Introduction

According to WHO Global status report on road safety 2018, road traffic injuries are now the leading killer of people aged 6–29 years old, and the number of annual road traffic deaths has reached 1,350,000*. It has become a major issue in society.

To minimize such accidents, it's vital to implement preventional measures.

This report aims at developing a model for Seattle government to understand the severity of accidents under different conditions. With this knowledge, the government could revise current rules and regulations, set necessary reminders and allocate resources to control the road traffic injuries.

- World Health Organization (WHO). Global Status Report on Road Safety 2018. December 2018. Available from URL: https://www.who.int/violence_injury_prevention/road_safety_status/2018/en/

Data

The data we are looking at was collected by SDOT Traffic Management Division of Seattle Government from 2004 to 2020. The original dataset contains 37 attributes and 194673 records. Information is labelled but unbalanced. Out of 37 attributes, 9 attributes that have the highest relevancy were picked.

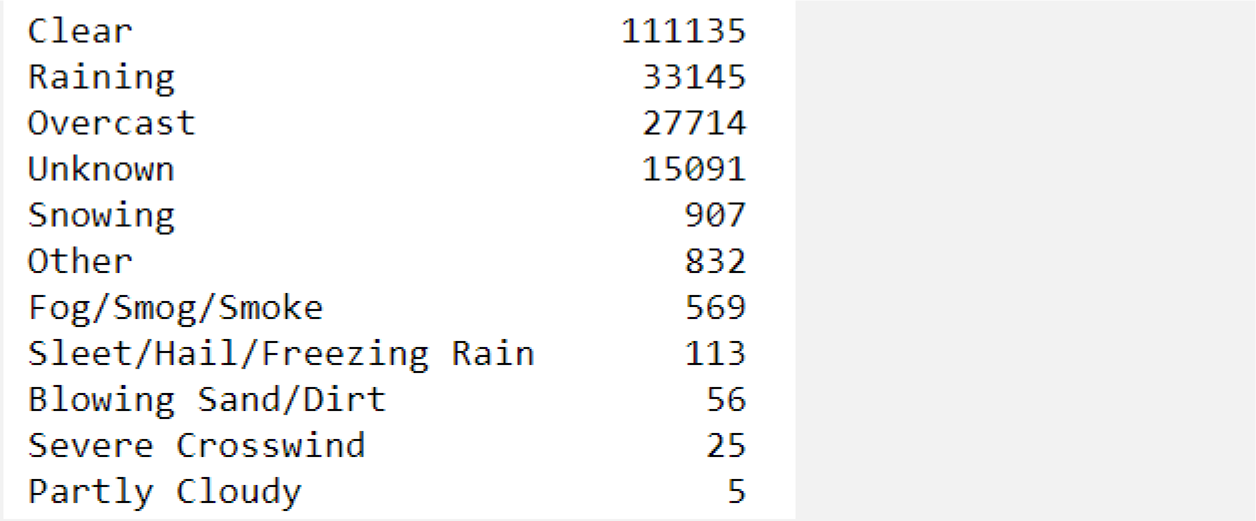
Our target attributes are severity. A code that corresponds to the severity of the collision:

- *3 — fatality*
- *2b — serious injury*
- *2 — injury*
- *1 — prop damage*
- *0 — unknown*

There are two levels recorded in the dataset, namely 1 and 2.

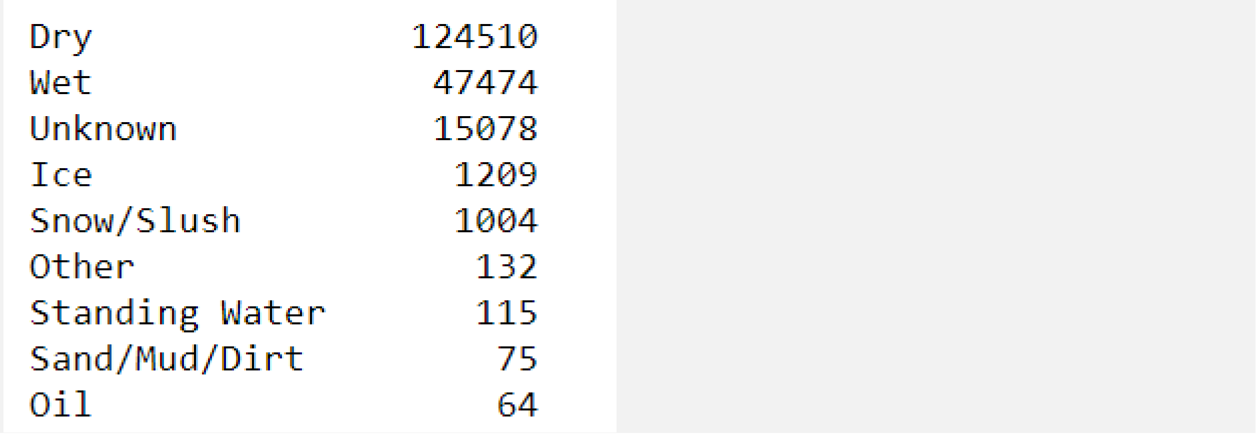
4 attributes, namely, weather, junction type, light condition and road condition are selected as the independent variables.

Weather — A description of the weather conditions during the time of the collision. 11 weather conditions were recorded.



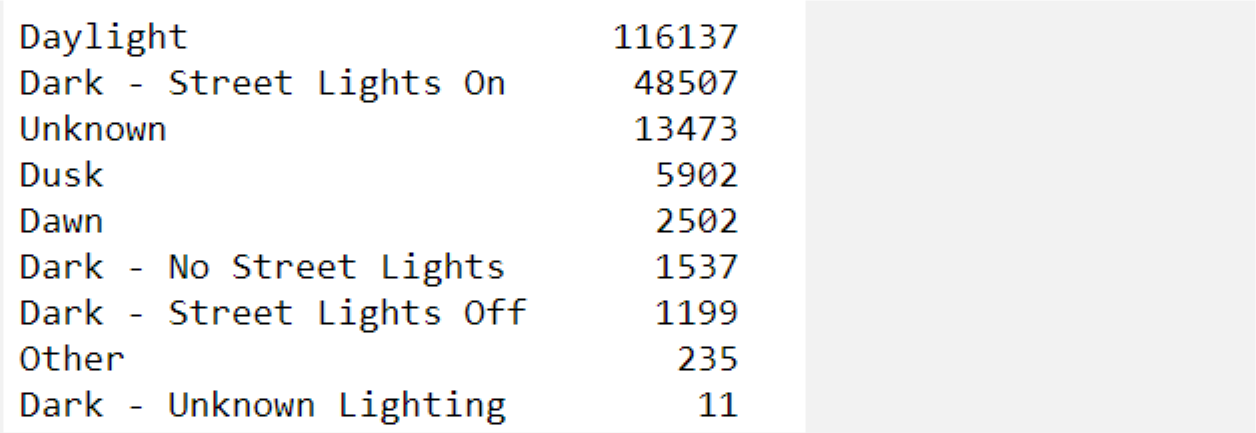
```
Clear                       111135
Raining                      33145
Overcast                     27714
Unknown                      15091
Snowing                        907
Other                          832
Fog/Smog/Smoke                 569
Sleet/Hail/Freezing Rain       113
Blowing Sand/Dirt               56
Severe Crosswind                25
Partly Cloudy                    5
```

Road Condition — The condition of the road during the collision. 9 road conditions were recorded.



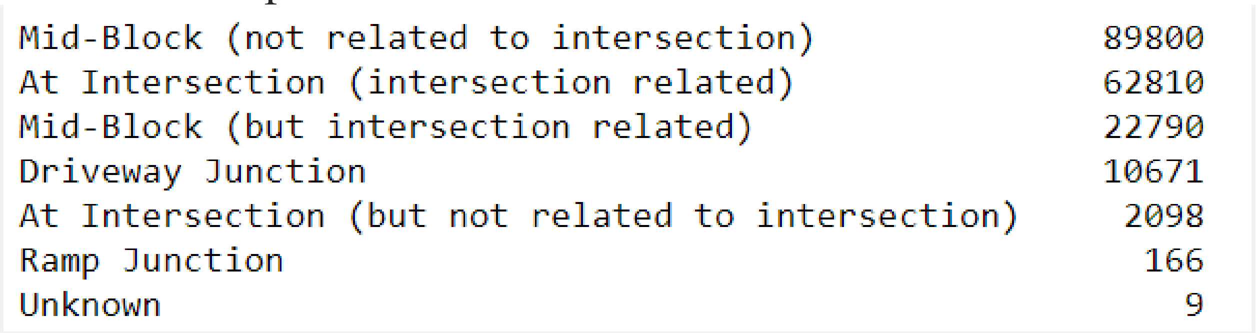
```
Dry               124510
Wet                47474
Unknown            15078
Ice                 1209
Snow/Slush          1004
Other                132
Standing Water       115
Sand/Mud/Dirt         75
Oil                   64
```

Light Condition — The light conditions during the collision. 9 light conditions were recorded.



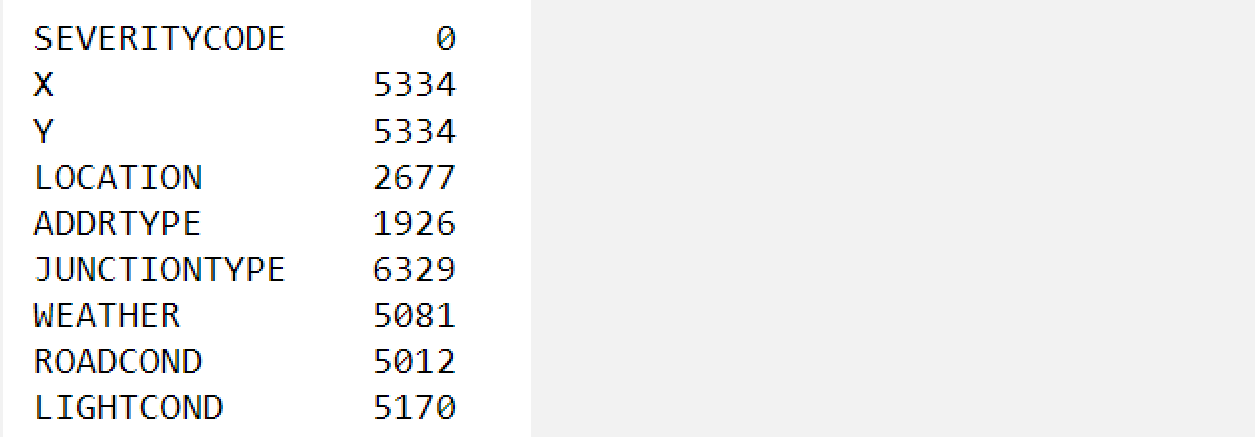
```
Daylight                    116137
Dark - Street Lights On      48507
Unknown                      13473
Dusk                          5902
Dawn                          2502
Dark - No Street Lights       1537
Dark - Street Lights Off      1199
Other                          235
Dark - Unknown Lighting         11
```

And lastly, junction type — Category of the junction at which collision took place.



```
Mid-Block (not related to intersection)              89800
At Intersection (intersection related)               62810
Mid-Block (but intersection related)                 22790
Driveway Junction                                    10671
At Intersection (but not related to intersection)     2098
Ramp Junction                                          166
Unknown                                                  9
```

By running the initial checking, there are missing values under each attribute.

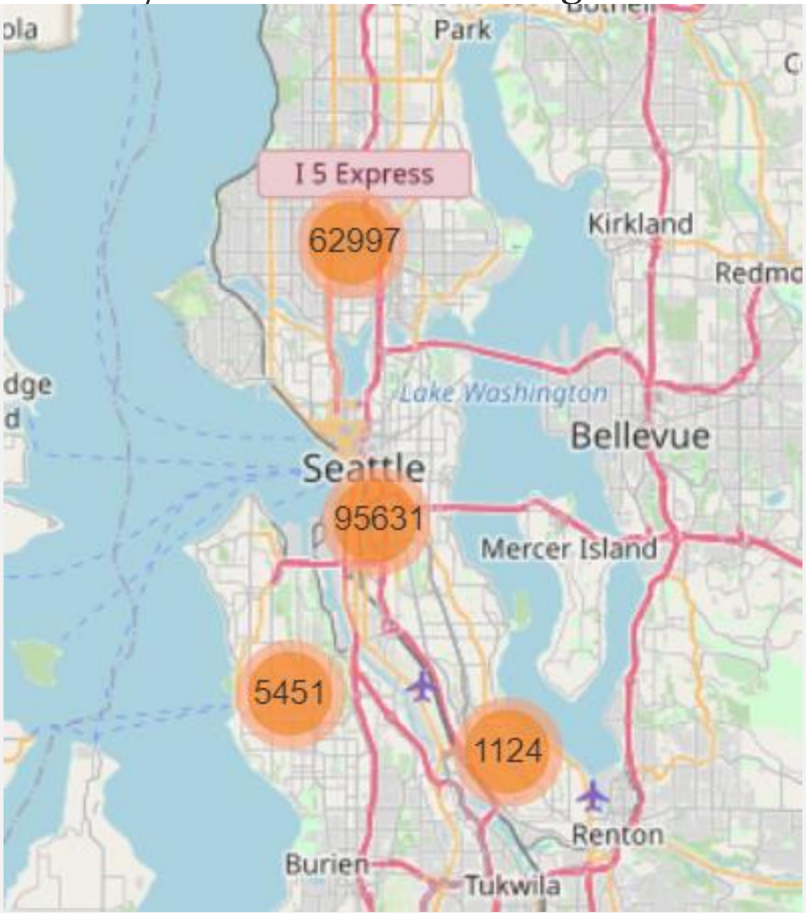


```
SEVERITYCODE        0
X                5334
Y                5334
LOCATION         2677
ADDRTYPE         1926
JUNCTIONTYPE     6329
WEATHER          5081
ROADCOND         5012
LIGHTCOND        5170
```

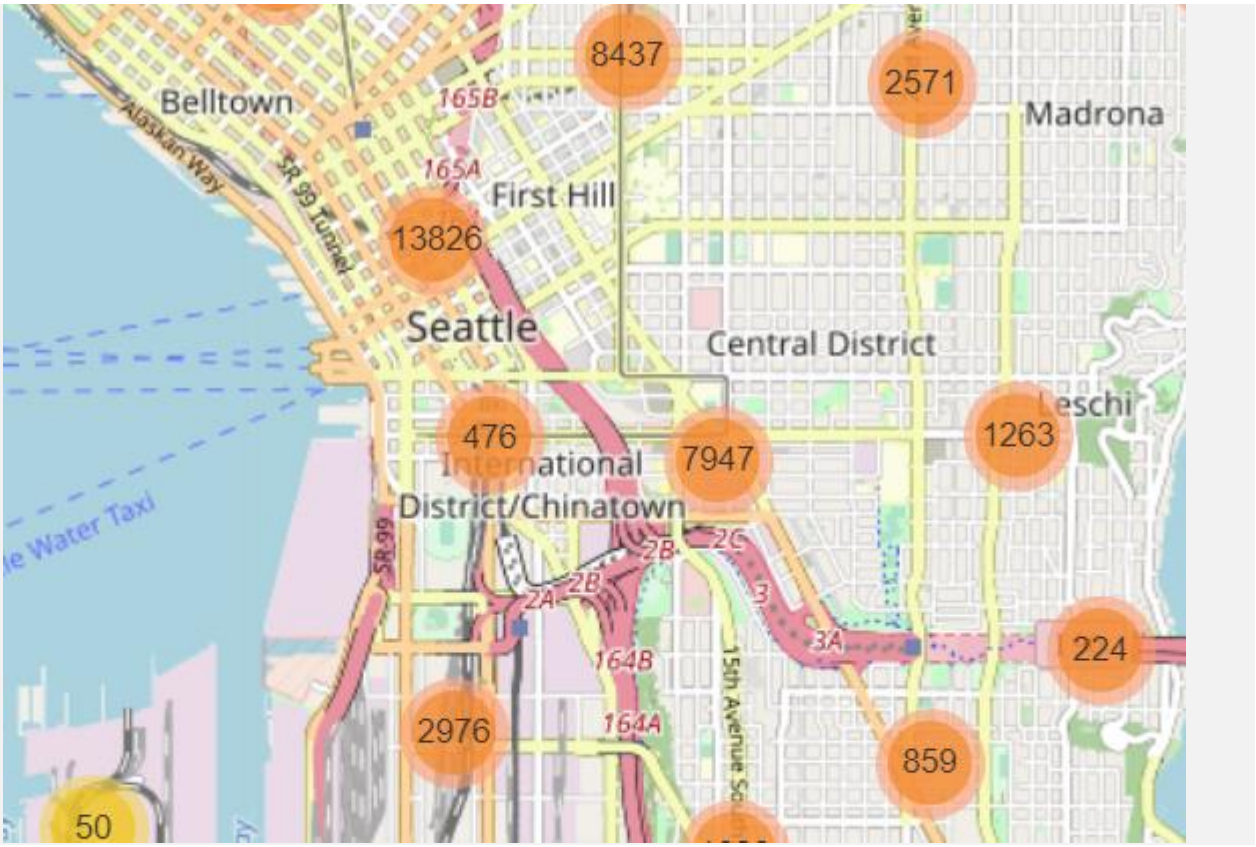
Missing values are off no significance in our analysis, hence rows with any missing values are dropped.

In addition, “unknown” records in each attribute provide no information on prediction, rows with any unknown records are deleted.

After clearing the data, a map marked with collisions were plotted with X, Y and locations. It gives a rough idea of which areas have the highest incident rate. Around the central district/International District/Chinatown has the highest incident rate.

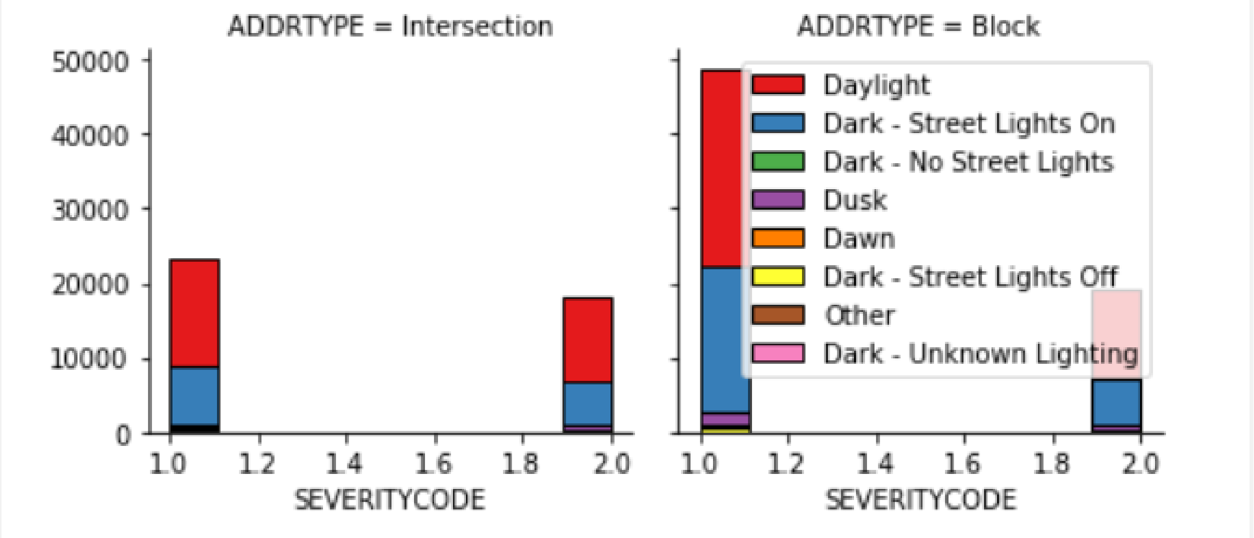


Distribution of collision incidents in Seattle.



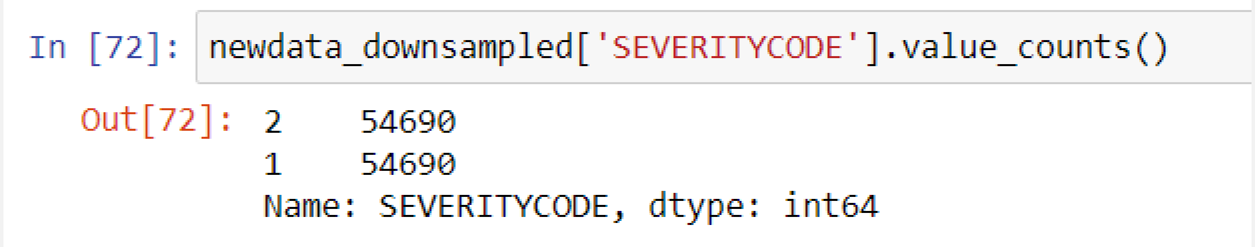
Zoom-in view of Distribution of collision incidents in Seattle.

Some columns were also plotted to understand data better. For collision happened at intersection and block, daylight and Dak with street lights on are the top frequent weather conditions.



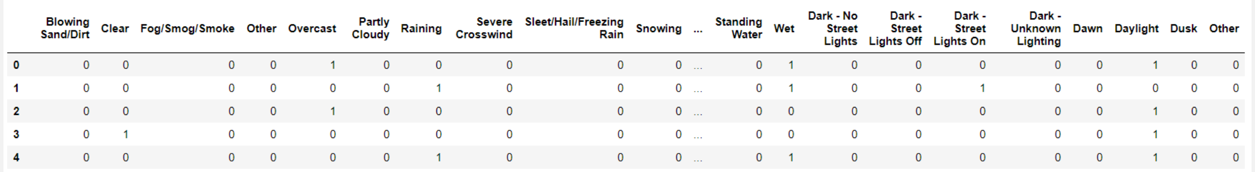
Conditions under which collisions happened at different address type. *Junction type were selected in the data analysis as it provides a more detailed categorization of incident locations.

To balance the data, a downsampling was performed. Each severity level now contains 54690 samples.



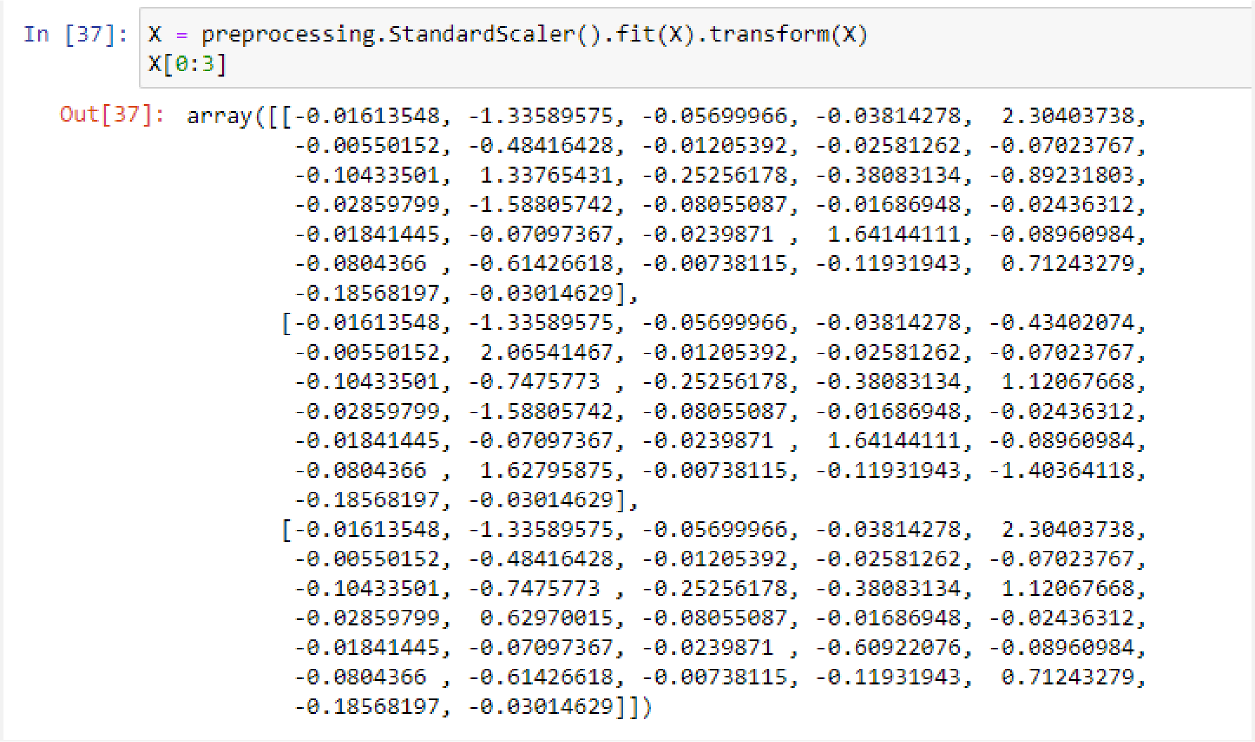
```
In [72]: newdata_downsampled['SEVERITYCODE'].value_counts()

Out[72]: 2    54690
         1    54690
         Name: SEVERITYCODE, dtype: int64
```

On to data preprocessing, to perform machine learning models, all categorical variables were converted to numerical values by performing on hot coding and normalized before carrying out any analysis.



	Blowing Sand/Dirt	Clear	Fog/Smog/Smoke	Other	Overcast	Partly Cloudy	Raining	Severe Crosswind	Sleet/Hail/Freezing Rain	Snowing	...	Standing Water	Wet	Dark - No Street Lights	Dark - Street Lights Off	Dark - Street Lights On	Dark - Unknown Lighting	Dawn	Daylight	Dusk	Other
0	0	0	0	0	1	0	0	0	0	0	...	0	1	0	0	0	0	0	1	0	0
1	0	0	0	0	0	0	1	0	0	0	...	0	1	0	0	1	0	0	0	0	0
2	0	0	0	0	1	0	0	0	0	0	...	0	0	0	0	0	0	0	1	0	0
3	0	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1	0	0
4	0	0	0	0	0	0	1	0	0	0	...	0	1	0	0	0	0	0	1	0	0

Feature data set — categorical data were converted to numerical data by one hot coding method.



```
In [37]: X = preprocessing.StandardScaler().fit(X).transform(X)
         X[0:3]

Out[37]: array([[-0.01613548, -1.33589575, -0.05699966, -0.03814278,  2.30403738,
                 -0.00550152, -0.48416428, -0.01205392, -0.02581262, -0.07023767,
                 -0.10433501,  1.33765431, -0.25256178, -0.38083134, -0.89231803,
                 -0.02859799, -1.58805742, -0.08055087, -0.01686948, -0.02436312,
                 -0.01841445, -0.07097367, -0.0239871 ,  1.64144111, -0.08960984,
                 -0.0804366 , -0.61426618, -0.00738115, -0.11931943,  0.71243279,
                 -0.18568197, -0.03014629],
                [-0.01613548, -1.33589575, -0.05699966, -0.03814278, -0.43402074,
                 -0.00550152,  2.06541467, -0.01205392, -0.02581262, -0.07023767,
                 -0.10433501, -0.7475773 , -0.25256178, -0.38083134,  1.12067668,
                 -0.02859799, -1.58805742, -0.08055087, -0.01686948, -0.02436312,
                 -0.01841445, -0.07097367, -0.0239871 ,  1.64144111, -0.08960984,
                 -0.0804366 ,  1.62795875, -0.00738115, -0.11931943, -1.40364118,
                 -0.18568197, -0.03014629],
                [-0.01613548, -1.33589575, -0.05699966, -0.03814278,  2.30403738,
                 -0.00550152, -0.48416428, -0.01205392, -0.02581262, -0.07023767,
                 -0.10433501, -0.7475773 , -0.25256178, -0.38083134,  1.12067668,
                 -0.02859799,  0.62970015, -0.08055087, -0.01686948, -0.02436312,
                 -0.01841445, -0.07097367, -0.0239871 , -0.60922076, -0.08960984,
                 -0.0804366 , -0.61426618, -0.00738115, -0.11931943,  0.71243279,
                 -0.18568197, -0.03014629]])
```

Normalized.

Modelling and Results

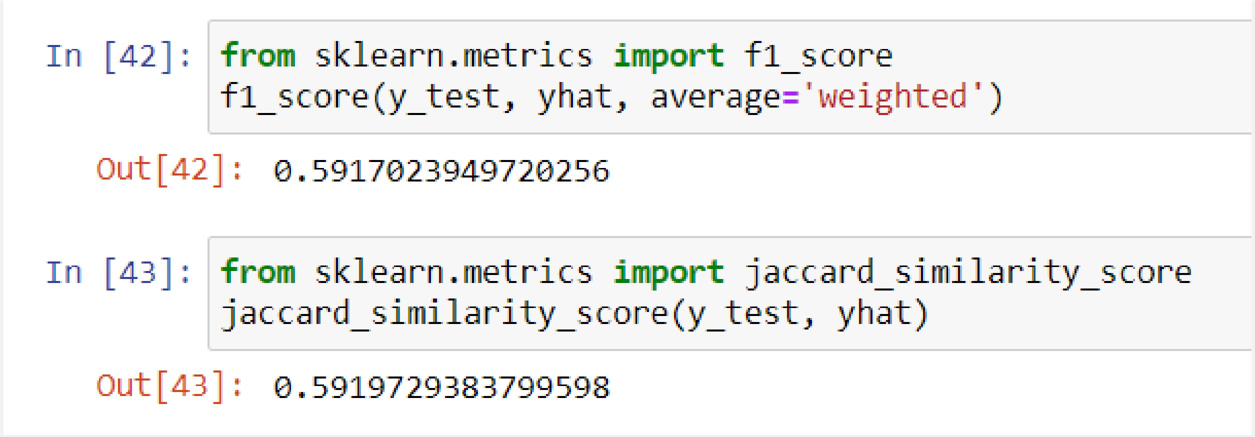
Two machine learning classification model

SVM

Logistic Regression

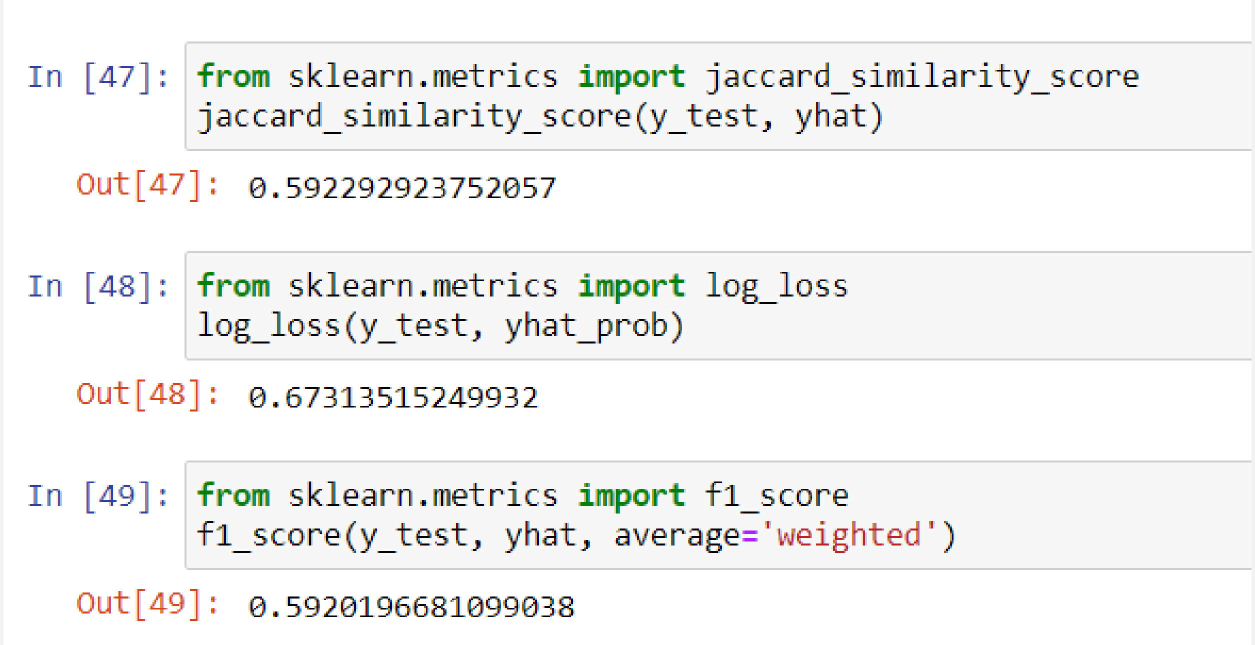
KNN

After fitting with the SVM model, the f1 and Jaccard similarity score are as follows:



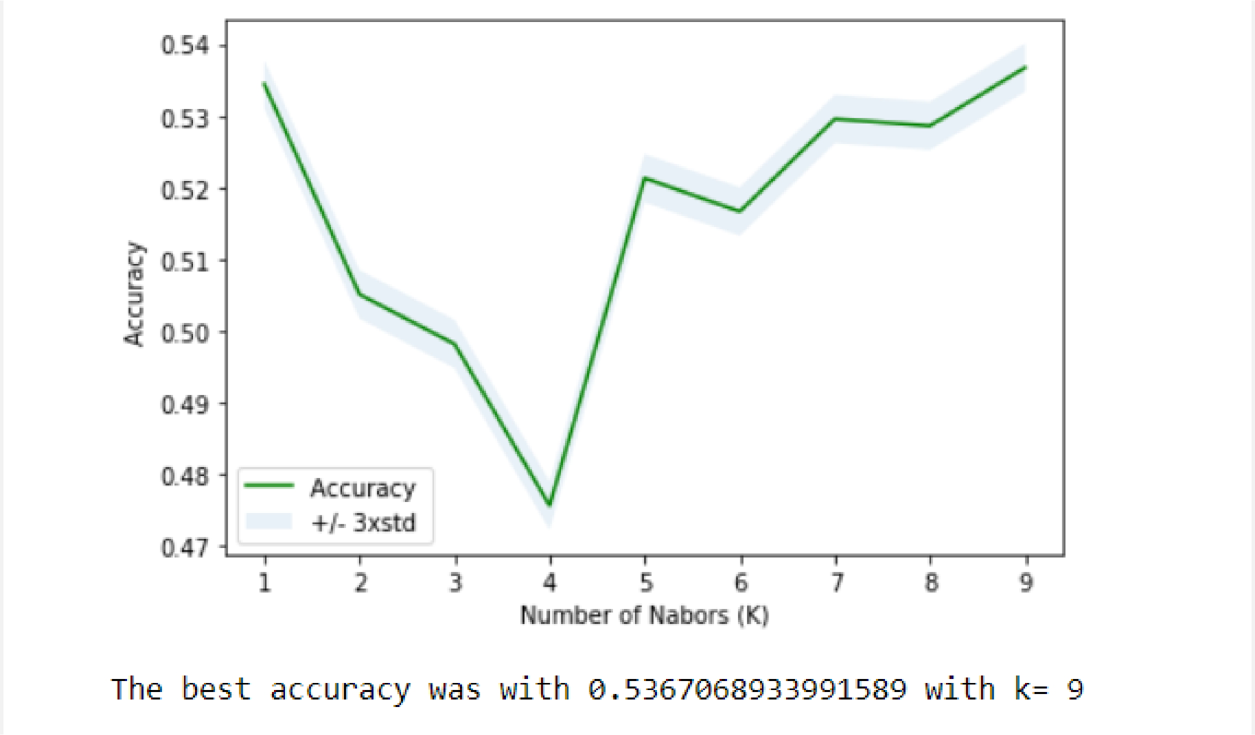
```
In [42]: from sklearn.metrics import f1_score
         f1_score(y_test, yhat, average='weighted')

Out[42]: 0.5917023949720256

In [43]: from sklearn.metrics import jaccard_similarity_score
         jaccard_similarity_score(y_test, yhat)

Out[43]: 0.5919729383799598
```

After fitting with the Logistic Regression model, the f1, Jaccard similarity score and log loss are as follows:



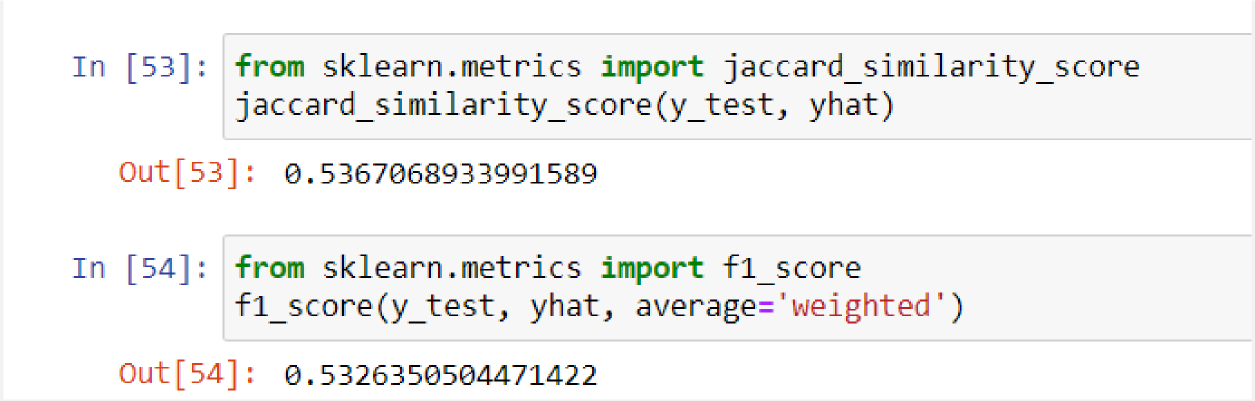
```
In [47]: from sklearn.metrics import jaccard_similarity_score
         jaccard_similarity_score(y_test, yhat)

Out[47]: 0.592292923752057

In [48]: from sklearn.metrics import log_loss
         log_loss(y_test, yhat_prob)

Out[48]: 0.67313515249932

In [49]: from sklearn.metrics import f1_score
         f1_score(y_test, yhat, average='weighted')

Out[49]: 0.5920196681099038
```

For KNN model, first, the best K was found to be 9.



```
The best accuracy was with 0.5367068933991589 with k= 9
```

After fitting with the SVM model, the f1 and Jaccard similarity score are as follows:



```
In [53]: from sklearn.metrics import jaccard_similarity_score
         jaccard_similarity_score(y_test, yhat)

Out[53]: 0.5367068933991589

In [54]: from sklearn.metrics import f1_score
         f1_score(y_test, yhat, average='weighted')

Out[54]: 0.5326350504471422
```

SVM: f1 = 0.5917, Jaccard= 0.592

Logistic Regression: f1= 0.5920, Jaccard=0.5923, logloss = 0.6731

KNN: f1=0.5326, Jaccard=0.5367

Discussion

According to the evaluation result, none of the models is accurate enough for predicting the severity of the collisions (if happened). To have a good model, the best f1 and Jaccard score should be close to 1 while the log loss value for the logistic regression model should be close to 0.

Possible reasons are: the data only has two severity level recorded, namely 1 and 2. The record may miss other severity levels which has significance in the data analysis.

In addition, to save evaluation time, downsampling was performed. Meaningful data may be lost in this process.

Besides, the data was collected in a period of 12 years. The situation in the city may have changed significantly, so the analysis should be performed by using data collected in recent years.

However, compare these two models, the KNN model is slightly better in predicting the severity of the collisions.

Despite this, by visualizing the data on the map, it is clear to see which area in Seattle is accident-prone.

In addition, the dataset itself could tell some interesting findings. For example, people tend to think that it’s more dangerous to drive in the dark, hence for light condition, the dark condition should have a higher incident rate. However, the dataset shows that daylight, in contrast, has a higher incident rate. It’s possible that people tend to driver more carefully during dark as they know it’s dangerous, but more recklessly during daylight as they tend to think it’s harder to get collisions during the day.

```
Daylight                    116137
Dark - Street Lights On      48507
Unknown                      13473
Dusk                          5902
Dawn                          2502
Dark - No Street Lights       1537
Dark - Street Lights Off      1199
Other                          235
Dark - Unknown Lighting         11
```

Conclusion

The dataset used here did not generate an accurate enough model to predict the severity of the collisions for different conditions. The data provides basic information on accident-prone areas in Seattle and conditions that frequently cause collisions in different street types.