1. **The Data**
   1. **Data Source**

The data set used in this project was provided by Coursera, but it could also be found (slightly different) at <https://data.seattle.gov/Land-Base/Collisions/9kas-rb8d>. In its original state the data set contains all collisions provided by the Seattle Police Department and recorded by traffic Records from 2004 to September 2019. This includes all types of collisions with a code to categorize the severity of the accident. In the original data set this can be one out of five different codes. The data set provided by Coursera, however, has already been pre-cleaned and altered a bit. One of the alterations is that the severity code now is either 1 for property damage only, or 2 for injury. The metadata can be found at <https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf>.

**2.1 Data Exploration and Cleaning**

The data set was available through a CSV file and was loaded into a Pandas dataframe object in a jupyter notebook file. The un-cleaned data set consisted of 37 unique columns (plus a duplicate column of the column containing the severity codes) and 194 673 rows. 37 columns and potential features are a lot, so the next step will be to determine which of these to use and which ones to drop.