Seattle Collisions (2003-2020) Data Analysis

AUGUST 2020

Authored by: Shane Mullins

Table of Contents

Introduction

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 report 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?

With insights such as those discussed in this report, 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 about.

Data Understanding and Preparation

Dataset

The dataset pertains to collisions that occurred in Seattle from 2004 – present. It was obtained from Kaggle.com and prepared by the state, which is of course deemed to be a reliable source. It contains an abundance of informative and representative data with 39 attributes and 220812 rows. Such attributes include location, pedestrian count, date of accident, weather conditions, injuries… Our target variable (which we aim to provide an accurate machine learning model for) is the category of severity attributed to each collision. Naturally, we start by determining the number of collisions per severity category as seen in figure 1.

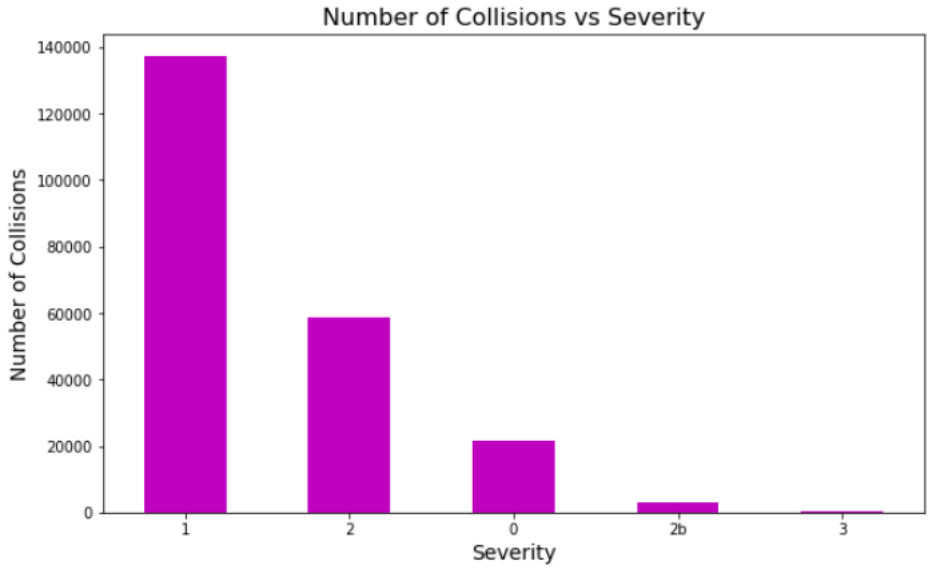

Figure 1: Severity categories: 0, 1, 2, 2b and 3 in increasing order of severity

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. 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. We can see that this is an unbalanced dataset due to the overwhelming amount of entries for class 1. This will need to be fixed when building our machine learning model. This is discussed in the following section.

Data Preparation

We augmented the dataset with several additional columns extracted from the data itself. This was done in order to accurately summarize certain characteristics and determine whether our new column would perhaps display a strong relationship between other attributes and/or the target variable. Such columns included total pedestrian and cyclist involvement and total serious injuries or fatalities associated with the collision. Additionally, the columns 'morning', 'afternoon', 'evening' and 'night' were added. These represent broader categories of the original date column. Certain rows contained useless data such as Na values or blanks. Such data was either removed or replaced by a reasonable value. An example of this was the attribute that described whether the driver was under the influence. It contained a large proportion of Na values which could not be otherwise interpreted. This column was dropped.

The following attributes were chosen for the model: 'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'INJURIES', 'SERIOUSINJURIES', 'FATALITIES', 'SDOT_COLCODE', 'INATTENTIONIND', 'UNDERINFL', 'SPEEDING', 'ST_COLCODE', 'MORNING', 'AFTERNOON', 'EVENING', 'NIGHT'. All attributes must be in numerical form and contain no Na/blank values. The new dataframe needed to be balanced so as not to introduce biases to the model. This was achieved by splitting the data by severity class and taking random samples of these groups. The size of these groups amounted to the size of the smallest class i.e. class 3 with roughly 350 entries. The data

was then accumulated into one dataframe and used for the model. This dataframe contained 1352 rows and 15 columns/attributes. This data was then normalized so as to diminish any unwanted weighting of larger values.

Methodology

EDA and Inferential Statistics

Exploratory data analysis (EDA) was executed, and inferential statistics tests were performed on the larger dataset in order to answer the questions we had about the data. Initially, we obtained some informative statistics on the dataset with the aim of gaining a better understanding of the dataset as a whole and to perhaps, unveil some useful facts.

- The average vehicle count per collision was 1.73 with a max of 15 vehicles. This is surprising as we would imagine that most collisions would involve greater than 1 vehicle.
- The average fatalities per collision were 0.0017 with a max of 5 people.
- The average injuries per collision were 0.374 with a max of 78 people, which tells us that fewer than 1 people were injured per collision.

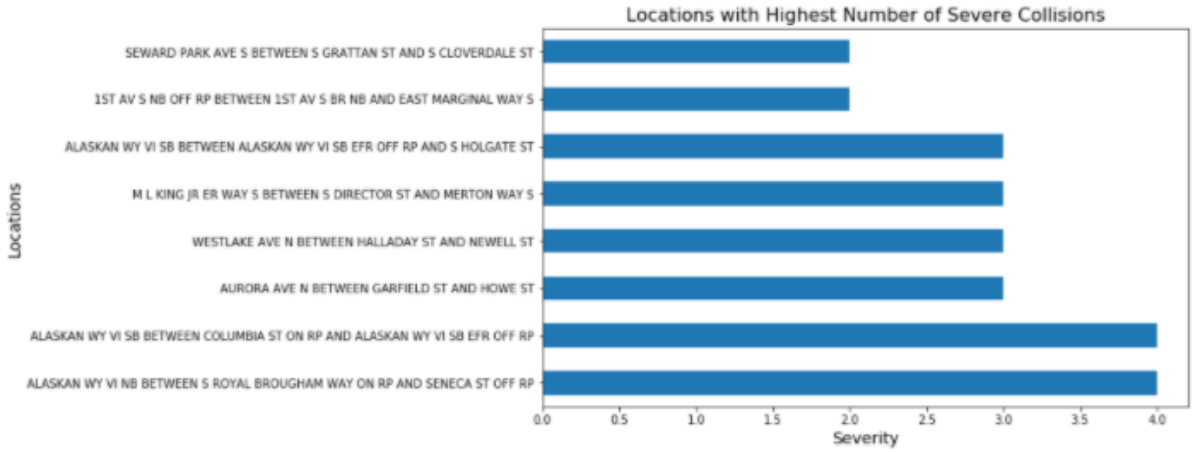

Figure 2: Seattle locations associated with the most collisions

We see in Figure 2 that the max number of collisions that occurred in one place was four. This is a low figure considering the dataset covers roughly sixteen years. This is most likely due to the fact that locations were split up into very specific areas such as streets or intersections.

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

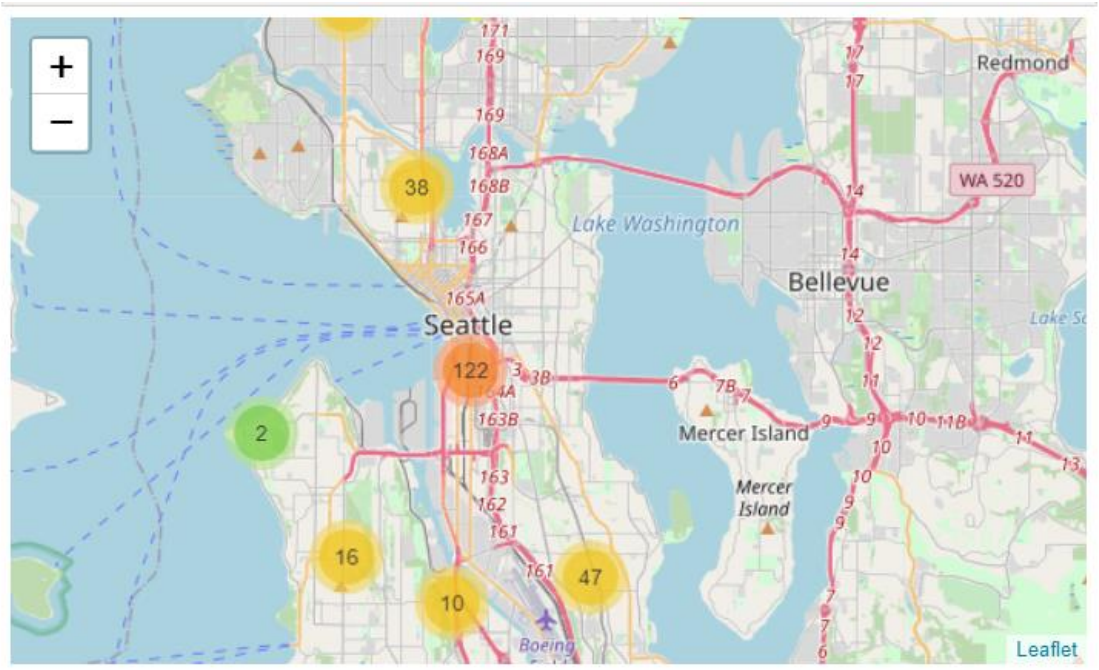

Figure 3: Map of collision clusters

It is no surprise that the largest cluster of collisions is located at the center of the city. A denser population means more cars, less space and a higher chance of collision. We can also note that other large clusters appear near the airport and near 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 of new information.

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 4 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.

Figure 4 details the collision type and those involved in the collision. Due to the variety of collisions, 'Other' describes a considerable amount of the data.

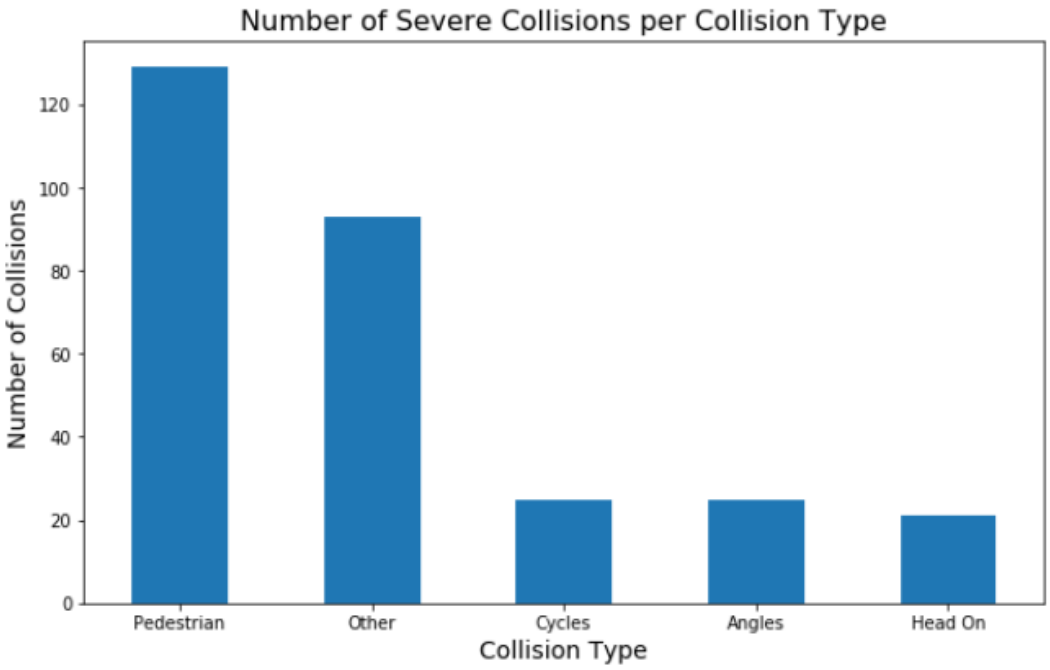

Figure 4: Proportion of collisions per collision type

We see the common types of collisions for the most severe accidents. 'Angles' refers to collisions that did not occur head on or from the rear. 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:

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

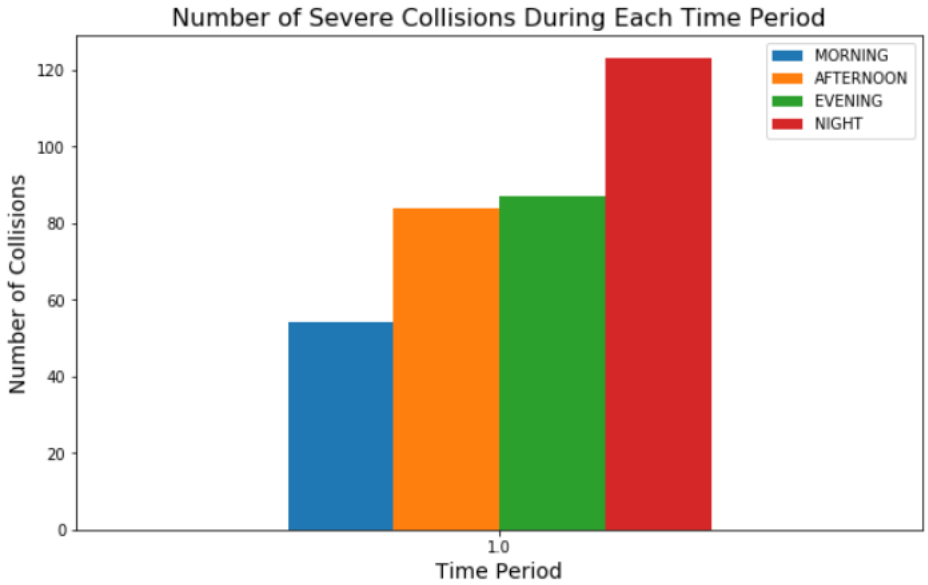

Figure 5: proportion of severe collisions associated with each time period

It is fascinating to see the correlation between the time of day and the number of (severe) collisions. We see from Figure 5, that there exists an upward trend of severe collisions throughout the day, perhaps due to tiredness or the nature of evening activities. This graph leads 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 Figure 6.

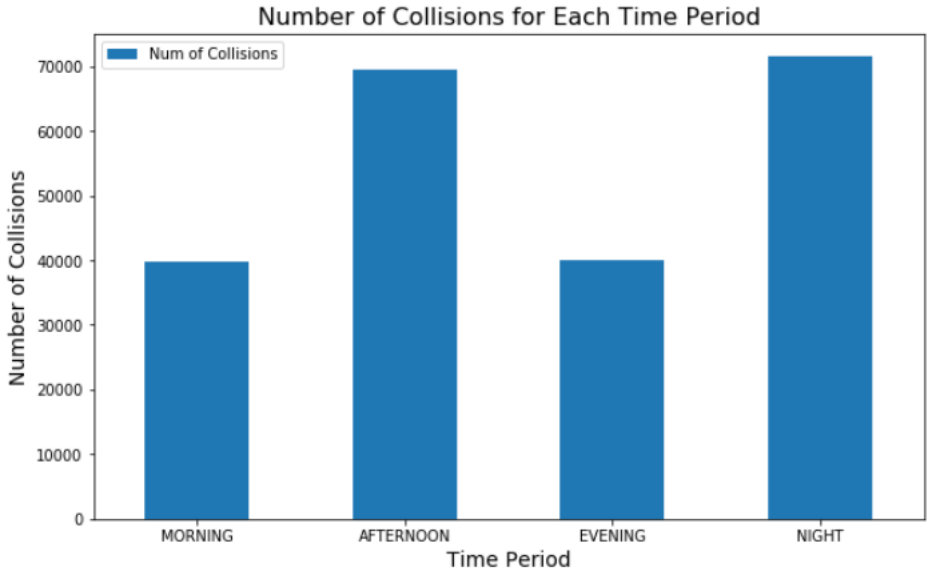

Figure 6: proportion of overall collisions associated with each time period

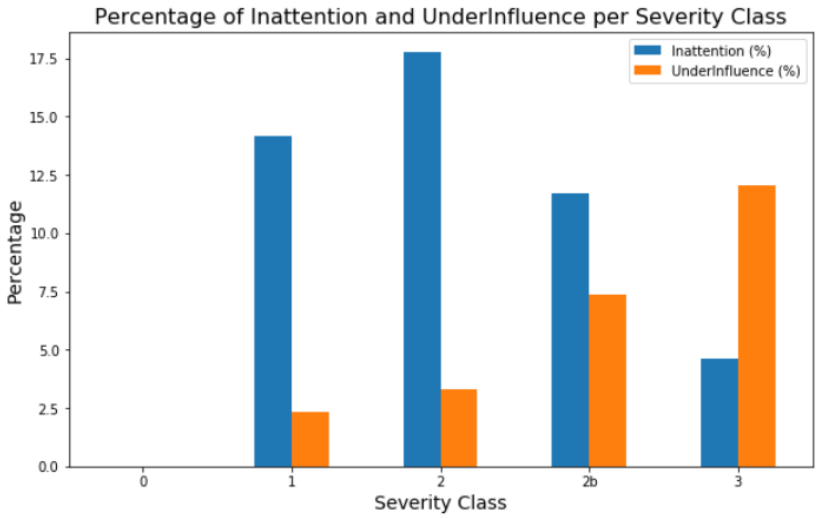

Figure 7: Proportion of drivers who were either inattentive or under the influence during a collision, per severity class.

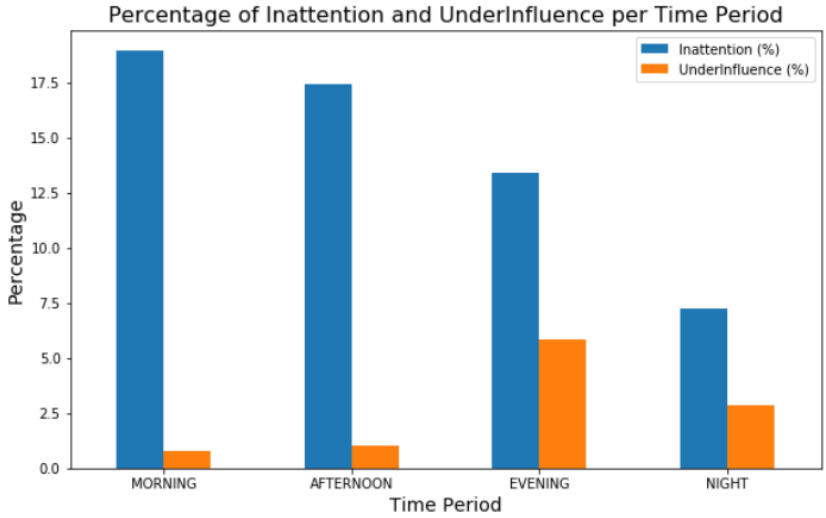

Figure 8: Proportion of drivers who were either inattentive or under the influence during a collision, per time period

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. The most common conditions for a collision are during clear and dry conditions in daylight. This is of course to be expected. One could gather statistics on the typical weather conditions year-round in Seattle and determine whether some of these conditions arise disproportionately during collisions.

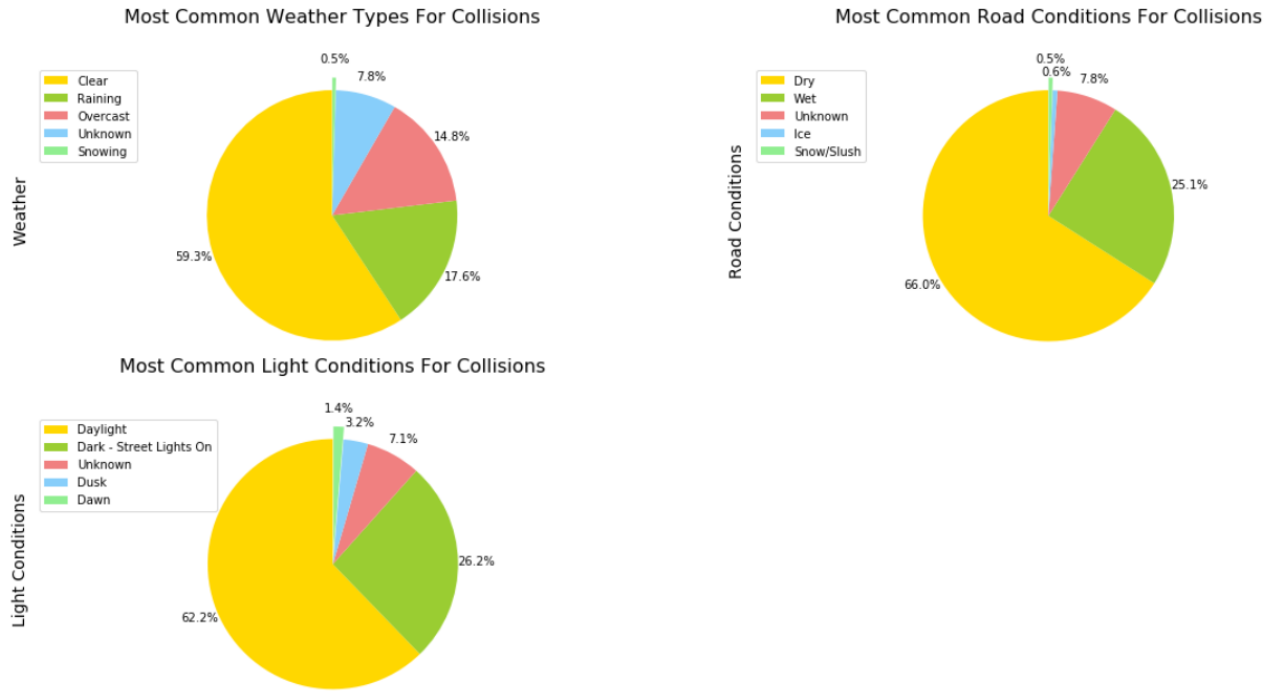

Figure 9: Most common weather, road and lighting conditions and their proportions for collisions

We see in Figure 10 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. This can comparatively be seen in major cities all around the world as cars become safer and more thought is put into our road systems and the hazards of dangerous driving.

Figure 10: Number of total collisions per year 2004-2019

Building A Model

We now have a suitable, ready-to-use dataframe to base our machine learning model on. Our goal is the accurately predict the severity class of a collision based on certain known attributes such as injuries, fatalities, period of day… There are numerous machine learning techniques to choose from, each boasting their own advantages and nuances. 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. If our model proves accurate enough then there would be no need to choose another approach.

Our dataset was split into testing (20%) and training (80%) partitions. Values of k were deployed ranging from 0-10. It transpired that k = 1 was often the most accurate value. It is of imperative to always test and evaluate the model so that we can ascertain its effectiveness. This was done using two common evaluation metrics: the Jaccard Index and F1-Score. After multiple iterations both accuracy scores hovered around the 91% mark, signifying a satisfactory result.

Results

It is clear that through some rudimentary EDA, inferential statistics and our machine learning model, we have built up a strong intuition of the dataset on collisions. We have answered key questions about car accidents and transportation in general. We have determined what types of collisions occur most frequently, which locations are more prone to collisions and the proportion of collisions associated with a lack of attention/being under the influence. We also looked at the specific conditions conducive of severe collisions and the overall trend of total collisions through the years. The introduction of smart cars, increased policing and public awareness are just some ways to combat the pitfalls of our transport systems. 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.

Discussion

As mentioned in the results section, we have gleaned plenty of useful information from the dataset. This is what I enjoy most about data science. We can take a huge set of seemingly trivial numbers, which carry very little meaning individually, organize them and transform them into tangible facts. Armed with these facts, we can shape our society and take on any challenge we deem worthwhile. The results of our analysis on the Seattle data has the potential to make up the foundations of further significant action. As mentioned in the introduction, the use of such insight can be valuable to nearly all of us as road-users. The work outlined in this report, despite its

high-level calculations and superficial nature, proves itself to be an example of the power and possibilities of data science.

Conclusion

Why do we collect and analyse data? Why do companies and other institutions hire data analysts? Why is the field of data science experiencing rapid growth and a new-found prominence? There are few tasks, entities or processes that cannot be strengthened by the mining of unique insights from data. Intelligent analysis of data is imperative when it comes to achieving success. 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. We are beginning to realise the importance of good data… and great data analysts.

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