Introduction where you discuss the business problem and who would be interested in this project.

Reducing traffic accidents is an important public safety challenge, therefore, accident analysis and prediction has been a topic of much research over the past few decades. The problem is to identify the possibility of getting in a car accident based on several different factors as given in the shared dataset. Say you are driving to another city for work or to visit some friends. It is rainy and windy, and on the way, you come across a terrible traffic jam on the other side of the highway. Long lines of cars barely moving. As you keep driving, police car start appearing from afar shutting down the highway. Oh, it is an accident and there's a helicopter transporting the ones involved in the crash to the nearest hospital. They must be in critical condition for all of this to be happening. Now, wouldn't it be great if there is something in place that could warn you, given the weather and the road conditions about the possibility of you getting into a car accident and how severe it would be, so that you would drive more carefully or even change your travel if you are able to.

In the shared dataset, there are 37 attributes (columns). The first column tells the severity of the accident. The remaining columns have different types of attributes. Some will be used to train the model. The label for the data set is severity, which describes the fatality of an accident. Some of the shared data has unbalanced labels which may create unbiased ML model so that will require balancing the data.

Target audience

- Vehicle Drivers
- Ambulance services

The drivers would care because they would not want to get in an accident. Ambulance services can use the model to predict accident severity and be ready to respond in a timely fashion.

Data

In the shared dataset, there are 37 attributes (columns). The first column tells the severity of the accident. The remaining columns have different types of attributes. Some will be used to train the model. The label for the data set is severity, which describes the fatality of an accident. Some of the shared data has unbalanced labels which may create unbiased ML model so that will require balancing the data.

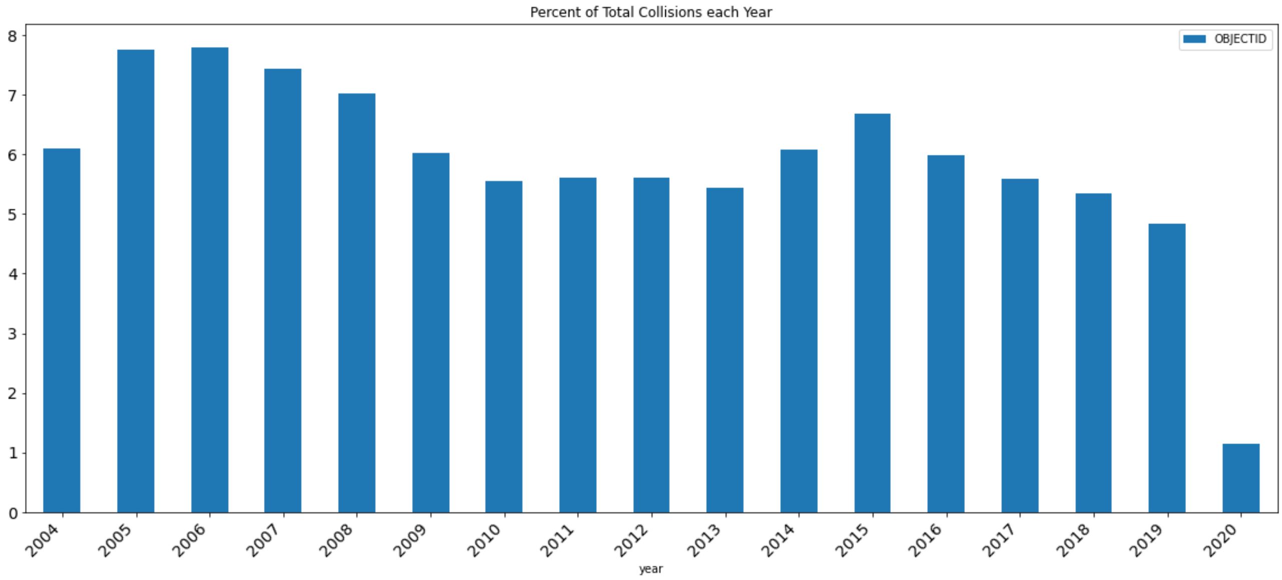
The data includes all types of collisions from 2004 to present. Collisions will display at the intersection or mid-block of a segment. Not all the attributes will be used. Only a select few attributes will be used to predict and train/test the model.

The target or label column is the ‘severity’. The remaining columns have different types of attributes e.g. collision type, person count, pedestrians involved, number of vehicles in the collision, weather, road condition, light condition, whether pedestrian had right of way etc.

3 attributes, PEDROWNNOTGRNT, SPEEDING, INATTENTIONIND could 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. 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.

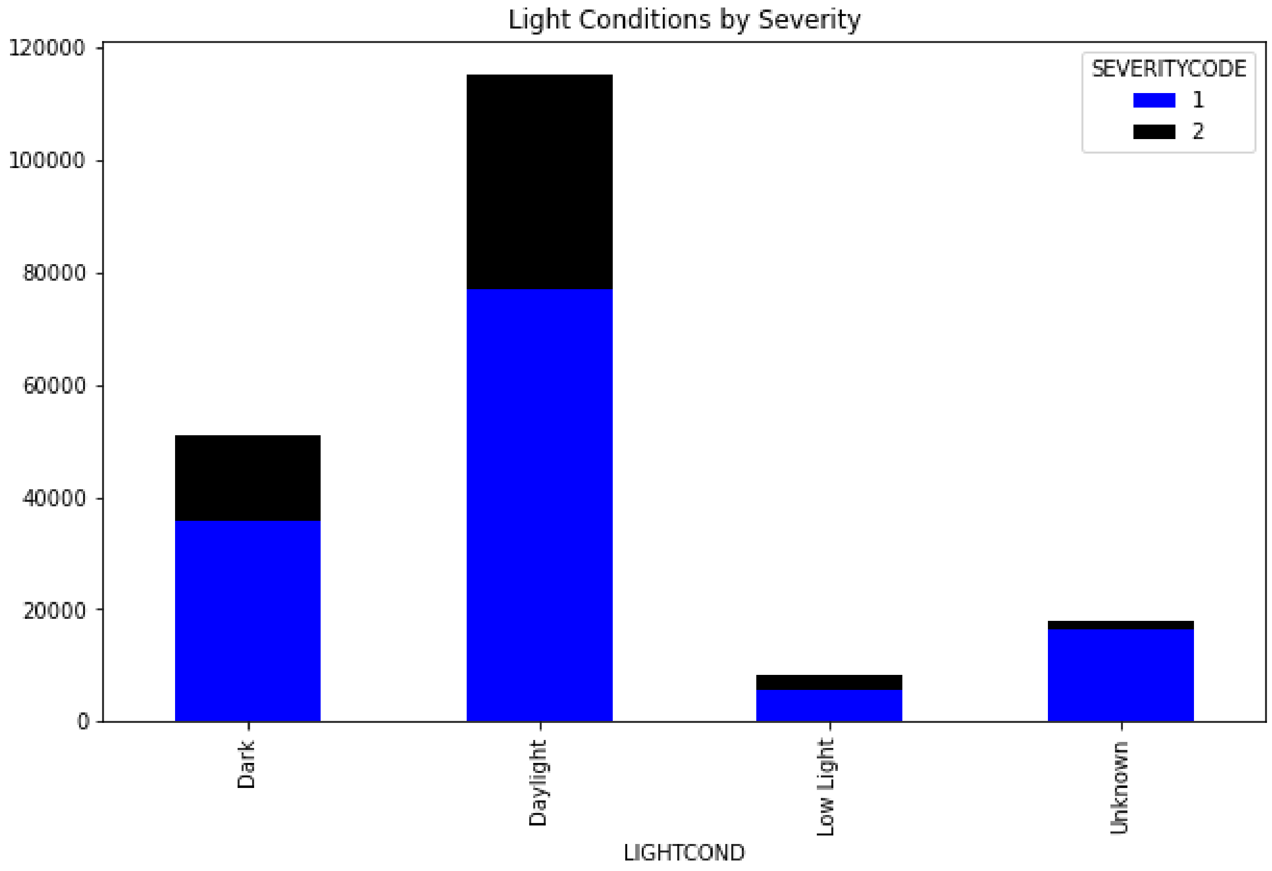
First a bar chart of data count from 2004 to 2020 was taken and it showed that 2020 had too few data so it was removed.



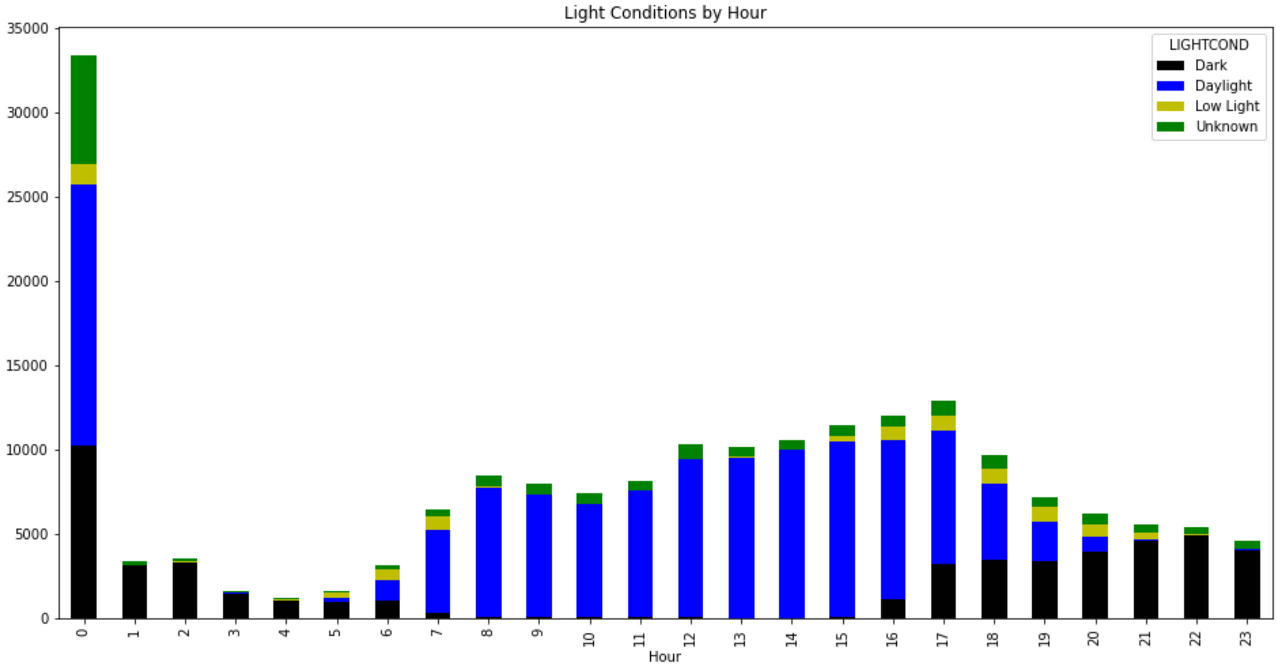
For categorical data, null values were replaced with “Unknown” or they were replaced with values matched on certain condition in data set, Other values were classified or binned to for a group. For example LIGHTCOND, day light was unchanged however Dark Condition were replaced with “Dark” and Dawn and 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 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

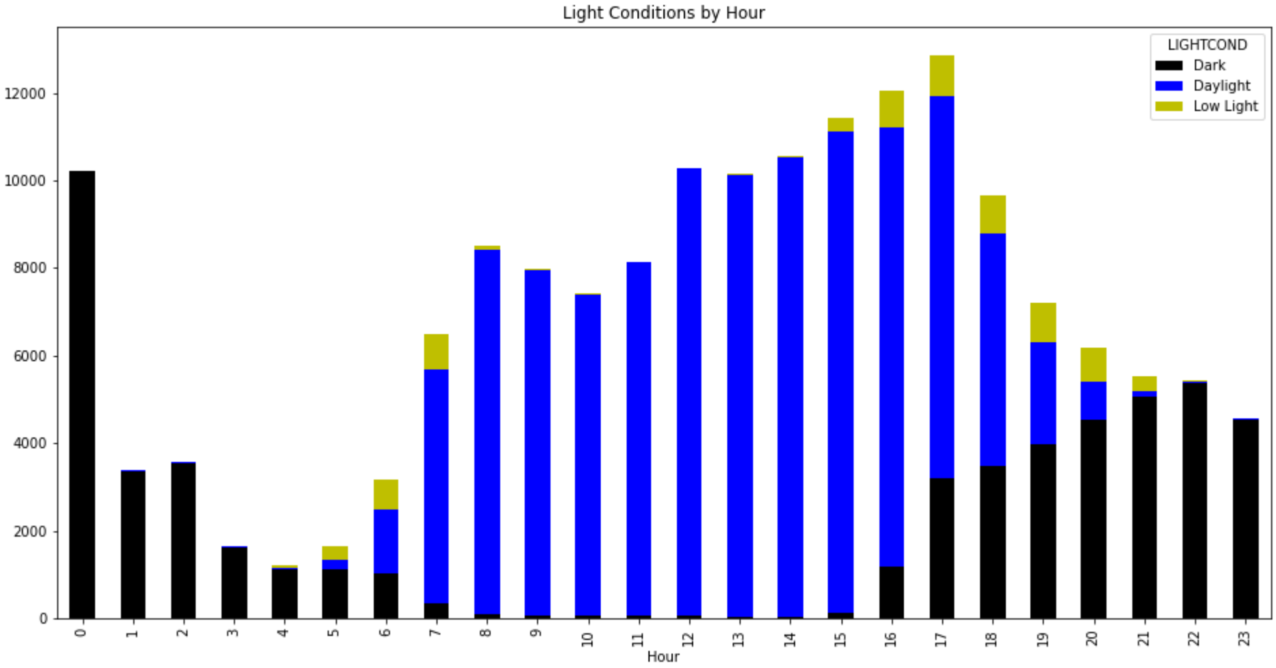
After categorization there were records with no data in light condition:



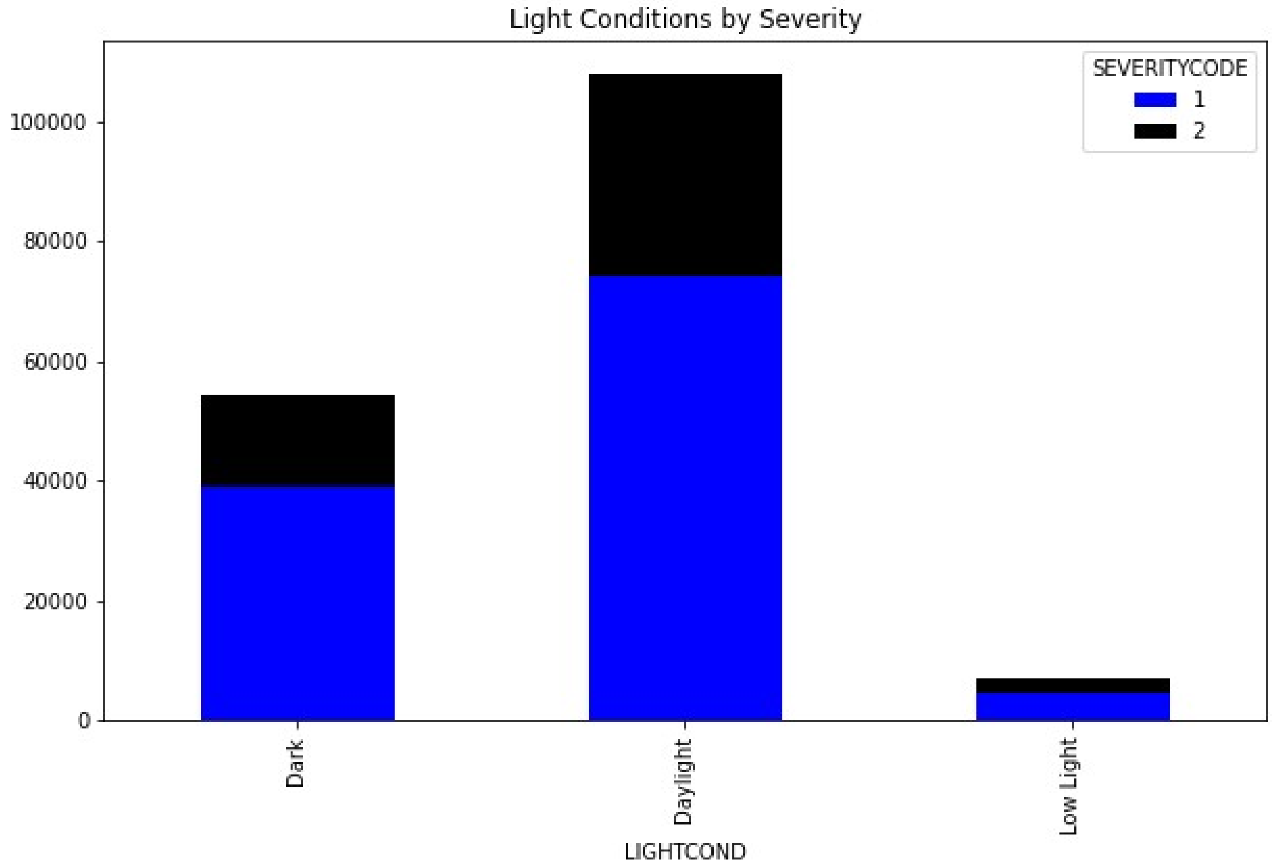
However, data has the hour attribute from INCDTTM, let’s see how light condition is spanned across hour of the day. This helped in in replacing the unknown values in light condition.



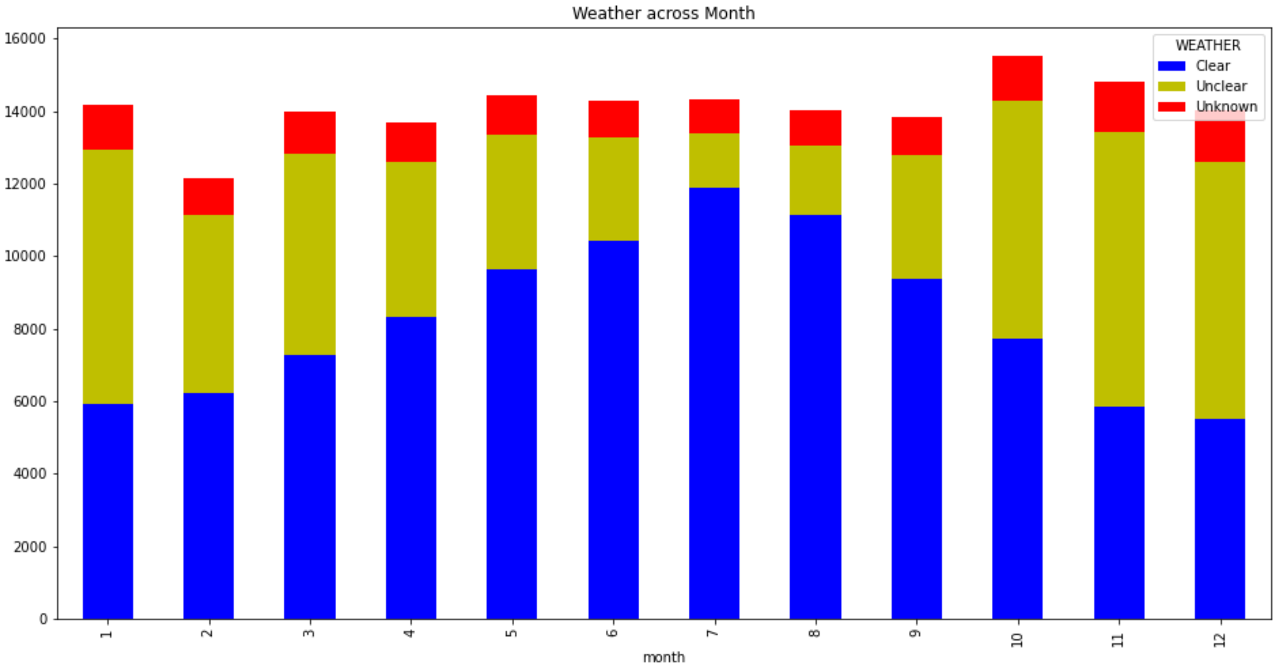
From the graph it can be visualized that there are bad records in INCDDTM columns as well. As there are records with light condition as “daylight” while time is midnight. Such records were dropped from the dataset. As it clear that from 8 AM to 4 PM mostly it is daylight in Seattle, similarly from 8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



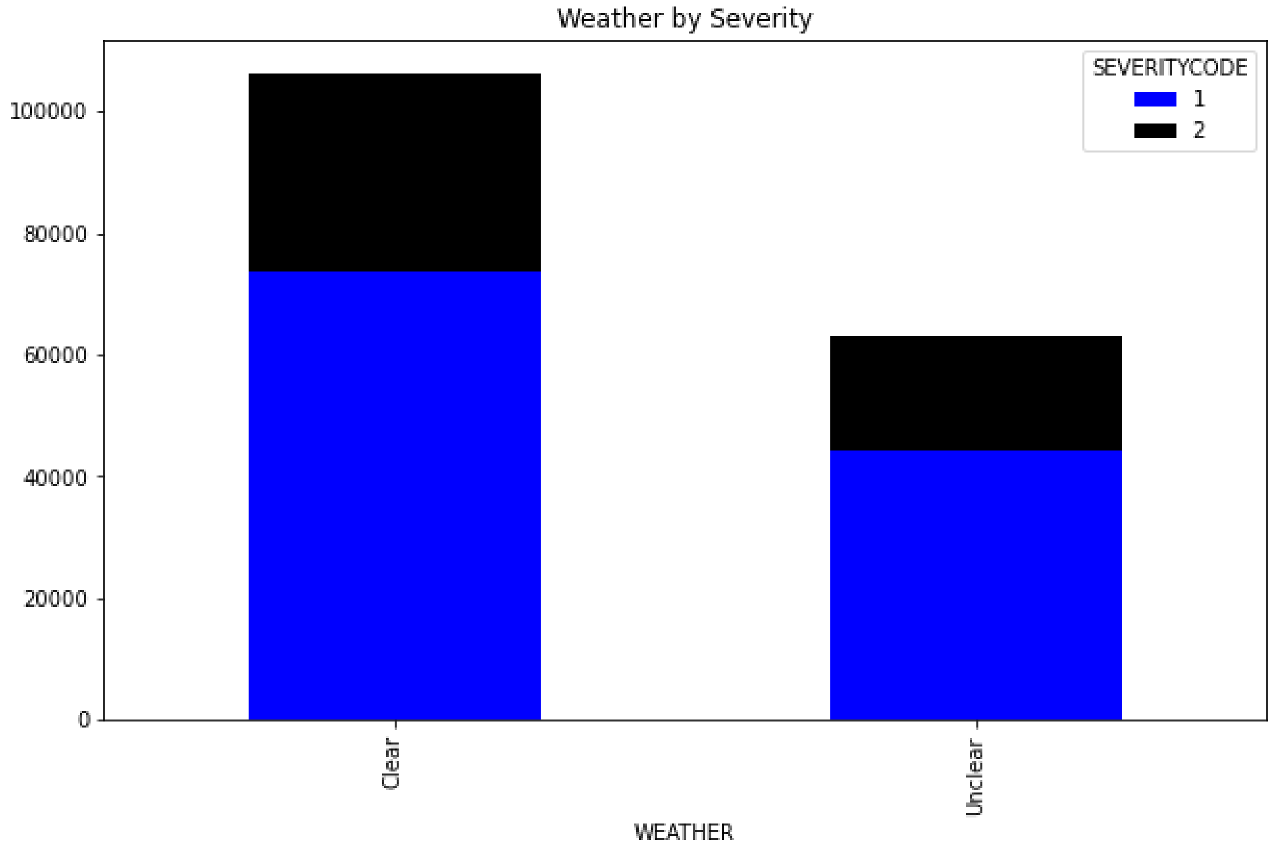
Mapping the Light Condition vs severity shows that light condition does effect the severity of a collision.



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.

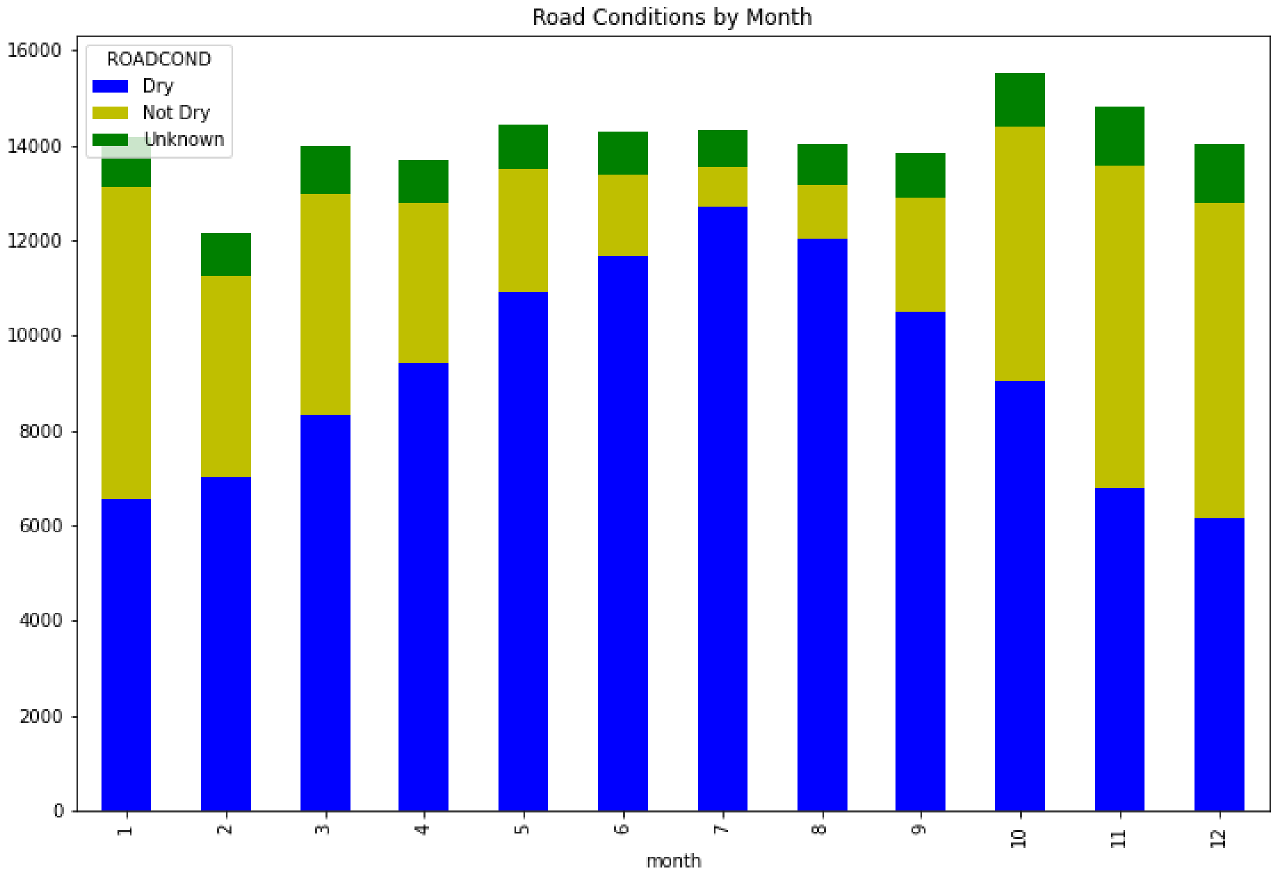


It can be deduced that weather is mostly clear in months of April to September. But weather can vary across the day and the data is spanned across 15 years, so it is wise to use INCDATE to fill missing values. It was assumed that weather was similar across Seattle on same date so if at a particular date 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.

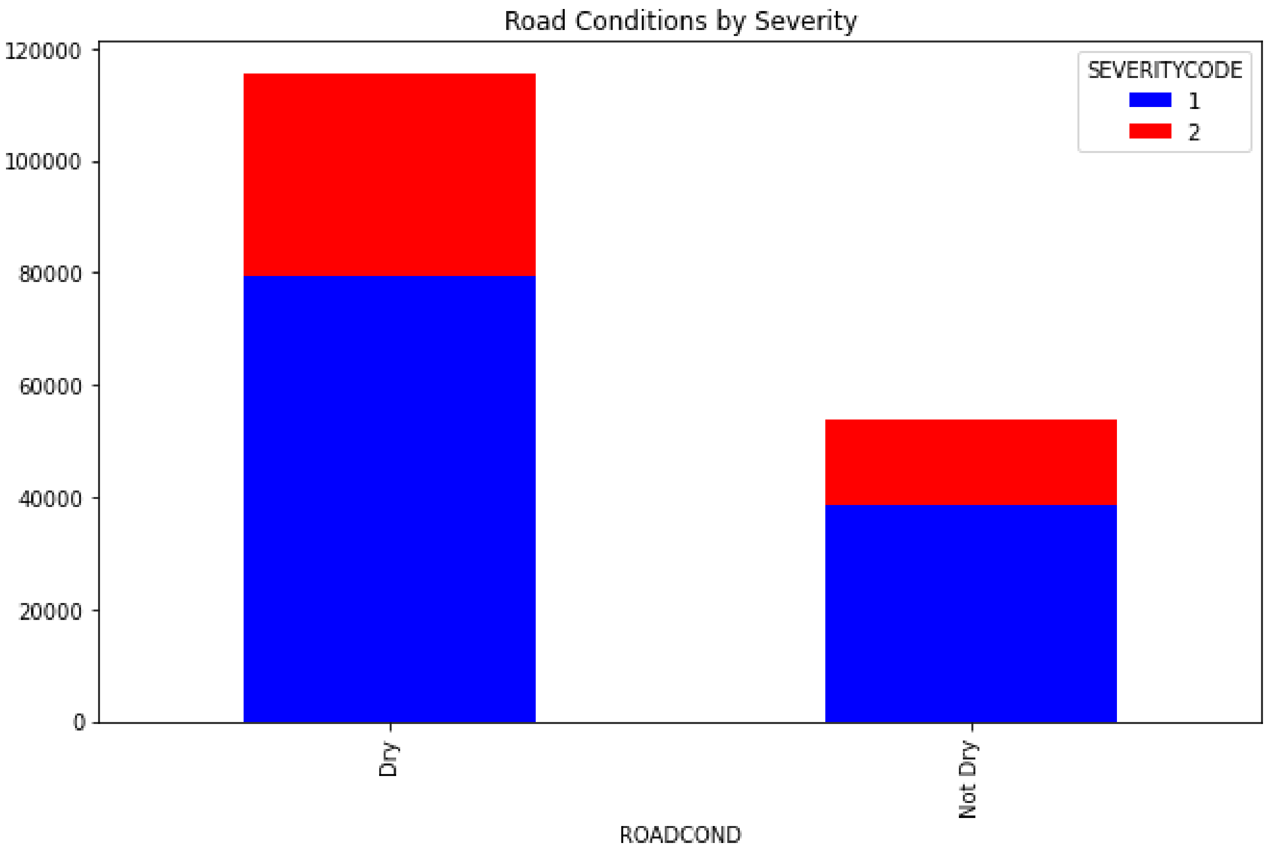


Road Condition has values which showed the condition of the road at site of the collision. Road Condition of Dry contributed more than 60% of the data, so other values were classified as “Not Dry”.

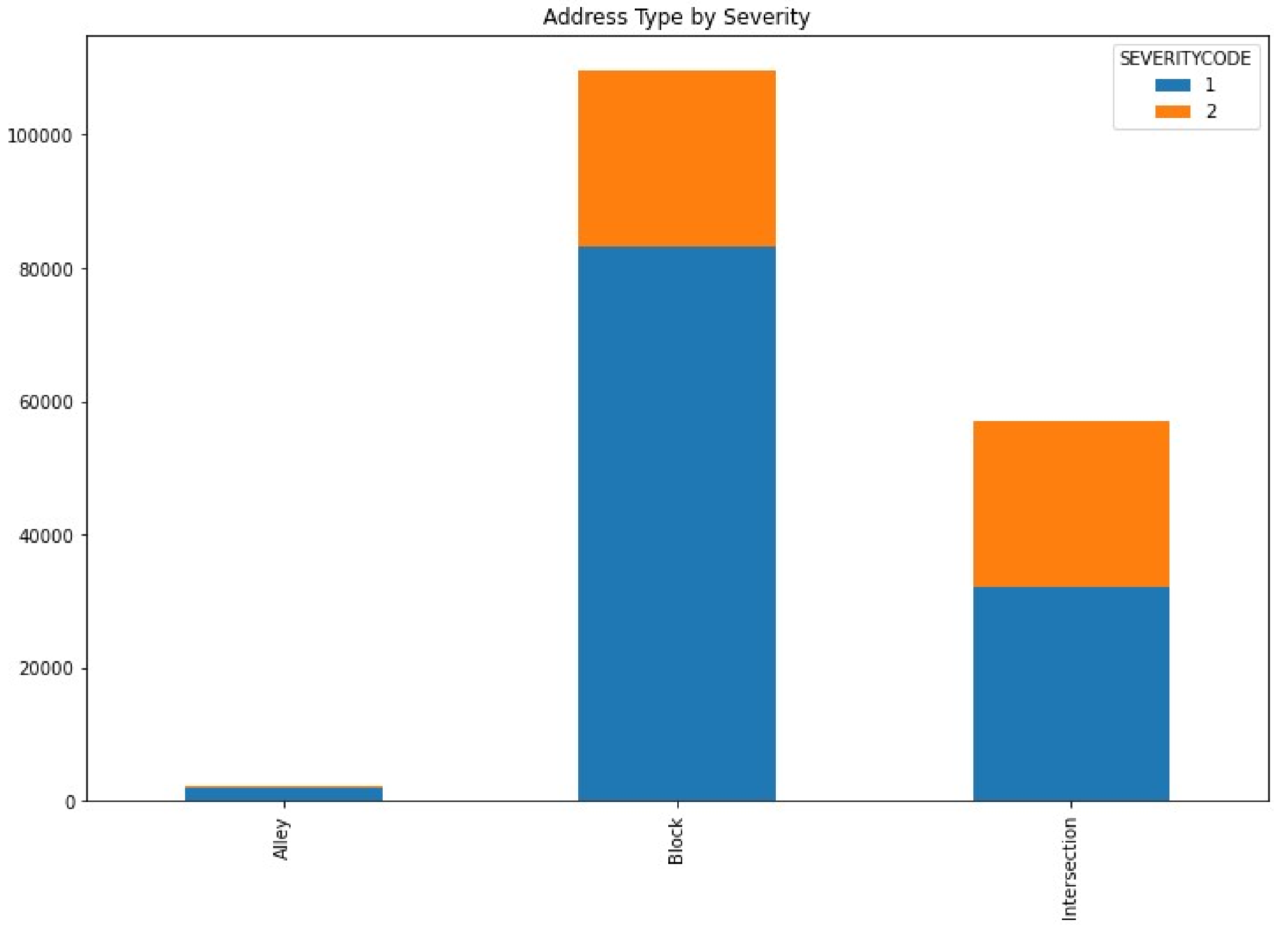
Around 10% of records have missing road condition. From the values it is quite evident that many of these values depends on the weather, the relation between road condition across months can be visualized from the graph below.



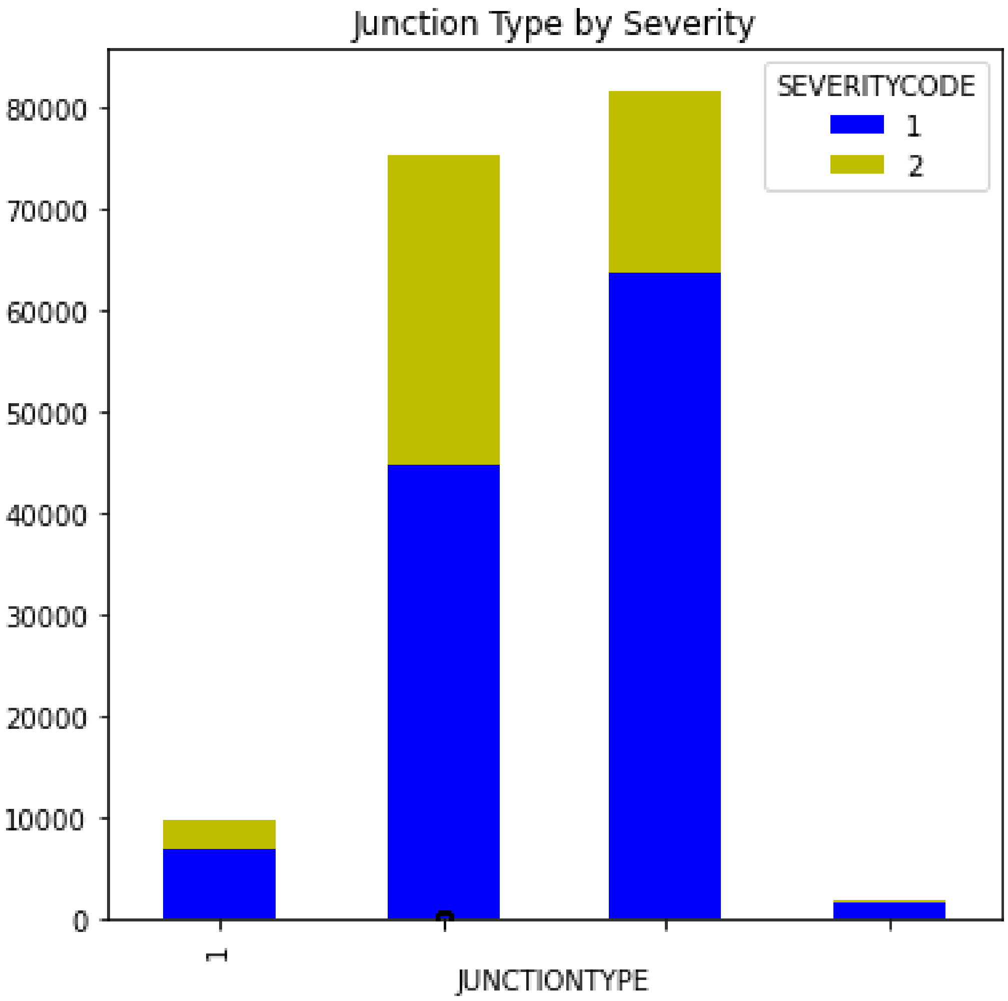
Graph shows that during summer the roads are dry compared to other months in Seattle. So the missing values for these months was set to dry and for rest months it was replaced with not dry.



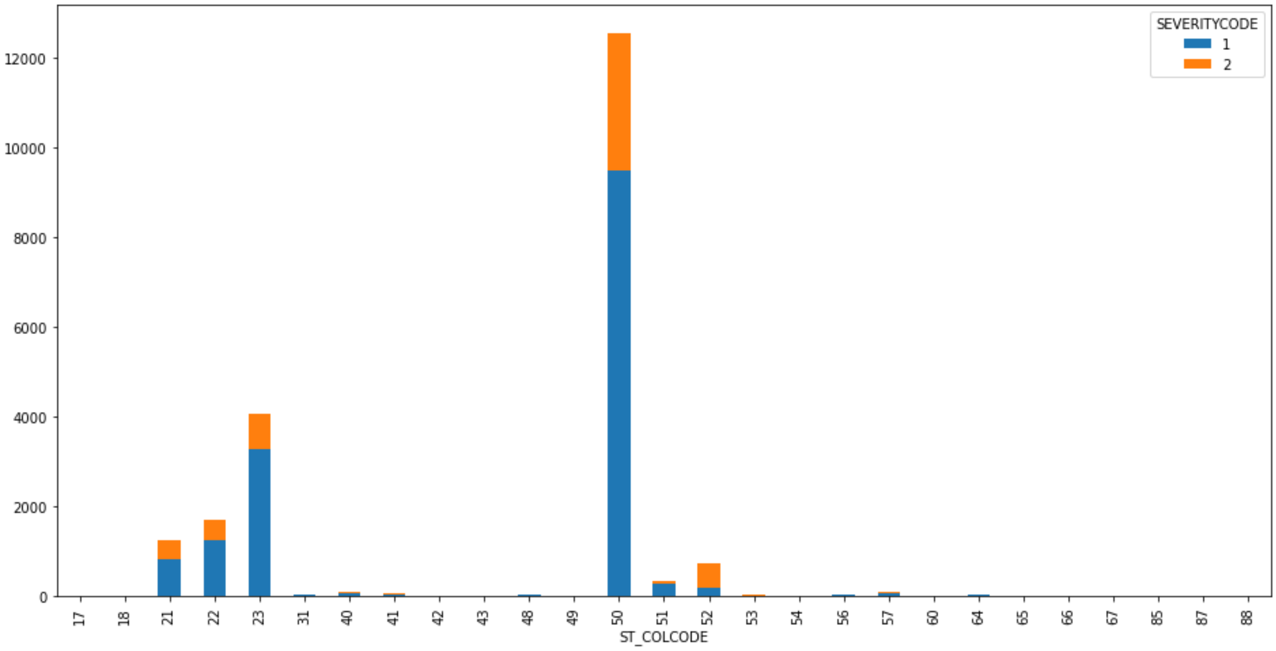
Address type determines the type of address of the collision site. Data has latitude, longitude and Location (street names) which determines the exact location of collision. However, ADDRTYPE is classified column so shall be useful in prediction. Null Values in ADDRTYPE columns were replaced with the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



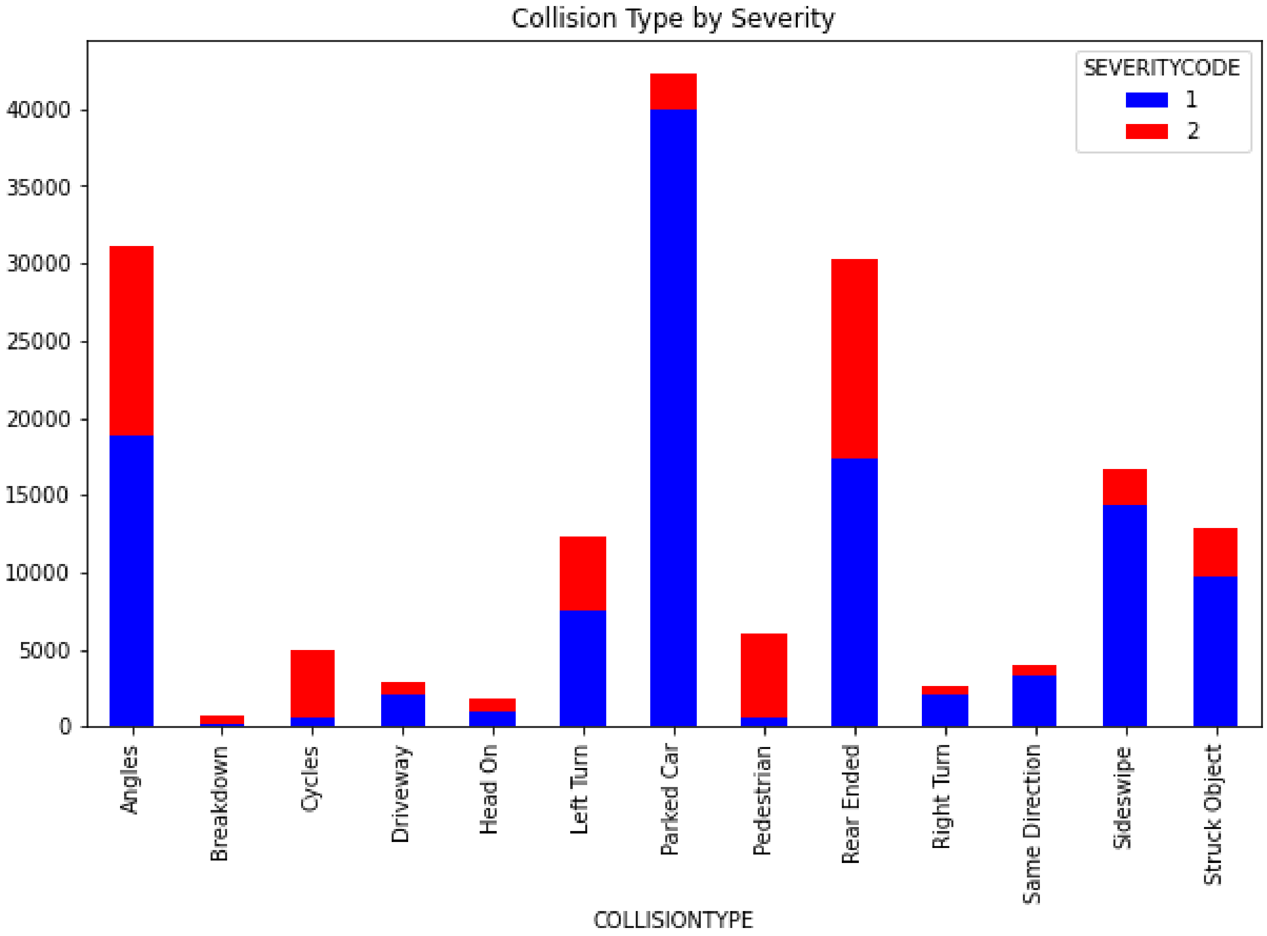
Junction type determines the type of junction at collision site. The values determine whether collision was it intersection, ramp or at a block.



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.



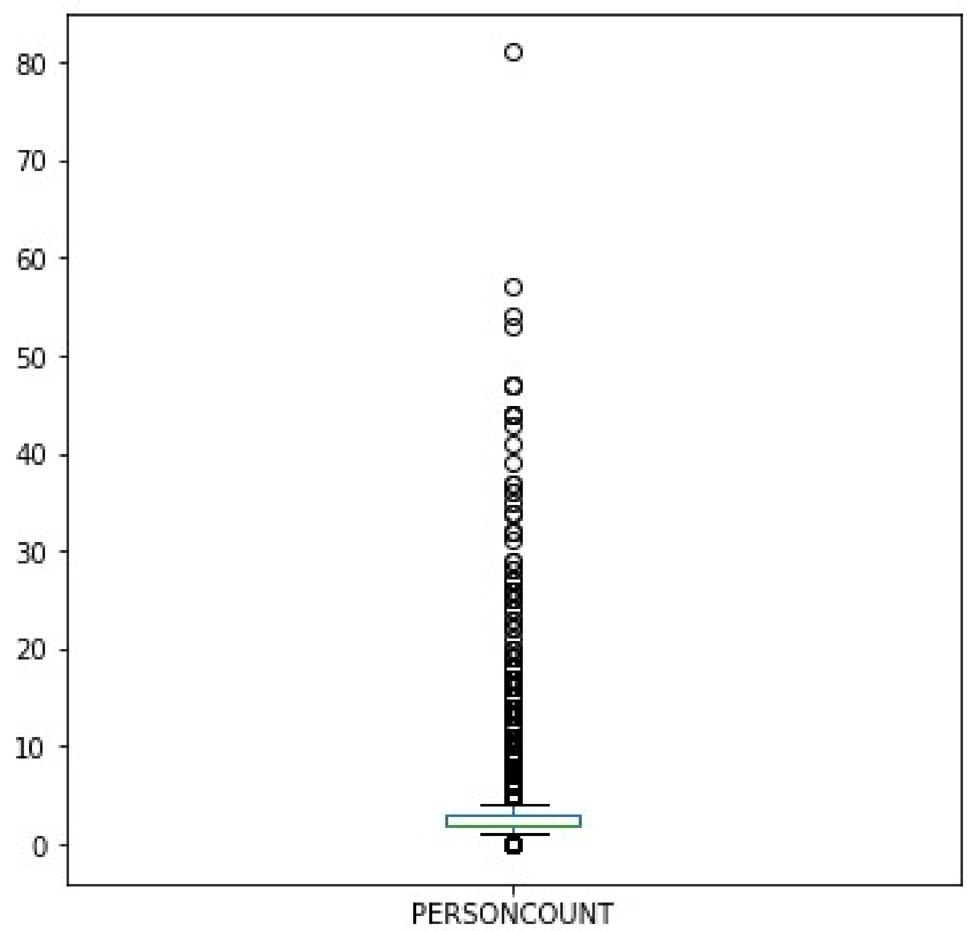
From the graph, it was determined that ST_COLCODE with values of 21, 22, 23, 50, 51, 52 are useful. Records with Collision Type s Other and values not in above list were dropped.



Methodology section

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.

The PERSONCOUNT 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.



check for P-values using Pearson’s coefficient to finalize the numerical features.

```
 Column Name  Pearson Correlation Coefficient     P-value of
0     SEVERITYCODE                        1.000000   0.000000e+00
1         PEDCOUNT                        0.251263   0.000000e+00
2      PEDCYLCOUNT                        0.217310   0.000000e+00
3         VEHCOUNT                       -0.088294  5.464418e-281
4   INATTENTIONIND                        0.040824   2.016460e-61
5        UNDERINFL                        0.043036   4.618621e-68
6    PEDROWNOTGRNT                        0.208156   0.000000e+00
7         SPEEDING                        0.038203   5.226122e-54
8     HITPARKEDCAR                       -0.106021   0.000000e+00
9            month                        0.004751   5.437663e-02
10       dayofweek                       -0.017850   4.910711e-13
11            Hour                        0.024883   7.009487e-24
```

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:

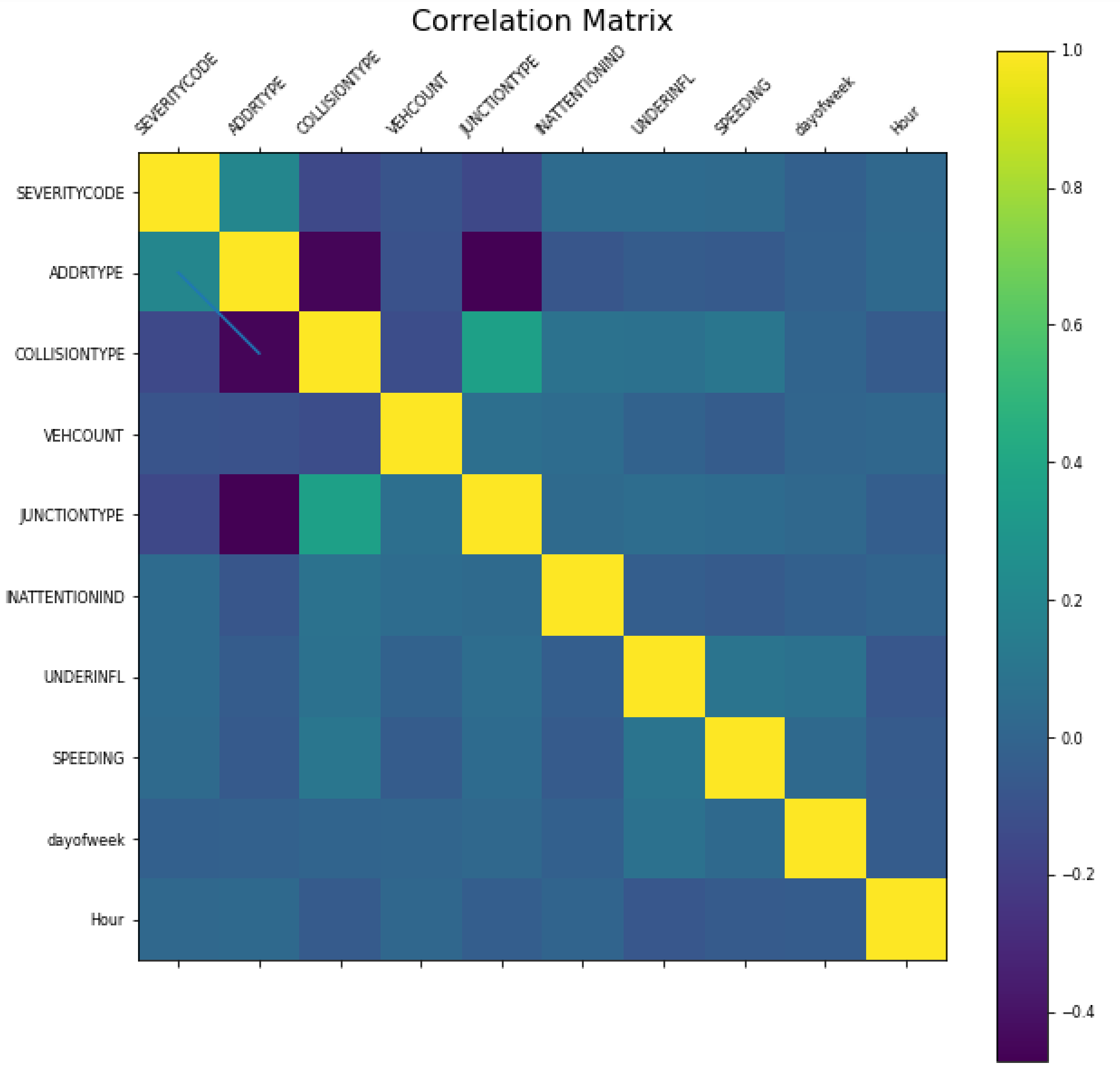
F-test score: ANOVA assumes the means of all groups are the same, calculates how much the actual means deviate from the assumption, and reports it as the F-test score. A larger score means there is a larger difference between the means.

P-value: P-value tells how statistically significant is our calculated score value.

large F test score showing a strong correlation and a P value of almost 0 implying almost certain statistical significance. expect ANOVA to return a sizable F-test score and a small p-value.

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

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



4 models were evaluated. The data set was split into test and train dataset. The train dataset was used to train 4 classification models. KNN, SVM, Logistic Regression and Decision Tree.

Results

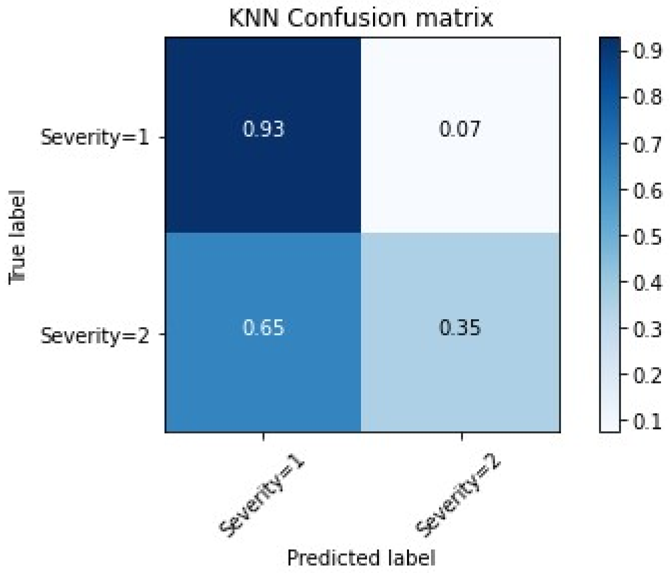
The confusion matrix showed that KNN and Decision Tree showed the highest precision when it came to accurately predicting .

For KNN results:

```
KNN Jaccard index: 0.721
KNN F1-score: 0.723
KNN LogLoss: 0.638
```

```
 precision    recall  f1-score   support
```

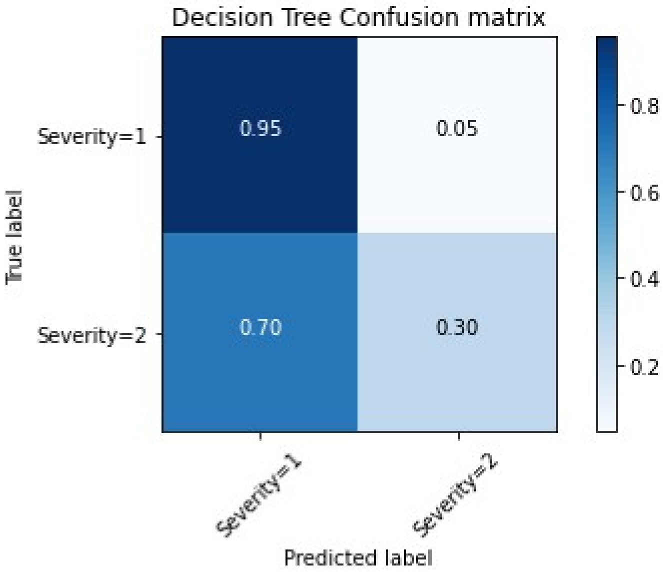
```
           1       0.76      0.93      0.84     28458
           2       0.68      0.35      0.46     12528

    accuracy                           0.75     40986
   macro avg       0.72      0.64      0.65     40986
weighted avg       0.74      0.75      0.72     40986

Normalized confusion matrix
[[0.93 0.07]
 [0.65 0.35]]
```



for Decision Tree results:

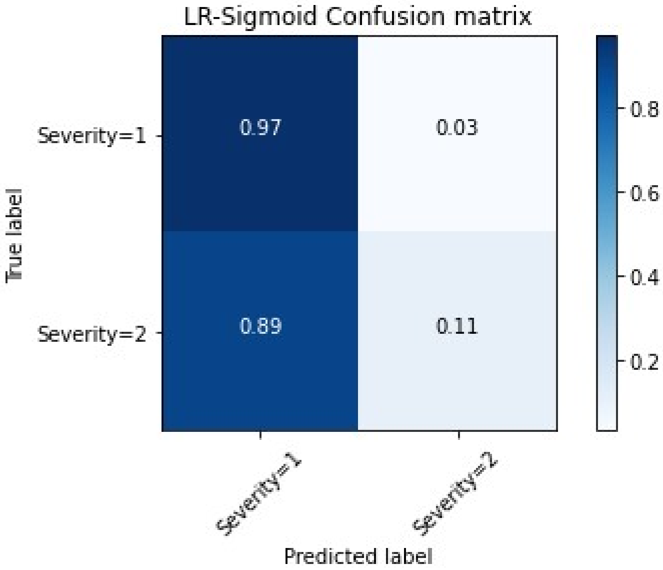
```
DT Jaccard index: 0.728
DT F1-score: 0.714
DT LogLoss: 0.487


 precision    recall  f1-score   support

           1       0.75      0.95      0.84     28458
           2       0.74      0.30      0.42     12528

    accuracy                           0.75     40986
   macro avg       0.75      0.62      0.63     40986
weighted avg       0.75      0.75      0.71     40986

Normalized confusion matrix
[[0.95 0.05]
```

```
 [0.7  0.3 ]]
```



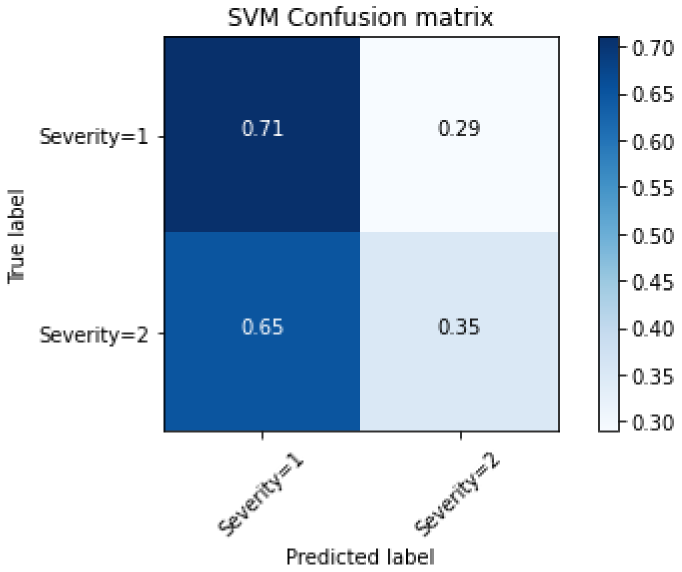
Logistic Regression results:

```
LR Jaccard index: 0.695
LR F1-score: 0.624
LR LogLoss: 0.583

 precision    recall  f1-score   support

           1       0.71      0.97      0.82     28458
           2       0.60      0.11      0.18     12528

    accuracy                           0.71     40986
   macro avg       0.66      0.54      0.50     40986
weighted avg       0.68      0.71      0.62     40986

Normalized confusion matrix
[[0.97 0.03]
 [0.89 0.11]]
```



Support Vector Machine results:

```
SVM Jaccard index: 0.552
SVM F1-score: 0.601

 precision    recall  f1-score   support

           1       0.71      0.71      0.71     28458
           2       0.35      0.35      0.35     12528

    accuracy                           0.60     40986
   macro avg       0.53      0.53      0.53     40986
weighted avg       0.60      0.60      0.60     40986

Normalized confusion matrix
[[0.71 0.29]
 [0.65 0.35]]
```



Discussion

Various collision parameters for a collisions 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.

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 medical personnel to access the collision and direct resources 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 a newer cars are more safe than old cars. If we find such data sets, could improve the performance of models significantly.