# Predicting the Severity of Car Accidents

## David Good

## Draft 1

## 1. Introduction

### 1.1 Background

In the United States, most fatalities are generated by road vehicles. In 2016, more than 37,000 Americans died in road collisions. Seattle is one the worst US cities to drive in ranking above average in traffic, driving in precipitation, and accident likelihood. In 2017, Seattle police reported 10,959 motor vehicle collisions on city streets. According to SDOT, 28 percent of collisions result in some kind of injury, even if not severe or fatal. The impact of road accidents further extends to higher costs to the city, higher traffic, and higher financial strain on families. Therefore, it would be advantageous to discover factors that increase accident severity and problematic roads, so people know when and where to be more careful when driving.

### 1.2 Problem

Historical data on prior accidents including location, injuries, weather, road conditions, and collusion type would be helpful in determining factors that are most likely to increase accident severity. This project aims to predict accident severity in Seattle based on these data.

### 1.3 Interest

The city of Seattle would be very interested in accurate prediction of accident severity, so they best utilize their resources to mitigate accident impact. Residents and visitors of Seattle would also be interested in knowing the most historically impactful risk factors increase their likelihood of a severe auto accident.

## 2. Data acquisition and cleaning

### 2.1 Data Source

The Seattle Department of Transportation (SDOT) maintains detailed records of motor vehicle accidents in the Seattle. The dataset is provided by SPD on an individual collusion basis from January 2004 to May 2020. This is the only dataset used in this analysis.

### 2.2 Data Characteristics

There are 37 attributes SDOT strives to record for every accident. This section will give a brief description of some the most important attributes. Accident location (with regards to junction), weather conditions and light conditions are part the standards characteristics used in this analysis. Accident severity is classified by a severity code with 5 categories: Fatality, serious injury, injury, property damage and unknown. Additionally, each accident has a collision code that describes the type of collision. Collisions are broken down by 10 categories which include “parked car, sideswipe, and rear ended”

### 2.3 Feature Selection

This analysis looks at 6 features to predict severity of accident: Location, collision type, junction, weather, and light conditions. These attributes were picked to make this analysis more accessible to the general public. These attributed can quickly be gathered and utilized on a daily basis to determine how much precaution to use when driving. For example, weather and road conditions were two similar attributes in this dataset. Weather was chosen over road conditions because it is easier to determine the weather at any given in time as opposed to the road conditions.

### 2.4 Data cleaning

Severity code will be the target variable to predict accident severity. As mentioned previously, there are 5 severity codes used by SPD. This analysis will only look at severity codes: property damage and injury. Thus, focusing this analysis to predict factors that increase the chances of injury (1 or 0). Location data is stored in a longitude, latitude format. To increase usability, each location will be classified to one of seven of Seattle’s neighborhoods. To make the models more accurate, accidents that had any of the key features missing or classified as unknown were removed (approximately 13% of total data). After data cleaning, there are 168,437 samples in the data.