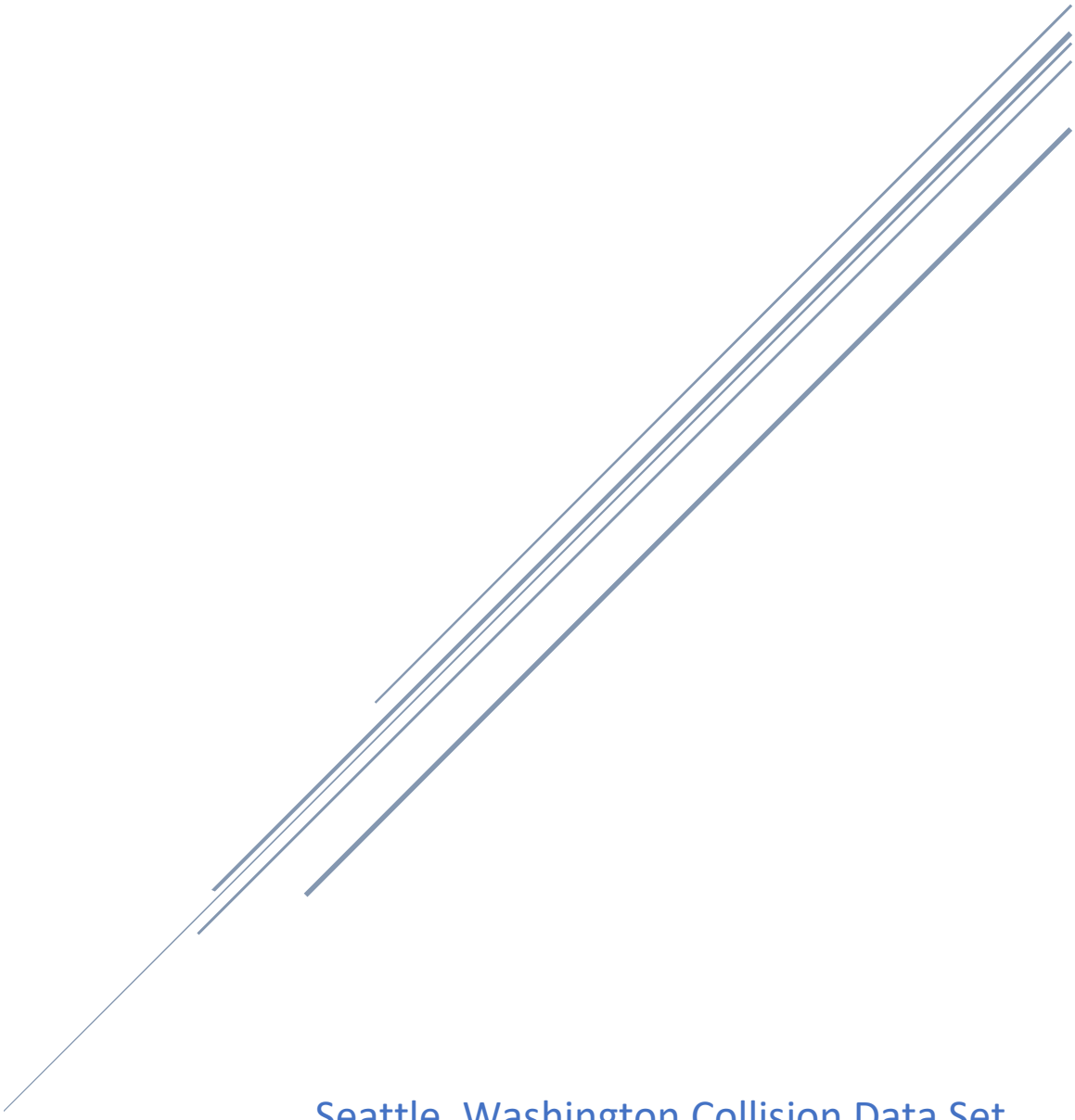
CAPSTONE PROJECT

IBM Data Science Profession Certificate



Seattle, Washington Collision Data Set

By: Austyn R. Matheson, Date: 05-August-2020

Table of Contents

1. Background

In North American cities, road accidents are a significant cause of fatality and injury each year. Cities and regions are likely to have comparable collision trends, owing to similarities in road rules, road infrastructure, and lifestyles. According to the most recent report on collision data for Alberta, Canada*, 290 fatalities and 17,186 injuries were reported in 2017 with a 7% yearly increase of traffic collisions (142,467) from 2016. In this same report, documented collision risk was higher based on several different factors:

- The highest number of **fatal collisions occurred in July**.
- The highest number of **injury collision occurred in November.**
- **Friday** was the most collision-prone day.
- **Afternoon rush hour** was the most collision prone time-period.
- Male drivers between the ages of **(M) 16-17** had the highest involvement rate in casualty collision.
- Following too closely, running of the road, and making a left turn across the path of an oncoming vehicle contributed the most to casualty collisions.
- Fatal collision occurred more frequently in rural areas.
- 15.4% of pedestrians involved in fatal collisions were impaired compared to 4.4 % of drives in the injury collisions.
- 10.2% of drivers in fatal collisions were impaired compared to 1.7% of drives in the injury collisions.

Out of these risk factors, data attributes including date/time, location (rural/local, cross-section type), and impairment were key factors for collisions involving injuries and fatalities. In the dataset and statistics for Alberta, specific road conditions (weather, lighting) were not included. These key parameters (date/time, location, and impairment) are likely similar throughout cities/regions in North America. In this study, the Data Collision for the Seattle, Washington region was used. In the Seattle dataset, each collision was assigned a **severity code** (1 or 2). For each collision, a detailed entries included:

1. Date/Time
2. Location Attributes (longitude, latitude, intersection type, intersection name, etc …)
3. Collision Attributes (number of people/cars/cyclists/pedestrians, collision type, inebriation, etc…)
4. Driving Condition Attributes (weather, road conditions, lighting conditions)
5. State Code Attributes

The Seattle dataset has been collected from 2004 until 2020. In this analysis, the contribution of attributes to collision severity will be analyzed by applying a machine learning model. Evidently, it would be desirable to present collisions and provide a 'risk index' to drivers such that they could appropriately prepare for conditions and avoid driving if warranted.

Therefore, the objective of this project was to develop a predictive model with the Seattle Washington collision data set to determine which attributes contribute to the risk of collisions.

This was achieved by the following three aims:

1.1 Data Analysis Aims:

1. Aim 1: Data Preprocessing and Preparation
2. Aim 2: Data Visualization and Understanding
3. Aim 3: Model Fitting and Model Evaluation

The end users of this model could be a variety of groups including; the government, insurance companies, individual drivers/cyclists/pedestrians Here are some of the ways end-users could implement the findings from this model

1.2 End Users:

1. The Government
    a. Design road infrastructure and driving rules.
    b. Design driver/bicyclist education programs.
    c. Use findings to reduce collisions and improve safety.
2. Insurance Companies
    a. Adjust premiums around driver habits
    b. e.g Increase premiums for drivers that commute during the afternoon rush hour.
3. Individual Drivers/Cyclists
    a. A mobile or web-based application could be adjusted (e.g. google map) could input these findings to show road conditions/roadways with high collision risk so drivers could proceed with caution.

2. Data Loading and Pre-Processing

For this data analysis, the Seattle Collision Data Set was used, which provides a well-managed and maintained historical data set and collision severity code. In the Seattle Collision Data set a severity index code is provided, which was our target variable in order to create a model for a 'risk index' to inform drivers prior to heading out on the road. The data set includes 194673 collision entries, collected from dates ranging from 2004 to 2020.

2.1 Data pre-processing

A. Missing Values

For data analysis, columns and rows with more than 50% null, nan, nas values were removed that included the factors listed in the table below.

Table 1. Columns Removed from the Dataset for Analysis

Column Name	Reason for Removal
INTKEY	129603 null
EXCEPTRSNCODE	109862 null
EXCEPTRSNDESC	189035 null
INATTENTIONIND	164868 null
PEDROWNOTGRNT	190006 null
SPEEDING	185340 null
X	5332 null-rows removed
REPORTNO	Unique – not predictive
OBJECTID	Unique – not predictive
STATUS	Matched vs. Unmatched
SEVERITYCODE.1	SEVERITYCODE duplicate
PERSONCOUNT	Column removed not predictive
VEHCOUNT	Rows with 0 removed
SEGLANEKEY	Column removed not predictive
CROSSWALKKEY	Column removed not predictive

B. Missing Categorical Values

To address categorical values, the following variables were converted **unknown categorial type**.

- **For example:** df["COLLISIONTYPE"].replace(np.nan, "Parked Car", inplace=True)

Table 2. Categorical variables most frequent type

Column Name	Most Frequent Variable
COLLISIONTYPE	Parked Car
JUNCTIONTYPE	Mid-Block (not related to intersection)
WEATHER	Clear
ROADCOND	Dry
LIGHTCOND	Daylight
ST_COLCODE	32
UNDERINFL	N – converted to 0 Y – converted to 1 NaN - converted to 0

C. Formatting Date and Time Variables

The variables INCDATE and INCDTTM were converted to datetime objects for easier analysis, rows were added for Year, Month, Day, and Hour.

3. Data Visualization

The data was first explored by visualization to improve the understanding of the variables with the greatest effect. As noted in the previous report for road collisions in Alberta (Canada), it was expected that month, day of the week, and time would be significant factors.

3.1 Date and Time

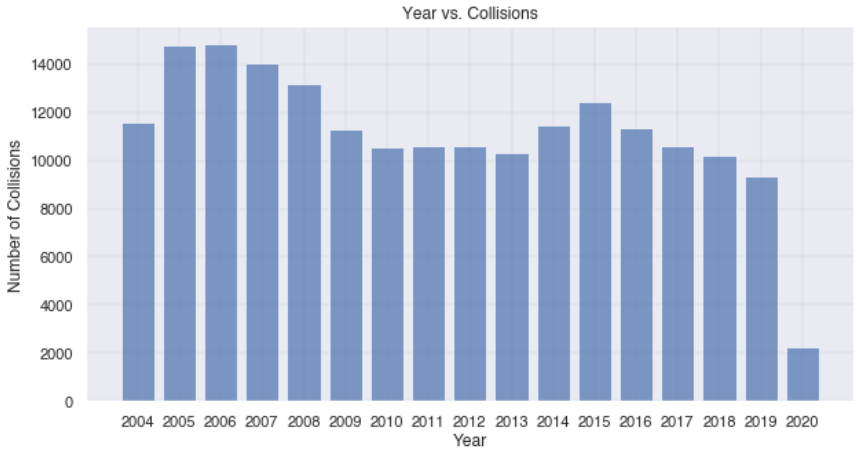


Figure 1. The collision data has been collected since 2004. After 2009 there was a apparent reduction in total collisions, possibly due to measures to increase driver, pedestrian, and cyclist safety. Interestingly, collisions are much lower in 2020. This is likely due to business shut down and stay home orders due to covid-19. However, the data was just recorded halfway through 2020.

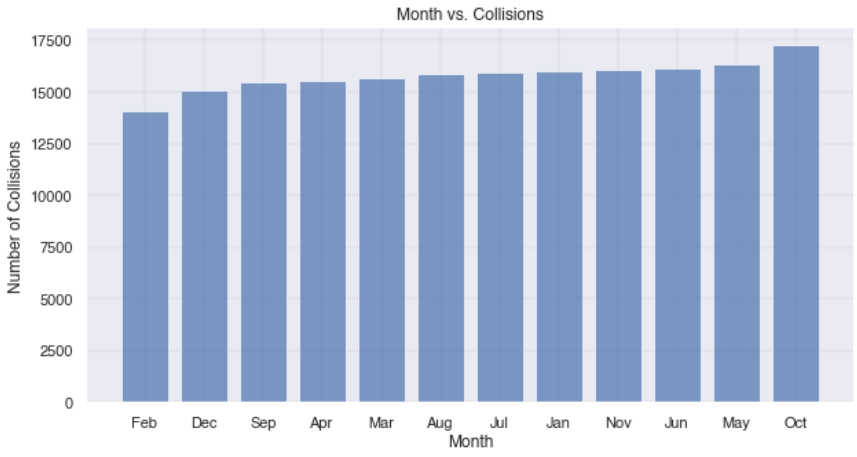


Figure 2. The lowest month for collisions is February and the highest month is October. The rest of the year appears to be consistent.

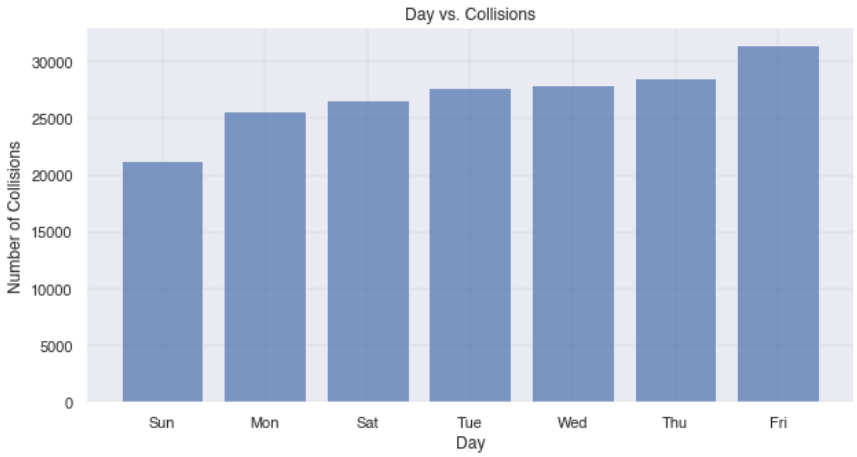


Figure 3. The day of the week also appears to be an important attribute with Friday being the day with the highest count of collisions, and Sunday is the day with the lowest.

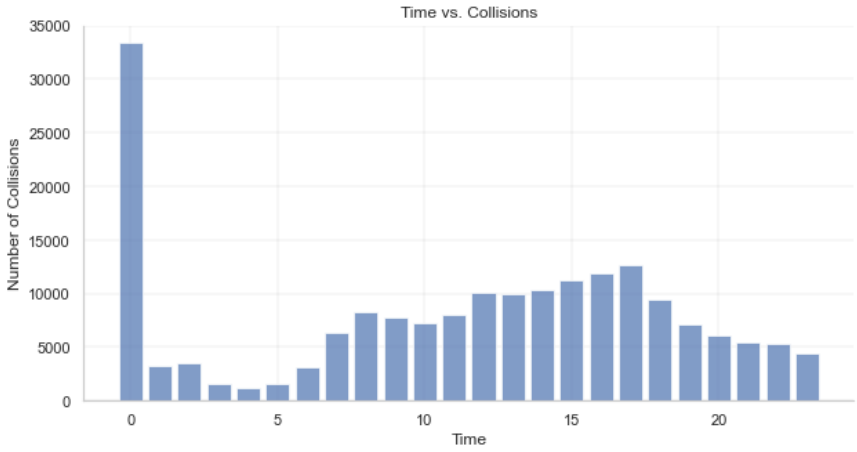


Figure 4. Many missing entries have been incorrectly put to 0 (24:00/12:00AM), which accounts for the last spike in this dataset. Therefore, it will be assumed here that no "jump/spike" occurred at 24:00 and the rows containing "0" for time were removed for the below graph.

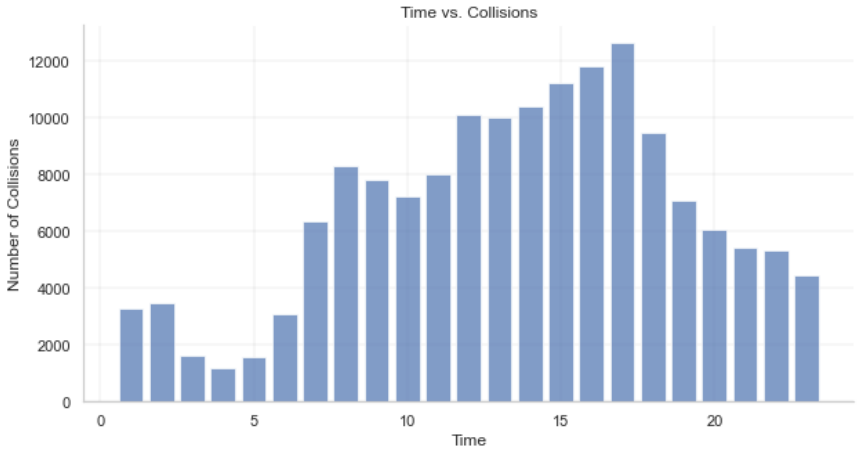


Figure 5. Collision occurrence spiked during the afternoon rush hour (as drivers are returning home from work) between 15:00 to 17:00 (3:00 PM to 5:00 PM).

3.2 Collision Type

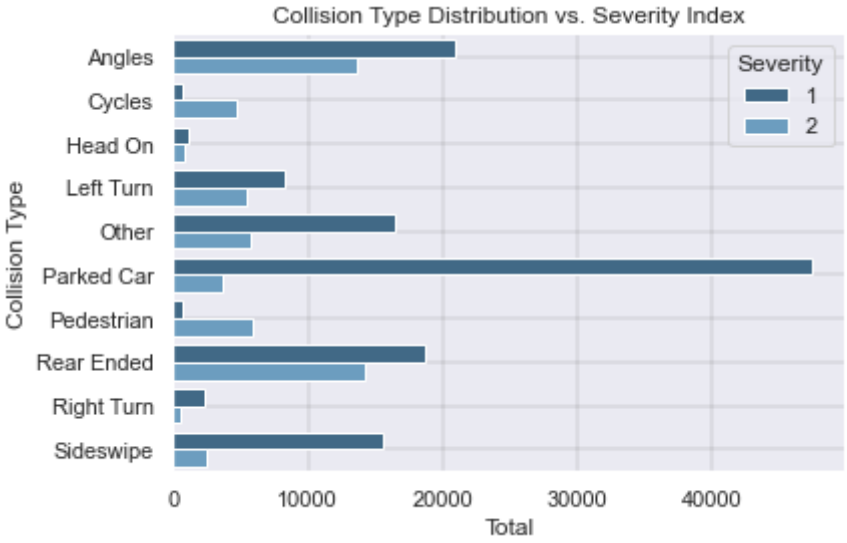
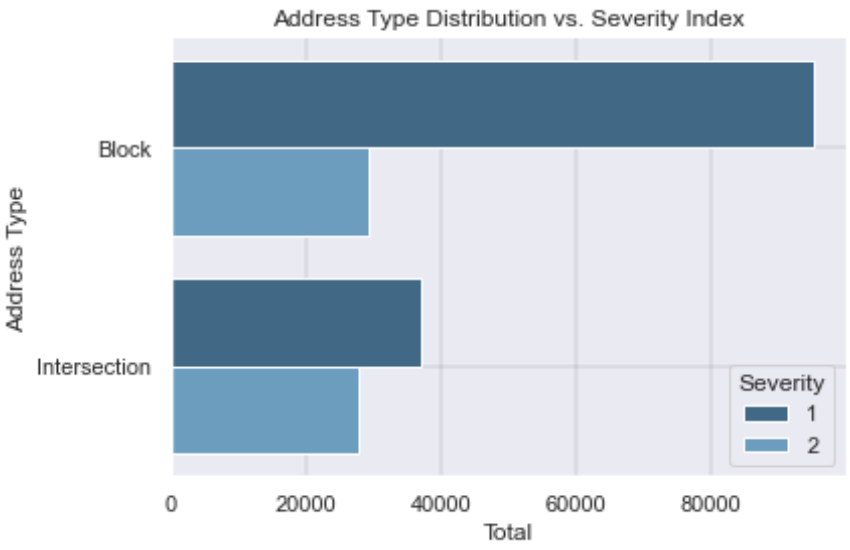


Figure 6. Collision type is an outcome attribute not a prediction attribute, therefore it will not be included in our model. It is interesting to note that the most frequent type of collision involved with parked cars, although these incidents are predominantly severity code = 1. The collision types with a greater quantity of severity code = 2 are Read-Ended and Angles type collisions.

3.3 Location Type



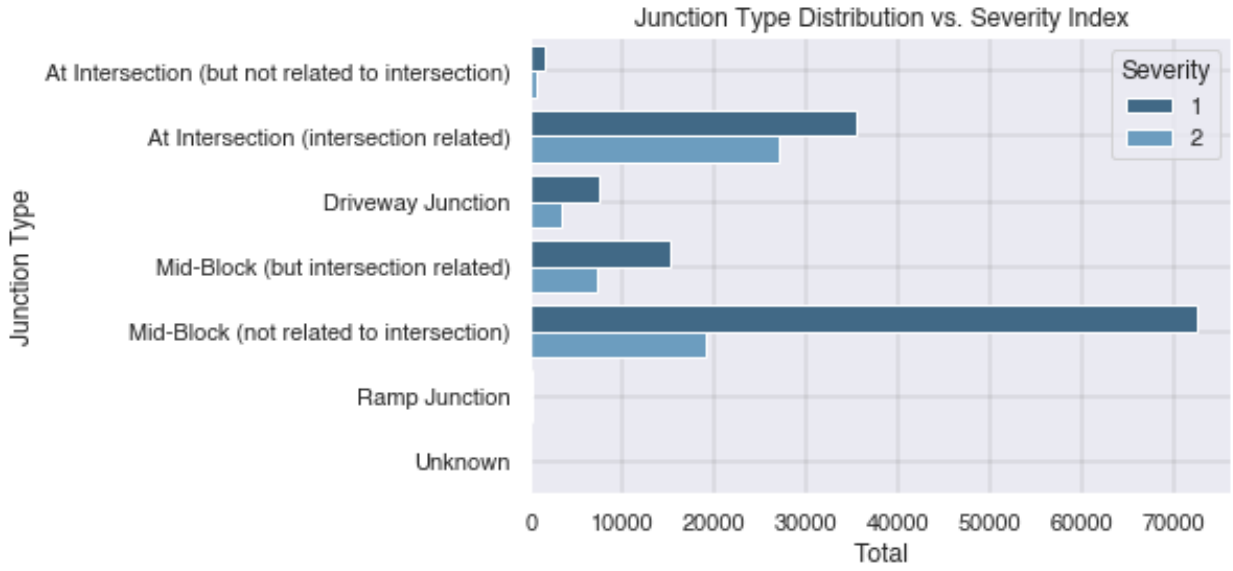


Figure 7. By visualizing the above two figures it is evident that block intersections (Address type vs. collision) and specifically mid-block junction-types (not related to intersection) are were the most collisions occur (severity code = 1). Still, the most severe collisions (severity code = 2) are related to collisions that occur at intersection.

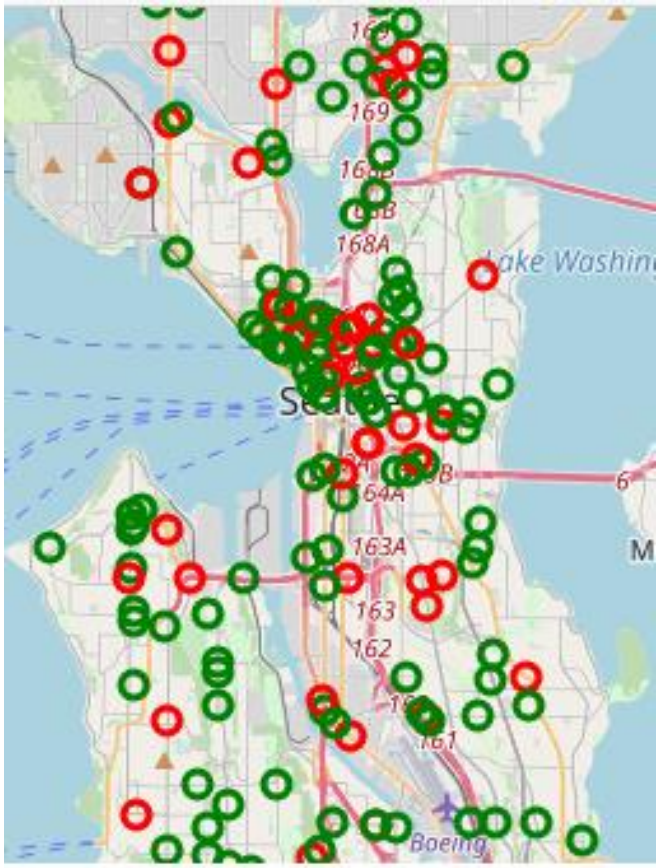
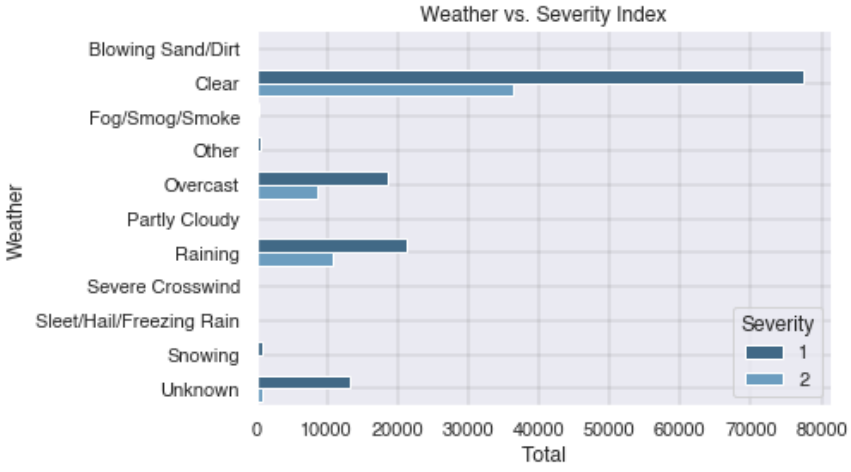


Figure 8. 200 data points were randomly selected. The majority of collisions occurred in the city-center as would be expected. There does not appear to be a pattern for severity code for collision set.

Legend Red: Severity Code 2, **Green:** Severity Code 1.

3.3 Weather, Road, and Light Conditions



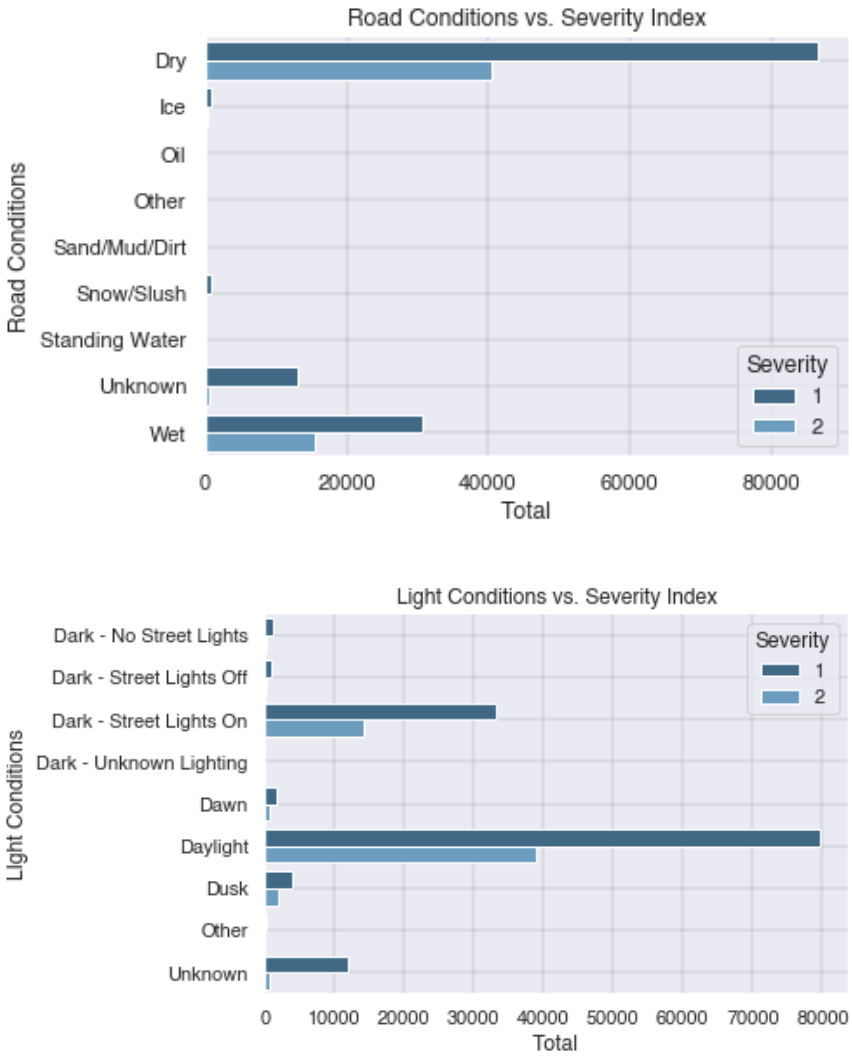


Figure 8. It appears that most collisions occur when the weather is clear. It is possible that more drivers are on the roads when conditions are better as opposed to when they are not favorable. Again, for road conditions the most accidents occur when the roads are Dry, followed by Wet. Perhaps indicative that drivers either drive more carefully when conditions are severe or drivers avoid driving at all. Again, for light conditions the most accidents occur when the roads are Daylight, followed by Dark - Street lights. Perhaps indicative that drivers are less likely to drive later in the evening and in the dark.

3.4 Inebriation and Collision Severity

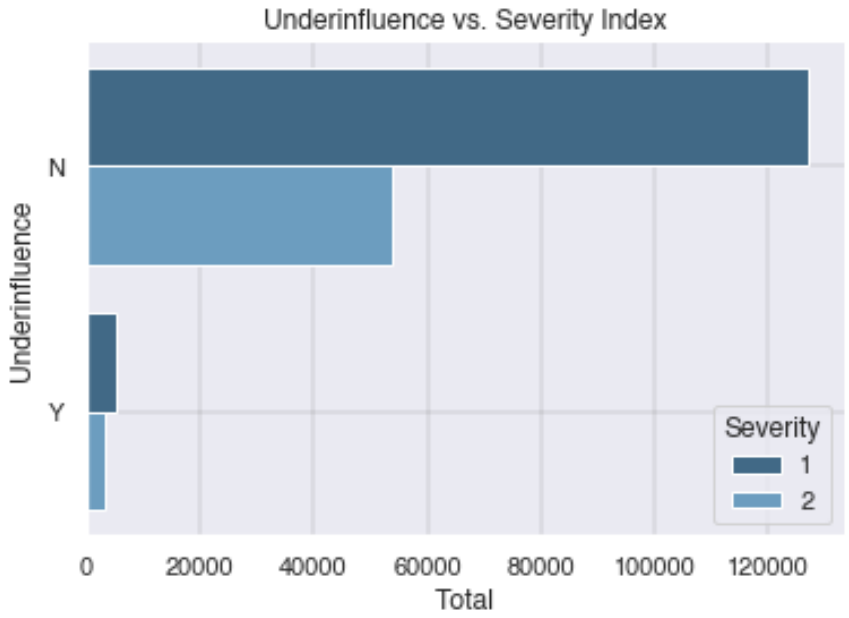


Figure 9. The role of inebriation does not appear to be a significant factor on collision severity.

4. Machine Learning Methodology

4.1 Logistic Regression

The logistic function is used for a binary output model. The output of the logistic regression is probability ($0 \leq x \leq 1$). The logit function predicts the binary 0 or 1 as an output (if $x<0.5$, output= 0, else output=1). In addition, the logistic regression assigns weights to the attributes which shoes the relative importance of factors.

For this analysis, first logistic regression was completed to:

- Determine the strength of the effect of important variables.

A. Steps

1. Load packages
2. provide categorical variables dummy variables
3. select the features for analysis
4. split the data into testing and training set

B. SMOTE*

The smote package was used to find and select with variables were important or redundant for logistic regression modelling.

- Column attributes with "FALSE" score were removed.
- The model was then fit using Logistic Regression
- Column attributes with $p<0.05$ were removed
- The model was then finally refit using Logistic Regression

The final X set including the following columns:

- cols=['Clear', 'Overcast', 'Raining', 'Severe Crosswind', 'Sleet/Hail/Freezing Rain', 'Snowing', 'Dry', 'Ice', 'Sand/Mud/Dirt', 'Standing Water', 'Wet', 'Dark - Street Lights Off', 'Dark - Street Lights On', 'Dawn', 'Daylight', 'Dusk']

C. Model Development and Evaluation

First a loop was run to determine the best setting for the Logistic Regression fit. Accurary was maximized when c = 0.001 and solver = liblinear.

The final model accuracy was evaluated on the Training set and Test Set

- Jaccard and F1 score
- The confusion matrix
- The ROC curve

*The work for using the SMOTE package was based off the article on Towards Data Science. The link for the report can be accessed by this URL https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8

For both SVM and KNN modeling, the same attributes (X) "Features" from logistic regression were used for subsequent analysis.

4.2 SVM & KNN

SVM is a supervised machine learning algorithm that is used for classification or regression problems. It applies a technique called the kernel trick to transform your data and then based on these transformations it finds an optimal boundary between the possible outputs.

K-nearest neighbors is a non-parametric method used for classification and regression. It is one of the easiest ML technique used. It is a lazy learning model, with local approximation.

* Both were abandoned for analysis because the data set is too large for the equipment owned by the data scientist. It is recommended that an Apache Spark based method is applied for classification.

5. Results

5.1 Logistic Regression

After running the logistic regression model, the factors that contributed to a higher severity index included:

Results: Logit

Model:	Logit	Pseudo R-squared:	0.005
Dependent Variable:	y	AIC:	255628.3568
Date:	2020-09-05 14:19	BIC:	255780.2994
No. Observations:	185258	Log-Likelihood:	-1.2780e+05
Df Model:	14	LL-Null:	-1.2841e+05
Df Residuals:	185243	LLR p-value:	1.3489e-252
Converged:	1.0000	Scale:	1.0000
No. Iterations:	5.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Clear	0.0489	0.0148	3.3006	0.0010	0.0199	0.0780
Raining	0.1307	0.0198	6.5865	0.0000	0.0918	0.1696
Severe Crosswind	-1.4454	0.6406	-2.2562	0.0241	-2.7010	-0.1898
Sleet/Hail/Freezing Rain	-1.2087	0.2750	-4.3952	0.0000	-1.7478	-0.6697
Snowing	-0.7595	0.0847	-8.9628	0.0000	-0.9256	-0.5934
Dry	0.1709	0.0257	6.6560	0.0000	0.1205	0.2212
Ice	-0.4240	0.0727	-5.8291	0.0000	-0.5666	-0.2814
Sand/Mud/Dirt	-0.5870	0.3092	-1.8983	0.0577	-1.1930	0.0191
Standing Water	-1.0874	0.2727	-3.9872	0.0001	-1.6220	-0.5529
Wet	0.1866	0.0273	6.8364	0.0000	0.1331	0.2401
Dark - Street Lights Off	-0.8064	0.0720	-11.2033	0.0000	-0.9475	-0.6653
Dark - Street Lights On	-0.2345	0.0244	-9.6271	0.0000	-0.2823	-0.1868
Dawn	-0.1391	0.0473	-2.9387	0.0033	-0.2318	-0.0463
Daylight	-0.0679	0.0234	-2.9007	0.0037	-0.1137	-0.0220
Dusk	-0.0981	0.0349	-2.8083	0.0050	-0.1666	-0.0296

Figure 10. Logistic regression model summary.

The coefficients of logistic regression can be used to calculate the Odds and Probability of an occurrence. The attributes clear, raining, dry, and wet had the greatest odds of contributing to an incident however the probabilities were all low, just slightly over 50% chance of being factors driving a collision.

Equations: Odds = exp(logit coefficient), Probability = odds/(1+odds)

Table 3. Odds and probabilities associated with the model collision data attributes.

Attribute	Odds	Probability
Clear	1	0.50
Raining	1.14	0.53
Severe Crosswind	0.24	0.19
'Sleet/Hail/Freezing Rain'	0.3	0.23
Snowing	0.47	0.32
Dry	1.19	0.54
Ice	0.65	0.39
Sand/Mud/Dirt	0.56	0.36
Standing Water	0.33	0.25
Wet	1.21	0.55
'Dark - Street Lights Off'	0.45	0.31
'Dark - Street Lights On'	0.8	0.44
Dawn	0.87	0.47
Daylight	0.93	0.48

Dusk	0.9	0.47

Table 4. Model Evaluation

Metric	Value
Jaccard Score	0.559
F1 Score	0.545

Table 5. Confusion Matrix

10655	17037
7432	20453

The confusion matrix tells us that our model has 10655+20453 correct predictions vs. 7432+17037 incorrect predictions. The accuracy (jaccard score), F1 score, and precision of this model are poor and perhaps another machine learning algorithm would be better. This is shown further by a ROC curve with a minimum surface area.

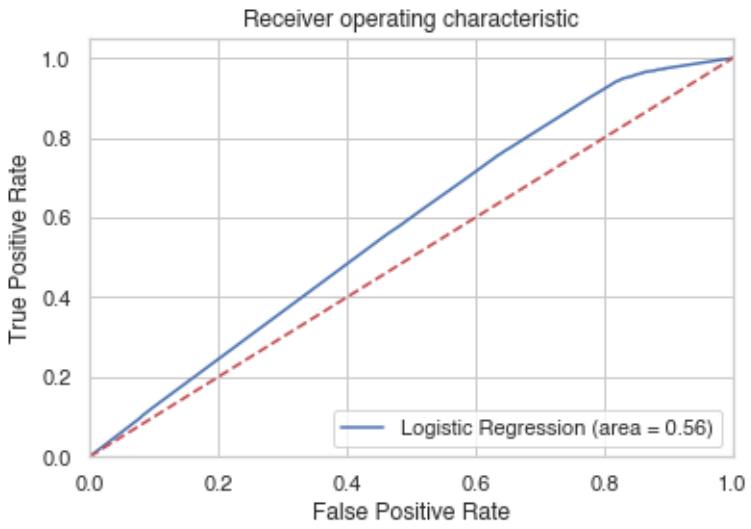


Figure 11. Receiver operating characteristic (ROC) curve.

6. Discussion

6.1 Model evaluation

The accuracy of this model is poor. Other machine learning algorithms including KNN and SVM may improve classification. Unfortunately, the processing power on the machine used for this data analysis was low and the methods used were limited.

SVM is recommended, SVM supports both linear and non-linear solutions (kernel trick) and it manages outliers effectively. However, for large data sets performance is reduced. Challenges with SVM include the multiple iterations required to find the optimal solver parameters.

KNN is inefficient, lazy, and slow when handling large data sets. KNN is easy to use and implement.

6.2 Model findings

Common sense may suggest that dangerous weather, road, or poor lighting conditions would increase the risk of collisions and increase severity overall. However, the findings in the model (supported by the exploratory analysis in the data visualization) demonstrate that collisions predominantly occur when conditions are clear, roads are dry, and lighting is normal. Contrastingly, attributes including month, day of the week and time appear to have a significant impact on collision count.

7. Conclusion

It is recommended that machine learning algorithms, such as ApacheSpark, which are better able to handle and manage large datasets be used for this type of model.