Predicting the Severity of Car Accidents in Seattle

**Yurgenis Rincón**

**October, 2020**

**1. Introduction**

**1.1 Background**

Seattle is the largest city of Washington state, its metropolitan area has a population of 3.98 million people, all this according to U.S Census data released in 2019, the city itself is hilly, though not uniformly so, it lies on seven hills like Capitol Hill, First Hill and more. Due to its location in the Pacific Ring of Fire, Seattle is in a major earthquake zone, on February 28, 2001, the magnitude 6.8 Nisqually earthquake did significant architectural damage. Seattle has a temperate climate, classified in the Mediterranean Zone by the main climatic classification*,* but some sources put the city in the Oceanic Zone. The Seattle area is the cloudiest region of the United States, due in part to frequent storms and lows moving in from the adjacent Pacific Ocean. With many more "rain days" than other major American cities.

**1.2 Problem**

Seattle residents get around by car, trolley, streetcar, public bus, bicycle, on foot, and by rail. With such bustling streets, it’s no surprise that Seattle sees car accidents every day. In 2015, a crash occurred in Washington [every 4.5 minutes](https://www.wsdot.wa.gov/mapsdata/crash/pdf/2015_Annual_Collision_Summary.pdf). Seattle recorded the highest number of car accidents in the state that year. There are many causes that provoke car accidents, and those car accidents have different kinds and levels of injuries, due to that, this project will predict the possibility of getting into a car accident and the severity of them, taking in mind the climate and the roads condition in Seattle.

**2. Data Acquisition and Cleaning**

**2.1 Data Sources**

The data to use in this project was obtained by SDOT Traffic Management Division, Traffic Records Group. The SDOT Traffic Management Division develops, maintains, and operates a transportation system that promotes the mobility of people and goods, and enhances the quality of life, environment, and economy of Seattle. The information of the data was collected between January, 2004 and May, 2020.

The data contain 194.673 rows and 38 columns. However there is a lot of lack of information, for example the column named *"PEDROWNOTGRNT*" has only 4667 rows of total data rows, and the data shows all necessary information related with collisions, car accidents, fatalities, injuries, weather and more details about those accidents.

**2.2 Data Wrangling**

As the first step, I visualized the data using the pandas’ command **.head**, **.shape**, **.columns**, to know the names of the attributes, what information there is in those attributes and to figure out what attributes will help me to achieve the objectives. In this step I noticed that there are several attributes that won’t fit with the purpose of this project, eventually I dropped those attributes, staying with only 19 attributes.

After that I looked for the missing values in the rows. Only 10 attributes had missing values, the most of them were attributes of *object* type, what I did was replace with the function **.replace** the missing values with the most common argument. However, there were other attributes that needed other types of treatment, for instance, the attributes **X, Y** and **LOCATION,** those attributes are geographical coordinates. The best way to take care of the missing values in those attributes is use the function **.dropna,** to eliminate all missing values of those specific attributes.