columns, data that can be used on a (potential) predictive basis related to collisions is:

PEDCOUNT PEDCYLCOUNT VEHCOUNT JUNCTIONTYPE INATTENTIONIND UNDERINFL WEATHER ROADCOND LIGHTCOND PEDROWNOTGRNT SPEEDING CROSSWALKKEY HITPARKEDCAR.

Most of these 12 attributes are text based and will need to be converted to numeric codes or columns as one-hot encoding in order to be used flexibly within machine learning models.

In addition, some of the data elements are more detailed than needed, for example the value counts for WEATHER are as follows:



Clear	111135
Raining	33145
Overcast	27714
Unknown	15091
Snowing	907
Other	832
Fog/Smog/Smoke	569
Sleet/Hail/Freezing Rai	113
Blowing Sand/Dirt	56
Severe Crosswind	25
Partly Cloudy	5

It may be beneficial to reduce this to just whether or not the weather was uneventful (clear or partly cloudy or overcast) vs eventful. A similar mapping may be needed for ROADCOND, which is also likely to be highly correlated to WEATHERCOND.