Leveraging Data Science to Predict Car Collision Severity

Theo Hayeck

September 2020

Contents

1 Introduction

1.1 Background

A traffic collision, also called a motor vehicle collision, car accident, or car crash, occurs when a vehicle collides with another vehicle, pedestrian, animal, road debris, or other stationary obstruction, such as a tree, pole or building. Traffic collisions often result in injury, disability, death, and property damage as well as financial costs to both society and the individuals involved.[0] When motor-vehicles were first introduced to the US in the early 20th century, they were few in numbers but resulted in a disproportionate number of roadside casualties compared to the traditional horse-drawn carriages at the time. Neither comprehensive traffic laws nor significant safety features in cars or on roads existed. Since then numerous improvements in safety features, manufacturing regulations, and traffic laws have been introduced. For comparison, in 2015 almost 5,000 pedestrians died in traffic accidents, whereas in 1937, 15,000 pedestrians were killed, when the US had far fewer cars and two-fifths of its current population.[1]

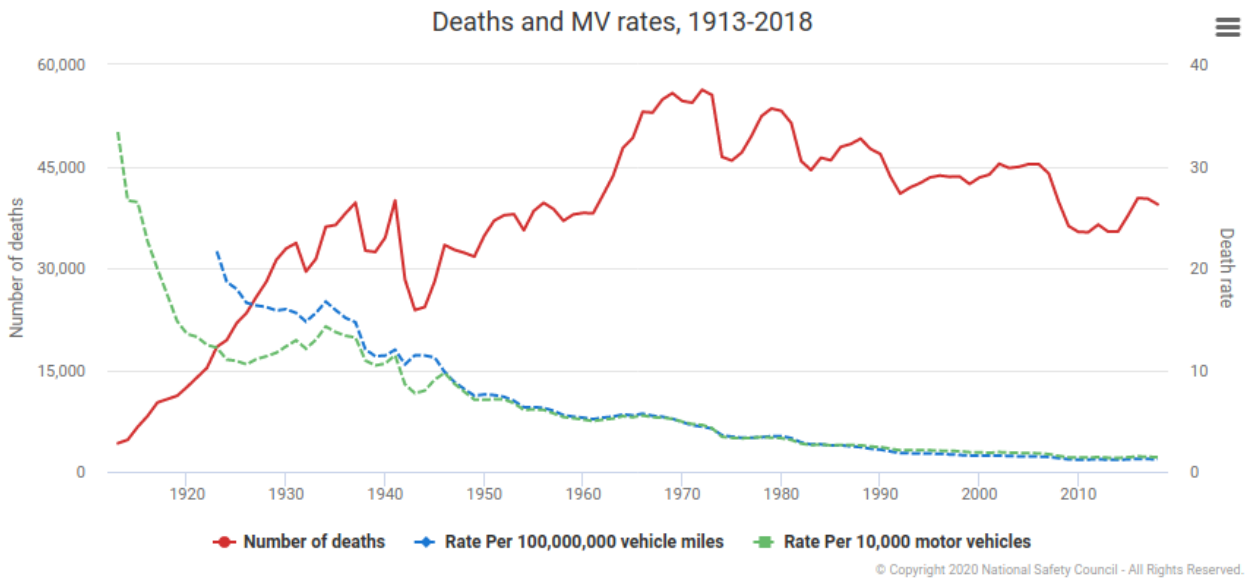

Figure 1: Deaths and MV rates 1913-2018

In 2015, the Seattle Department of Transportation (SDOT) released a 10-year plan for providing safe and sustainable transportation infrastructure for the city. A "Vision Zero" goal was adopted to eliminate serious and fatal crashes by the year 2030. However, since the release of the plan, the annual total number of collisions has only decreased by 27.6%, with the total number of severe collisions (where a collision results in an injury or a fatality) has decreased by only 18.4% [2], and making up a larger proportion of all recorded incidents annually. These findings suggest that although the initiatives implemented by the SDOT to reduce collisions were successful to a point, there are still improvements that could be made by more accurately targeting the causes of such collisions; such as: weather conditions, speeding regulations and light conditions.

In this report, we will be answering the question "Can we build a Machine Learning Model that than predict the collision severity of a crash given a set of characteristics?". This will be vital to the Seattle Department of Transportation in order to better allocate resources to limit these collisions, and to other road users who may chose to use this model in order to be alerted when they are entering potentially dangerous locations/conditions.

2 Data Understanding

2.1 Label Selection

As a case study, this report will use the collision data collected by the SDOT Traffic Management Division, Traffic Records Group (from 2004 to present) for the city of Seattle [3]. In this section, we will be undertaking an exploratory analysis on the dataset to determine the potential features that should be used in the machine learning model to predict the car collision severity.

The dataset consists of 194,673 incidents recorded by the SDOT, with 38 columns corresponding to various attributes about each collision. A SEVERITYCODE is assigned to each incident, classifying each as:

- 1 = property damage only
- 2 = injury collision

This will be used as our dependent variable and label for each collision. The data shows that the majority of all collisions (136,485) are of type 1, and the remaining 58,188 are of type 2. This is important to note because this imbalance of class labels will result in a bias to the machine learning model if it is not addressed later.

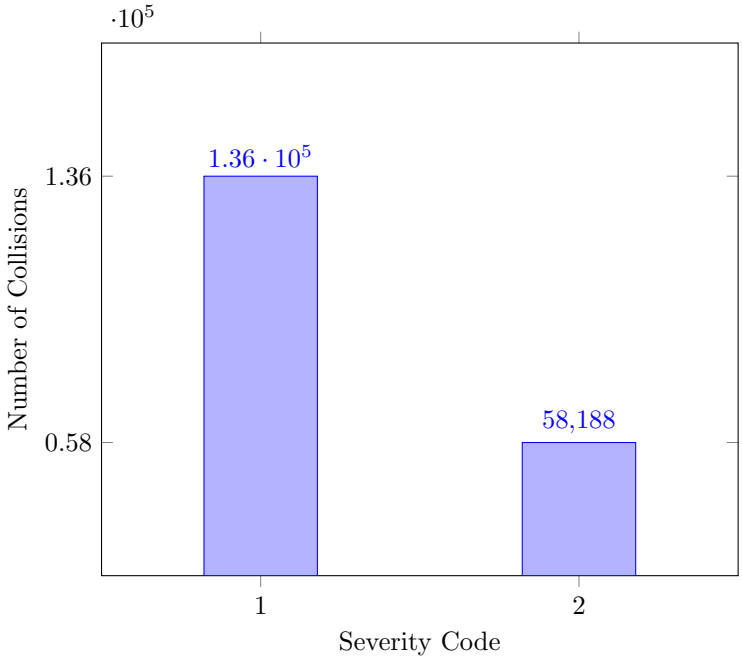

Figure 2: Collision Severity (Inbalanced)

The information about the label can be taken either from the metadata or from the column SEVERITYDESC.2. Thus, SEVERITYCODE.1 being a duplicate of SEVERITYCODE can be dropped along with SEVERITYDESC and COLLISIONTYPE.

2.2 Feature Selection

In this subsection we will explore what features should be used in our machine learning model that contribute to predicting the severity of a car collision.

Firstly, the dataset contains numerous other unique identification keys that could be used to identify each collision or it's location. These include:

- OBJECTID
- INCKEY
- COLDETKEY
- INTKEY
- SEGLANEKEY
- CROSSWALKKEY
- REPORTNO

However, none of these can uniquely identify the *severity* of the collision and therefore cannot act as the dependent variable. Furthermore, none contribute relevant information that could be used find the collision severity. Therefore, all of these columns can be dropped from the feature set as they hold no value to the machine learning model.

The remaining columns provide information regarding the conditions relating to the collision (such as weather and road conditions), or the resulting impact of the collision to it's surroundings (such as the number and type of person or vehicle affected). Such columns include:

- ADDRTYPE
- LOCATION
- JUNCTIONTYPE
- COLLISIONTYPE
- WEATHER
- ROADCOND
- LIGHTCOND
- SPEEDING
- UNDERINFL
- PERSONCOUNT
- PEDCOUNT
- PEDCYLCOUNT
- VEHCOUNT
- PEDROWNOTGRNT
- HITPARKEDCAR
- STATUS
- EXCEPTRSNCODE
- EXCEPTRSNDESC
- SDOT_COLCODE
- SDOT_COLDDESC
- SDOTCOLNUM
- ST_COLDDESC
- X & Y Coordinates of the collision

ADDRTYPE, LOCATION, and X & Y coordinates all provide information relating to the location of the collision. Since X & Y coordinates provide an exact location of the crash, ADDRTYPE and LOCATION in comparison to X & Y coordinates provide less specific information, and can be considered dispensable. Therefore, these shall be dropped from the feature set.

Furthermore, PEDROWNOTGRNT and HITPARKEDCAR provide information relating to whether or not the pedestrian right of way was not granted and whether or not the collision involved hitting a parked car, respectively. Although these provide information relating to a potential cause and effect of the collision, they both do not affect the *severity* of the collision. Therefore, these too will be dropped from the feature set.

EXCEPTRSNDESC, COLLISIONTYPE, SDOT_COLCODE, SDOT_COLDDESC SDOTCOLNUM and ST_COLDDESC provide information describing the collision, such as the angle of the collision. Individually, these can be relevant to predicting the severity of the crash. However, as mentioned later in this report, I have chosen to use information provided by the state rather than by the SDOT due to the state's ST_COLCODE providing more reliable and reproducible descriptions. Therefore, these columns will also be dropped from the feature set.

Finally, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, and VEHCOUNT all provide a continuous measurement of the number of people, pedestrians, cyclists and vehicles involved in the crash. However, the number of participants involved in the collision does not directly affect the severity of the crash. These therefore do not contribute any relevant information needed by the machine learning model and therefore will also be dropped from the feature set.

This leaves the remaining columns as features for the machine learning model to be developed later on in this report. The table below shows each feature to be used along with a description of what information it provides. These features provide information that is relevant to predicting the severity of the collision.

Feature	Description
SEVERITYCODE	Code that corresponds to severity of the collision
X	The longitudinal coodinate of the collision:
Y	The latitudinal coodinate of the collision
JUNCTIONTYPE	category of junction at which the collision took place
WEATHER	Description of the weather conditions during the collision
ROADCOND	Condition of the road during the collision
LIGHTCOND	Light conditions during the collision
INDTTME	Date and time of the incident
INATTENTIONIND	Whether or not the collision was due to inattention
UNDERINFL	Whether or not driver involved was under the influence of drugs or alcohol
SPEEDING	Whether or not speeding was a factor in the collision
ST_COLCODE	A code issued by the state corresponding to a description of the collision.

- **X & Y:** The location of the collision is important because there could be a possible relation between the location of the collision and its severity. For example, a collision on a motorway is more likely to be a type 1 than on a quite road.
- **JUNCTIONTYPE & ROADCOND:** The type and condition of the road at the collision can also be a factor in determining the severity of the collision. It is commonly known that some junctions are more dangerous than others, and road conditions can affect a car's handling and stopping distance. These therefore may determine how severe a collision is.
- **WEATHER & LIGHTCOND:** These environmental factors may impede a driver's ability to see, this could therefore affect the driver's response time and result in a more severe collision is vision was impaired.
- **INATTENTIONIND, UNDERINFL & SPEEDING:** These human factors can result in the driver being unable sufficiently control their vehicle, leading to a greater probability of a more harmful crash due to a greater speeds and lack of control.
- **INDTTME:** The date and time of a collision is also significant when predicting the severity of a collision. For example, despite 60% less traffic on the roads, more than 40% of all fatal car accidents occur at night. [4]
- **ST_COLCODE:** The type of crash can have a significant affect on the severity of the collision. For example, a head on collision with another vehicle will have a higher probability of being more severe than a collision by two vehicles both moving in the same direction.

All other columns in the dataset not aforementioned can be assumed dropped. This is because of information pertaining to these columns already being provided (duplicate features) and/or are more accurately represented by other selected features.

3 Data Preparation

The purpose of this section is to clean and format the selected features to be used in the machine learning model. Examples of data preparation, often called pre-processing, include the handling of missing or NaN values (cleaning) and converting variables into machine-legible data types (formatting). Data preparation also involves balancing labels to ensure an unbiased model, this will also be discussed at the end of this section.

3.1 Cleaning

The features X & Y provide information relating to the longitude and latitude of the collision location. In the dataset, there are 5,334 missing values between them. Since longitude and latitude are both continuous variables, a common strategy of dealing with missing values would be to replace them with either the median or mean of X & Y. In this context, using a median value of X & Y would simply just point to the center of the X*Y map of Seattle (See Figure 3 below). However, using the mean would point to an area of the map where collisions happen most frequently. X has a mean value of -122.330518439041, and Y having a mean value of 47.6195425176886. These values shall be used to fill in X & Y's missing values, respectively.

JUNCTIONTYPE, WEATHER, ROADCOND and LIGHTCOND are all categorical variables that describe the surrounding conditions present at the time of the collision. To handle missing values for these type of variables, a most frequent value for each variable called the mode is chosen to replace the missing values.

- **JUNCTIONTYPE**, has 6,329 missing values. The mode, "Mid-Block (not related to intersection)", is chosen to replace these values. The JUNCTIONTYPE "Unknown" is also replaced with the mode, since this alone does not contribute any meaningful weight to the machine learning model. Afterwards,the different categories are converted into dummy variables for the model, and the original variable JUNCTIONTYPE is dropped.
- **WEATHER**, has 5,081 missing values. The mode, "Clear", is chosen to replace these. Furthermore, since there is no explanation of what "Other" weather is, it is merged with the category "Unknown". "Unknown" is renamed to "Unknown Weather", this is to distinguish it from other "Unknown" values given in other features. Additionally, "Partly Cloudy" is synonymous to "Overcast", these are therefore merged together as "Overcast". Again, the feature header WEATHER is also dropped in place for its dummy.
- **ROADCOND**, has 5,012 missing values. The mode, "Dry", is chosen to replace these. Similar to WEATHER, "Other" is merged with "Unknown" and renamed as "Unknown Roadcond".
- **LIGHTCOND**, has 5,170 missing values. These are replaced by its mode, "Daylight".

3.2 Formatting

- **INATTENTIONIND** has 29,805 observations given as "Y", we will therefore assume that the missing 164,868 observations are "N". To use these observations in our machine learning model, we need to format them in such a way that the model can easily interpret. We will therefore encode them as 1 and 0, respectively.
- **UNDERINFL** observations are given as either "Y", "N", "1", or "0". Following the logic we have used for INATTENTIONIND, we can safely assume that "1" and "Y" are synonymous. Likewise, for "0" and "N". We will therefore convert any occurrence of "Y" and "N" to their respective number.
- **SPEEDING** has 9,333 "Y" observations. We can again assume that the remaining missing 185,340 values are to be "N". These values will also be encoded as 1 and 0 respectively.
- **INCDTTM** We shall be obtaining the day of the week, and the hour of the day of each collision using the date-time string found in this column. The INCDTTM feature will then be discarded.

One-Hot Encoding: Where any feature is expressed as a categorical variable (such as WEATHER, LIGHTCOND and WEATHERCOND) we will perform one-hot encoding, such that the feature is expressed as a matrix of 1's and 0's. With 1 signifying that the feature column was present, and 0 signifying that the feature was not. For example, if the weather at a collision was "Dry", a 1 will be present in the newly created, "Dry", feature column.

3.3 Balancing the Dataset

As mentioned previously, the dataset contains 58,188 collisions of class 2 and 136,485 collisions of class 1. If left as it is, this could result in a bias to the machine learning model. This is because it is more challenging for a machine learning model to learn the characteristics of examples from the minority class due to the comparatively less data it has for it. This makes it harder for the model to distinguish members of the minority class from members of the majority class.

The two commonly used strategies to balance classes in a dataset are random over-sampling (ROS) and random under-sampling (RUS). ROS is the process of supplementing the dataset with multiple, randomly chosen copies of cases from the minority class, until the number of samples match the majority class. RUS randomly deletes samples from the majority class until the number of samples matches the minority class.Both methods come with advantages and disadvantages. While ROS may inflate or exaggerate underlying patterns in the minority class, RUS may potentially discard important samples of majority class and distort its underlying patterns. A rule of thumb is to use ROS when the given dataset is small and RUS when the given dataset is large. As the collision dataset is sufficiently large, this report will employ the random under-sampling method to balance the dataset and thereby reduce class 1 collisions to 58,188 samples.

3.4 Exploratory Data Analysis

3.4.1 Relationship between location and collision severity

It is well known that some locations result in a greater number of crashes than others. But does this mean that there is a relationship between the *severity* of a collision and the location in which it occurred?

Figure 3 shown below, displays the location of each major hotspot where at least 100 collisions (class 1 and 2) have taken place. Where those of just class 1 are depicted with yellow and blue circles, and hotspots containing at least 1, class 2 collision, as red circles. Since there are zero yellow and blue circles on the figure, it is clear that out of each hotspot where there have been more than 100 collisions, every spot has had at least 1, class 2 (serious injury) collision. Furthermore, the locations of the hotspots are often on major motorways in Seattle, such as Rainier Ave S. We can therefore conclude that these locations are particularly dangerous, and that there is a relationship between location and collision severity.

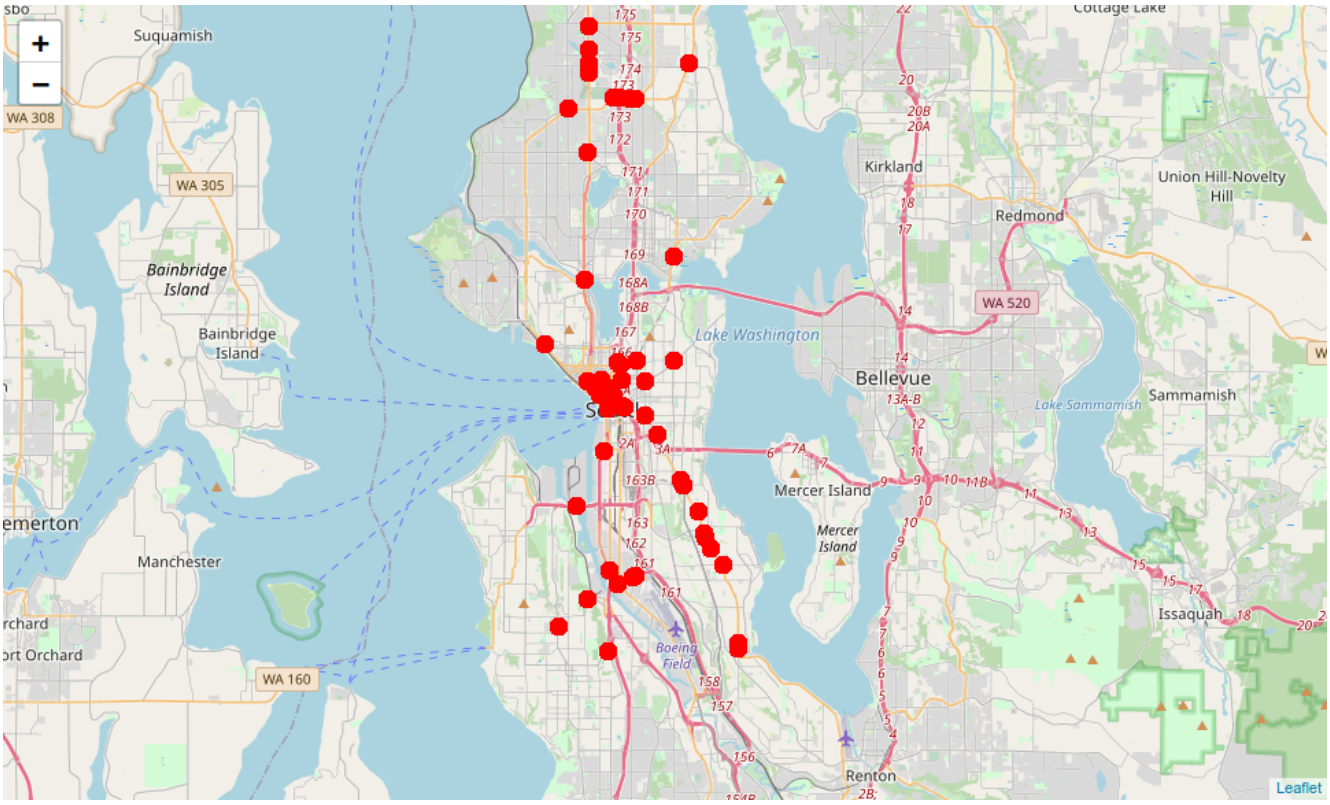

Figure 3: Locations in which more than 100 collisions (Class 1 & 2) have ocurred

3.4.2 Relationship between junction type and collision severity

It is often recognised by drivers that there are harder types of junctions to navigate than others. However, does this mean that these junctions result in a higher proportion of collisions than others?

It is shown in Figure 4 below that the most number of collisions occur at a Mid-Block. There is also a higher ratio of class 2 collisions to class 1 at "At Intersection (intersection related)" than any other type of junction. However, we would need to normalize the total number of each type of junction in Seattle before we are able to draw any conclusions about which type is more likely to result in a Class 1 or 2 type collision.

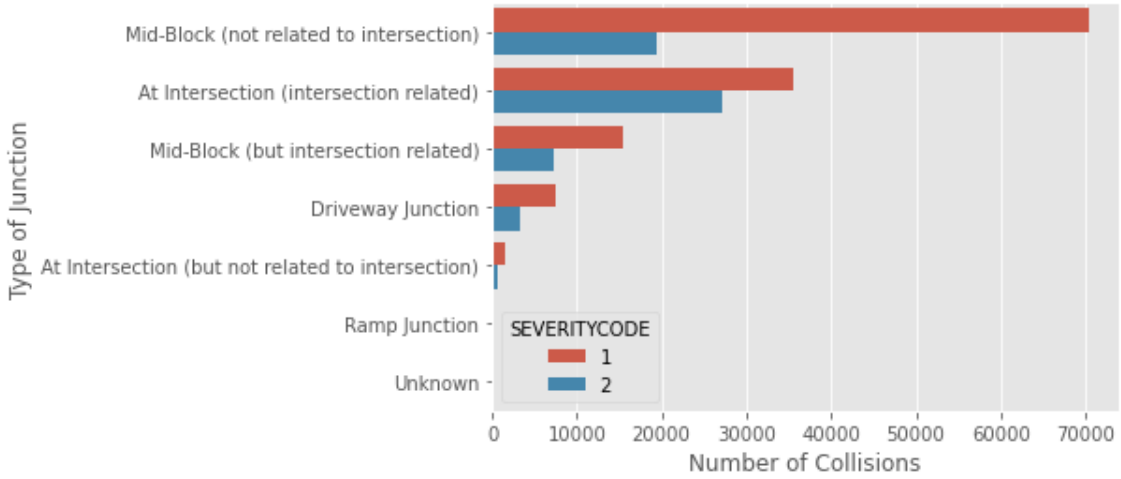

Figure 4: Bar chart to show relationship between junction type and the number of collisions

3.4.3 Relationship between weather and collision severity

Rain and snow often cause hazardous driving conditions by reducing the handling and increasing the stopping distances of vehicles on the road. It is therefore wise to investigate any relationship between weather conditions and the severity of a collision.

As you can see from Figure 6 below, the majority of collisions occur during times when the weather is "Clear". This is to be expected, since the weather is clear in Seattle most of the time. This pattern is expressed for the rest of the dataset, where the greatest number of collisions occur during the most common weather conditions.

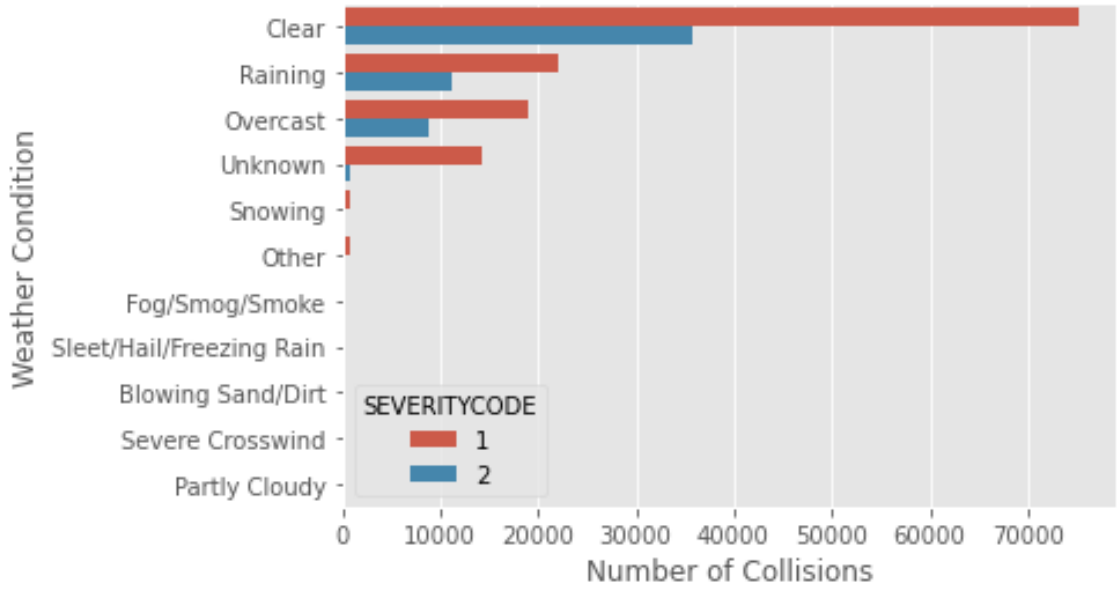

Figure 5: Bar chart to show relationship between weather conditions and the number of collisions

3.4.4 Relationship between road conditions and collision severity

It is well known that poor road conditions can reduce a driver's handling of their vehicle and increase the stopping distance should they need to react to a hazard. However, does this necessarily impact the severity of a collision?

It is shown by Figure 6 that oil on the road results in the greatest chance of a class 2 collision occurring out of all the road conditions, with 40 resulting collisions being of class 1 and 24 being of class 2, the largest class 1:2 ratio out of all road conditions at 0.6, however, class 1 collisions are still the predominant type of collision for all road conditions.

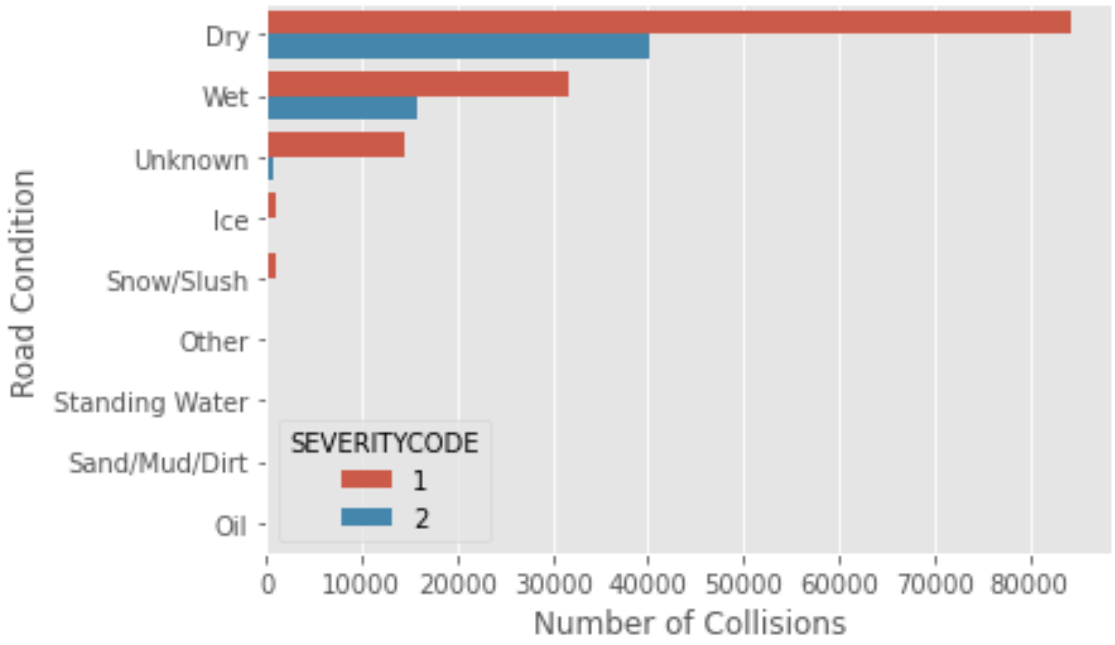

Figure 6: Bar chart to show relationship between road conditions and the number of collisions

3.4.5 Relationship between light conditions and collision severity

In lower light conditions it is often harder for drivers to see. However, does this necessarily relate to an affect on the collision severity of a crash?

In the figure below, you can see again that the most common lighting condition, "Daylight", results in the most collisions. However, "Dusk" results in the greatest chance of a class 2 collision occurring out of all the light conditions, with a ratio of 0.491. It is to be noted that Dark Unknown did have a class 1:2 collision ratio of 0.571, however, due to it being unknown, it is less reliable to use.

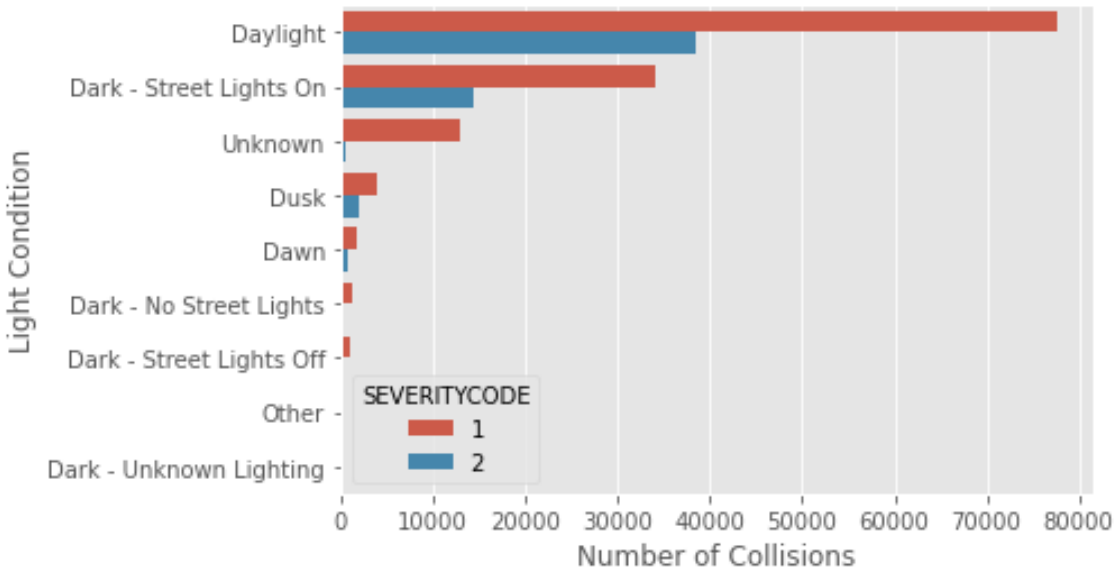

Figure 7: Bar chart to show relationship between light conditions and the number of collisions

3.4.6 Relationship between driver inattention and collision severity

A drivers lack of attention to their surroundings is often the cause of many collisions on the road. However, based on this data we can see that, the majority of the time, driver were paying attention leading up to the collision. Furthermore, we can see that out of those drivers who were not paying attention, a larger proportion of collisions were of class 2. This could suggest that there is a relationship between a lack of attention and an increased collision severity.

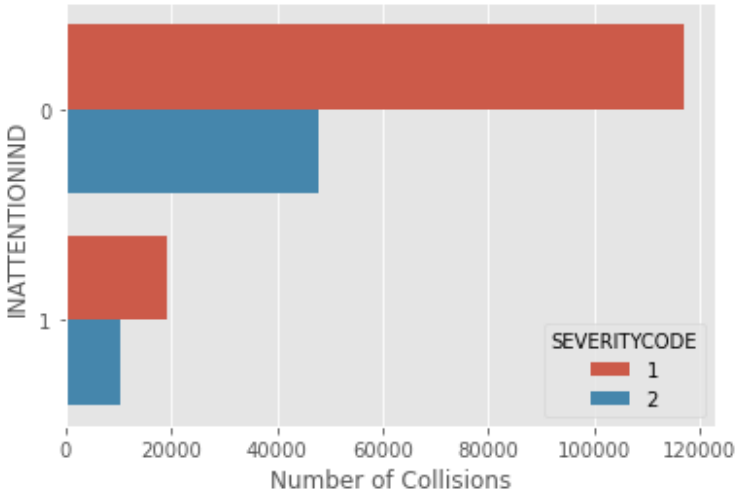

Figure 8: Bar chart to show relationship between a driver's attention and the number of collisions

3.4.7 Relationship between the influence of drugs or alcohol and collision severity

A similar trend is again seen when comparing drivers under the influence of drugs or alcohol during the time of the collision, where the majority of collisions occur when the driver is sober and not under the influence of anything; but with a higher proportion of severe, class 2, collisions occurring when the driver was under the influence.

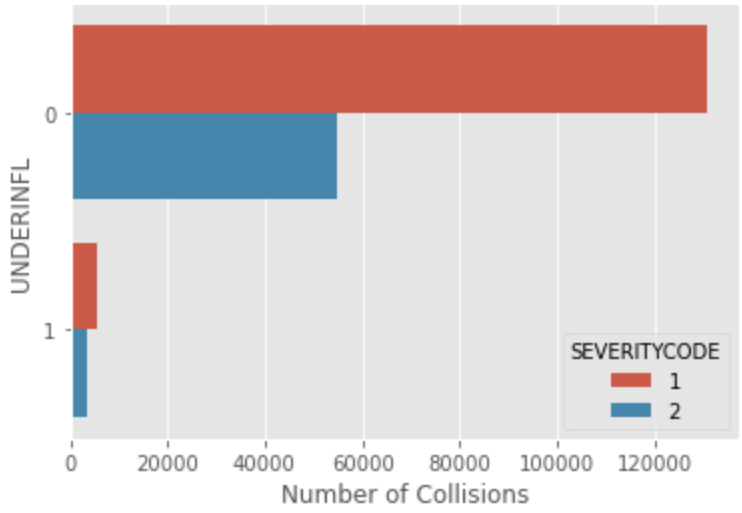

Figure 9: Bar chart to show relationship between drivers being under the influence and the number of collisions

3.4.8 Relationship between a driver speeding and collision severity

The trend again follows through when comparing drivers who were speeding during the time of the collision, where the majority of collisions occur when the driver was driving within the speed limit, but with a higher proportion of severe, class 2, collisions occurring when the driver was speeding.

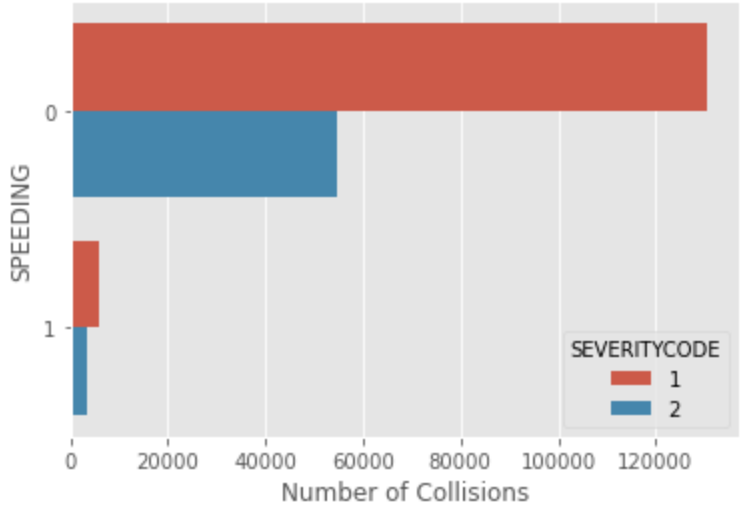

Figure 10: Bar chart to show relationship between drivers speeding and the number of collisions

3.4.9 Relationship between the time and date and collision severity

The assumption about the time and date of the collision is that during certain times, such as during week days or rush-hour, people are less attentive, for instance, due to exhaustion from work. As shown in figure(s) 11 and 12 below, this assumption holds. In figure 12 you can see that a bell shaped distribution is present, centered around 17:00 having the most number of collisions (of both class 1 and 2) in the day. Furthermore, in Figure 12 you can also see a gradual increase of collisions per day, from the start (Monday) to the end (Friday) of the working week. With a decrease in the number of collisions over the weekend. These findings reiterate that belief that when people get more tired during the day and week, more collisions occur. However, it is also to be noted that these patterns might arise as a result of there simply being more cars traveling on the roads during these times, where most people are likely to be travelling to and from work at around 8:00 and 17:00, respectively. Similarly, this is shown over the weekend, where a significant decrease in the number of collisions is seen; perhaps due to there being less cars on the road as people do not need to travel to work on the weekends. It is also important to note that in the event that a collision was recorded and the SDOT did not provide a time in which the collision occurred, just the date was inputted into the INCDTTM feature column. This has resulted in all of these events being assigned a time of 00:00, the start of the day; explaining the disproportionately large number number of collisions that have been shown to happen on the 0^{th} hour of the day in Figure 11.

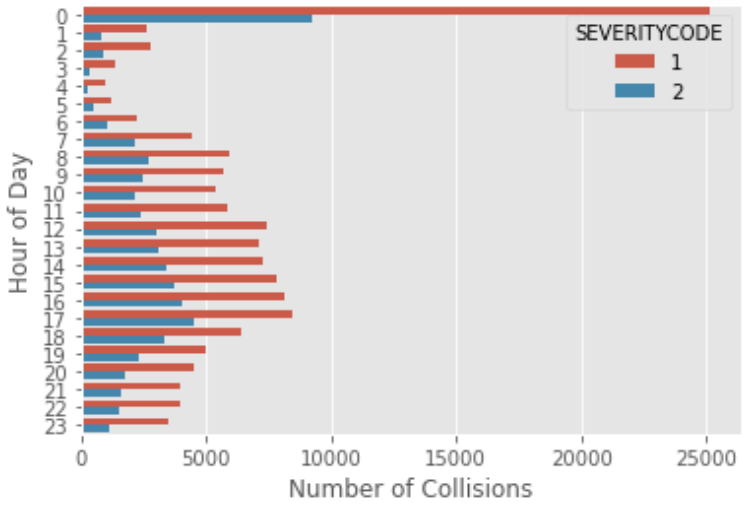

Figure 11: Bar chart to show relationship between the hour of the day and the number of collisions

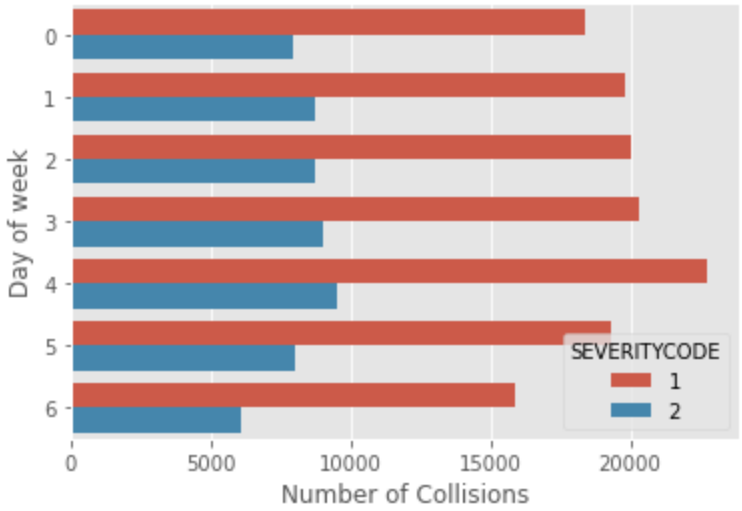

Figure 12: Bar chart to show relationship between the day of the week and the number of collisions

3.4.10 Relationship between ST_COLCODE and the number of collisions

Finally, it is has been proven that the type of collision a vehicle is in can affect its severity. As shown in Figure 13 given at the end of this report, you can see that there are specific ST_COLCODE's that result in more class 2 type collisions than class 1. These include ST_COLCODEs: 45, 0, 1 and 2. This again proves a relationship between the type of collision a vehicle is in and its severity.

4 References

- [0] https://en.wikipedia.org/wiki/Traffic_collisions
- [1] https://injuryfacts.nsc.org/motor-vehicle/historical-fatality-trends/deaths-and-rates/
- [2] https://www.seattle.gov/visionzero
- [3] https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv
- [4] https://seriousaccidents.com/legal-advice/top-causes-of-car-accidents/nighttime-driving/

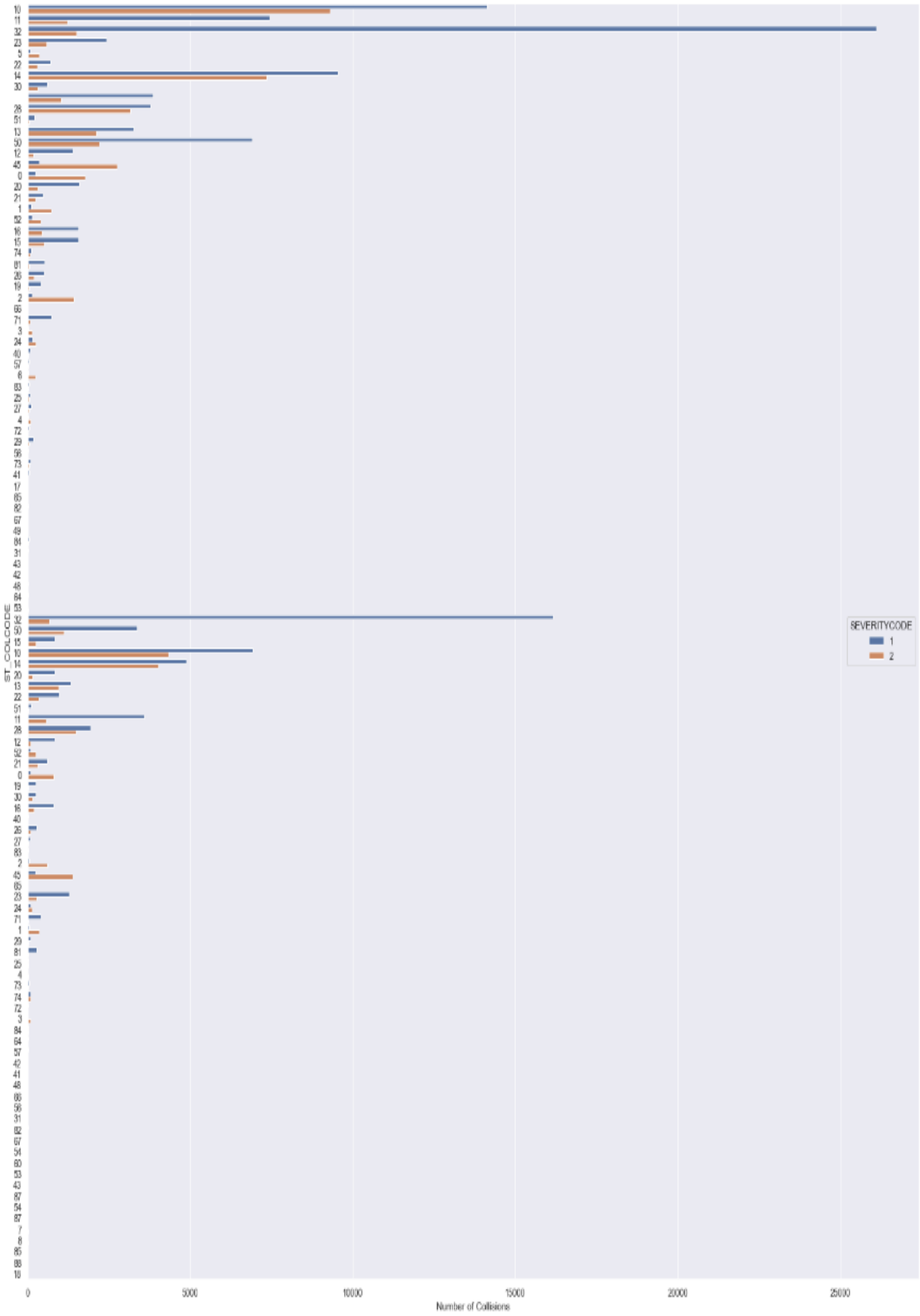

Figure 13: Bar chart to show the frequency of each type of state defined collision code.