Predicting Car Accident Severity: A Machine Learning-Based Approach

Maarten de Graaf

26 september 2020

1. Introduction

This section outlines the problem context. Although every licensed driver will have been exposed traffic accidents somewhere down the line, we describe the problem background and interest here. Furthermore, we formalize the problem within a problem statement

1.1. Background

With the increasing use of motorized vehicles, came a vast pressure on the network of highways. In large cities such as Seattle, this often leads to traffic congestion and as a result, accidents. Those incidents may leave other road users in peril, requiring action from emergency services to restore safety. Each accident is different in terms of severity, it is therefore often unknown how many emergency vehicles should be sent out. In high-severity cases, emergency aircrafts and fire brigade involvement may be required too. It is advantageous for those institutions to accurately predict how many vehicles and manpower are required to mitigate the impact of car accidents – or to not assign resources at all. A pre-emptive assessment of car accident severity leads to more complete information on the means required to escort each casualty to safety and restore the infrastructure.

1.2. Problem

This study aims to predict the severity of a car accident according to historical data. Emergency services often receive incomplete information when informed about collisions. Therefore, they struggle to assess how many resources they should allocate to mitigate the accident. Based on historical data, machine learning tools may aid them in resource allocation by predicting collision severity according to data provided by civilians. This study aims to bridge the gap between past accidents and future emergency assistance.

1.3. Interest

Emergency services would reap large benefits from early information on the severity of incoming accidents. These benefits mainly apply to more efficient resource planning, optimization of personnel use and more effective mitigation. Furthermore, road users are stakeholders as well, since they benefit from faster accident resolution and experience enhanced safety when involved in accidents themselves.

2. Data

This section describes the data requirements for this study. The data set of choice is introduced, as well as the steps for data cleaning and feature selection. The latter consists of a list of features that serve as input to predictive modelling, to be used by machine learning algorithms later.

2.1. Data source

Data on collisions within the city of Seattle is available through an Open Data initiative hosted by Seattle GeoData (Collisions, n.d.). It is noted this dataset is the example dataset provided for this capstone project. Since it provides an extensive set of attributes, it was decided to use it as main data source for this project.

2.2. Data cleaning

The data retrieved from Seattle's Open Data bank contained an abundance of null values, along with various other issues. It was chosen to drop every row containing null values for the attributes chosen for feature selection. Due to insufficient record keeping, it was undesired to drop all rows with null values for each attribute, as this led to dropping rows that may include values we can use for model building.

Several issues exist within the data set. The main problem is the large variance in accident size. As where the mean of people and cars involved is around two, maximum values extend up to 80 for people and 15 for cars. For vehicle count, collisions between seven cars or more amounted for <0.05% of the total data and were therefore dropped. Likewise, collisions involving more than 10 people were also relatively rare and were therefore dropped as well (<0.2% of the total dataset).

Besides null values, various columns include 'Unknown' entries. For some attributes, the occurrence of missing values, labelled as unknown, is as high as 12000. However, it was chosen to not drop these values. This mainly stems from the fact that emergency services may not always receive perfect information ahead of coming to the rescue. In order to increase predictability of the model for cases with limited information, unknown values were chosen to be left in place.

Secondly, the inclusion of categorical data required a variety of data transformation operations. Attributes such as junction type, weather, road condition and light condition hold categorical values, rather than numerical ones. In order to prepare those attributes for machine learning algorithms, they were transformed through one-hot encoding.

Thirdly, speeding data contained a significant amount of null values, as only 9339 out of 188617 rows were filled out. A first exploration of the attribute resulted in finding no entries were recorded in case speeding did not occur. That is, speeding data only concerns cases in which speeding occurred, rather than stating 'no' if speeding did not apply. In order to surmount this issue, each NA value was replaced by 'No', as to represent speeding did not occur.

Finally, collision location was transformed into a feature representing whether the accident occurred at a high risk location. Originally, the location attribute yields a description of the collision location. The top 20 most occurring descriptions were labelled as a high risk location, as where others were not.

2.3. Feature selection

Data cleaning resulted in a data set with 188,617 rows and 40 attributes. Only a smaller subset of those features are viable for machine learning algorithms. First, the data set includes metadata for each recorded collision such as an incident key and a unique identifier. Those were all dropped. Features regarding date and time were dropped, as they incorporated several problems. Date and time were often not fully filled out, causing time to often be incomplete. If one were to be interested in investigating the relationship between time of day and the likelihood of collisions, this would not be possible with this data. As where it would be interesting to investigate which day of the week, the transformation from dates to weekdays was deemed too expensive for this particular research.

Since emergency services are the problem owner for this study, we assume the machine learning model will have to forecast collision severity based on incoming civilian reports. Therefore, attributes related to codes and descriptions provided by officials are not viable. Furthermore, civilians are unable to distinguish between injuries, serious injuries and fatalities. Therefore, they are not included within the selected features.

Some attributes were redundant taking into consideration other features. For example, including the amount of people involved, renders the count of pedestrians and bicycles rather obsolete. Those features are significantly narrow focused. Here, a more general approach to the amount of people involved is assumed.

The data set yields a multitude of binary features. This includes whether the collision occurred due to speeding or inattention, whether the person was under the influence of drugs or alcohol or whether a parked car was hit. It was chosen to include speeding and intoxication/drug use as features, since data exploration unveiled those cases are prone to higher accident severity. Severe cases require more extensive effort and resources by emergency services. It is therefore paramount our model encapsulates such information. Intoxication or drug use may not always readily be assessed by civilians, yet it is assumed high levels of substance abuse reflect in driver behaviour.

The final set of features includes nine attributes, which yield the following characteristics: 1) no expert view is required to assess their value and 2) they are not highly correlated to other features within the selection. Table 1 provides an overview of the feature selection process.

Kept features	Dropped features	Reason for dropping features
High risk location, JUNCTIONTYPE	X, Y, LOCATION, CROSSWALKKEY	Aggregated into binary high risk location (Y/N), included JUNCTIONTYPE as additional predictor.
PERSONCOUNT, VEHCOUNT	PEDCOUNT, PEDCYLCOUNT, INJURIES, SERIOUSINJURIES, FATALITIES	Dropped features that cannot be assessed by civilians. Aggregated people into PERSONCOUNT.
WEATHER, ROADCOND, LIGHTCOND	-	Included to assess environmental conditions.
-	INCDATE, INCDTTM	Incomplete, too expensive to transform.
SPEEDING, UNDERINFL	INATTENTIONIND, PEDROWNOTGRNT, HITPARKEDCAR	Opted for two binary variables that imply higher severity.
-	COLLISIONTYPE, SDOT_COLCODE, SDOT_COLDESC, SDOTCOLNUM, ST_COLCODE, ST_COLDESC, SEGLANEKEY, CROSSWALKKEY	(Meta)data provided by Seattle's emergency services, excluded as civilians cannot assess them.

Table 1: Feature selection

3. Methodology

This section addresses data exploration, results from inferential statistical testing and an outline of the machine learning techniques that were used during experimentation.

3.1. Exploratory data analysis

3.1.1. Target variable

The goal of this research is to predict accident severity. Therefore, the target variable is "SEVERITYCODE", which entails a scheme for indexing the severity of an accident according to a number. Two types of severity exist within the data set. Code 1 represents property damage caused by the accident, code 2 indicates the collision led to injuries. Figure 1 displays the occurrences per

severity code. It can be seen non-injury accidents are incrementally more prevalent throughout the data in comparison to collisions harming civilians.

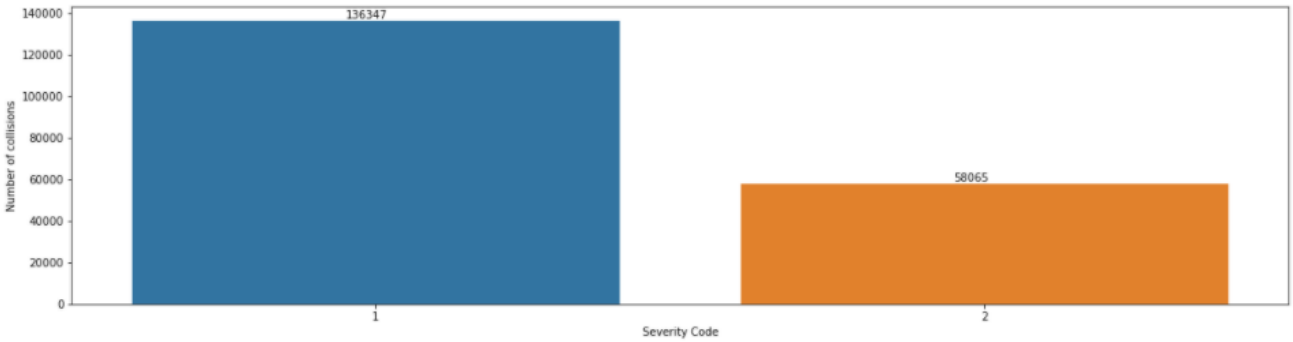

Figure 1: Collisions per severity code

3.1.2. Descriptive statistics

Descriptive statistics provide a compelling overview of numerical data within a dataset. Statistical metrics aid in identifying outliers and inferring the distribution of attribute values. In this particular case, non-categorical attributes included during feature selection were investigating in terms of statistics. Table 2 holds an overview of the results.

	Number of people	**Number of vehicles**
mean	2,444427322	1,920779975
std	1,345928746	0,631046688
min	0	0
25%	2	2
50%	2	2
75%	3	2
max	81	12

Table 2: Descriptive statistics for numerical variables

It can be seen large variance exists within the numerical, independent variables. As where mean values for both the number of vehicles and people is around two, maximum values range up to 12 and 81. This is problematic, as such outliers may largely offset model behaviour. Large-size accidents are disproportionally prevalent throughout the dataset in comparison small accidents.

3.1.3. Distribution of attributes

A variety of visualization techniques may unveil more characteristics of the underlying data. As descriptive statistics hinted at large variance within the dataset, further investigation of the skewness of the data is necessary. The distribution of values for the number of vehicles and people involved is displayed in Figure 2.

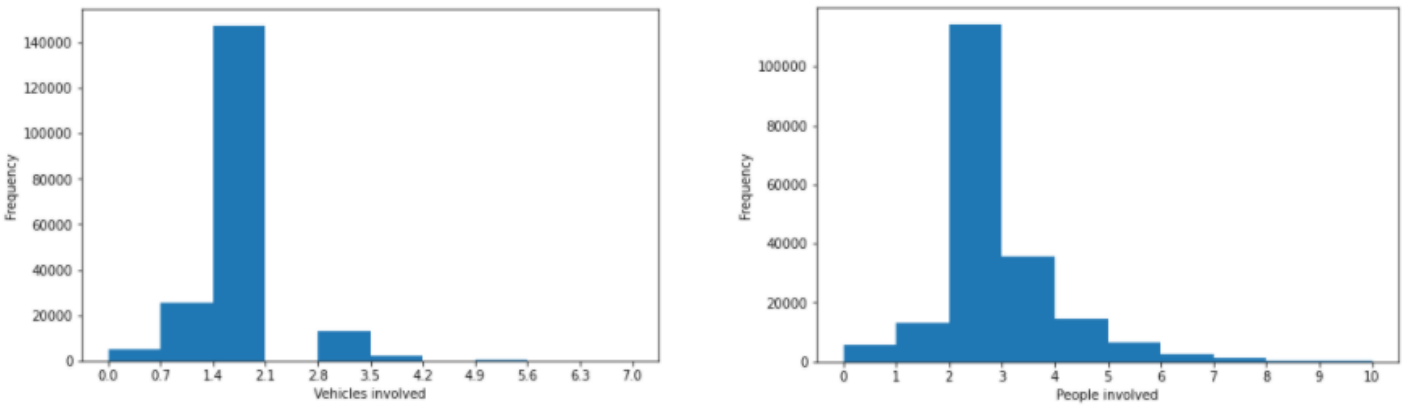

Figure 2: Histogram of vehicles involved (left) and people involved (right)

The histograms displayed in Figure 2 unveil the number of vehicles and people involved is significantly skewed. For each attribute, a certain range of values outnumbers any other range by a large amount. One may conclude that most accidents recorded within the dataset are small and only involve one or two cars, with, usually, two people involved.

We observe similar trends for weather conditions. Figure 3 shows clear weather is far more prevalent than any other weather type, with some weather types not evidently showing from the graph. Most observations took place under clear weather conditions. Yet again, data is skewed into this direction.

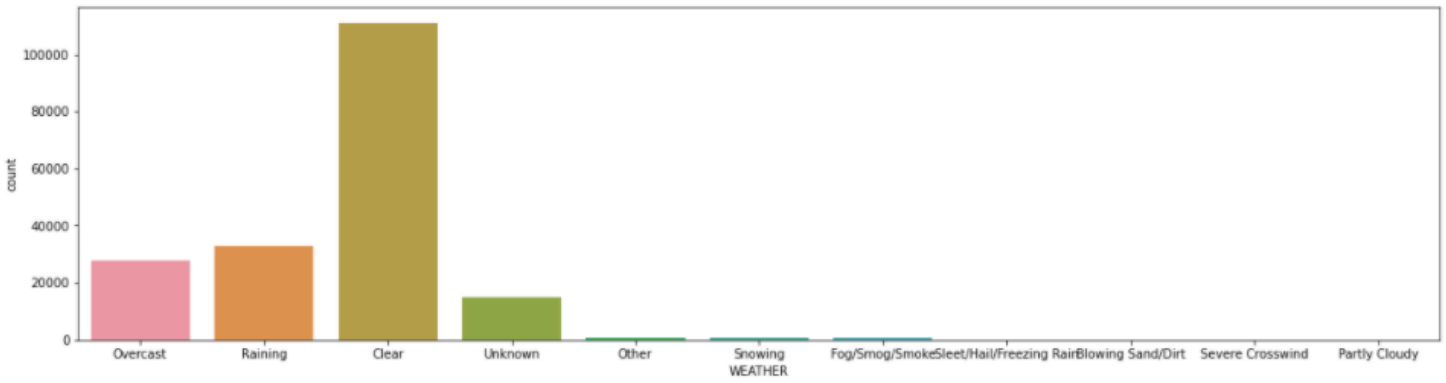

Figure 3: Bar chart of weather conditions

Similar trends can be distinguished for lighting conditions, road conditions and junction types. Appendix I contains an overview of the numbers for those attributes.

3.1.4. Relationships between attributes

Ahead of introducing the methods utilized for predictive modelling, the relationship between attributes selected as features needs to be investigated. With data skewness as high as shown in the previous paragraph, it is likely data skewness issues persist through to the relationship between attributes.

The number of people involved is disproportionally centred on 2, showing almost no difference between accidents severities with regard to the number of people involved in the accident. Figure 4 displays the boxplot for this relationship.

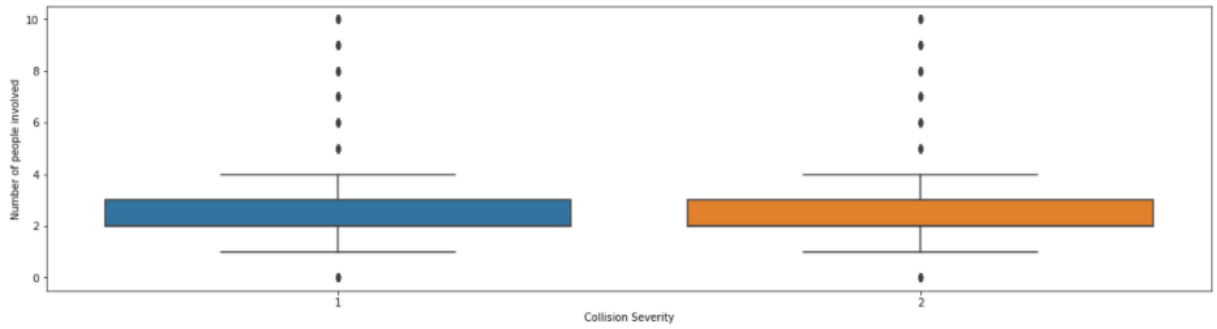

Figure 4 Box plot of collision severity and number of people involved

Similar characteristics pertain to the relationship between weather conditions and the number of people involved. Since we consider the number of people involved as a proxy for accident severity and expect certain weather conditions to increase the occurrence of collisions, one would expect to observe a relationship between weather conditions and the number of casualties. Figure 5 shows this is not the case.

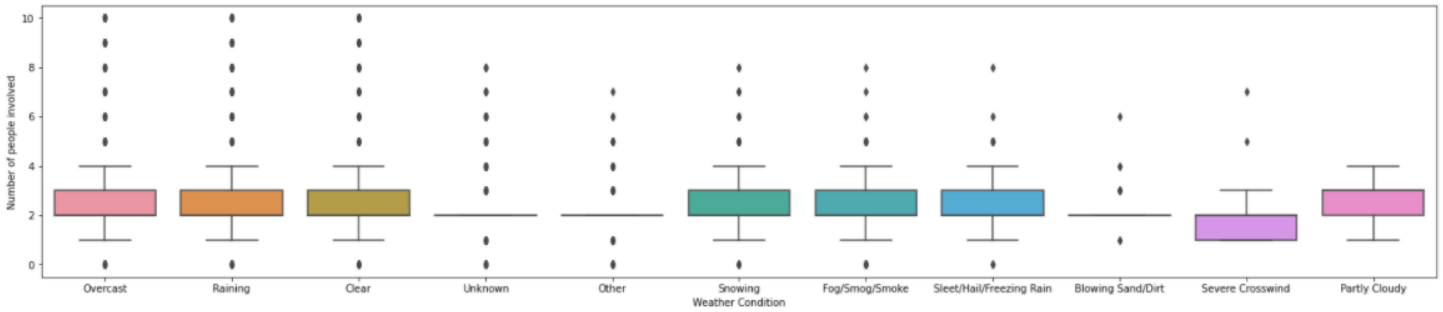

Figure 5 Box plot of weather conditions versus the number of people involved

The observed patterns may be attributed to either one out of two of the following arguments:

- Data skewness eliminates variance within the Interquartile Range. Therefore, other observations are labelled as outliers. Despite removing large values for both vehicle count and the number of people involved, the skewness remains to pose issues. This leads to concluding the data set is heavily biased towards small accidents.
- Person count and vehicle count are not appropriate as proxies for collision severity. As the definition of collision severity solely concerns the level of harm caused, rather than providing an absolute score for the number of casualties, the issue here may supersede simple classification.

Based on this data exploration, we adopt the view that the data set is skewed, yet ample motivation is present to apply machine learning models. Data bias towards small accidents is not regarded troublesome, as this reflects real-world behaviour. It is noted we strive to construct a general model, rather than a model tailored to a certain type of collision.

3.2. Machine learning techniques

In order to predict collision severity, a range of machine learning tools may yield insightful results. In this particular case, classification tools are most appropriate to address the problem. The target variable stipulates the prediction of a binary code, indicating collision severity. Classification approaches are viable for this type of problem, especially since the range of options is limited to two.

From the range of tools available within the realm of classification algorithms, Logistic Regression, Decision Trees and K-nearest neighbour are the most viable. Although the latter is rather slow in processing big data sets, computation times are rendered sufficiently small to be considered. Logistic regression and decision trees are generally suitable for big data sets. With the number of observations approximating 200,000, it is clear memory-efficient algorithms should be opted for here. Therefore, it was chosen to omit SVM from predictive modelling, as this algorithm is typically computationally inefficient on large data sets.

4. Results

The application of classification models was rather straightforward, considering the target variable was already structured according to a binary score. The number of samples involved in predictive modelling was equal to 188,617. As noted before, SVM was not viable for modelling, given the computationally intense process of incorporating approximately 200,000 observations. Nonetheless, logistic regression, decision tree and k-nearest neighbours proved viable for modelling. Three performance metrics were utilized to assess model performance; Jaccard index, F1-score and Log Loss. It is noted the latter only applies to logistic regression, as the other algorithms do not yield probability scores. The scores are presented in Table 3.

Algorithm	Jaccard	F1-score	LogLoss
KNN	0.71	0.69	NA
Decision Tree	**0.72**	**0.70**	NA
Logistic Regression	0.69	0.65	0.57

Table 3 Performance of classification models. Best performance labelled in red.

Among the individual models, the decision tree performed best. Logistic regression yields lower scores for both Jaccard and F1, with a rather undesirable probability. The Jaccard score indicates the decision tree model is able to correctly categorize up to 72% of the actual labels, as where the F1-score indicates the ability of the model to predict the correct class, at the expense of False Positives. The decision tree achieves the best performance on both metrics and is therefore the desired model of choice.

5. Discussion

The results pose ample room for discussion. First of all, the target variable may not accurately represent the information needed by emergency services. The severity code merely indicates whether property damage or injuries occurred, both the size of the property damage and the number of injuries is left uncovered by the target variable. One may obtain more compelling results by transforming the severity code into a severity score, incorporating the number of vehicles, people, size of the property damage and casualties into the grading scale. As for now, the model only provides a general indication of the accident, rather than offering extensive insights into the resources required. This flaw stems from a combination between the data available and the modelling technique of choice.

The data set is undeniably biased towards small accidents, as they occur more often in proportion to larger accidents. It was deliberately chosen to consider large accidents as outliers and focus on smaller accidents. This presumably leads to the model performing worse when assessing the severity

of large accidents. Again, the target variable obscures the assessment of traffic hazard size, yet it is noted the model has been trained and tested according to the biased data set.

Furthermore, we opted for a general model here. That is, model training and testing predominantly concerned a sizable dataset with large variance in calamity size, environmental conditions and casualties. Therefore, it is not tailored to specific conditions. For example, it would be possible to manipulate the data set as to only keep records with extreme environmental conditions. This would result in the model being able to predict such cases more accurately.

5.1. Recommendations

A few recommendations for emergency services apply, based on the results described above. First of all, classification results divulge that the severity of collisions can be predicted relatively well according to the number of people and vehicles involved and a set of environmental conditions. Emergency services are therefore advised to take note of those factors when receiving notifications on traffic calamities, as they could use them to assess the gravity of the issue at hand.

The type of collisions that generally occur are of low severity with limited vehicles and people involved. Nonetheless, it remains necessary to distinguish between property damage and actual injuries in order to assess which resources should be allocated. The model aids in doing so.

Furthermore, including locational data in the form of labelling a set of highly prevalent locations as high risk, did not significantly improve model performance. Therefore, emergency services are advised to look at what the calamity entails, rather than where it occurred.

In order to build onto the model presented here, emergency services and the city of Seattle are advised to extend their data collection beyond the current set of attributes. If severity codes were transformed into a severity score, incorporating data on the size and gravity of the accident as well, predictions could be more compelling.

6. Conclusion

This study analyses the relationship between collision severity and several environmental and accident-specific factors. Environmental conditions such as weather, lighting and road conditions were identified as important features, as where the number of people and vehicles involved served as proxy for accident severity. The model presented here may be increasingly useful to aid emergency services in acting upon incoming calamity notifications and optimizing their resource allocation.

It is noted the model is subject to a set of constraints. First of all, the data set is largely biased towards small-size accidents involving a low number of both people and vehicles. Therefore, the generalized model produced here achieves best performance when applied to similar cases. As where one could regard the high prevalence of small accidents as positive, this general-purpose model may come short in predicting the resources required for large accidents. It is suggested here to transform severity from a mere code to a score, encompassing information on the absolute size of the accident – something it currently fails to represent.

On a final note, no increased performance was achieved through infusing the model with locational data. This rejects the notion of high risk junctions being of influence on collision severity.

7. Opportunities for future research

Although the best performance model achieves approximately 72% accuracy, it does not account for a significant portion of the variance existent within the data set. This mainly arises from data

skewness issues, as well as the target variable being an inaccurate proxy for collision severity. Therefore, future research is advised to transform the severity code into a severity score by incorporating more accident-specific data. For this research however, devising such a score was rendered out of scope. If such a score were to be designed, the preferred modelling approach would shift from classification, to numerical prediction, i.e. regression. This would lead to more compelling predictions for emergency services.

As where devising a new severity score could be a direction for future research, improving data collection could also lead to new insights. More data on, for example, the number of casualties and the size of the property damage could propel the prediction of collision severity into unprecedented accuracy. With such historical data readily available, future endeavours may be able to develop models to fully automate resource allocation to hazardous traffic situations.

Appendix I: Tables for Junction Type, Road Conditions and Lighting Conditions

Junction Type	Frequency
Mid-Block (not related to intersection)	89689
At Intersection (intersection related)	62719
Mid-Block (but intersection related)	22744
Driveway Junction	10660
At Intersection (but not related to intersection)	2096
Ramp Junction	166
Unknown	9

Road Conditions	Frequency
Dry	124315
Wet	47415
Unknown	15078
Ice	1207
Snow/Slush	1001
Other	131
Standing Water	115
Sand/Mud/Dirt	75
Oil	63

Lighting Conditions	Frequency
Daylight	115943
Dark - Street Lights On	48455
Unknown	13473
Dusk	5896
Dawn	2497
Dark - No Street Lights	1534
Dark - Street Lights Off	1198
Other	235
Dark - Unknown Lighting	11

References

Collisions. (n.d.). Retrieved September 27, 2020, from https://data-seattlecitygis.opendata.arcgis.com/datasets/collisions