IBM Data Science Capstone Project

Predicting Car Crash Severity

by

Satyabrata Choudhury

September 2020

Table of Contents

Introduction/Business Problem

Traffic collisions continue to be a serious problem. Roads safety is pressing concern for many countries, where road crash fatalities and disabilities is gradually being recognized as a major public health concern. According to World Health Organization (WHO); nearly 1.25 million people die in road crashes each year, on average 3,287 deaths a day. In addition, road traffic crashes rank as the 9th leading cause of death and account for 2.2% of all deaths globally.

Collisions are financial burden on government and society. Prediction of severity of collision helps local transport authority and emergency responders to manage traffic and avoid loss of life and property.

This project uses collision data of Seattle, WA. The aim of this project is to use data science methodology and machine learning to gain an understanding of the problem and predict the severity of collision and develop prevention mechanisms the same.

The target audience of the project is local Seattle government, police, rescue groups, and last but not the least, insurance institutes. The model and its results are going to provide some advice for the target audience to make insightful decisions for reducing the number of accidents and injuries for the city.

Data

Data Source

For this project, we used the data provided by the Washington State Dept. of Transportation (WSDOT). The dataset was hosted by Coursera as part of the Data Science Professional course. To understand the data, a supplementary metadata was also provided.

Data Cleaning

After importing dataset into Jupiter Notebook, a quick analysis showed data with missing values. From definition of the columns many of these attributes like speeding, inattention indictor could be used in prediction so it's better to clean such attributes.

The top 3 attributes can have possible values of Y & N as per the metadata. These attributes had only 'Y' in them to it was assumed that the null data was N. As prediction model use int the Y & N(null) were replaced with 1 & 0 respectively.

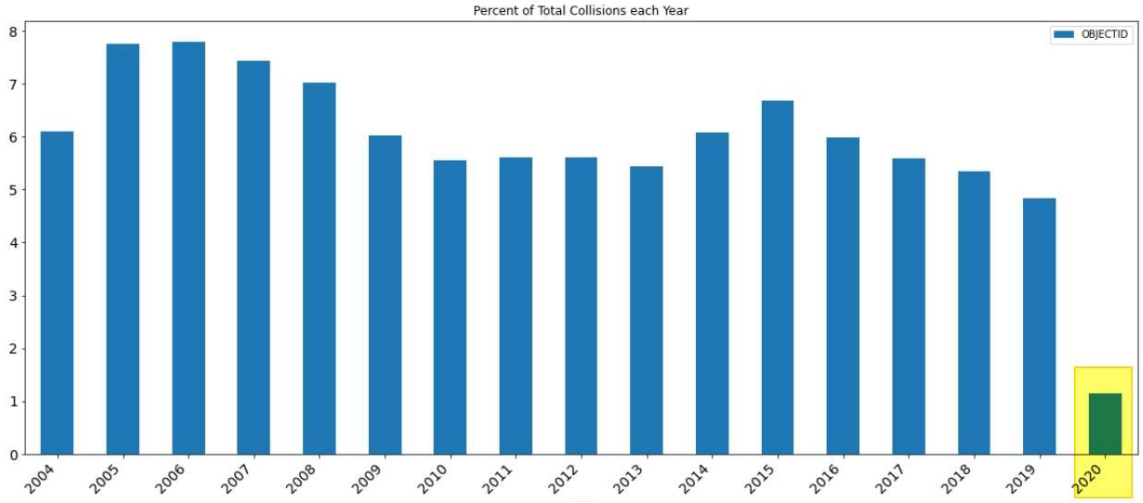
	data_type	percent_missing_values	total_unique_values
PEDROWNOTGRNT	object	97.60	1
SPEEDING	object	95.21	1
INATTENTIONIND	object	84.69	1
JUNCTIONTYPE	object	3.25	7
Y	float64	2.74	23839
X	float64	2.74	23563
LIGHTCOND	object	2.66	9
WEATHER	object	2.61	11
ROADCOND	object	2.57	9
COLLISIONTYPE	object	2.52	10
UNDERINFL	object	2.51	4
LOCATION	object	1.38	24102

Based on metadata it is quite evident that many of the attributes like UNDERINFL, HITPARKEDCAR are indicators (Y or N), a similar operation was performed on these attributes.

Column Name	Description	Values	Clean-up Action
PEDROWNOTGRNT	Whether or not the pedestrian right of way was not granted.	Y/N	Null Data was replaced with 0 and "Y" was replaced with 1. Column as type casted to int
SPEEDING	Whether or not speeding was a factor in the collision.	Y/N	Null Data was replaced with 0 and "Y" was replaced with 1. Column as type casted to int
INATTENTIONIND	Whether or not collision was due to inattention.	Y/N/0/1	Null Data was replaced with 0 and "Y" was replaced with 1. Column as type casted to int

UNDERINFL	Whether or not a driver involved was under the influence of drugs or alcohol.	Y/N/0/1	As data contained multiple parameters, Y, N, 1, 0 it was streamlined to 1 & 0
HITPARKEDCAR	Whether or not the collision involved hitting a parked car.	Y/N/0/1	Null Data was replaced with 0 and "Y" was replaced with 1. Column as type casted to int

The data spanned across from 2004 to 2020. For accurate prediction it is very necessary to remove noise from the data or inaccurate data. A simple bar chart shows data is distributed across year by percentage. It is quite evident that data from 2020 is not enough to train the model, hence data from 2020 was removed from the dataset.



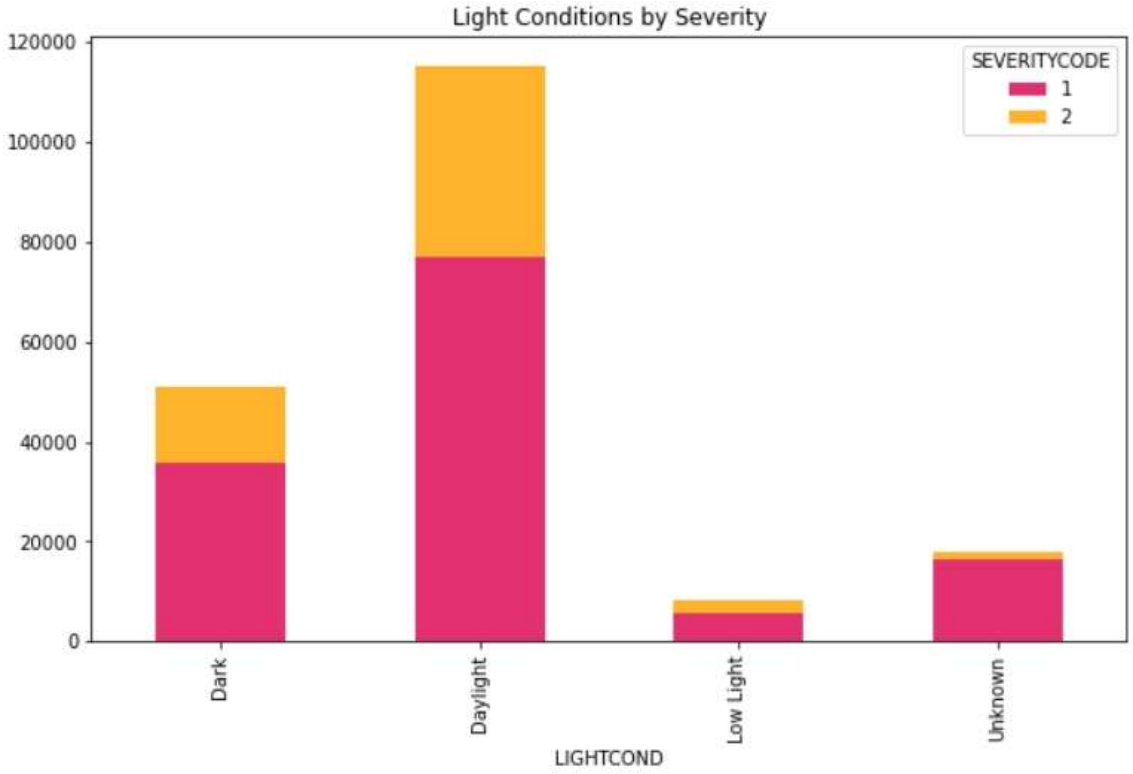
For categorical data, null values were replaced with "Unknown" or they were replaced with values matched on certain condition in data set, Other values were classified or binned to for a group. For example, LIGHTCOND, day light was unchanged however Dark Condition were replaced with "Dark" and Dawn and Dusk with "Low Light". Such classification will help in fitting the model correctly. As we have date attributes present in data set, it helps us in deriving missing records.

Light Condition (LIGHTCOND)

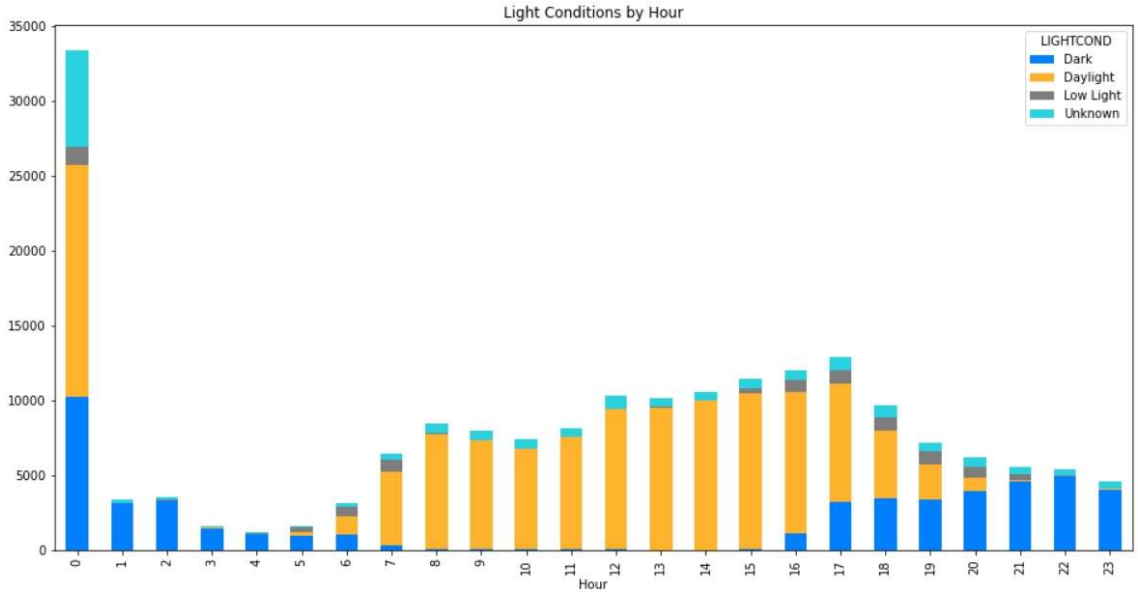
Light Condition can be effective to determine the severity of collision, it had multiple values which can be categorized into 3 main buckets, i.e. Dark, Daylight and Low Light as below.

Clean-up Action
Dark - Street Lights On -> Dark
Dark - No Street Lights -> Dark
Dark - Street Lights Off -> Dark
Dark - Unknown Lighting -> Dark
Dusk -> Low Light
Dawn -> Low Light

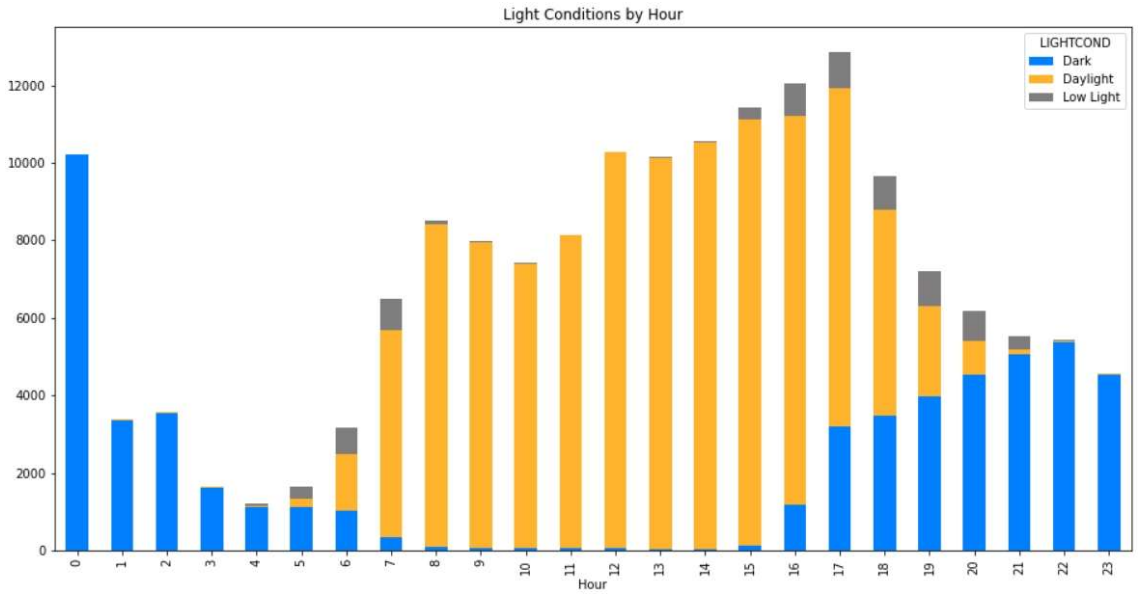
After categorization there were records with no data in light condition, bar with label as "Unknown" in figure below.



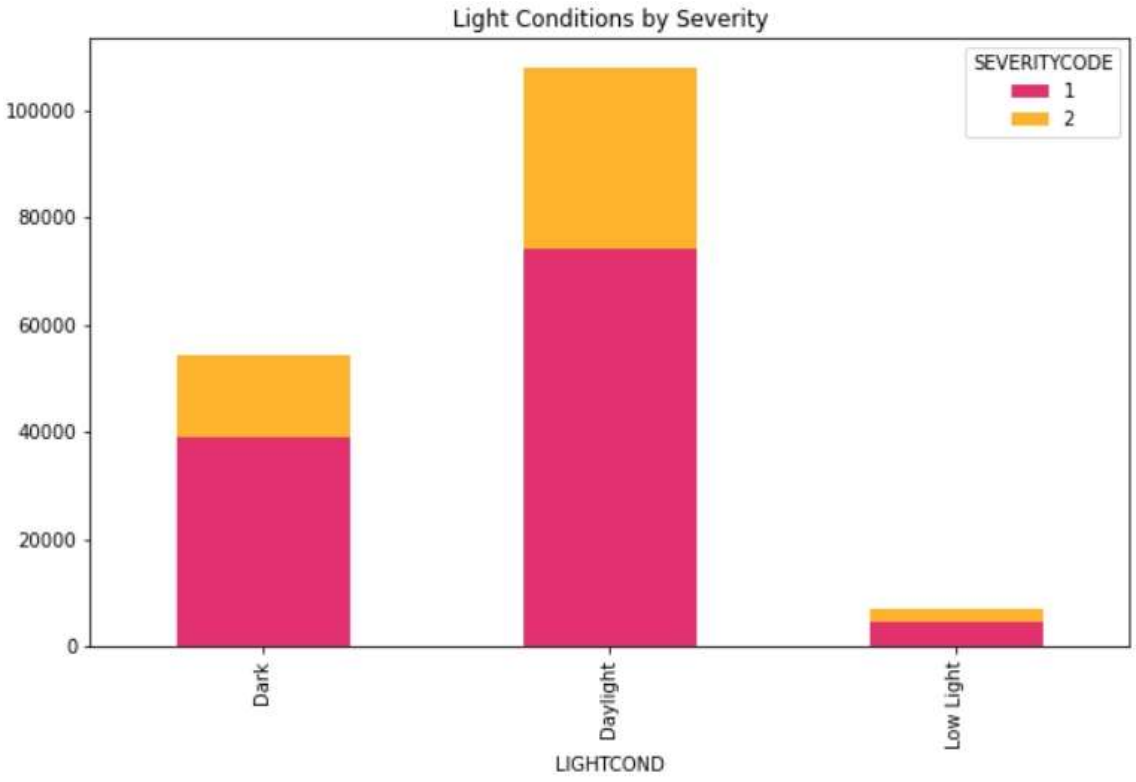
However, data has the hour attribute from INCDTTM, let's see how light condition is spanned across hour of the day. This helped in in replacing the unknown values in light condition.



From the graph it can be visualized that there are bad records in INCDDTM columns as well. As there are records with light condition as "daylight" while time is midnight. Such records were dropped from the dataset. As it clear that from 8 AM to 4 PM mostly it is daylight in Seattle, similarly from 8 PM to 5 AM it is night. Based on this observation light condition with null values were replaced with Daylight or Dark.



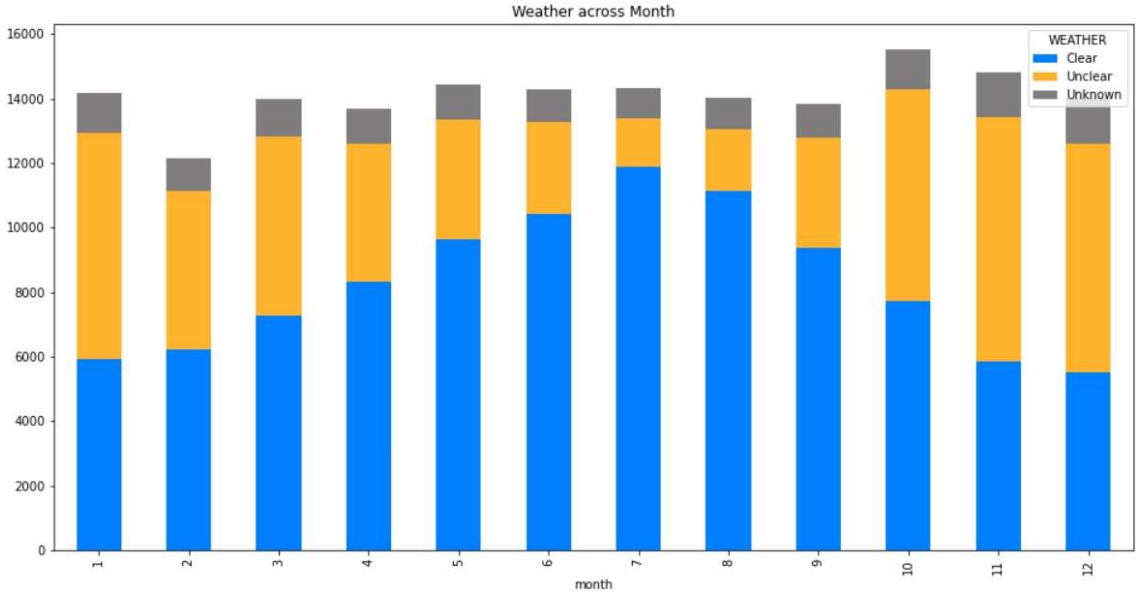
Mapping the Light Condition vs severity shows that light condition does affect the severity of a collision. This will be helpful in feature selection.



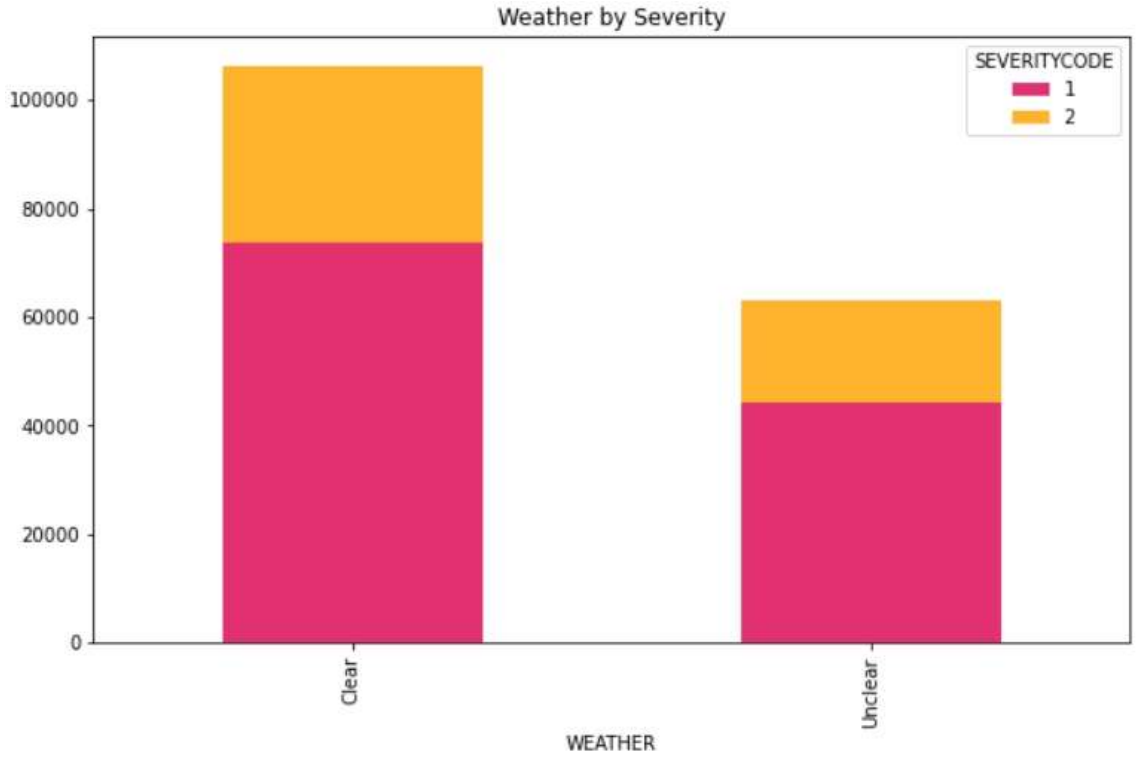
Weather (WEATHER)

Weather data had multiple attributes however certain values very less significant, to simplify the records it was classified weather into 2 attributes, clear and unclear. However, there are still null values in data.

Clean-up Action
Blowing Sand/Dirt -> Not Clear
Overcast -> Not Clear
Raining -> Not Clear
Severe Crosswind -> Not Clear
Sleet/Hail/Freezing Rain -> Not Clear
Snowing -> Not Clear
Fog/Smog/Smoke -> Not Clear



It can be deduced that weather is mostly clear in months of April to September. But weather can vary across the day and the data is spanned across 15 years, so it is wise to use INCDATE to fill missing values. It was assumed that weather was similar across Seattle on same date so if at a particular date weather is available in data, it was used to replace the null weather records. After replacing the null values, stacking severity code across weather shows that weather can used in determining severity of the collision.

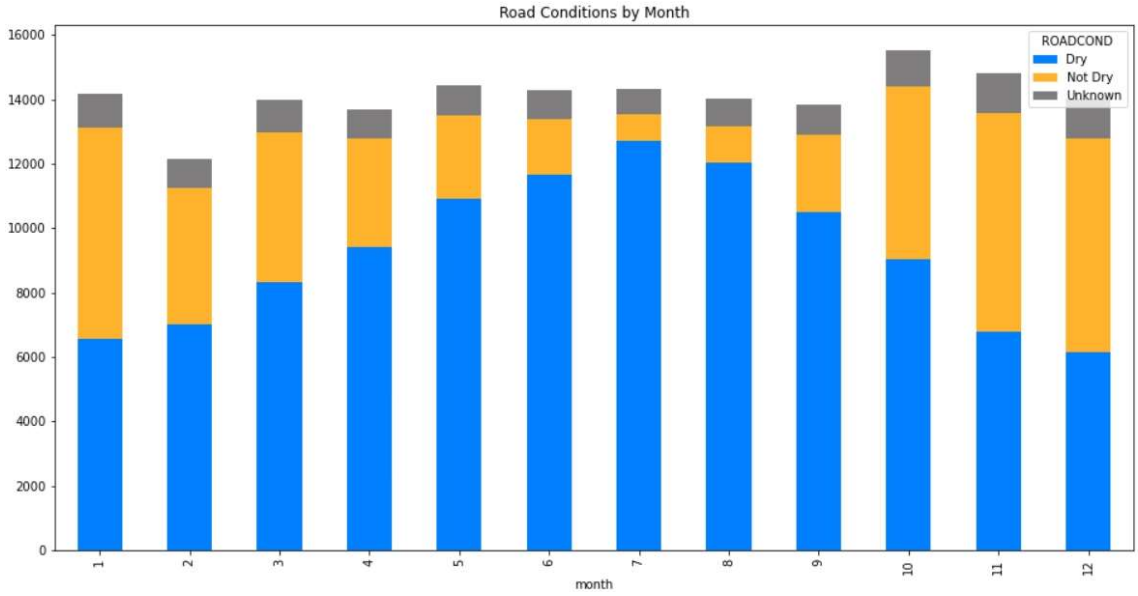


Road Condition (ROADCOND)

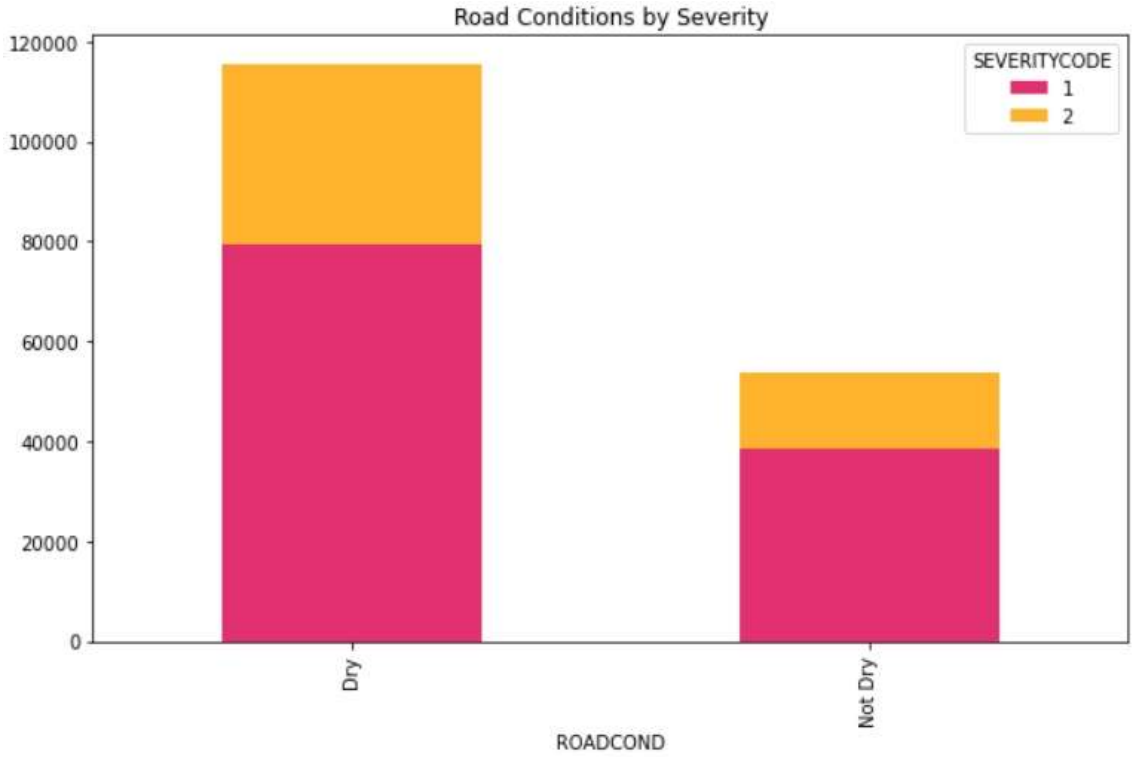
Road Condition has values which showed the condition of the road at site of the collision. Road Condition of Dry contributed more than 60% of the data, so other values were classified as "Not Dry" as below.

Clean-up Action
Wet -> Not Dry
Ice -> Not Dry
Oil -> Not Dry
Sand/Mud/Dirt -> Not Dry
Standing Water -> Not Dry
Snow/Slush -> Not Dry

Around 10% of records have missing road condition. From the values it is quite evident that many of these values depends on the weather, the relation between road condition across months can be visualized from the graph below.

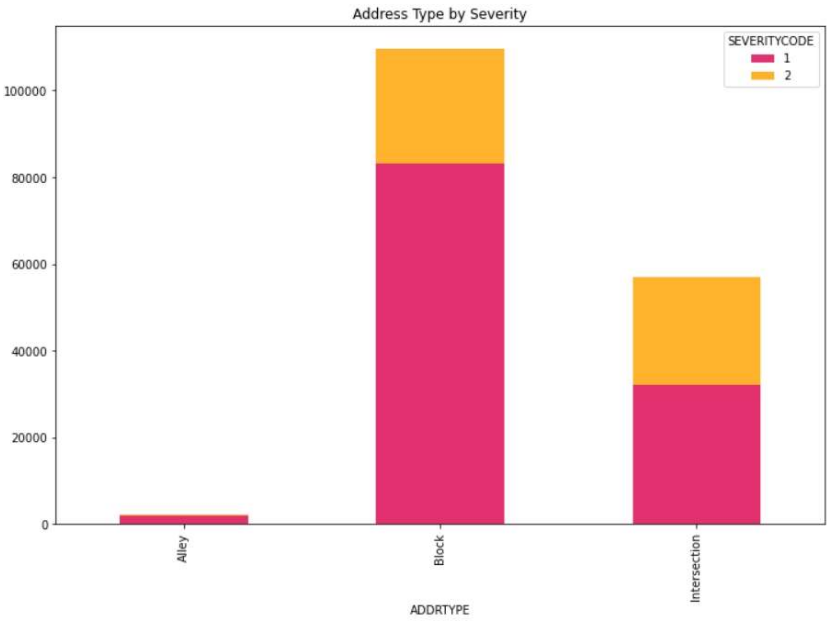


Graph shows that during summer the roads are dry compared to other months in Seattle. So, the missing values for these months was set to dry and for rest months it was replaced with not dry. After cleaning up the road condition was classified into 2 bins, this will very useful in feature selection.



Address Type (ADDRTYPE)

Address type determines the type of address of the collision site. Data has latitude, longitude and Location (street names) which determines the exact location of collision. However, ADDRTYPE is classified column so shall be useful in prediction. Null Values in ADDRTYPE columns were replaced with the values where ADDRTYPE was available at same location.

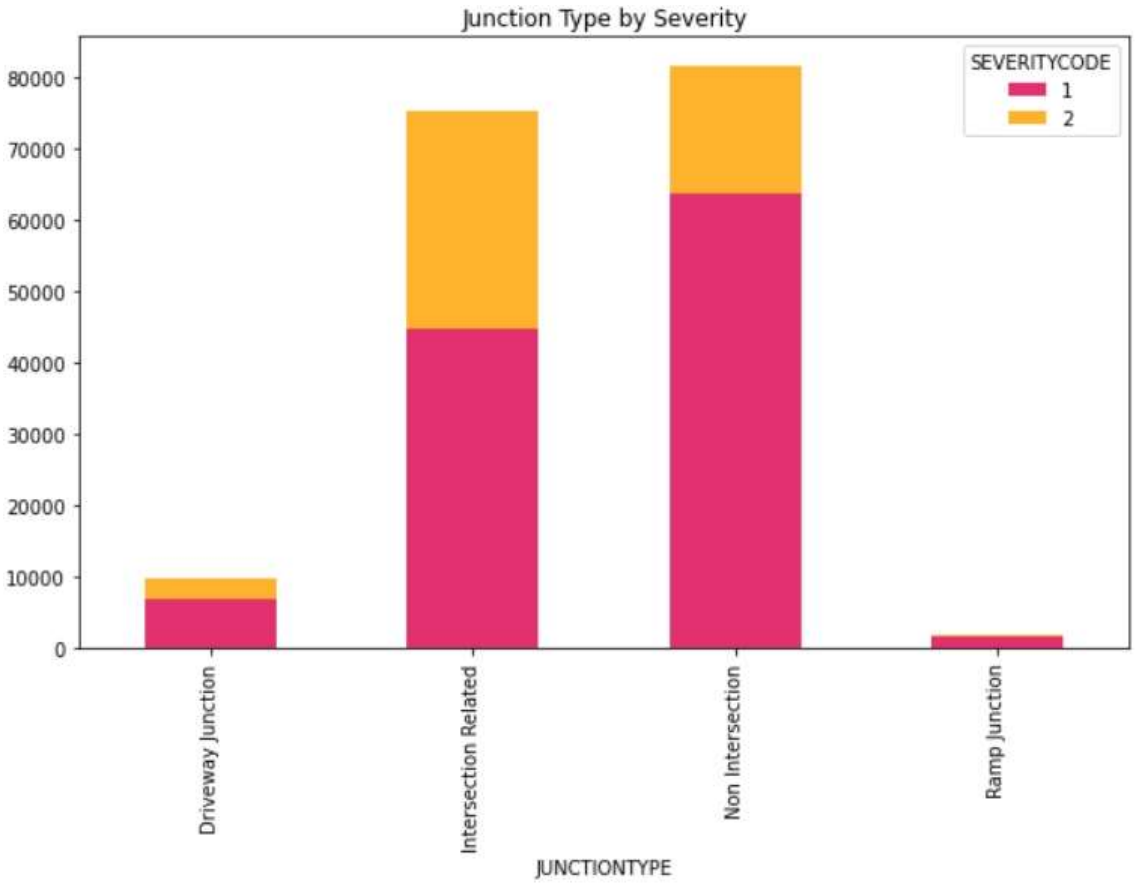


Junction Type (JUNCTIONTYPE)

Junction type determines the type of junction at collision site. The values determine whether collision was it intersection, ramp or at a block.

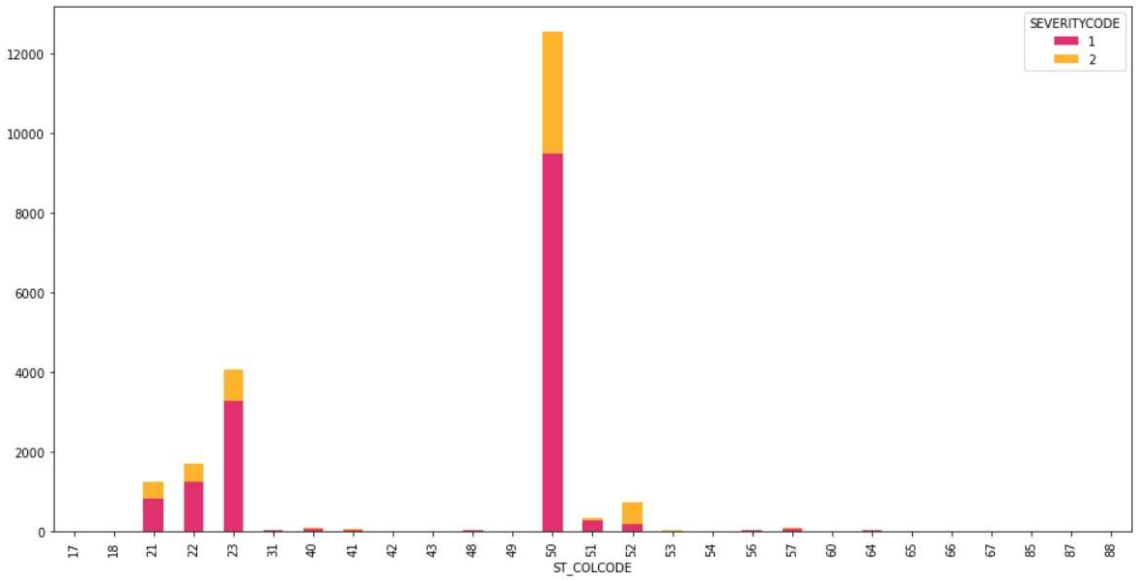
Clean-up Action
Mid-Block (not related to intersection) ->Non Intersection At Intersection (intersection related) -> Intersection Related Mid-Block (but intersection related) -> Intersection Related At Intersection (but not related to intersection) -> Non Intersection

Like ADDRTYPE, LOCATION column can be used to determine junction type if missing.

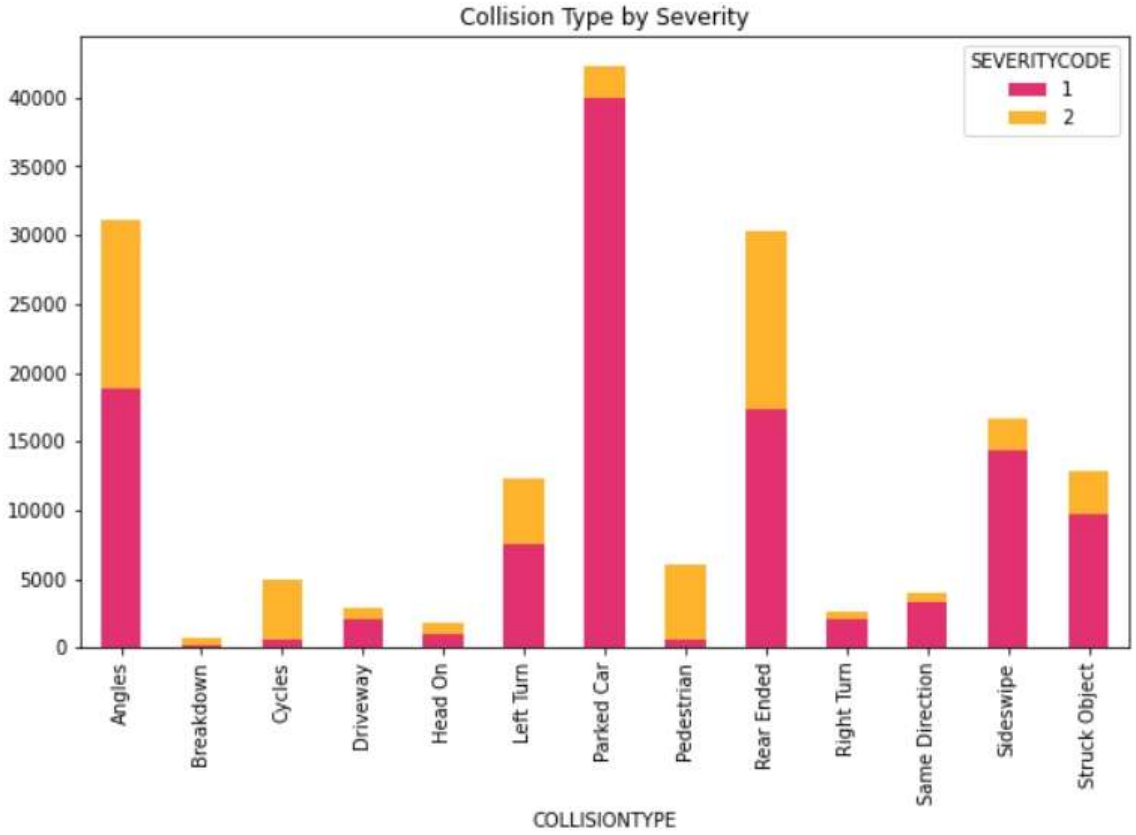


Collision Type (COLLISIONTYPE)

Collision type determines the how the accident occurred like if it was rear ended or due to pedestrians. Around 10% of the records were classified as other. We can determine ST_COLCODE to derive the collision type for missing values. Visualizing the ST_COLCODE for collision type as "other" shows that the data is significant and should be corrected.



From the graph, it was determined that ST_COLCODE with values of 21, 22, 23, 50, 51, 52 are useful. Records with Collision Types Other and values not in above list were dropped. Using the metadata, the collision type was determined.



Methodology

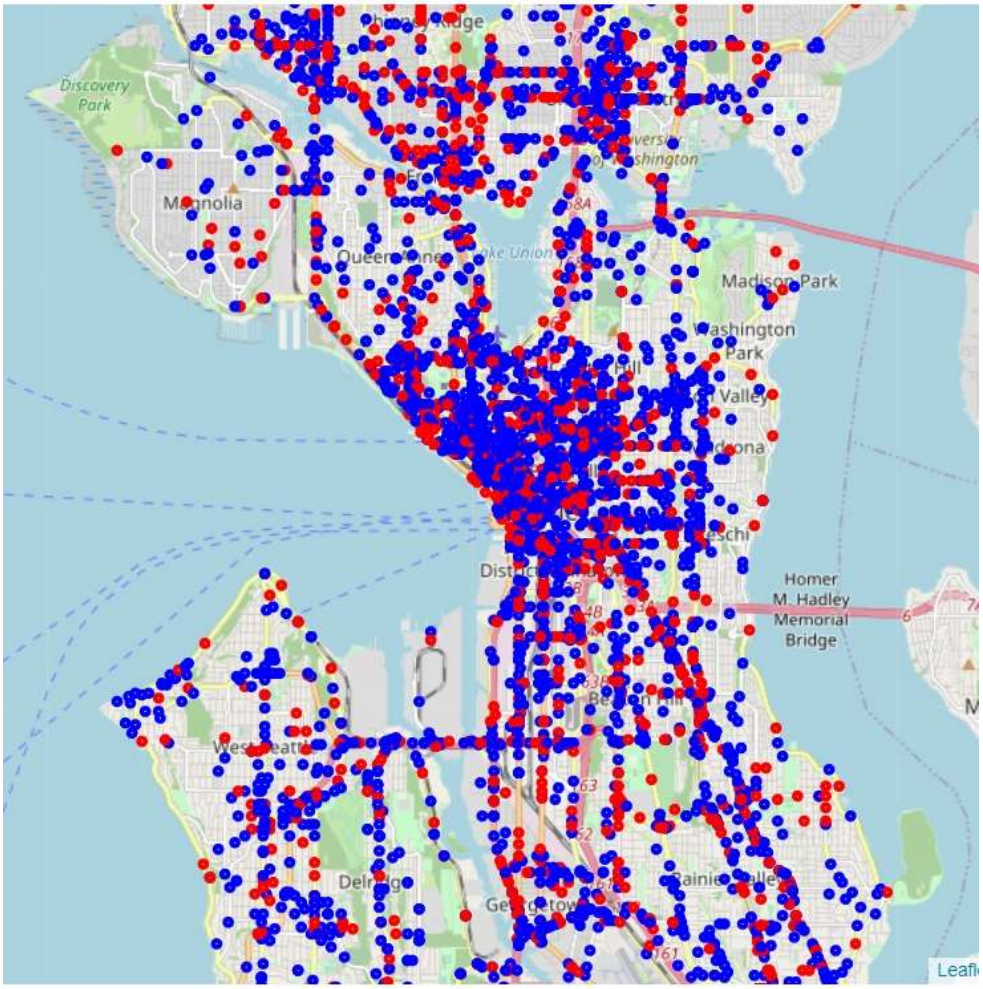
As data is cleansed, the next step is to select the attributes for model. Dataset was split into 2 types.

Numerical Features

All numerical attributes with datatype of number were selected. Year and Day were dropped from the selection same action cannot happen on same date of next year. However, day of the week, hour and month were split for feature determination.

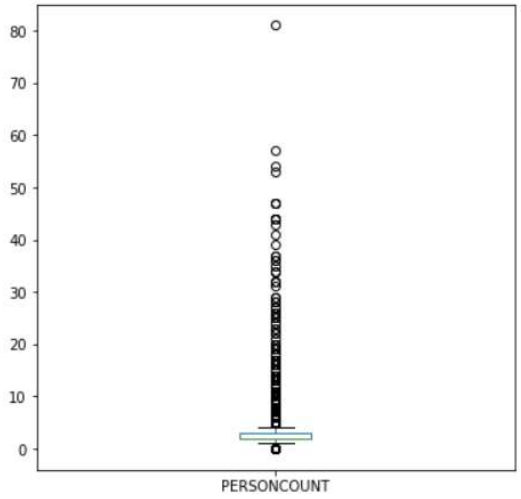
X, Y (Lat, Long)

Does location determine severity of a collision? Using folium data was mapped for months of 2016. From the map it was determined that at same location collision can be of hight or low severity. Red dot indicates high severity. As we have other attributes like address type, junction type lat long were dropped from feature.

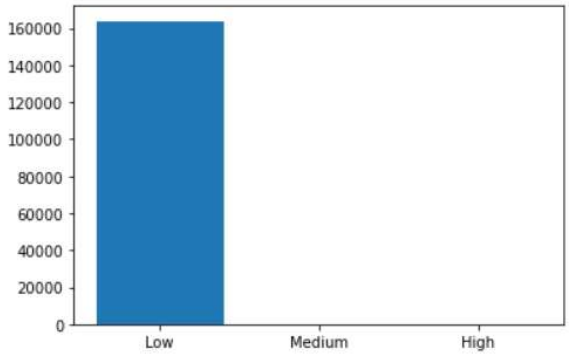


Person Count (PERSONCOUNT)

This attribute determines how many persons were involved in the collision. A box plot shows how data is distributed. There are lots of outliers and scattered across, binning the data shows that this feature can be dropped.

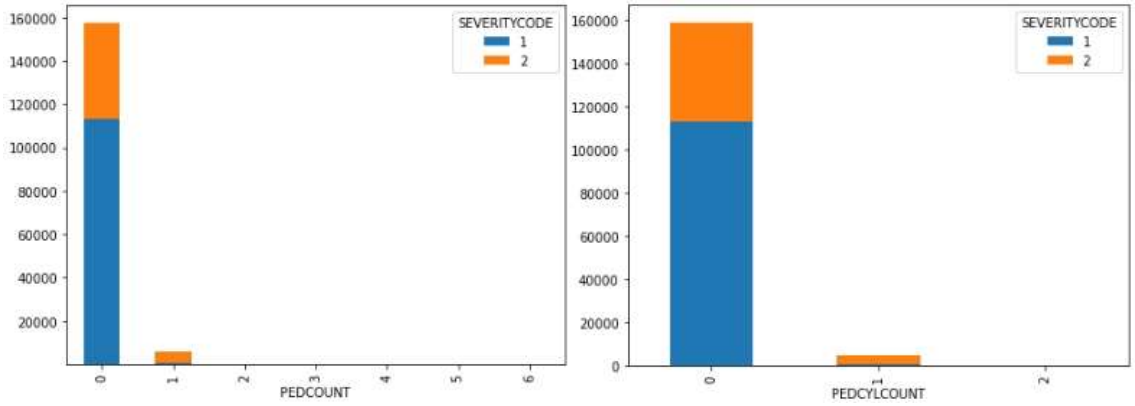


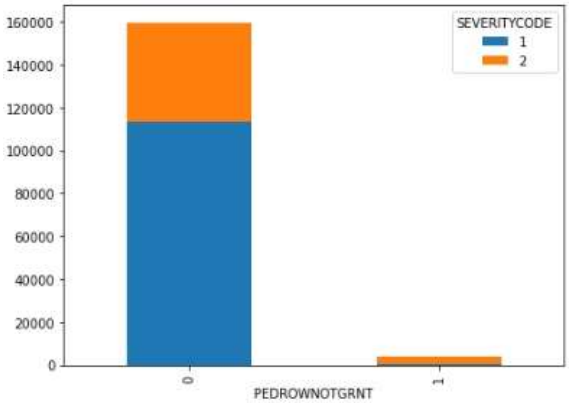
SEVERITYCODE	1	2
PERSONCOUNT-binned		
Low	113640	50272
Medium	23	5
High	1	1



PEDCOUNT, PEDROWNOTGRNT & PEDCYLCOUNT

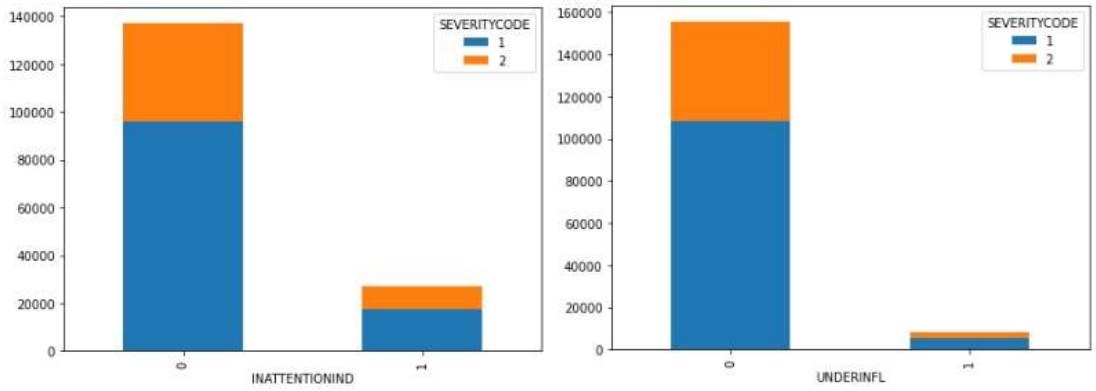
These 3 attributes determine how pedestrians impact the severity. All these three attributes show that the relationship is not quite strong.





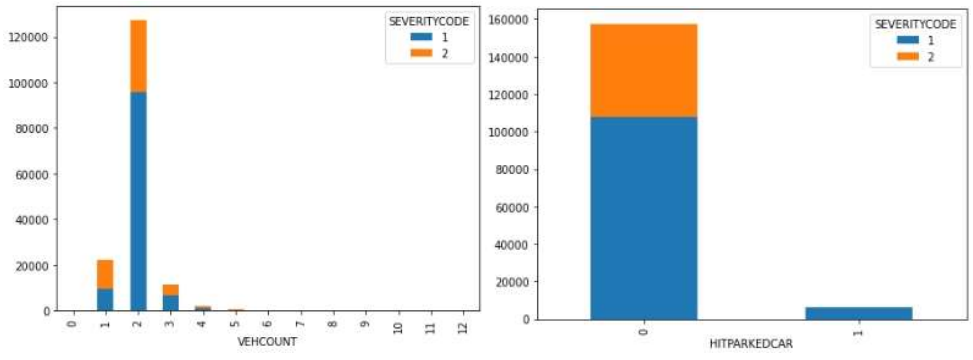
INATTENTIONIND & UNDERINFL

Does alcohol or drugs impact the severity of a collision? How about not paying attention, these both factors have strong relation to severity.



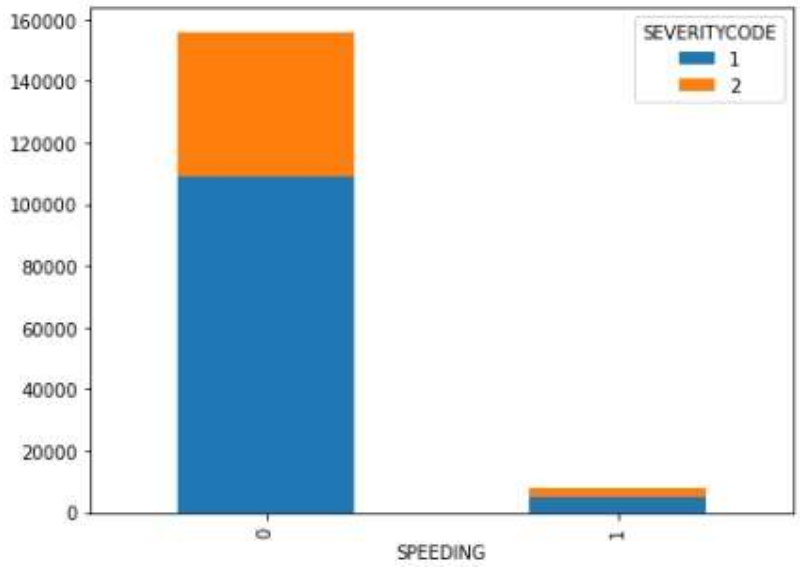
VEHCOUNT & HITPARKEDCAR

How does other vehicle impact the severity? Hitting a parked car does not impact the severity of collision but if the number of vehicles involved are more the chance of severity being high also increase.



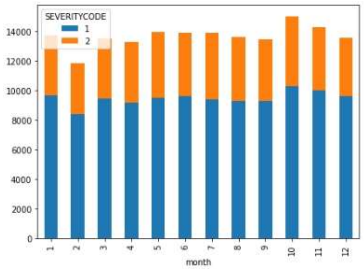
SPEEDING

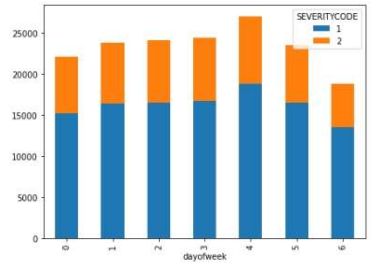
Is speeding a useful feature? Even if the vehicle was not speeding the chances of severity high. From the visualization it seems speeding does affect the severity, the collision are stills ever if vehicles are not speeding. This feature will be helpful in reducing false positives.

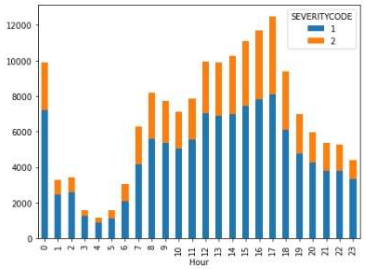


Time Attributes

How about time parameters? Clearly month is not useful as the severity ratios seems to same across months, but hour has strong relationship. The accidents occurring during day can be more severe than once in night.







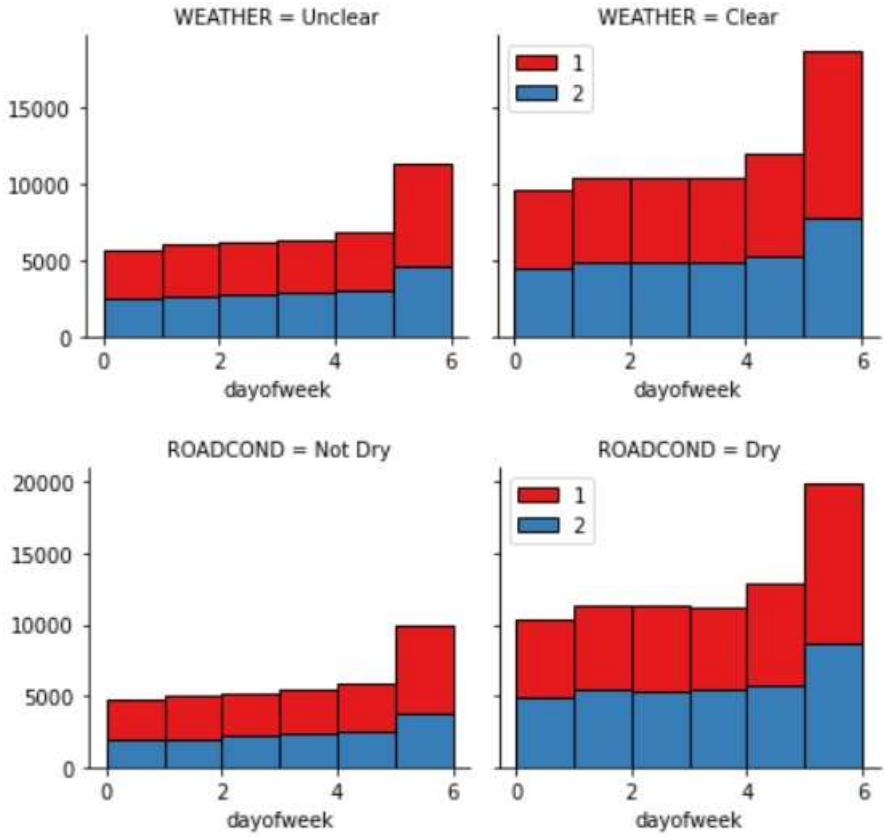
Pearson's Coefficient

After visualizing it's better to check for P-values using Pearson's coefficient to finalize the numerical features. The table below shows pearson's coefficient for various attributes. The visualization matches with P-values. HITPARKEDCAR, Month, PERSONCOUNT do not have strong relationship with target variable as there P values are 0 or >- 0.001.

Column Name	Pearson Correlation Coefficient	P-value of
SEVERITYCODE	NaN	NaN
PEDCOUNT	0.251263	0.000000e+00
PEDCYLCOUNT	0.217310	0.000000e+00
VEHCOUNT	-0.088294	5.464418e-281
INATTENTIONIND	0.040824	2.016460e-61
UNDERINFL	0.043036	4.618621e-68
PEDROWNOTGRNT	0.208156	0.000000e+00
SPEEDING	0.038203	5.226122e-54
HITPARKEDCAR	-0.106021	0.000000e+00
month	0.004751	5.437663e-02
dayofweek	-0.017850	4.910711e-13
Hour	0.024883	7.009487e-24

Categorical Features

Features with type of object were explored separately. As we saw above week is related to severity, these features were mapped against day of the week. As categorical features are visualised before the below example shows how weather and road condition are related to severity of a collision



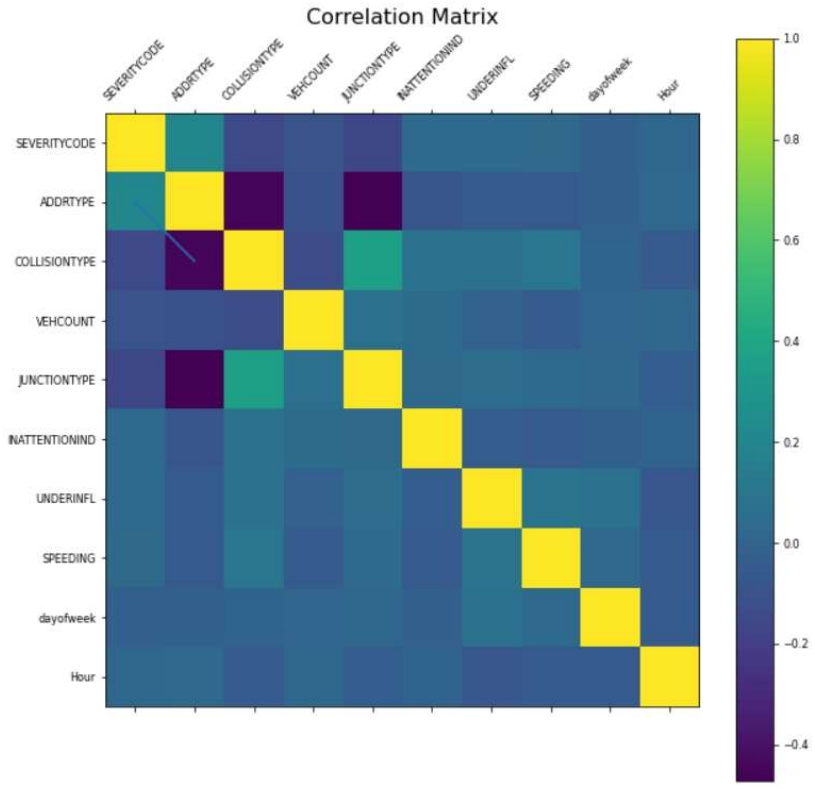
ANOVA: Analysis of Variance

The Analysis of Variance (ANOVA) is a statistical method used to test whether there are significant differences between the means of two or more groups. ANOVA returns two parameters. F-Test score: ANOVA assumes the means of all groups are the same, calculates how much the actual means deviate from the assumption, and reports it as the F-test score. A larger score means there is a larger difference between the means. P-value: P-value tells how statistically significant our calculated score is value. large F test score showing a strong correlation and a P value of almost 0 implying almost certain statistical significance.

Column Name	F Value	P value
LIGHTCOND	113.296989	6.756411e-50
ADDRTYPE	6929.297391	0.000000e+00
COLLISIONTYPE	3892.866144	0.000000e+00
JUNCTIONTYPE	2216.275852	0.000000e+00
WEATHER	14.132951	1.703928e-04
ROADCOND	119.806061	7.132669e-28

Features

Based on the P values the final set of features are: ADDRTYPE, COLLISIONTYPE, VEHCOUNT, JUNCTIONTYPE, INATTENTIONIND, UNDERINFL, SPEEDING, dayofweek, Hour. A correlation visualisation helps in understanding the relationship in better way.



Results

As the target values is Boolean, a classification model will be best fit here. 4 models were evaluated.

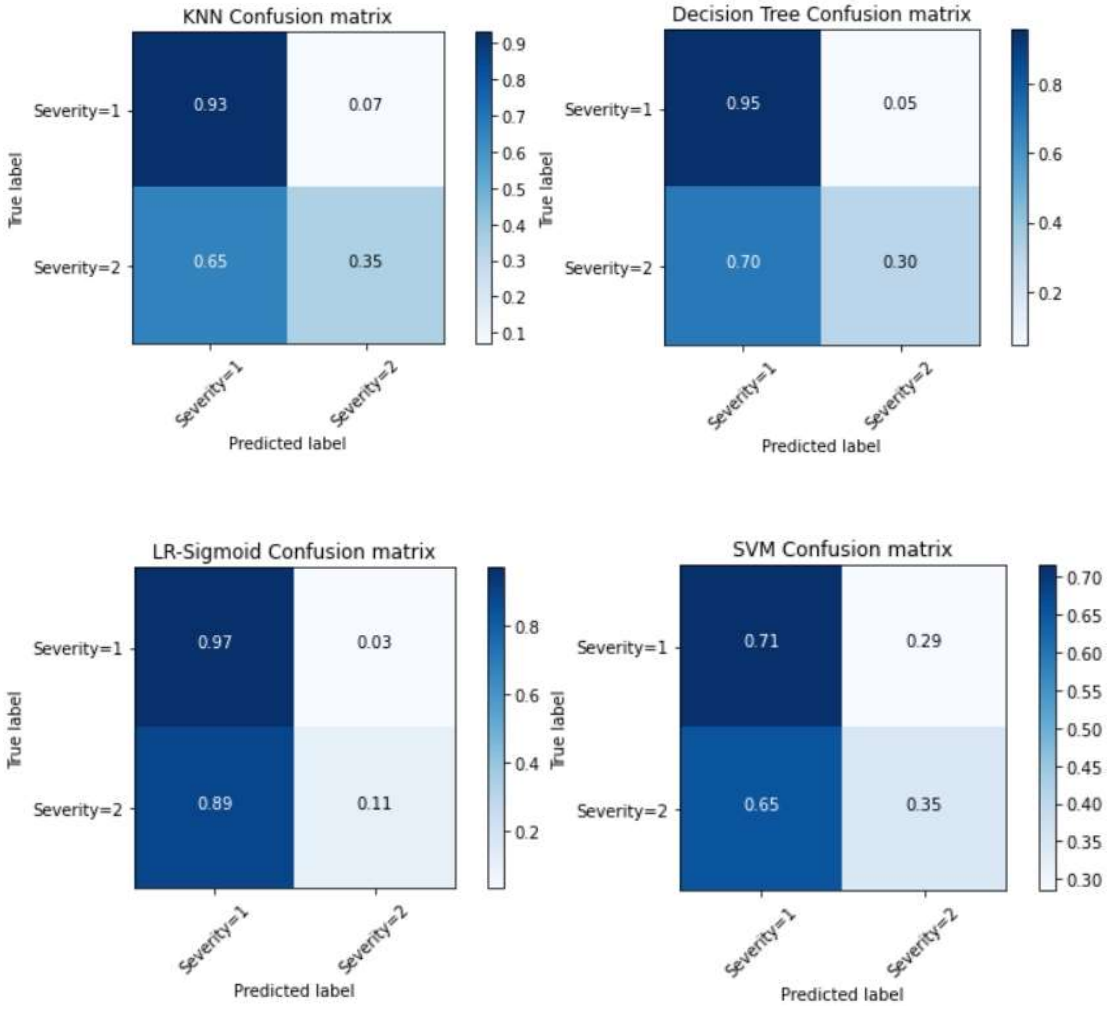
Model Training

The data set was split into test and train dataset. The train dataset was used to train 4 classification models as below:

- K Nearest Neighbour – with K as 14
- Support Vector Machine – using sigmoid function
- Logistic Regression – with solver as CG newton
- Decision Tree Matrix 0 with max depth of 6

Model Evaluation

The Model was tested with random data from original dataset. The models were evaluated based on the Jaccard Index, LogLoss, F1 Score, Precision, Recall & confusion Matrixes.



Performance of Classification Models:

Model	Jaccard index	F1-score	Log Loss	precision	recall	TN	TP	FN	FP
KNN	0.752	0.724	0.649	0.74	0.75	0.93	0.35	0.65	0.07
DT	0.753	0.715	0.487	0.75	0.75	0.95	0.3	0.7	0.05
LR	0.705	0.625	0.584	0.68	0.71	0.97	0.11	0.89	0.03
SVM	0.603	0.603	NA	0.6	0.6	0.72	0.35	0.65	0.29

Model Selection

Based on the confusion matrixes and the performance parameters, K-Nearest Neighbour will be the best fit here. In this case Higher True Positives and lower False Negatives are more important. This leads a tie between SVM and KNN model, but the precision, Jaccard Index of KNN is better than that of SVM.

Discussion

For this project, various collision parameters for a collision were used to predict the severity of collision in city of Seattle. Based on the relationship of dependent and independent variables and feature selection it is evident that collision type, number of vehicles involved affects the severity of a collision. Location does not affect the severity of the collision, but the type of location does. Contrary to the belief that severity of collision might be higher at night time the collision severity is more during the daylight.

This model should also be trained & evaluated with data from other cities. Machine learning is an iterative process which refines itself based on more training datasets. There are other classification models like gradient boost, random forest, Native Bayes which were not used for this exercise. Using the dataset on other model might yield in better performance.

Lots of missing data were filled using the patterns or occurrences. At some places generic assumptions were applied. If these models are evaluated against more accurate dataset it might yield in better accuracy. One should not hesitate in feature selection as well.

Conclusion

Based on the dataset the model selected yielded accuracy of 74%, which is good but still can be improved. Using this model one can closely predict the severity of collisions against various parameters. This model will be helpful to first responders to access the collision and direct appropriate department to reach on the collision site. The models are based on external factors however there are other feature which shall be used like age of drivers, car type, car age can change the outcome of prediction. For example, newer cars are more safe than old cars. If we find such data sets, could improve the performance of models significantly.