# VEHICLE COLLISIONS IN SEATTLE

IBM CAPSTONE | BY ARTEM SHEVCHUK

# Contents

## 1. INTRODUCTION

Road accidents have become very common nowadays. As more people are buying automobiles, the incidences of road accidents are just increasing day by day. According to The Annual Global Road Crash Statistics report published by **www.asirt.org**, approximately 1.35 million people die in road crashes each year, and on average 3,700 people lose their lives every day on the roads. An additional 20-50 million suffer non-fatal injuries, often resulting in long-term disabilities.

Considering the importance of this problem, we have to try our best to understand what factors affect the probability of being injured during the car crash the most. To achieve this, we will use the modern tools data science has to offer.

This research is based on collisions data from Seattle; however, I believe that the derived conclusions can serve as a benchmark for all other cities and counties.

This report will be of interest to anyone who drives cars or works in traffic regulation authorities. Understanding the roots of this problem may bring us one step closer to making our cities a safer place to live and may save the lives of those readers who take this issue seriously.

## 2. A DESCRIPTION OF DATA

Every data-driven solution requires one most important thing: data. In our research, we will use the Traffic Collisions Dataset compiled and maintained by the Seattle Department of Transportation (SDOT). This dataset has been referred to within the IBM capstone project, and therefore we have no doubts regarding its authenticity and quality.

Below is the **quick summary** of our dataset:

a. Range: 2004 to Present
b. 37 Attributes
c. Almost 195,000 rows (observations)

The key intention of our research will be an attempt to measure the **impact of various characteristics on the severity** of the accident. It will be helpful to mention what severity codes correspond to:

| Code | Meaning |
|---|---|
| 3 | Fatality |
| 2b | Serious injury |
| 2 | Injury |
| 1 | Property damage |
| 0 | Unknown |

Our dataset will contain only the following severities:

- ✓ 1 (Property Damage, 136 485 observations)
- ✓ 2 (Injury, 58 188 observations)

After investigating the dataset, I decided to keep only the most important attributes, and the resulted data frame after basic pre-processing looks the following way:

| | SEVERITYCODE | ADDRTYPE | UNDERINFL | WEATHER | ROADCOND | LIGHTCOND | SPEEDING |
|---|---|---|---|---|---|---|---|
| 0 | 2 | Intersection | N | Overcast | Wet | Daylight | NaN |
| 1 | 1 | Block | 0 | Raining | Wet | Dark - Street Lights On | NaN |
| 2 | 1 | Block | 0 | Overcast | Dry | Daylight | NaN |
| 3 | 1 | Block | N | Clear | Dry | Daylight | NaN |
| 4 | 2 | Intersection | 0 | Raining | Wet | Daylight | NaN |

Let us have a quick look at what each column represents and what actions we might need to take to prepare it for analysis:

| Column | Description |
|---|---|
| SEVERITYCODE | Already described in the previous table |
| ADDRTYPE | Whether an incident occurred at an intersection or block. Requires 0/1 encoding |
| UNDERINFL | Whether a driver was under influence of drugs or alcohol. Requires 0/1 encoding. Contains NA values which will be filled with 0 by default |
| WEATHER | Weather conditions during the accident. The column includes 11 types of conditions. For efficiency, they will be converted to FAVORABLE (0) and UNFAVORABLE (1) |
| ROADCOND | Road conditions during the accident. The column includes 9 types of conditions. For simplicity, they will be encoded to FAVORABLE (0) and UNFAVORABLE (1) |
| LIGHTCOND | Light conditions during the accident. The column includes 9 types of conditions. For simplicity, they will be encoded to DAY (0), OTHER (1) |
| SPEEDING | Whether an incident occurred while the driver was obeying or disobeying the speed limit. Requires 0/1 encoding. |

Other important points:

- ✓ The data has unbalanced labels (more property damage in comparison to injuries). To avoid bias, we will have to trim it and keep an equal amount of labeled observations for both categories.
- ✓ Some categories like weather, road, and light conditions contain “unknown” observations. To avoid guessing, we will just remove those observations entirely. I believe this will not have a significant statistical impact on the quality of research, as the overall number of recorded rows is still very large.

I will not include the long list of all processing actions taken here; they will be available in my Git repository in a separate workbook. After processing the data this is the final table we arrived at:

| | SEVERITYCODE | ADDRTYPE | UNDERINFL | WEATHER | ROADCOND | LIGHTCOND | SPEEDING |
|---|---|---|---|---|---|---|---|
| **0** | 2 | 1 | 0 | 1 | 1 | 0 | 0 |
| **1** | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| **2** | 2 | 1 | 0 | 1 | 0 | 0 | 0 |
| **3** | 2 | 0 | 0 | 0 | 0 | 1 | 0 |
| **4** | 2 | 1 | 0 | 0 | 0 | 0 | 0 |

- ✓ The shape of data: 7 columns (1 label, 6 categorical variables)
- ✓ All categorical variables are encoded in 1/0 format
- ✓ The total number of rows is 96 662
- ✓ Both Severity Codes 1 and 2 have an equal amount of observations (48 331 each)

We can now move to our next “Methodology” section, where we will make a quick visual exploratory analysis, and then try to use some Machine Learning techniques to predict what conditions can be linked to severity codes.