Seattle Collisions (2003-2020) Data Analysis

By Shane Mullins on Aug 20, 2020

Each year, approximately 1.35 million people are killed on roadways around the world. Road traffic accidents are the leading cause of death for people between the ages of 5-29 years worldwide. It is clear that our transportation systems and attitude towards driving are important areas to focus on. Over the past 16+ years, the Seattle Department of Transportation (SDOT) has been gathering data on collisions in the city. The purpose of this article is to unveil some interesting truths about the nature of collisions in Seattle. With this enhanced understanding, the hope is to extrapolate these insights to large cities around the globe and to implement effective societal and governmental change to curtail the severity and frequency of such collisions.

Some interesting questions we look to answer are:

- What proportion of accidents take place whilst a driver is under the influence?
- During what period of the day do the most collisions occur?
- Is there an increasing or decreasing trend in the number of collisions per year?
- What is the relationship between road conditions and the severity of collisions?
- How accurately can we predict the severity of a collision given certain pieces of information?

In order to determine the correlation between the attributes of our dataset (time of day, fatalities, collision type…) and the predictability of severity class, a **K-Nearest Neighbour (KNN) machine learning model was built**. A KNN algorithm was deemed to be more appropriate over others - such as Support Vector Machine (SVM) or Logistic Regression - due to the fact that we were dealing with multiclass classification. Using cross-validation techniques, the optimal k value was chosen. The model predicted the severity of each collision, with approximately 91% accuracy, given certain information about the collision itself. This accuracy was measured using the Jaccard Index and F1-Score metrics.

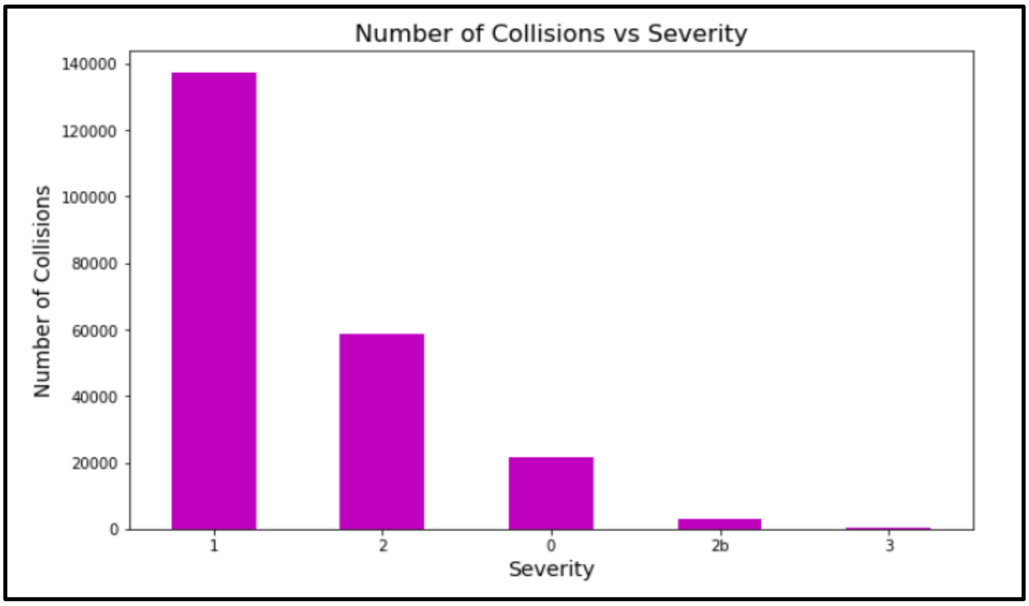

Figure 1: Severity categories: 0, 1, 2, 2b and 3 in increasing order of severity

It is important to be clear about what the data is telling us. We must make the distinction between inferences about the topic of interest and inferences about the dataset and its categorisation.

Naturally, we start by determining the number of collisions per severity category as seen in figure 1. This helps us to understand the dataset we are working with and to gauge the severity of typical collisions in Seattle. One might surmise that most of the collisions were innocuous due to the dominance of category 1. However, we must be aware that this could also be a consequence of a broadly defined category 1. One such definition could be: ‘collisions which result in no injuries and minor damage to the car’. Interestingly, the least severe crashes occur less frequently than one might expect. One other basic, but valuable, question to ask is: which types of collisions occur most often? According to the World Health Organization (WHO), **pedestrians, cyclists and motorcyclists make up almost 50% of road-related deaths**. Figure 2 supports this statistic. The following table provides additional statistics on those involved in collisions of varying severity.

Severity Class	Ped/Cyc Involvement	Serious Injuries/Fatalities
0	0	0
1	0.010	0
2	0.196	0
2b	0.457	1.052
3	0.511	1.363

Table 1: Shows typical number of 1) pedestrians and/or cyclists and 2) serious-injuries/fatalities involved in a collision.

The graph below details the collision type and those involved in the collision. Due to the variety of collisions, 'Other' describes a considerable amount of the data.

Another attribute took the form of a single integer number/code which provided a more detailed description of the collision type. It was discovered that code 11 was associated with 91,749 collisions, 14 with 59,092 collisions and 0 with 19,133 collisions. For the dataset containing only the most severe collisions (severity = 3), 24, 11 and 28 were the most frequent codes. Below are their descriptions:

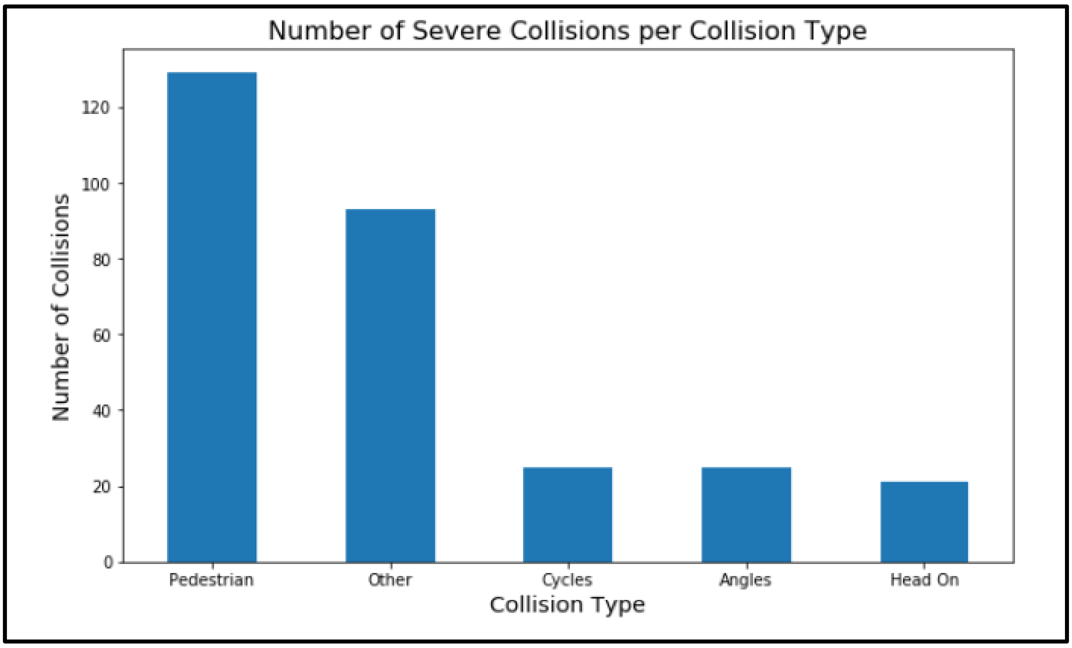

Figure 2: Proportion of collisions per collision type

0: Vehicle Going Straight Hits Pedestrian
11: From Same Direction -Both Going Straight-Both Moving- Sideswipe
14: From Same Direction - Both Going Straight - One Stopped - Rear End
24: From Opposite Direction - Both Moving - Head On
28: From Opposite Direction - One Left Turn - One Straight

We can learn a lot about the nature and cause of collisions from the locations t which they occur. Below is a map of Seattle displaying clusters of the most severe collisions.

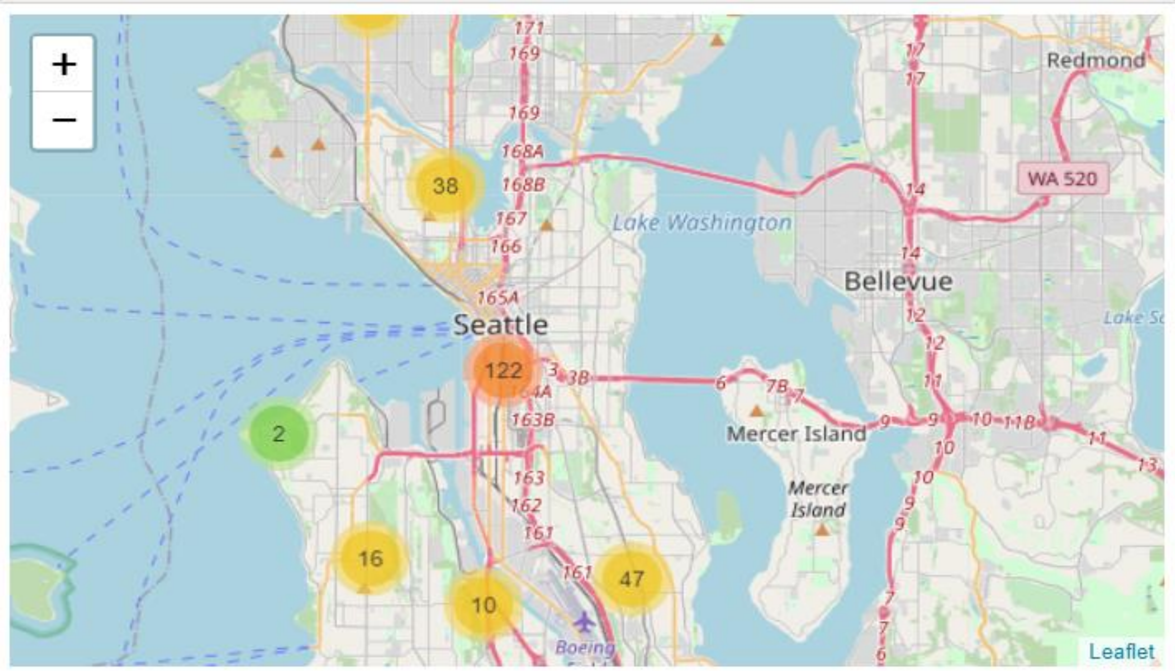

Figure 3: Map of collision clusters

It is no surprise that the largest cluster of collisions is located at the centre of the city. A **denser population means more cars, less space and a higher chance of collision**. We also note that other large clusters appear near the airport and main roads. This all points to the same obvious fact: more cars -> more collisions. This statistic is more of a confirmation of what we already expected rather than an introduction to new information.

It is fascinating to see the correlation between the time of day and the number of (severe) collisions. We see an **upward trend of severe collisions throughout the day**, perhaps due to tiredness or the nature of evening activities.

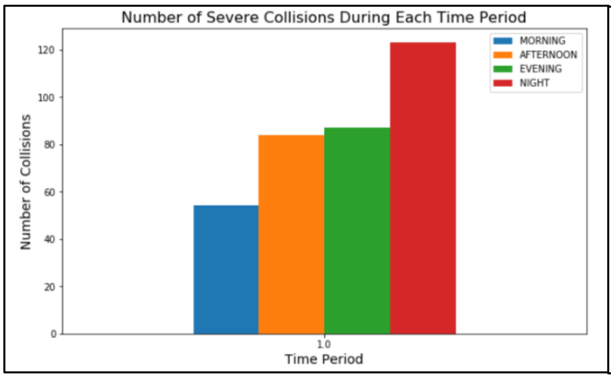

Figure 4: proportion of severe collisions associated with each time period

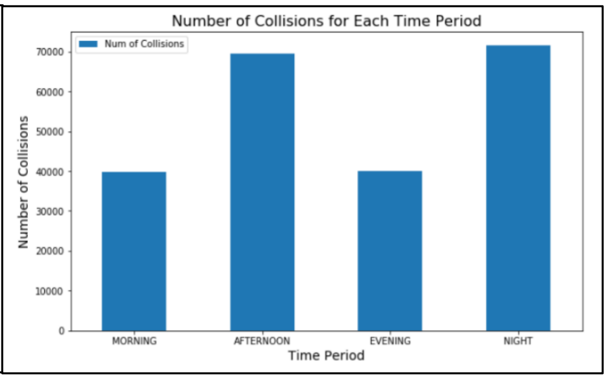

Figure 5: proportion of overall collisions associated with each time period

These graphs lead us to believe that more attention and resources need to be focused towards controlling the dangers of driving at night. This conclusion is supported also by the following bar charts.

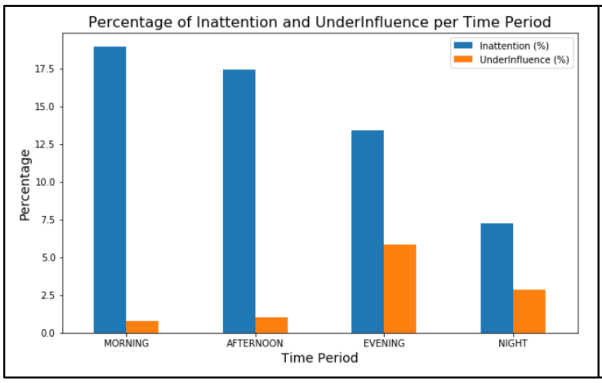

Figure 6: Proportion of drivers who were either inattentive or under the influence during a collision, per time period.

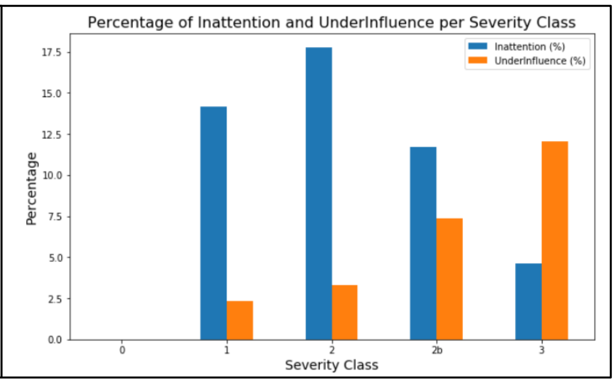

Figure 7: Proportion of drivers who were either inattentive or under the influence during a collision, per severity class.

We see a rise in ‘lack of attention’ and ‘under the influence’ driving throughout the day. Approximately **12.5% of all severe crashes occur with a driver who was under the influence**. Clearly, this is an issue that warrants our increased attention. We also see that inattention exhibits a downward trend throughout the day.

Pie charts have their limitations. However, in this case, they provide us with valuable information on typical weather, road and lighting conditions for a collision. This information is useful for all road users to compare current conditions to those that give rise to a higher probability of collision.

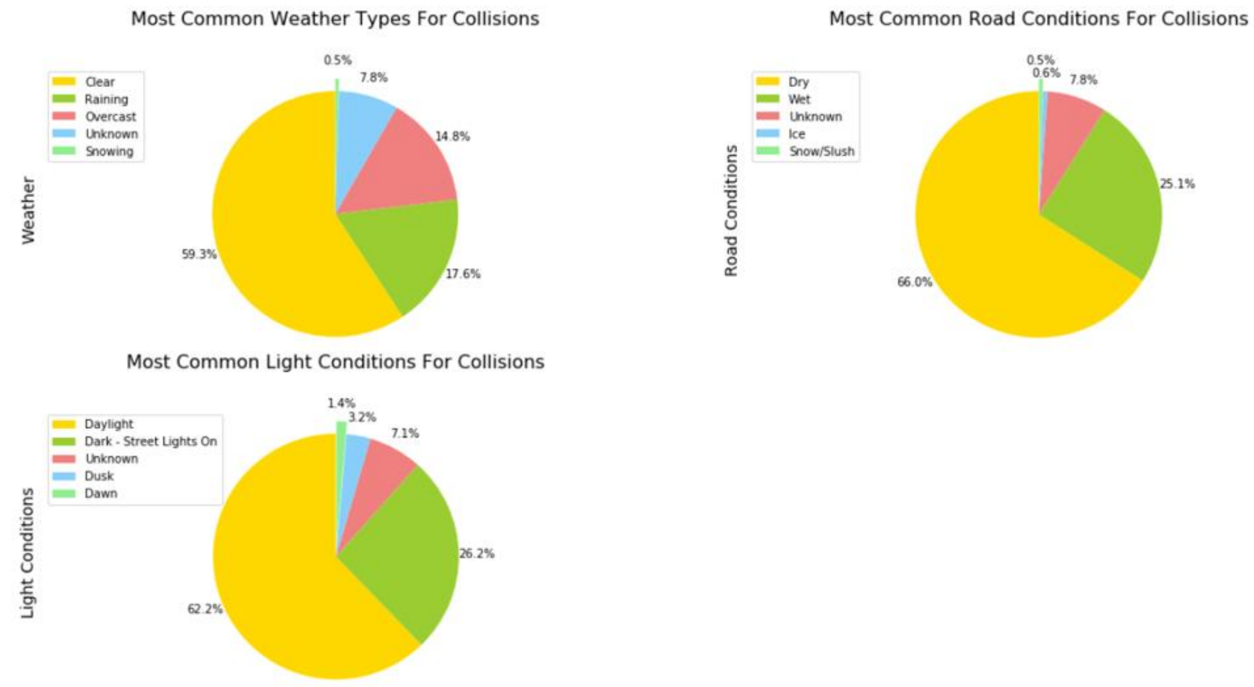

Figure 8: Most common weather, road and lighting conditions and their proportions for collisions

And finally, we end on a more positive note! We see in Figure 9 that the **number of collisions per year is decreasing considerably, from roughly 16000 in 2005 to 11500 in 2019**. In early 2015, an initiative called 'Vision Zero' (see References for further information), was launched with the aim of alleviating all traffic deaths/serious-injuries by 2030. This was most likely driven by the surge in such events between 2010 and 2015 as seen in the graph. This initiative seems to have had a positive impact on the overall number of collisions in Seattle after 2015. With insights such as those discussed in this article, road users can make life-changing decisions about their travel, governments can intelligently allocate their spending on roads and infrastructure, and car manufacturers can focus on the right details to improve the overall safety of their cars. Smarter use of government funding, a reduced negative impact on the environment and a quantifiable increase in lives saved are just a few of the possible benefits that such changes could bring.

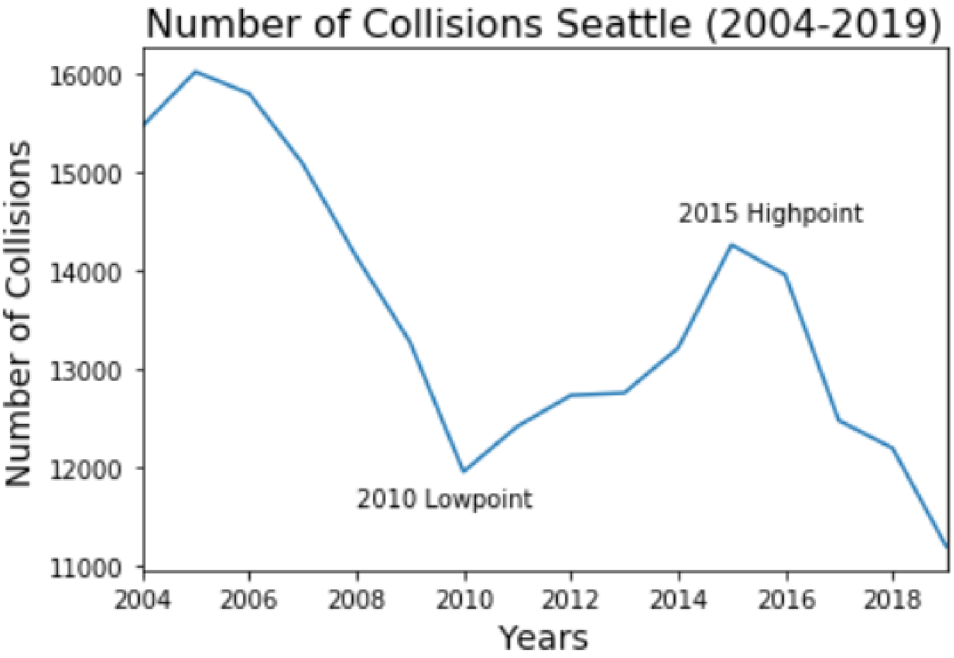

Figure 9: Number of total collisions per year 2004-2019

The 'Vision Zero' initiative is a great example of how the informed decisions we make can have a considerable impact on our progression towards a safer society. It is often difficult to find the relationship between policies/initiatives implemented and measurable societal improvement. The goal for governments and companies alike, is to manipulate the former to bring about positive-sum progression.

Intelligent analysis of data is imperative when it comes to achieving such progression. Often, its power lies in confirming what was previously suspected. It facilitates the replacement of the anecdotal with the empirical. It is this strengthening of our presuppositions that allows us to strive forward in the pursuit of our objectives.

References

Kaggle - https://www.kaggle.com/jonleon/seattle-sdot-collisions-data

WHO - https://www.who.int/gho/road_safety/mortality/traffic_deaths_distribution/en/

Vision Zero - https://www.seattle.gov/visionzero