# An Exploratory Analysis of Accident Severity and Environmental Conditions in Seattle, Washington

## 1. Introduction

It has long been recognized across that weather and subsequent road conditions play a major factor in road accidents (e.g., Andrey, 2001; Edwards, 1998, 1999; Malin et al., 2019; Pisano et al., 2008). Pisano et al., (2008) reported that 25% of all US public road accidents are weather-related, translating into over 1.5 million crashes annually resulting in hundreds of thousands of injuries and thousands of deaths. The type of adverse road condition and current weather influences the driver's perception of the overall risk, with weather conditions that impact visibility (e.g., rain and snow) causing an increase of driver caution over just wet road conditions (Pisano et al., 2008). Malin et al., (2019) cite snowy and icy road conditions as having the highest accident risk compared to a clean road surface along with a corresponding increase in the risk for fatal accidents during slushy conditions. As such, when road conditions deviate from bare road conditions and good weather with high visibility, the prevailing wisdom is the risk for an accident increases which includes a potential for injuries or fatalities.

One potential mitigation strategy is dissemination of information that conditions are hazardous with a high potential for an accident to both drivers and emergency responders. This would warn drivers to take extra caution and allow for extra time when on the road or consider postponing their trip if possible. For emergency responders, this would provide information when to expect a potentially higher number of automobile incidents and accidents, allowing them to prepare to have resources available to respond to increased incidents based off weather forecasts and other information sources. The objective of this work is to predict severity of an automobile accident based on weather and road conditions. The goal of this work is *not* to determine which specific intersections or locations are the most dangerous nor likelihood of an accident, but what overall conditions would lead to different severity of accidents with the type of location factored into the analysis.

## 2. Data

### *2.1 Overview of the Data Used*

The data used in this work is the accident severity data provided within the Capstone project. The data describes characteristics of accidents from January 1, 2004 to May 20, 2020 in Seattle, WA, USA including a severity index and numerous columns describing ambient conditions surrounding the incidents. This section describes the initial processing and culling of the features of the dataset to allow for complete data analysis in later sections.

*2.2 Feature Selection*

Given the objective of modeling potential accident severity based on the road and overall ambient conditions; a number of the initial 37 features were filtered as redundant or non-explanatory within the context of this analysis. Features with less than 25% data coverage were investigated more closely to understand why this was the case (Figure 1, Table 1). Of these four, only EXCEPTRSNDESC was determined to be not useful since no information about it was in the metadata. The other three columns only contained data if the condition were true; the missing data were filled with false values to build a continuous feature set for that column header. Of the other columns, three had data coverage of under 95% (Table 1). EXCEPTRSNCODE was removed as no information about this column was in the metadata; SDOTCOLUMN doubles the OBJECTID column as an incident identifier, and INTKEY is a code for the intersection that does not provide more detail to the overall ambient conditions contributing to the incident that is not available in other features.

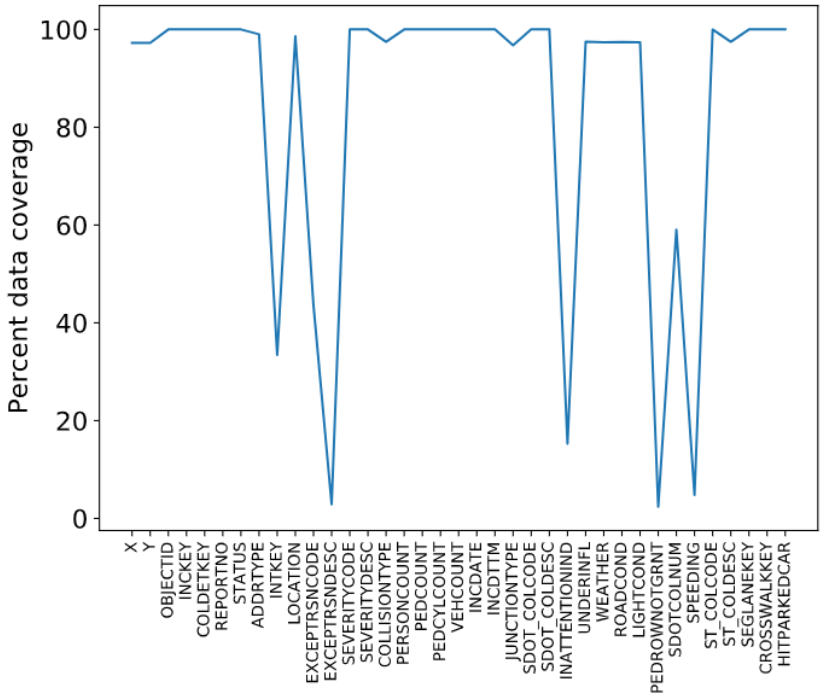


**Figure 1.** Percent of data coverage for each column header within the initial dataset. Percentage was determined by dividing the count of each column by the total length of the dataset.

A number of different identifiers of location and incident type is provided within the dataset, of these, only one provides a unique identifier (INCKEY) to match specific incident. Otherwise, other incident identifier coders were filtered out. Features describing the general location were kept (ADDRTYPE), the rest were filtered as well (Table 2). Other redundant columns for the type of collision, either an integer code or text description were removed to reduce confusion of which system is being used (Table 2). The SDOT_COLCODE was used as the description of the collision as it was relatively complete and had clear descriptions of each incident type. This left 18 features including the SEVERITYCODE being used as the dependent variable on which to base the model predictions. All others passing the initial filtering are potential explanatory variables. Any further filtering will occur within the initial data exploration described in Section 3.

**Table 1.** Listing of columns with less than 25% and 95% data coverage and their total coverage.

| **Description** | **Column Header** |
|---|---|
| Less than 25% data coverage (Percentage of data available) | EXCEPTRSNDESC (2.9%), INATTENTIONIND (15.3%), |

| | |
|---|---|
| | PEDROWNOTGRNT (2.4%), SPEEDING (4.8%) |
| Less than 95% data coverage (Percentage of data available) | EXCEPTRSNCODE (44%), SDOTCOLUMN (59%), and INTKEY (33%) |

Two extra features were generated as part of this analysis to include a time-based component within the dataset. The incident dates were converted to calendar day of year as adverse conditions are more comment during the winter months and around potential holidays due to increased travel. Also, the hour of the incident was split into its own features as a 24-hour clock to add an extra dimension to give a better idea of what parts of the day are more likely to have a collision. In a bigger city, the assumptions is during the morning and evening commutes accidents will be more likely. There are some datetimes recorded that do not include the time as part of the feature's record, so these were removed. The other issue was it appeared that no 00:00-01:00 (12am-1am) incidents were recorded or the way the data were input and stored, it removed references to this timeframe. Details in Section 3.1.

**Table 2.** Listing of features in the entire dataset and if they were removed for analysis or retained for the final analysis. More details on why different columns were removed are contained within the text.

| **Description** | **Column Header** |
|---|---|
| Removed for redundant location and object identifiers | X, Y, OBJECTID, COLDETKEY, INTKEY, LOCATION, SEGLANEKEY, INCDTTM, REPORTNO, JUNCTIONTYPE, SDOTCOLUMN, CROSSWALKKEY |
| Removed for redundant collision severity and type columns | ST_COLCODE, ST_COLDESC, SDOT_COLDESC |
| Removed for lack of metadata | STATUS, EXCEPTRSNDESC, EXCEPTRSNCODE |
| Retained for analysis | INCKEY, ADDRTYPE, SEVERITYCODE, COLLISIONTYPE, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, SDOT_COLCODE, UNDERINFL, WEATHER, ROADCOND, LIGHTCOND, HITPARKEDCAR, INATTENTIONIND, SPEEDING, PEDROWNOTGRNT, INCDAT |

*2.3 Data Processing and Cleaning*

The dataset required a few instances of data processing and cleaning during the initial data filtering and feature selection phase (Section 2.2). The columns not retained (Table 2), were dropped from the pandas dataframe for ease of use and reduced computational complexity. As shown in Figure 1; not every column retained was continuous due to lack of complete coverage. The three columns with minimal data indicating only a positive value were filled with an ‘N’ to indicate a negative to generate a continuous feature column. After these values were filled, every column being used had at least 97% of the labels fully filled with data across all the features. To make sure all labels had data in every feature, labels that contained any missing data across the features were removed. This left 187,504 labels with full data coverage across all features, representing 96.31% of the original labels to be used for the analysis to answer the objective of the work. A deeper exploration of the dataset will be accomplished in Section 3 as part of the model development.

*2.4 Final Feature Usage and Dependent Variable Description*

The SEVERITYCODE (SC) will be used as the main dependent variable on which the analyses and model(s) will be built. There are only two SCs that are available within the dataset, '1' or '2' indicating a non-injury collision (1) or a collision with injuries (2). Due to this limitation, the overall severity of the collisions will be limited to a binary state, either involving an injury or not regardless of seriousness of the injury. There are a number of variables that interact with each other that could lead to a collision and all will be related back to the SC. The initial data exploration will determine the level of correlation between the SC and the features being used to guide model building. Likelihood of accident will not be possible to determine as there is no data on the total amount of miles driver per incident in which to compare. The final output will be a determination that if there is a collision occurs given a set of road and weather conditions what the potential that the collision will involve an injury or not and with what external factors (speeding, intoxication, etc.,) may contribute to the collision.

## 3. Methodology and Exploratory Data Analysis

3.1 *Methodology Used*

Data analysis was completed in two major steps, a basic data exploration step looking specifically at the prevalence of incidents under different road, weather, and light conditions, the variables broadly outside of the drivers' control. The other incidental variables were checked but a full discussion of these is beyond the scope of this work plus many had minimal 'positive' values in that they were marked as factors involved with the incident. Given the binary nature of the severity code, there is minimal initial correlation or regression that can be completed since it is only a single value; most of the exploratory analysis will be direct comparisons to the severity of the accident with the different potential drivers. This analysis involved normalizing the percent of incidents by the total under the different categories listed in the feature sets. Also, during the initial phase, the creation of the time variables allowed for a look at what time-of-day and time-of-year were associated with more total and more severe incidents. Simple linear or multi-linear regression was not used as the dependent variable is a discrete value with the independent variables as non-continuous values.

The second major analysis step involved attempting to model the data using logistic regression and K-nearest neighbors (KNN) by splitting the data into a train and test set. A check

on the appropriate number of neighbors to use for KNN was run from 1 neighbor to 10. Both methods will be discussed briefly along with a note about the data being used. Logistic regression was initially chosen as the dependent variable is a discrete value. A model was built and brief notes on the results will be mentioned but as the independent variables are not all continuous, in fact most are discrete values, the model will likely not work well hence the move to a KNN analysis.

*3.2 Exploratory Data Analysis*

Overall, 130634 (70%) property or non-injury and 56870 (30%) injury accidents were reported within the data set. So, initially, of the accidents reported, the likelihood of it not involving an injury was 2.33 to 1. Starting with timing of incidents across the year, the highest frequency of incidents occurred generally around late October to early November, with a drop off at the end of the year (Figure 2). However, the number of incidents per day from percentage of the total was relatively consistent across the year, with minimal separation between the severity of the incident.

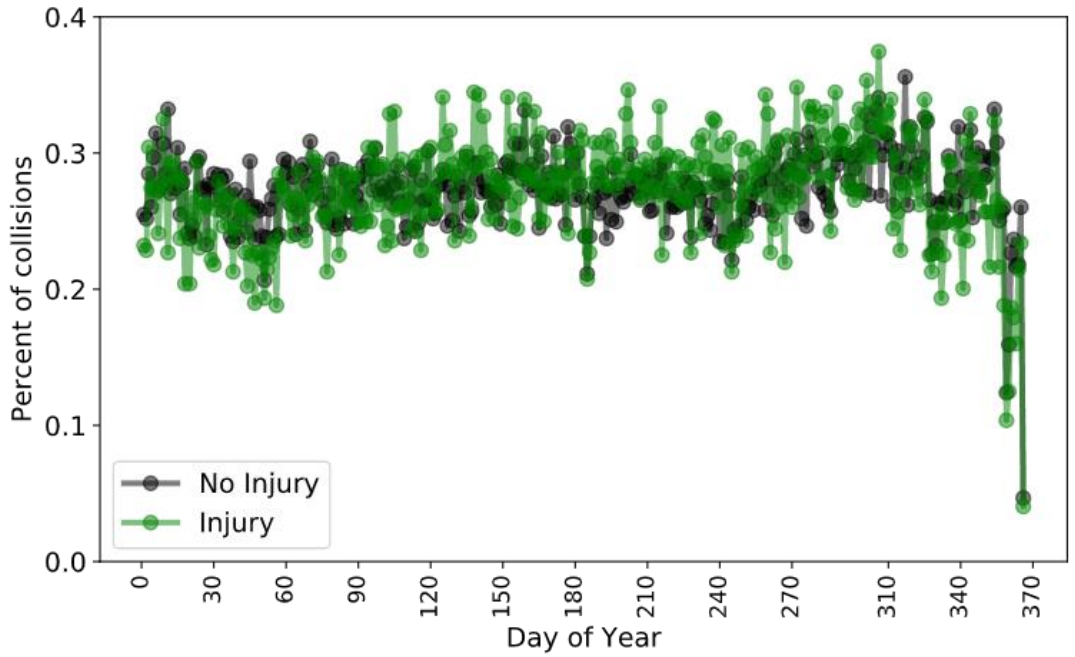


**Figure 2:** Percentage of incidents reported on every day of the year of the entire time period as normalized by total number of reports separated by non-injury and injury reports.

Time of day is a factor with the highest percent of collisions occurring during the 17:00-18:00 hour range with a sharp drop-off as commuters arrive home or their destination after workhours (Figure 3). Collisions start ramping up again around 07:00 in the morning when the first commuters start heading to work with a slight dip mid-morning before a steady increase until peaking at 17:00, the start of rush hour home from work. A number of the reports did not include the time of incident nor were listed as between 00:00-01:00 in the morning and were left off the plot though the total incidences remained as part of the percentages. It was unknown if reports with a time included were during that 00:00-01:00 hour or were missing the time of day for the report Again, there was little separation between the injury and non-injury reports apart from 15:00-18:00, which even then was relatively minor.

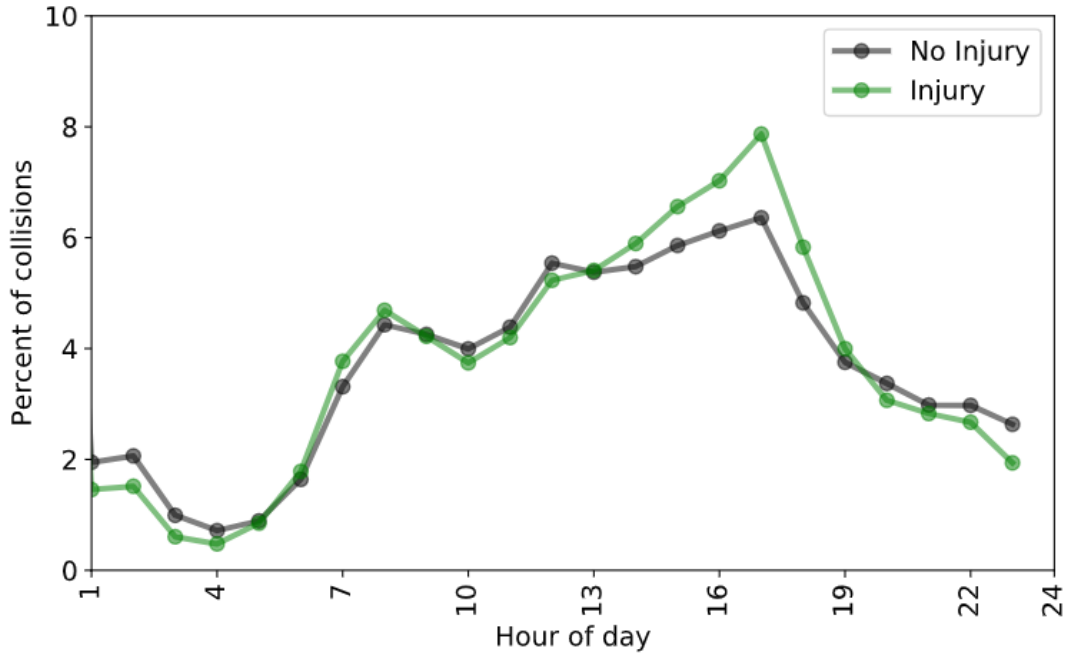


**Figure 3.** Percentage of reports generated in each of the day as normalized by the total number of reports over all years. Reports are classified within the hour that they were noted so a report at 12:59 was grouped with the 12:00 reports. Data is split between injury and non-injury incidents and normalized accordingly.

The analysis continued with the major environmental conditions (weather, road conditions, and lighting conditions). Weather apart from “clear” played a factor in approximately 40% of the collisions and no difference between the injury and non-injury collisions occurred across the different categories (Figure 4). Raining and overcast account for the majority of the remaining collisions though no notes on the severity of the rain is indicated in the dataset.

Further data here would be useful to help breakdown the type of weather conditions that contribute to incidents. Heavier rain can lead to issues with visibility that are not included in the LIGHTCOND feature set as well as standing water on the roads influence traction of the vehicle with the road.

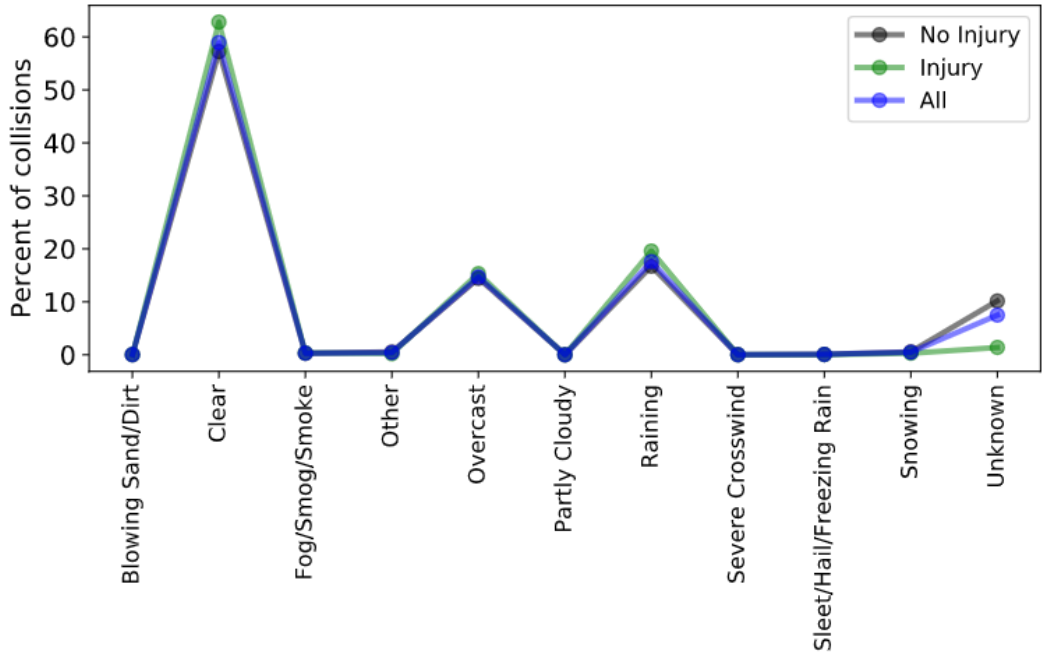


**Figure 4.** Breakdown of the weather conditions across all reports and no injury and injury reports. Each category is normalized by the total incidents within that category.

Looking at light conditions; it is a similar situation; most collisions occurred with decent to good light conditions with the vast majority occurring either in daylight or in dark but with streetlights on. Only during the UNKNOWN light conditions was there much separation between the different incidents (Figure 5).

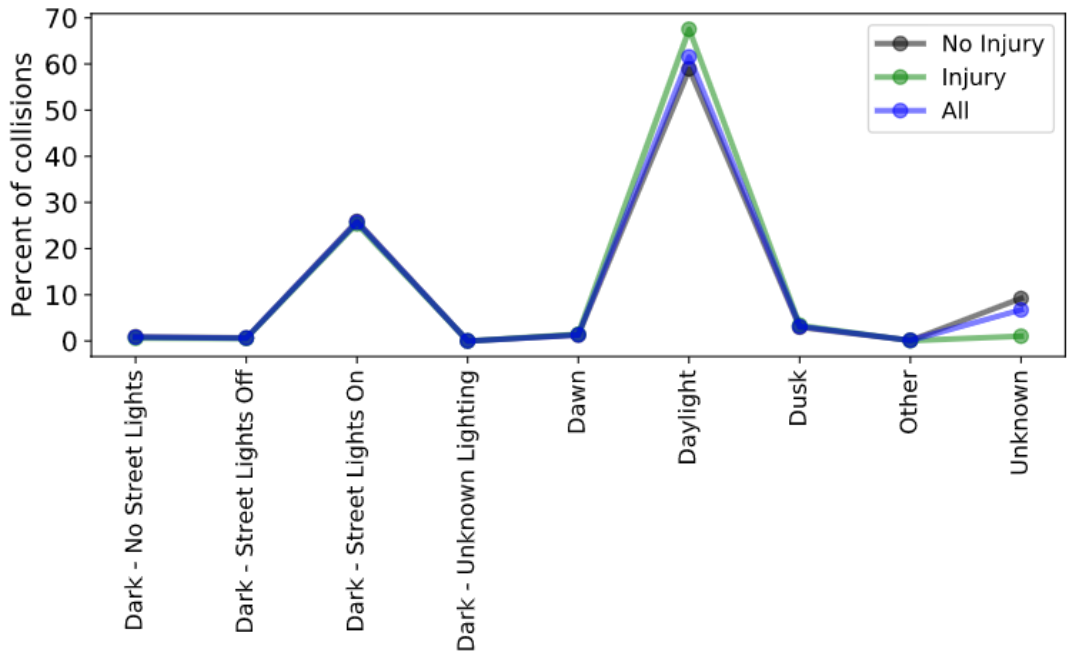


**Figure 5.** Breakdown of the light conditions across all reports and no injury and injury reports. Each category is normalized by the total incidents within that category.

Similarly, there is minimal difference with the road conditions as well; most collisions occurring under dry conditions with wet in second place and unknown in third. This tracks with the percent breakdown in the weather conditions. This makes sense for Seattle since the winter months of the city do not get too cold to where there are not many days with snow and frozen conditions. Given these similarities, the differences in collisions could be coming from other factors than general road conditions yet there is still the missing piece of how frequent a collision occurs compared to total amount of miles traveled or overall severity, up to and including fatalities, given the restrictions of the dataset (Figure 6).

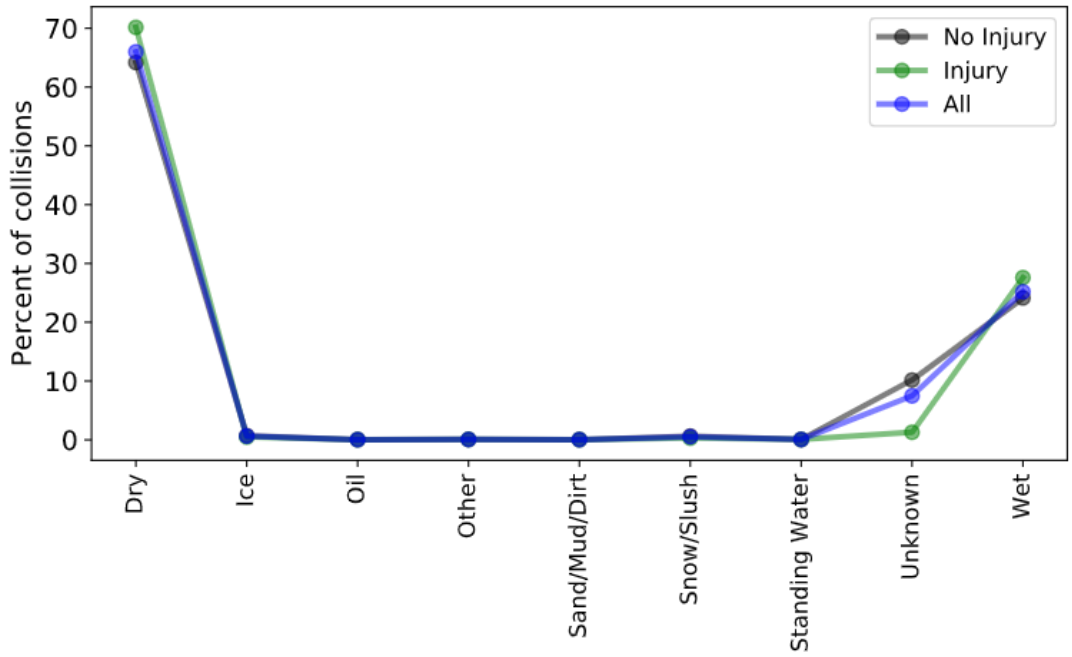


**Figure 6.** Breakdown of the road conditions across all reports and no injury and injury reports. Each category is normalized by the total incidents within that category.

*3.3 Machine Learning Modeling*

The train/test split was initially set to 70%/30% but was also tested with a 50%/50% split. This created only negligible changes in the results, so the initial 70/30 split results are used here. The logistic regression was unable to create a satisfactory confusion matrix, placing all the test set data in the non-injury category meaning it was unable to determine a distinction between the two severity classes (Figure 7). The logistic regression predicted probability for any given point was higher in the non-injury category, so it returned predictions of the crash severity for the entire train set as non-injury. Because of this, the model was correct approximately 70% of the time due to the ratio between non-injury and injury reports. Running the model multiple times with different train and test sets would converge on this result as it is a feature of the dataset itself, not emergent from the information contained within.

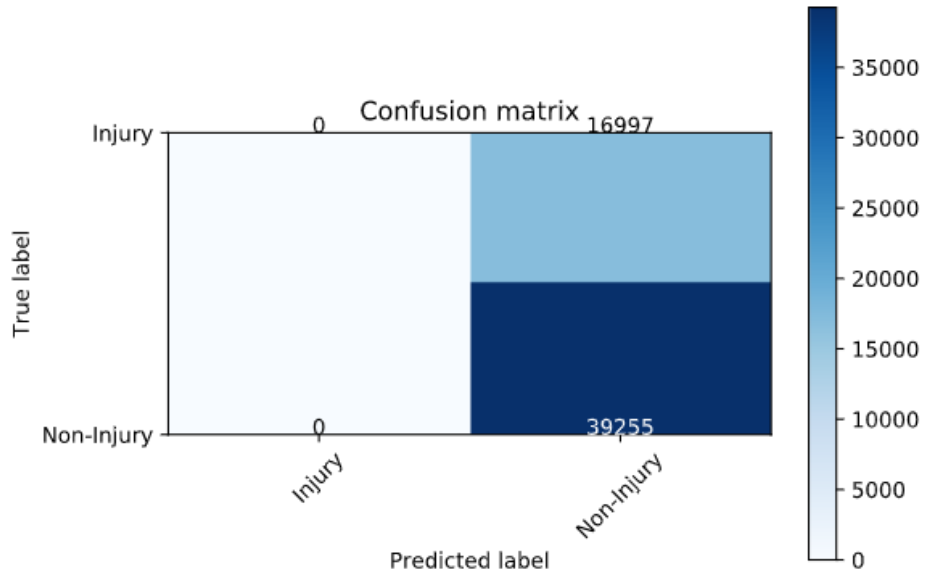


**Figure 7.** Confusion matrix for the train set from the logistic regression model used in this analysis.

Results from the tests of how many nearest neighbors to use are in Figure 8. Using KNN with the number of neighbors as 2; a tradeoff between test set accuracy and train set accuracy was made. Any higher neighbor value used would provide minimal benefit in accuracy with a loss in the initial train set accuracy. The train set accuracy was at approximately 80% while the train set accuracy was 67% for this particular model but from Figure 8; the train set and test set started to converge around 65-70%, which is the ratio between the non-injury and injury listed incidents. This leads toward a conclusion that with the similarities of the predictability between the two models that the data does not lend itself well to a machine learning application without further data engineering or expansion of the data set. The separation of the independent variables is not large enough to be able to distinguish between the two severity codes with accuracy above knowing the ratio between the two report types.

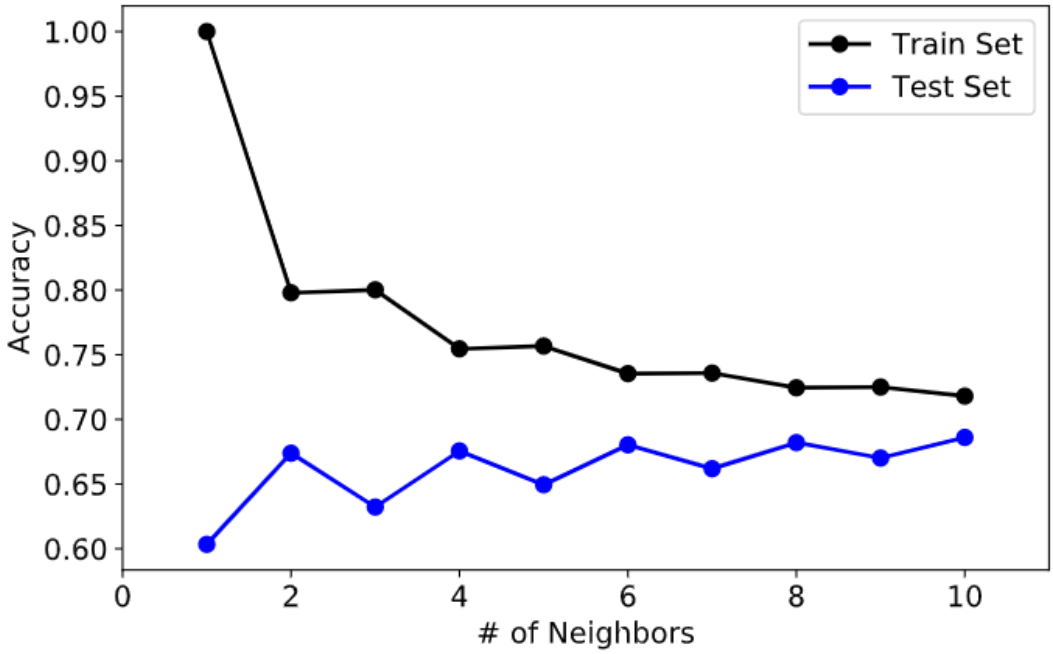


**Figure 8.** Accuracy results of the test for the appropriate number of nearest neighbors to be used in the KNN analysis. Note the data converges at approximately 70%.

**4. Discussion and Conclusions**

The severity of the accidents within this dataset does not hold a strong dependence on the weather, road, or light conditions. The likelihood of an accident being more severe than involving just property was effectively 2.33:1 regardless of the environmental conditions. The incidents involving other, non-environmental factors were too low or did not have fine enough gradations in their datasets to differentiate between the severity types. With more severity types or data regarding total miles driven per collision, or more granulated condition information, an answer to the question of what conditions are the most dangerous to drive in or are related to the most (severe) accidents would be able to be provided. The weather and road conditions most related with accidents in general were CLEAR and DAYLIGHT, respectively. The time of day is a bigger indicator of potential accidents but is likely related to the number of cars on the road during the evening rush hour home from work.

The similarities between the two severity codes within this dataset are great enough that it is not possible to do a machine learning application because there isn't enough separation to be able to classify the conditions around one or the other report type. With more information then it could be possible, but it would require an expanded dataset for more severity codes or finer-

scaled information about the environmental conditions. The K-nearest neighbor clustering did provide some power with a model, reaching upward of 70% accuracy with the test set while maintaining a high level of accuracy within the train set but this is as good as if just assuming the ratio between report types.

To avoid a more serious collision, the data used in this specific instance does not give a clear answer though based on total incidents per day, the likelihood is to be involved in an incident during the mid- to late afternoon. As for time of the year, there is a dip toward the end of February and around the holidays at the end of the year (late December to early January), but otherwise, across the year, the distribution of collisions is relatively consistent. To dive deeper into the question of what conditions create the most serious accidents, data about the most serious accidents is needed, including fatalities, as well better information about the different weather road conditions.

## 5. References


Andrey, D. J. (2001). *Weather Information and Road Safety*. 39.

Edwards, J. B. (1998). The Relationship Between Road Accident Severity and Recorded Weather. *Journal of Safety Research*, *29*(4), 249–262. https://doi.org/10.1016/S0022-4375(98)00051-6

Edwards, J. B. (1999). The temporal distribution of road accidents in adverse weather. *Meteorological Applications*, *6*(1), 59–68. https://doi.org/10.1017/S1350482799001139

Malin, F., Norros, I., & Innamaa, S. (2019). Accident risk of road and weather conditions on different road types. *Accident Analysis & Prevention*, *122*, 181–188. https://doi.org/10.1016/j.aap.2018.10.014

Pisano, P. A., Goodwin, L. C., & Rossetti, M. A. (2008). Highway crashes in adverse road weather conditions. *Proceedings of the American Meteorological Society 88th Annual Meeting. Retrieved September 23, 2009 from Http://Ams.Confex.Com/Ams/88Annual/Techprogram/Session_20929.Htm*.