# Analyzing of Seattle road collisions and predicting their severity

**Tristan MOUYNA-HAINRY**
**September 19, 2020**

## 1. Introduction

### 1.1. Background

Accidents and collisions are the burdens of transportation, even if they tend to diminish in time – with the development of new technologies – they still exist and can lead to major problems of property destruction to serious injuries and death.

For example, in the city of Seattle (population of 609k in 2010), 483k people were involved in a collision from 2004 and 2019, i.e. 4 inhabitants in 5 (if people are not involved twice). During this period, around 190 collisions a week happened, of which 3 led to serious injuries or deaths.

Road accidents are thus an important problem for public authorities, people, and other stakeholders.

### 1.2. Problem

The goal of this project is to analyze Seattle road collisions and try to predict their severity. The data that might contribute to determine this might include the location of the collision, its type, whether it involves pedestrians, cyclists, other vehicles, when it occurred, and other driving conditions.

### 1.3. Interest

Public authorities (especially the City of Seattle) would definitely be interested in understanding in what kind of circumstances theses collisions happened, to better target how to prevent them – especially deadly ones – in a time of scarce resources. Others who can be also interested are insurers, using these insights to modify premiums accordingly, or simple citizens to be more aware of the most dangerous circumstances.

# 2. Data acquisition and cleaning

## 2.1. Data sources

I used the data from the City of Seattle website (https://data.seattle.gov/Land-Base/Collisions/9kas-rb8d). They are quite thorough and include most of the information sought.

## 2.2. Data cleaning

Since the objective is to predict the severity of a collision, I first removed all the instances for which the severity code was missing or equal to 0, as I was not able to label them.

Then I created a test set with a stratified shuffle split on the severity code, as the data were very imbalanced. I implemented the following data cleaning and feature selection in a pipeline so that I was able to repeat the process on the test set.

First, I replaced some values for specific features. Indeed, the ***under-influence*** feature had *'Y', '1', 'N', '0'* values. I thus replaced the *'Y'* by *'1'* and the *'N'* by *'0'*. Similarly, I replaced the *'Y', 'N'* values of the ***hit-a-parked*-*car*** feature by *'1',* and *'0'*.

Second, I binned the continuous features. These features are the coordinates ***X*** and ***Y*** of a collision. I binned them so that the collision could be located in a 100m$^2$ square (10m x 10m).

Third, I created new features.

Mostly, the new features are a grouping of different values for an existing feature.
Thereby, I created a ***good-weather*** feature dividing the ***weather*** feature into *'1'* (for *'Clear', 'Overcast', 'Severe Crosswind', 'Partly Cloudy'*) and *'0'* (for *'Raining', 'Snowing', 'Fog/Smog/Smoke', 'Sleet/Hail/Freezing Rain', 'Blowing Sand/Dirt', 'Blowing Snow'*), and *NaN* (for *NaN* and *'Other'*).
I did the same for the ***road condition*** feature, for which I created a ***good-road*** feature (with a *'1'* when *'Dry'*, a *'0'* when *'Wet', 'Ice', 'Snow/Slush', 'Standing Water', 'Sand/Mud/Dirt', 'Oil'*, and a *NaN* when *NaN, 'Unknown', 'Other'*).
I created a new ***light-condition*** feature in which I regrouped some of the old one's values *('Dusk'* and *'Dawn'*; *'Dark - No Street Lights'* and *'Dark - Street Lights Off'*; *'Dark - Street Lights On'* and *'Dark - Unknown Lighting'*).
I created a ***with-pedestrian*** feature, using the ***pedestrian-count*** feature, and taking the value *'0'* when no pedestrian is involved in the collision, and *'1'* otherwise.
I did the same with the ***with-cyclist*** feature.
I did similarly with the ***with-vehicle*** feature, except that this new feature is taking a *'0'* when no vehicle is involved, a *'1'* when only 1 vehicle is involved, a *'2'* when only 2 vehicles are involved and a *'3'* otherwise.
I also created a ***time***, ***month*** and ***day*** features, respectively giving the hour of the collision (from *'0'*h to *'23'*h), its month (from *'1'* to *'12'*) and the day of the week it happened (from *'0'*: Monday to *'6'*: Sunday).

Fourth, I dealt with missing values.

The feature with the most important number of missing values was ***time***. As it was more than 15% of the values, I decided to drop all the missing values for that feature.

As it was quite similar for ***good-weather***, ***good-road***, and ***light-condition***, I choose to drop all their missing values.

The ***under-influence*** feature also had missing values. As it was less than 3% and well distributed among the different severity-type collisions, I choose to replace them by the most common value.

As it was the same thing for ***X*** and ***Y*** binned, ***junction-type***, ***collision-type*** features, I choose to replace their missing values by their median or their most common values.

For the ***with-pedestrian***, ***with-cyclist*** and ***with-vehicle*** feature, the missing values are when the three of them equal 0. In that case, I choose the most common values to replace them (i.e. ***with-vehicle*** equals 2, and the others two equal 0).

Finally, I created dummies variables for ***light-condition, with-vehicle, junction-type and collision-type*** features.

## 2.3. Feature selection

The data were initially composed of 40 features. Some of them were collisions identification numbers for the city or the state (***OBJECTID***, ***INCKEY***, ***COLDETKEY***, ***REPORTNO***, ***SDOTCOLNUM***), or meant nothing or were empty (***STATUS***, ***EXCEPTRSNCODE***, ***EXCEPTRSNDESC***), or were composed of too many missing values (***INTKEY***: >67%, ***INATTENTIONIND***: >86%, ***SPEEDING***: >95%, ***PEDROWNOTGRNT***: >97%, ***CROSSWALKKEY***: >98%, ***SEGLANEKEY***: >98%). I, therefore, dropped all these features.

The rest of the features could be split into 6 different groups: space-location, time-location, type of collision, exterior conditions, intrinsic conditions, and of course severity features.

Obviously, the severity-related features (***SEVERITYCODE***, ***SEVERITYDESC***, ***INJURIES***, ***SERIOUSINJURIES***, ***FATALITIES***) were only kept as a target feature in the following computation, and I only used the ***SEVERITYCODE*** feature for that.

Thereby, the features I kept for the analysis are the following:

- Space-Location: ***X, Y***, ***junction-type***
- Time-Location: ***time***, ***month***, ***day***
- Type of collision: ***collision-type***, ***number-of-persons-involved***, ***with-pedestrian***, ***with-cyclist***, ***with-vehicle***, ***hit-a-parked-car***
- Exterior conditions: ***good-weather***, ***good-road, light-condition***
- Intrinsic conditions: ***under-influence***

I dropped ***ST_COLCODE***, ***ST_COLDESC*** and ***SDOT_COLCODE, ST_COLDESC*** because they gave the same information as the type of collision's ones, had more missing values and were not completely coherent with one another.

I also dropped the ***ADDRTYPE*** and ***LOCATION*** features because they gave the same information as the space-location's ones.