Predicting severe occurrences of automobile collisions in Seattle, WA

1 Introduction

1.1 Description of Problem

There are over 200 million licensed drivers in America. With that many people on the road, accidents are going to happen. The United States Department of Transportation estimates about 7 million car accidents happen nationwide each year. No matter how safe a driver you are, there's a good chance you'll get into at least one accident in your lifetime.

As for the Emerald City, Seattle, it ranks as one of the nation's worst cities to drive in and own a car, according to a newly released study. The report, by personal finance website WalletHub, found that Seattle is the 10th worst U.S. city for drivers.

1.2 Background

The purpose of this report is to predict the severity of automobile collisions in Seattle with the application of machine learning models. These models should assist the Public Development Authority of Seattle focus limited resources towards the variables that have the highest impact on reducing the number of collisions and improve public safety.

2 Data Description

2.1 Data Source & Attributes

The data used for this project originated from the City of Seattle Open Data Portal website, which consists of vehicle collision incidents reported in the city of Seattle from 2004 to September 22, 2020 (the date dataset was downloaded). This input dataset is .csv format and consists of 221,526 incidents with 40 variables. Metadata regarding the variables is available from the portal website.

2.2 Pre-Processing

The pre-processing of the dataset is to prepare it for exploratory data analysis (EDA). Although this dataset is downloaded from a single source, it was probably concatenated from multiple sources.

2.2.1 Variable Redundancy

There are multiple variables that contain the same or similar information, which make them redundant and will be removed from the dataset.

OBJECTID has a unique value for each incident, which corresponds with the dataset index value.

INCKEY, COLDETKEY, and REPORTNO each have unique values for each collision incident.

SEVERITYCODE and SEVERITYDESC are duplicate information. Therefore, SEVERITYCODE will be dropped since SEVERITYDESC has more information.

INCDATE and INCDTTM both provide the date of the incident. Only INCDTTM is required since it includes both the time and date and will be used for time-series analysis.

LOCATION is a catagorical field that can be substituted by latitude and longitude variables (X and Y variables respectively).

SDOT_COLCODE and SDOT_COLDESC correspond to the same information. SDOT_COLCODE is the type of collision, and SDOT_COLDESC has the description of each type of collision. Therefore, we can drop SDOT_COLCODE since SDOT_COLDESC has more information.

2.2.2 Missing Values

The dataset contains 221,525 incidents (rows) and 40 variables (columns). It contains multiple variables with a significant amount of missing data. A large number of missing values could cause noise and bias in the results, therefore will need evaluated (yellow variables in Table 1).

#	Column	Count	Dtype
0	X	214050	float64
1	Y	214050	float64
2	OBJECTID	221525	int64
3	INCKEY	221525	int64
4	COLDETKEY	221525	int64
5	REPORTNO	221525	object
6	STATUS	221525	object
7	ADDRTYPE	217813	object
8	INTKEY	71936	float64
9	LOCATION	216935	object
10	EXCEPTRSNCODE	101122	object
11	EXCEPTRSNDESC	11779	object
12	SEVERITYCODE	221524	object
13	SEVERITYDESC	221525	object
14	COLLISIONTYPE	195212	object
15	PERSONCOUNT	221525	int64
16	PEDCOUNT	221525	int64
17	PEDCYLCOUNT	221525	int64
18	VEHCOUNT	221525	int64
19	INJURIES	221525	int64
20	SERIOUSINJURIES	221525	int64
21	FATALITIES	221525	int64
22	INCDATE	221525	object
23	INCDTTM	221525	object
24	JUNCTIONTYPE	209551	object
25	SDOT_COLCODE	221524	float64
26	SDOT_COLDESC	221524	object
27	INATTENTIONIND	30188	object
28	UNDERINFL	195232	object
29	WEATHER	195022	object
30	ROADCOND	195103	object
31	LIGHTCOND	194933	object
32	PEDROWNOTGRNT	5195	object
33	SDOTCOLNUM	127205	float64
34	SPEEDING	9929	object
35	ST_COLCODE	212112	object
36	ST_COLDESC	195212	object
37	SEGLANEKEY	221525	int64
38	CROSSWALKKEY	221525	int64
39	HITPARKEDCAR	221525	object

Table 2.2.2-1. The number of events located in each variable and its format. Yellow indicates variables that may have too many missing entries to be used. Green indicates variables with only a few missing entries

INTKEY has too many missing entries to be used.

EXCEPTRSNCODE and EXCEPTRSNDESC will be dropped since they have too many missing values to be used in a model.

INATTENTIONIND has a significant amount of missing data with only 30,188 values, which all have a value of 'Y' and pertains to the driver not paying attention while driving and should be dropped.

PEDROWNOTGRNT also has a significant amount of missing data with only 5,195 values, which all have a value of 'Y' and pertain to pedestrian right of way was not granted.

SPEEDING has a significant amount of missing data with only 9929 values, which all have a value of 'Y' and pertain to whether or not speeding was a factor in the collision.

2.2.3 Variable Reformat and Category Merging

To conduct time series analysis, INCDTTM needs to be converted from a string format to the pandas data-time format.

2.2.4 Variables with Skewed Distribution

SEGLANEKEY contains 2,101 unique values with the value '0' dominating with 218,489 of the observations. The distribution of unique values is highly skewed indicating that SEGLANEKEY should be dropped.

CROSSWALKKEY contains 2,343 unique values with the value '0' dominating with 217,283 of the observations. The distribution of unique values is highly skewed indicating that CROSSWALKKEY should be dropped.

2.2.5 Missing Values Matix and Heat Map

Using the *missingno* library to plot and identify where the missing values are located in each column and correlations between missing values across different columns (white lines in the figure 1). Note the high correlation with variables COLLISIONTYPE, UNDERINFL, WEATHER, ROADCOND, LIGHTCOND, and ST_COLDESC.

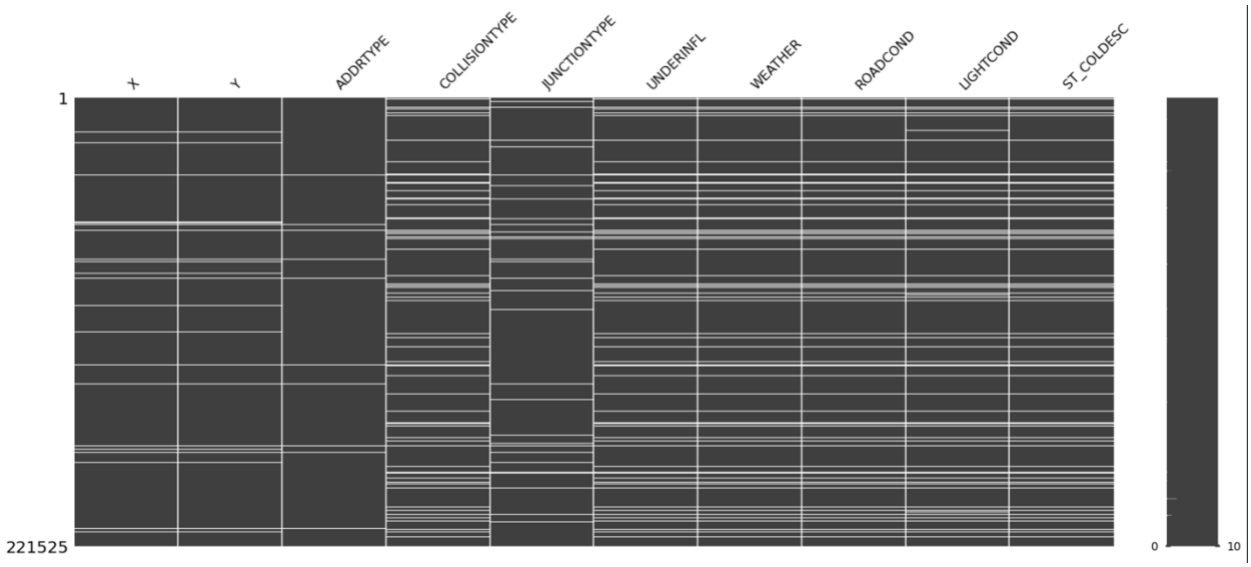

Figure 2.2.5-1. Missing Value Matrix. Visualization of missing values across all variables in the dataset.

The heat map confirms the correlation of the missing data for the same rows along with an approximation of its correlation. For example, variable JUNCTIONTYPE has about a 30% correlation with variables COLLISIONTYPE, UNDERINFL, WEATHER, ROADCOND, LIGHTCOND, and ST_COLDESC.

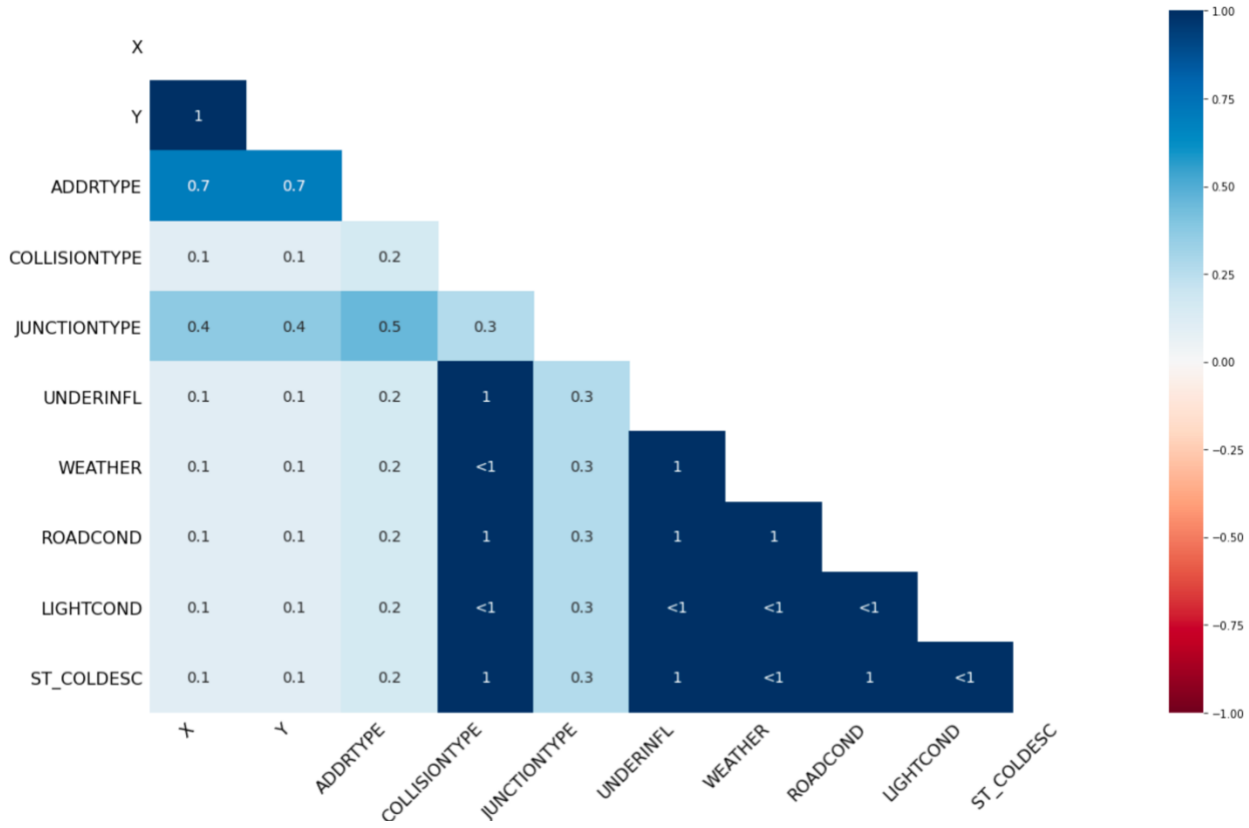

Figure 2.2.5-2 Missing Value Heat Map. Visualization of missing values across all variables in the dataset.

Another strong positive correlation is ADDRTYPE with the X and Y variables (Latitude and Longitude respectively) for about 70% of the rows.

After identifying which variables have missing values it needs to be determined if the variable should be dropped completely, remove the incidents that are empty, or to save for possible upscaling.

For X and Y, approximately 3 percent of the dataset had no values, so they were dropped. This also correlated directly with the missing data in ADDRTYPE, such that there are no longer any missing values.

2.2.6 Variable Reformat and Category Merging

UNDERINFL (under the influence) is an example of a variable that originated from multiple sources. This is a binary variable that has two different binary designations. First is 'N' or 'Y'. The second is '0' or '1'. The '0' corresponds to 'N' and '1' is 'Y'. Therefore, replace all '0' values with 'N' and '1' with 'Y'.

INCDTTM (date and time). There will be many factors to analyze with regards to the date and time of events. Therefore, the INCDTTM timestamp were parsed into new fields year, month, day, hour, minute and weekday.

Merging categories within some variables were necessary to simplify the information for the models. For example, WEATHER has 10 categories (including 'Other' and 'Unknown'). Combining similar weather conditions, the number of categories was reduced to six.

Clear	112508
Unknown	38212
Raining	32898
Overcast	27962
Snowing	906
Other	801
Fog/Smog/Smoke	561
Sleet/Hail/Freezing Rain	115
Blowing Sand/Dirt	50
Severe Crosswind	25
Partly Cloudy	10
Blowing Snow	1

Table 2.2.6-1. WEATHER before combining similar categories.

Clear or Partly Cloudy	112518
Unknown	39013
Raining	32898
Overcast	27962
Snowing	906
Severe Conditions	75

Table 2.2.6-2. WEATHER after combining similar categories.

This was also performed on variables ROADCOND, LIGHTCOND, and SEVERITYDESC

2.3 Exploratory Data Analysis

At this stage many variables have been identified as not required for the model, missing values resolved, reformatted, and some had their categories simplified. Now, with the SEVERITYDESC as the dependent variable, it is

compared to each independent variable to identify relationships that would be deemed beneficial for the models.

2.3.1 Map View

Begin analysis with a geographic distribution of the severe incidents compared to the not-severe incidents. This shows the density of severe incidents are concentrated in the downtown region, whereas, the distribution of non-severe incidents, is spread-out the Seattle metropolitan area.

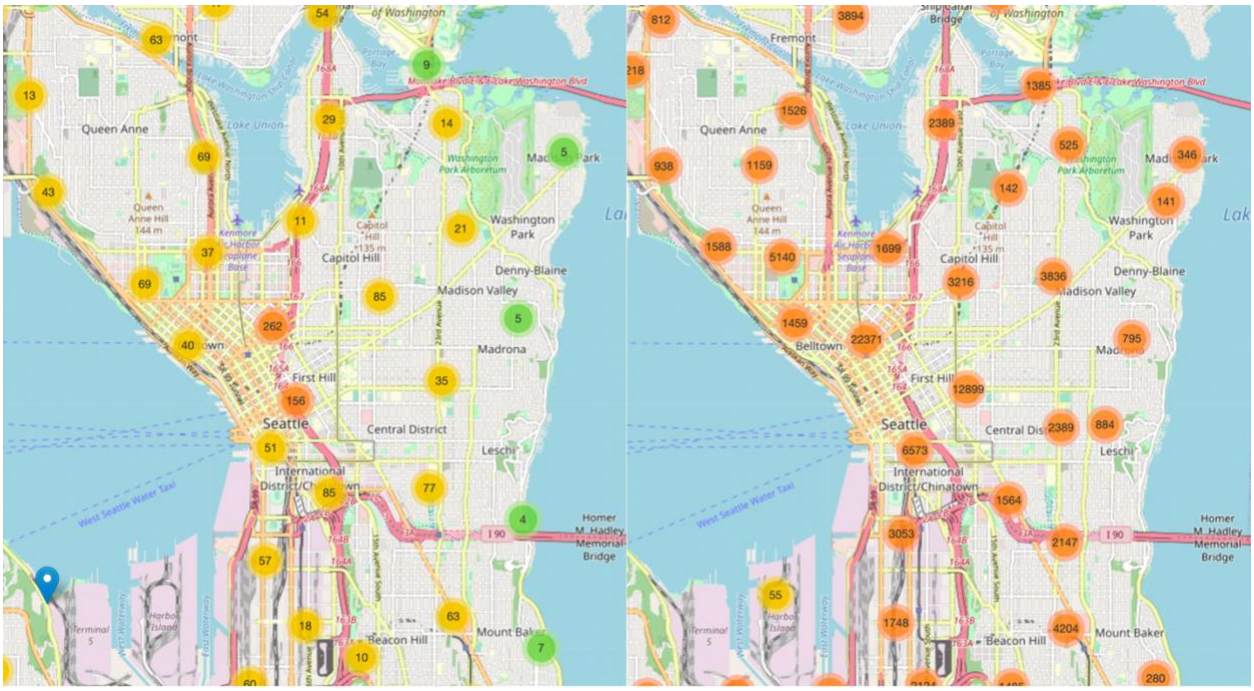

Figure 2.3.1-1. Severity map. Severe incidents (left) and non-severe incidents(right).

2.3.2 Weather

Collisions dominate during 'Clear or Partly Cloudy' conditions for both severe and not-severe incidents. The remaining conditions remain similar for both incidents.

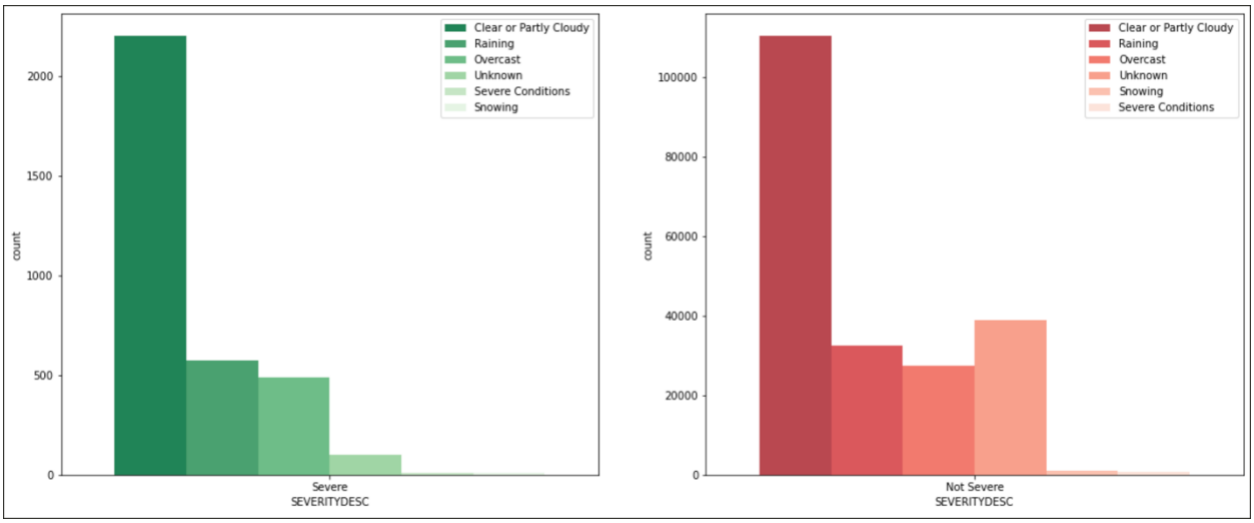

Figure 2.3.2-1 Weather conditions vs number of collisions.

2.3.3 Road Conditions

Collisions dominate for 'Dry' conditions for both severe and not-severe incidents. The remaining conditions remain similar for both incidents. This corresponds with the weather variable.

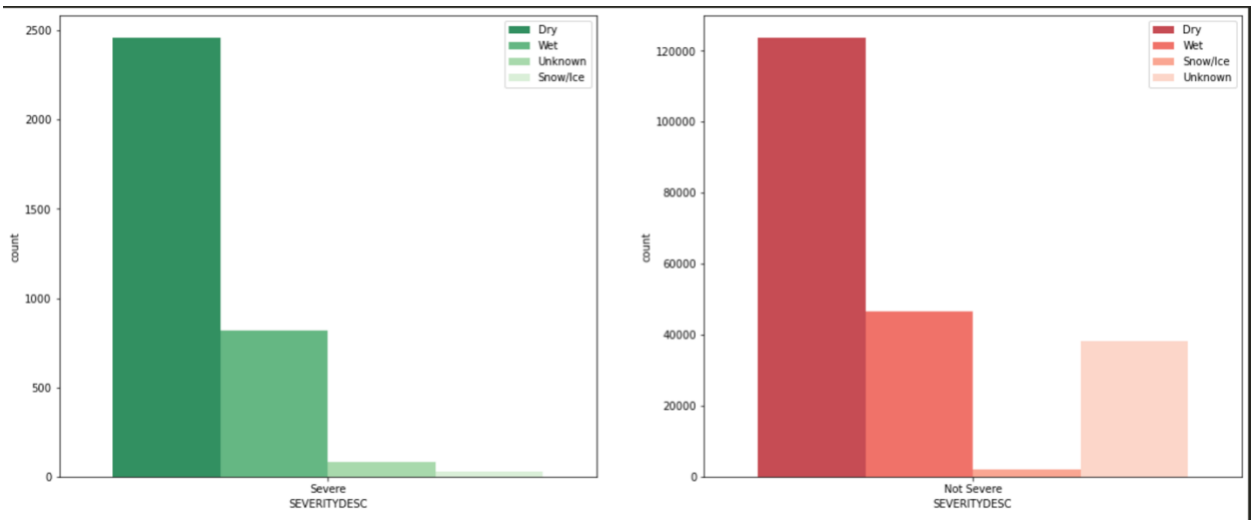

Figure 2.3.3-1 Road conditions vs number of collisions.

2.3.4 Light Conditions

Collisions dominate during ‘Daylight’ for both severe and not-severe incidents. The remaining conditions remain similar for both incidents.

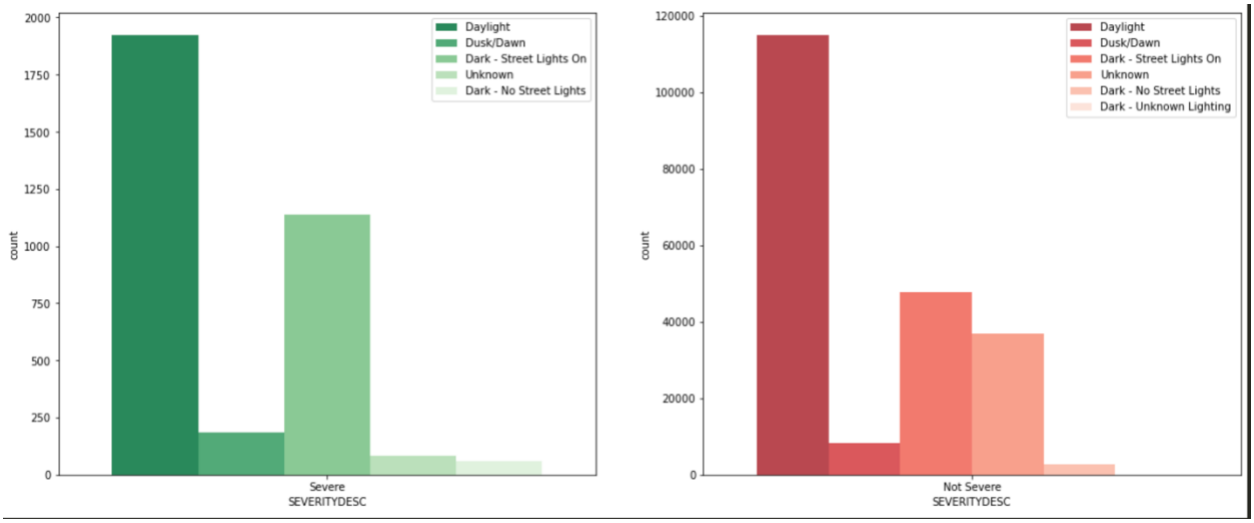

Figure 2.3.4-1 Light conditions vs number of collisions.

2.3.5 Annual

There is a decrease the number of collisions from 2005 to 2010 and trends slightly higher before decreasing again at 2015. The general trend from 2004 to 2020 is decreasing.

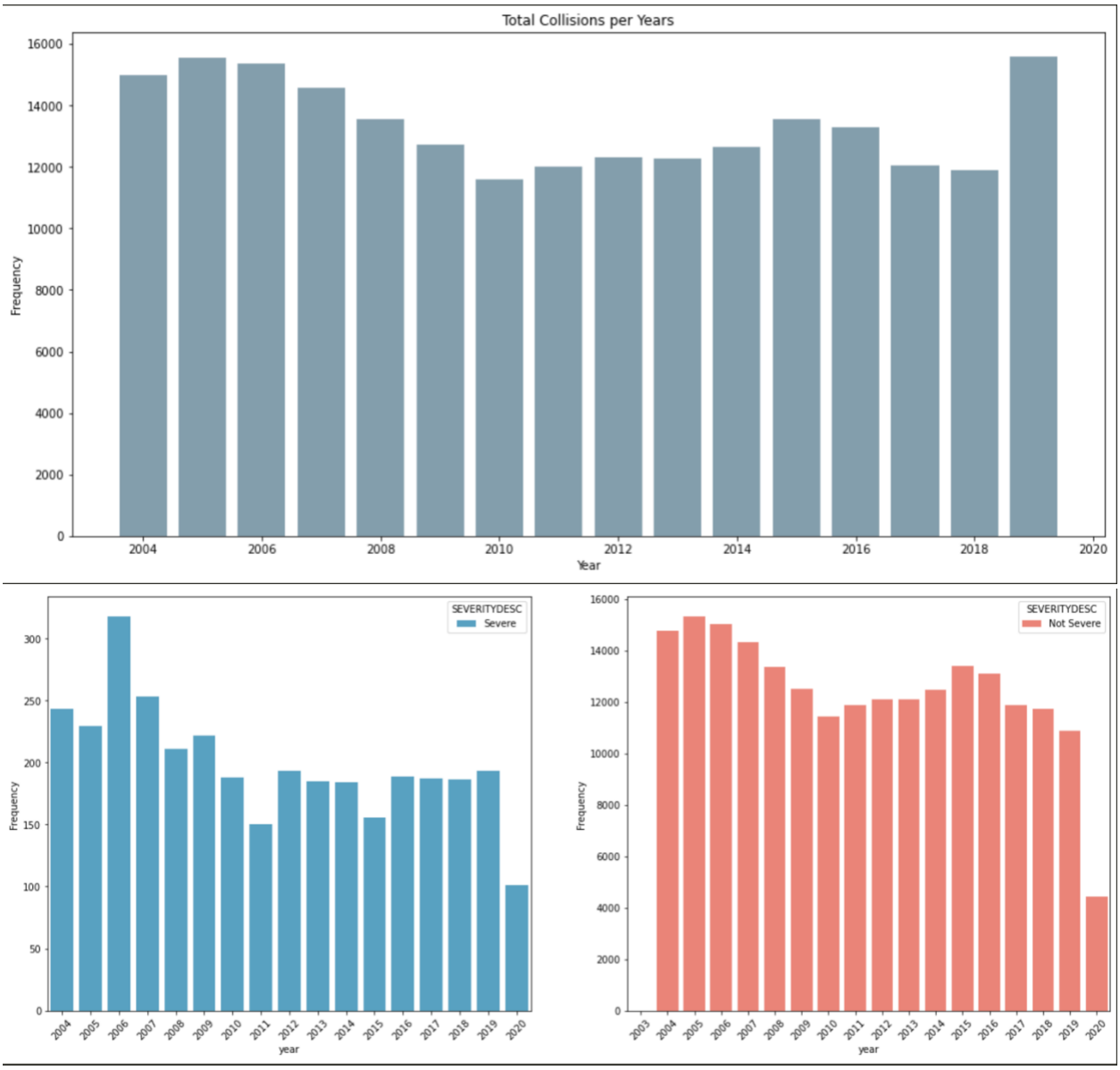

Figure 2.3.5-1 Annual collisions. Total (top); Severe (bot. left); Not-severe (bot. right).

2.3.6 Monthly

There is a peak of severe incidents that occur during the summer months of July – August.

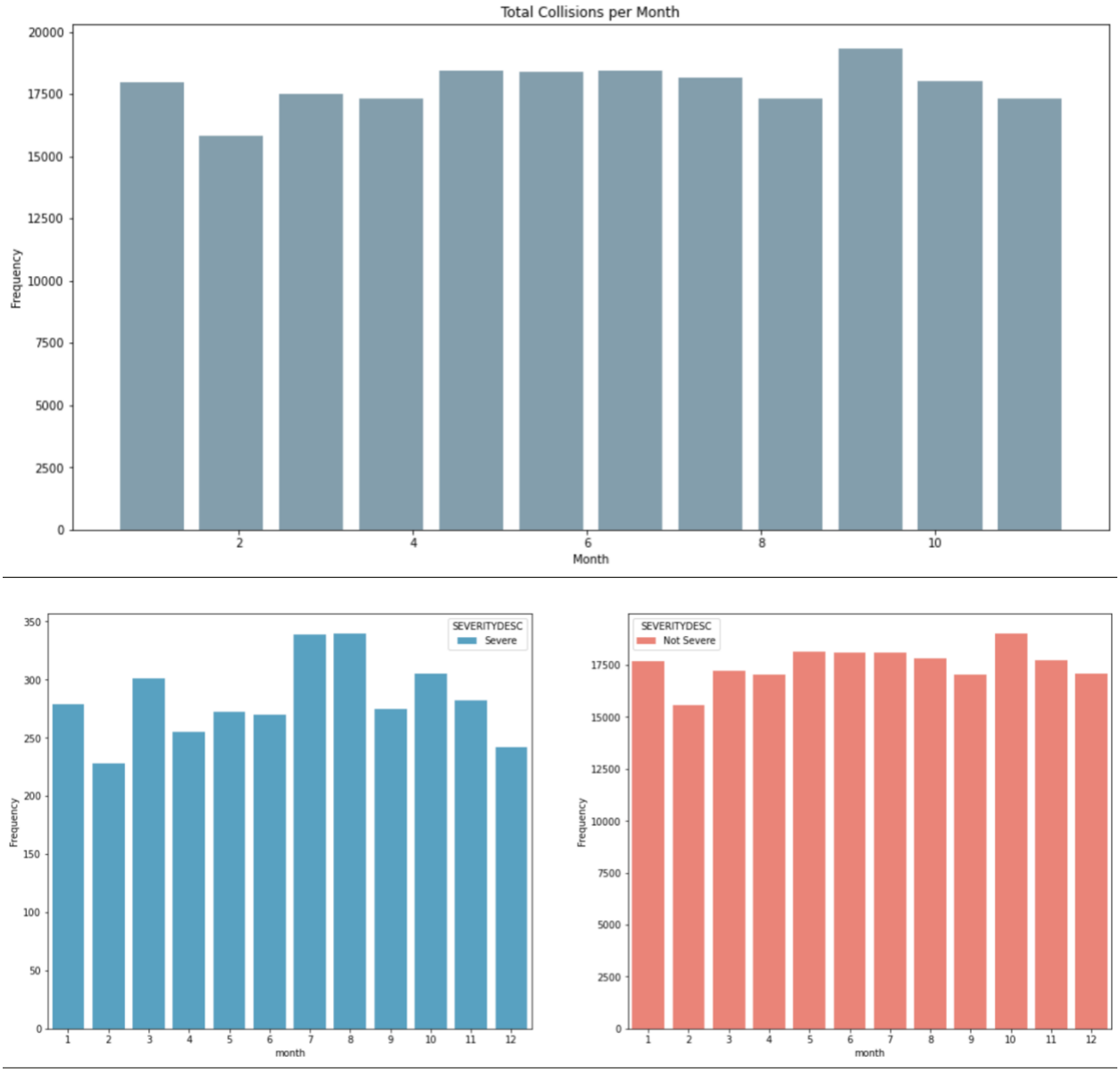

Figure 2.3.6-1 Monthly collisions. Total (top); Severe (bot. left); Not-severe (bot. right).

2.3.7 Hour

Collisions correspond with rush-hour peaking at 5pm and minimum during early morning hours from 3 to 5.

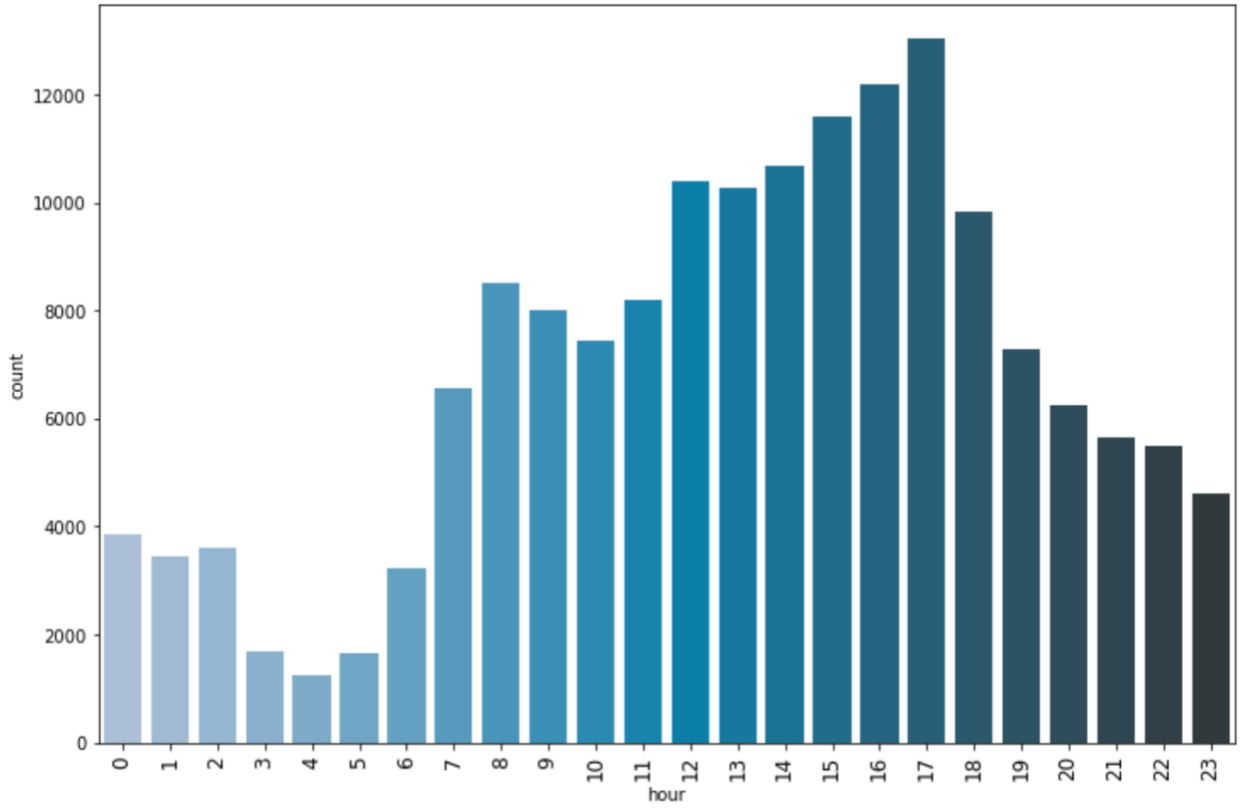

Figure 2.3.7-1 Monthly collisions. Total (top); Severe (bot. left); Not-severe (bot. right).

2.4 Modeling

2.4.1 Drop Unnecessary Variables

The critical variables remaining that correspond to having a correlation with the number of severe collisions are as follows:

- Address Type
- Junction Type
- Weather
- Road Conditions
- Light Conditions
- Month
- Year
- Hour
- Weekday

2.4.2 Convert Catagorical data types

The machine learning models require numerical values. Therefore, it was necessary to convert the catagorical variables to integers with the LabelEncoder Python function.

- Address Type
- Junction Type
- Weather
- Road Conditions
- Light Conditions
- Severity Description

	ADDRTYPE	SEVERITYDESC	JUNCTIONTYPE	WEATHER	ROADCOND	LIGHTCOND	month	hour	weekday
0	1	0	1	0	0	3	1	9	6
1	0	0	4	2	3	4	4	18	0
2	0	0	4	0	0	1	3	2	6
3	1	0	1	2	3	1	1	17	0
4	0	0	4	0	1	1	12	19	4

Table 2.4.2-1 Variables after conversion to integer values.

Our data is now ready to be fed into machine learning models.

We will use the following models:

- Decision Tree
- K-Nearest Neighbor (KNN)
- Logistic Regression

2.4.3 Train Test Split

There now consists of two datasets: explanatory variables (X) and the target variable (y). The training dataset will consist of 70% of the original data. Then then the remaining 30% will be the test dataset.

2.4.4 Oversampling

The dependent variable SEVERITYDESC is highly skewed, with the Severe category having approximately 1.6% of the data, and the remaining data in the Not Severe category. To overcome this disparity, the Severe category will be oversampled with the algorithm SMOTE (Synthetic Minority Oversampling TEchnique).

2.4.5 Decision Tree

Decision Tree Classification model uses the Decision Tree Classifier from the scikit-learn library. The criterion chosen for the classifier was 'entropy' and the max depth was '6'.

	precision	recall	f1-score	support
0	0.99	0.65	0.78	63157
1	0.03	0.54	0.05	1058
accuracy			0.65	64215
macro avg	0.51	0.60	0.42	64215
weighted avg	0.97	0.65	0.77	64215

Table 2.4.5-1 Decision Tree classification report

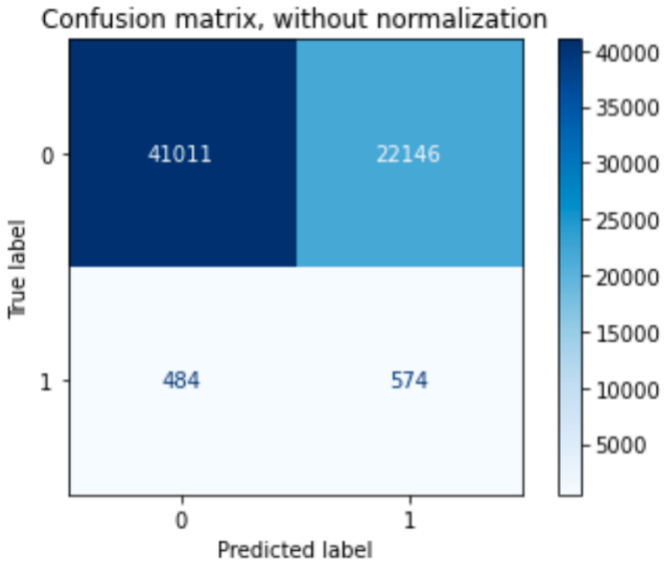

Table 2.4.5-2 Decision Tree confusion matrix

2.4.6 K-Nearest Neighbor (KNN)

K-Nearest Neighbor Classification Model uses the K-Nearest Neighbor classifier from the scikit-learn. The optimal K for the model exists is at 2.

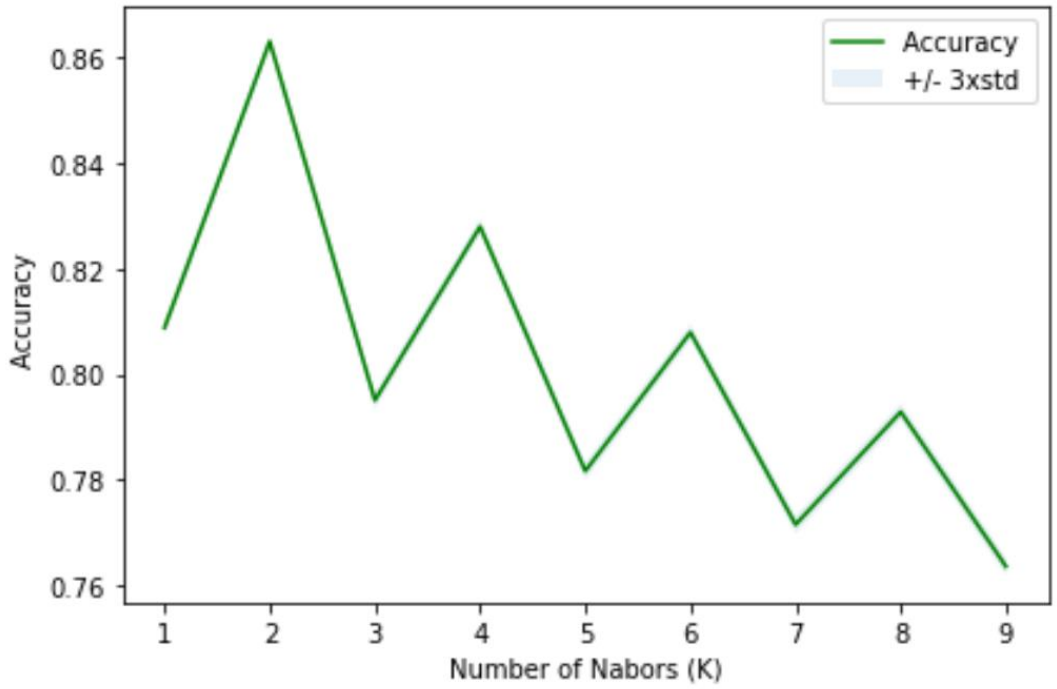

Table 2.4.6-1 Test to determine optimal k-value

	precision	recall	f1-score	support
0	0.98	0.87	0.93	63157
1	0.02	0.17	0.04	1058
accuracy			0.86	64215
macro avg	0.50	0.52	0.48	64215
weighted avg	0.97	0.86	0.91	64215

Table 2.4.6-2 k-Neural Network classification report

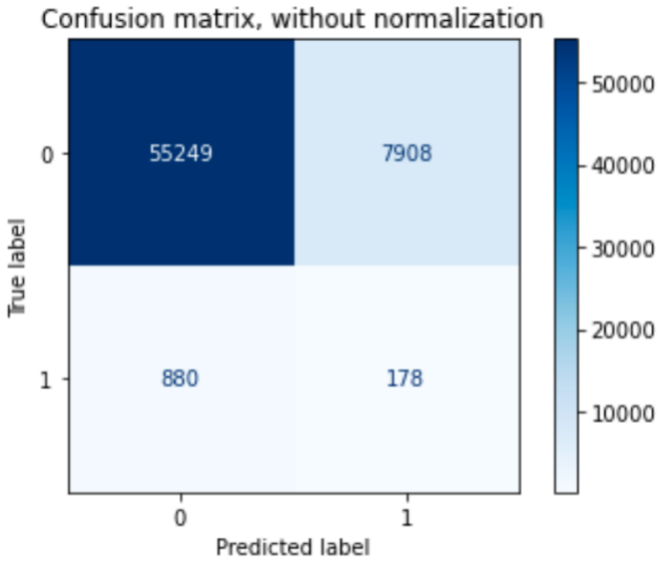

Table 2.4.6-3 k-Neural Network confusion matrix

2.4.7 Logistic Regression

The Logistic Regression Classification Model uses the Logistical Regression classifier from the scikit-learn. The C used for regularization strength was '0.01' and the solver used was 'liblinear'.

	precision	recall	f1-score	support
0	0.99	0.56	0.71	63157
1	0.02	0.62	0.04	1058
accuracy			0.56	64215
macro avg	0.51	0.59	0.38	64215
weighted avg	0.97	0.56	0.70	64215

Table 2.4.7-1 Logistic Regression classification report

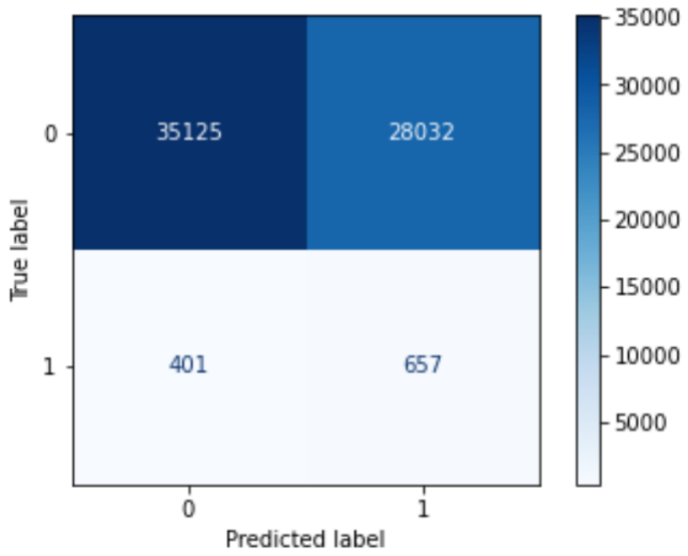

Table 2.4.7-2 Logistic Regression confusion matrix

2.4.8 Model Performance

Classifier	Precision	Recall	Avg. F-1 Score
Decision Tree	0.51	0.60	0.42
k-Nearest Neigbor	0.50	0.52	0.48
Logistic Regression	0.51	0.59	0.38

Precision

Precision refers to the percentage of results which are relevant, in simpler terms it can be seen as how many of the selected items from the model are relevant. Mathematically, it is calculated by dividing true positives by true positive and false positive.

Recall

Recall refers to the percentage of total relevant results correctly classified by the algorithm. In simpler terms, it tells how many relevant items were selected. It is calculated by dividing true positives by true positive and false negative

F1-score

F1-score is a measure of accuracy of the model, which is the harmonic mean of the model's precision and recall. Perfect precision and recall are shown by the f1-score as 1, which is the highest value for the f1-score, whereas the lowest possible value is 0, which means that either the precision or the recall is 0.

2.4.9 Conclusion

When comparing all the models by their f1-scores, Precision and Recall, we can have a clearer picture in terms of the accuracy of the three models individually as a whole and how well they perform for each output of the target variable. When comparing these scores, we can see that the f1-score is highest for k-Nearest Neighbor at 0.48. However, later when we compare the precision and recall for each of the model, we can see that the k-Nearest Neighbor model performs poorly in the precision of 1 at 0.02. The variance is too high for the model to be selected as a viable option. When looking at the other two models, we can see that the Decision Tree has a more balanced precision for 0 and 1. Whereas, the Logistic Regression is more balanced when it comes to recall of 0 and 1. Furthermore, the average f1-score of the two models are fairly close but for the k-Nearest Neighbor is higher by 0.06. It can be concluded that the both the models can be used side by side for the best performance.

2.4.10 Recommendation

After assessing the data and the output of the Machine Learning models, a few recommendations can be made for the stakeholders. The developmental body for Seattle city can assess how much of these accidents have occurred in a place where road or light conditions were not ideal for that specific area and could launch development projects for those areas where most severe accidents take place in order to minimize the effects of these two factors. Whereas, the car drivers could also use this data to assess when to take extra precautions on the road under the given circumstances of light condition, road condition and weather, in order to avoid a severe accident, if any.