**Gathering Insights from and Predicting the Severity**

**of Traffic Accident in Seattle**

Fredrik F. Ulltang

1. **Introduction**
   1. **Background**

As with any large city in a developed country, the traffic situation in the city of Seattle is a complex one. Even though there’s been a continues decrease in serious injuries on the Seattle Streets, there’s still a long way to go to reach the city’s goal of zero traffic fatalities and serious injuries by 2030. Not all traffic accidents lead to injury however, and some result in property damage only. It could be of valuable information if one could identify the relevance of different factors surrounding a traffic accident and use these to predict whether a traffic accident will result in injury or property damage.

**1.2 Problem**

There might be several different attributes which play an important role in the severity of a car accident. Looking through these and analyzing them could provide a deeper understanding into what factors are present when accidents happen and which of these play a more significant role. So, can we identify what factors are present when traffic accidents happen, and the magnitude of these? Can we separate the accidents leading to injury from the ones leading to property damage, based on patterns among these factors?

* 1. **Stakeholders**

Indeed, if one could retrieve such insights and predict the severity of traffic accidents, this would be very beneficial for road users, the emergency services and city of Seattle as a whole. Road users could use this information to make assumptions on when and where to be extra cautious, and which routes are more exposed than others and should be avoided at certain times or when certain conditions are meet. The emergency services could allocate their resources and adapt their preparedness with this information as a reference point. The government could implement measures and take actions to prevent or mitigate such predictable behaviors and the consequences of these in the future.

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.

**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 was to determine which of these to use and which ones to drop.

First, the data type of each column was explored. There were two columns containing data on the date of the accident. Both of these was cast as the object data type and were converted to the datetime data type. Also, these two columns were almost identical, the only difference being that one of them contained the time of the accident as well.

Before looking at missing values the first round of dropping unnecessary features was done.