Image by Thomas Klinder / EyeEm / Getty Images

Analysis of Car Accident Severity in the City of Seattle

CAPSTONE PROJECT

Ángel Garrido Castañeda | IBM Data Science Professional Certificate | 07/10/2020

Introduction

In this project the data of Seattle traffic accidents will be explored in order to discover the variables with the strongest influence to traffic collisions.

According to statista.com (https://www.statista.com/topics/3708/road-accidents-in-the-us/) 12 million vehicles were involved in crashes in the United States.

There are multiple benefits from traffic accident prediction. Reducing traffic accidents is both an important public safety challenge and critical to prevent traffic congestions. Prediction of traffic accident will help us to improve public transportation, to design cost-effectively transportation infrastructure and to enable safer routes.

Data

In this project data from Seattle GeoData will be used. You can download it from https://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0/data?selectedAttribute=LOCATION.

The data was collected by the Seattle Police Department and Accident Traffic Records Department from 2004 to present. It consists of 194,673 collisions and 37 independent variables. The target variable, “severitycode” has two outputs in our data set: 1 - Property Damage Only Collision and 2 – Injury Collision.

The following attributes will be used in order the create the model:

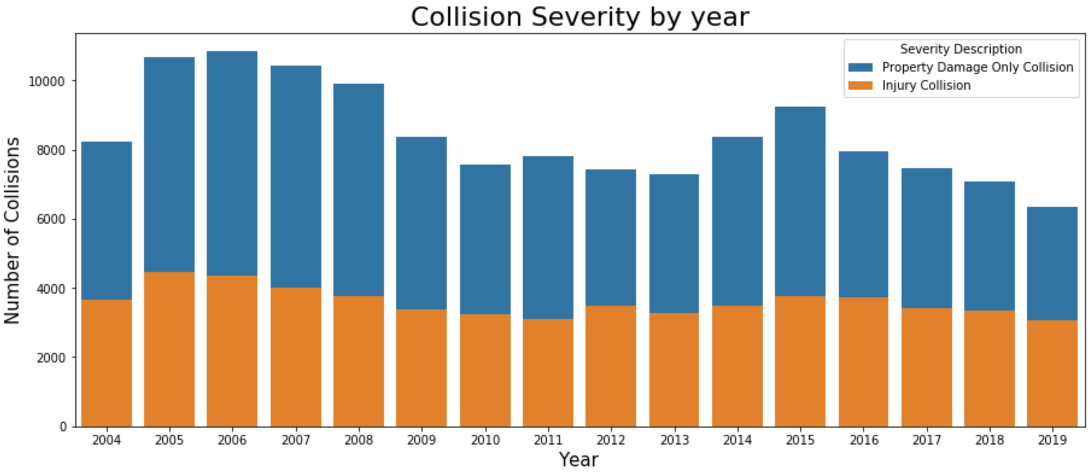
- personcount - number of people involved in collision.
- pedcount - number of pedestrians involved in collision.
- pedcylcount - number of cyclists involved in collision.
- vehcount - number of vehicles involved in collision.
- Incdttm – date and hour of collision.
- junctiontype - 7 types describing collision at intersection, mid-block, driveway and whether collision is related to intersection.
- sdot_coldesc - description of Seattle collision codes.
- weather - A description of the weather conditions during the time of the collision.
- roadcond - The condition of the road during the collision.
- lightcond - The light conditions during the collision.

Data preparation has consisted in:

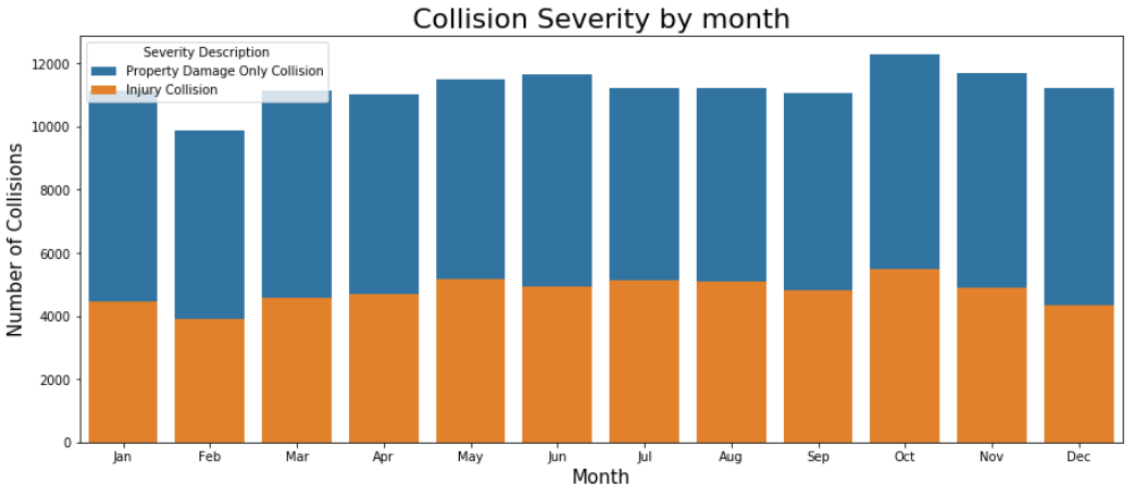
- Lowercase column names to be able to work with them in an easier way.
- Study each feature and select the ones which are interesting to work with. Several features have been not selected due to lack of enough data.
- Remove values with no hour in the incdttm column. When plotting the correlation between data and hour of the day, a great amount of accidents is distinguished at 00:00 am. When looking at accidents in the previous and latter hours, these values seems wrong. After checking the data, a great amount of data has no hour stated in the column incdttm and pandas assign it to 00:00 am. In order to study the influence of the hour in car accident severity, these values must be removed.
- Remove null values from these columns.
- Remove unknown these values.
- Convert data types to integers in order to run the models correctly.

In the following section I will use graphs to gain insight into the data. The features used in the model will be examined in more detail.

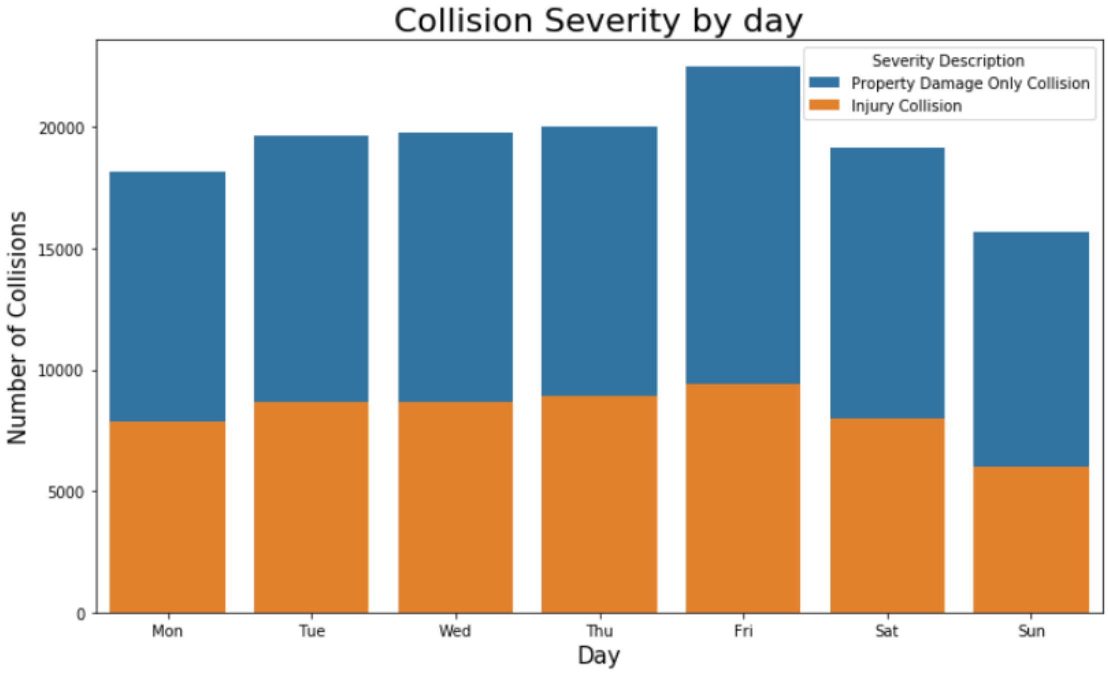
Firstly, number of collisions will be studied along time: yearly, monthly, daily and hourly:



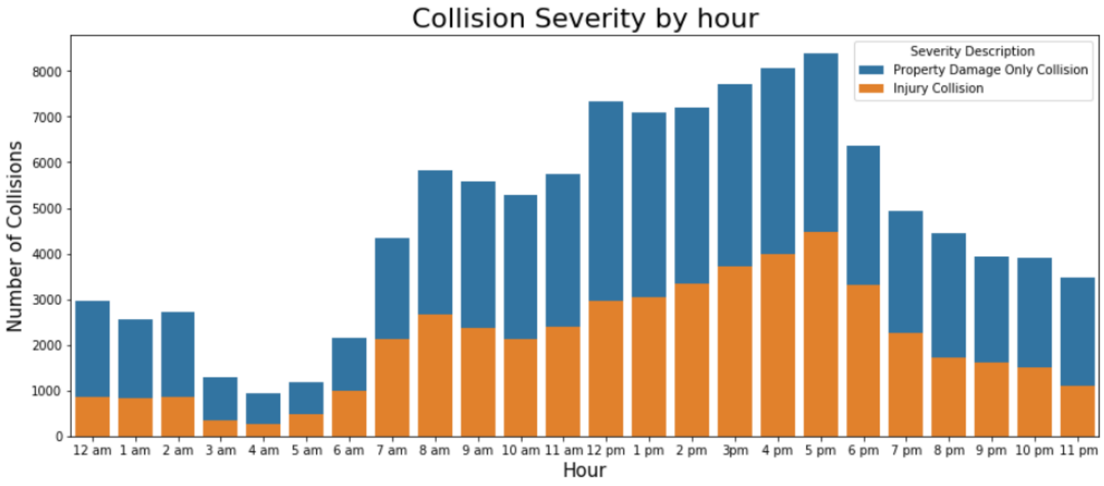
Traffic accidents have decreased over time. Although two rising periods can be distinguished between 2005 and 2008 and 2014 – 2016.



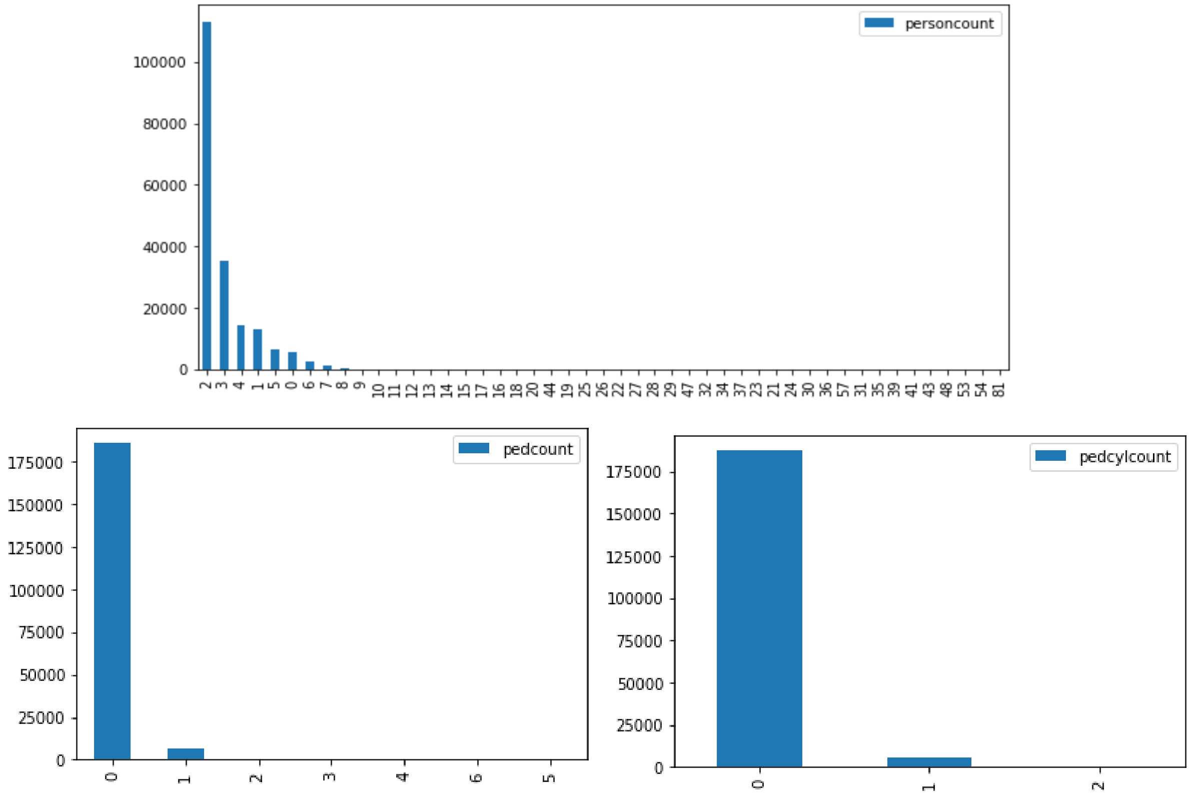
February is the month with the lowest number of collisions, and these are less severity. On the contrary, October is the worst month.



In the daily graph it can be observed that Friday is the worst day and Sunday is the one with fewest accidents.

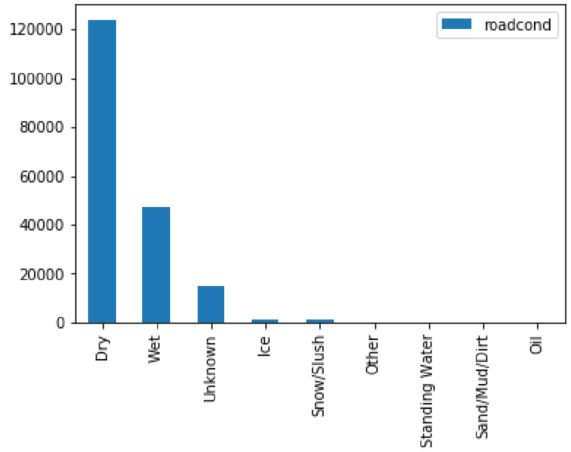


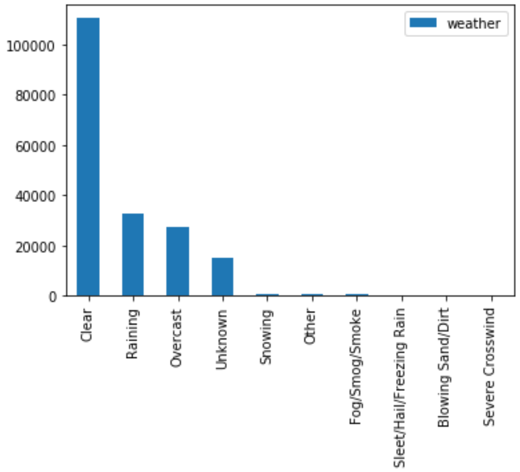
Here it can be noted that collisions go up throughout the day, starting at 7 am and having the peak at 5 pm. Night hours have less traffic accidents than daytime.

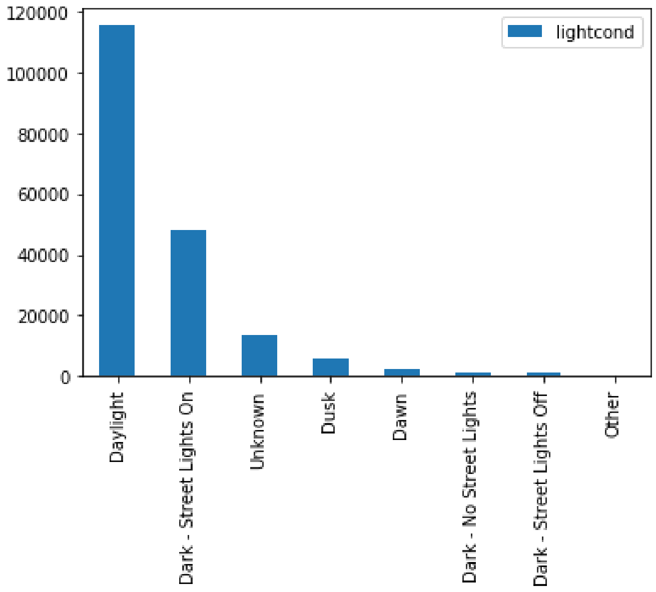


Most traffic accidents involve 2-3 people. Normally, there is no bicycle or pedestrian involved

Analyzing weather, road and light condition, these are usually good conditions for driving:

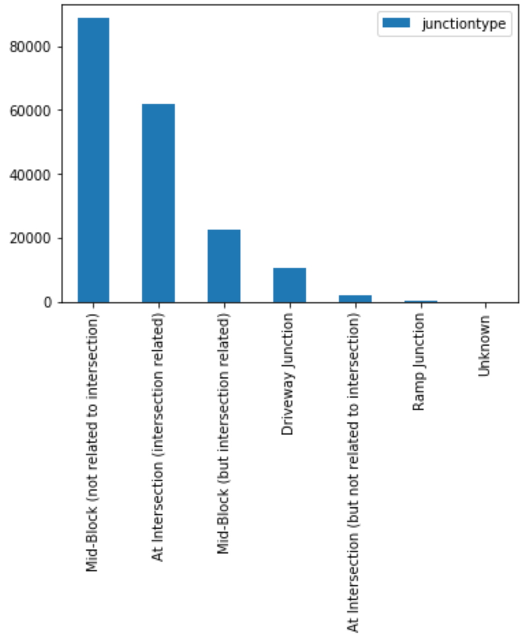


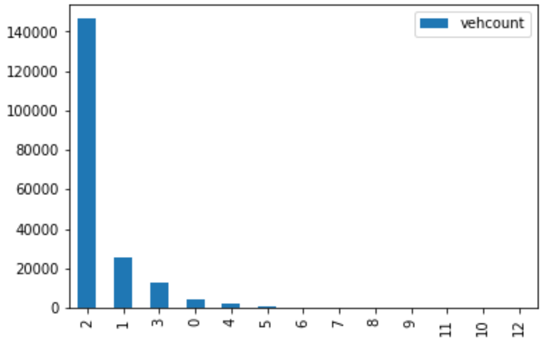


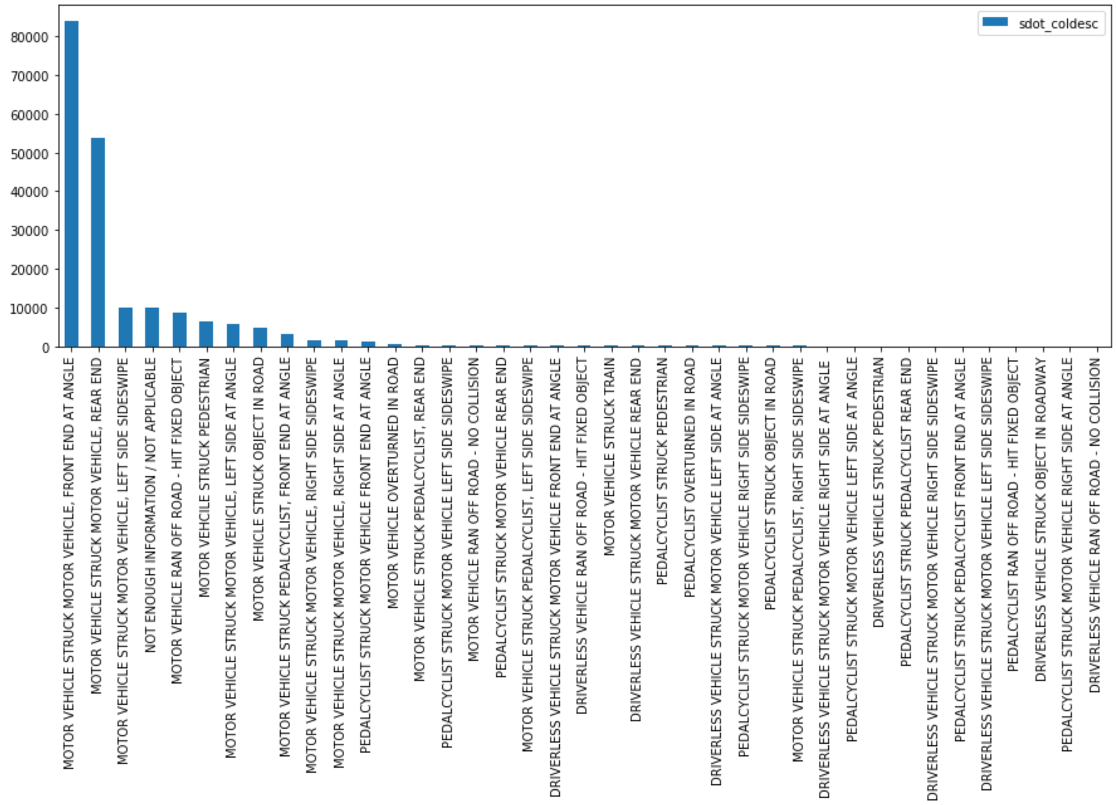


Lastly, it has been studied in which type of junction or intersection the accidents tend to occur and the description of the accident. Most accidents are caused by two vehicles, impacting at the rear or front end.

Lastly, it has been studied in which type of junction or intersection the accidents tend to occur and the description of the accident. Most accidents are caused by two vehicles, impacting at the rear or front end.

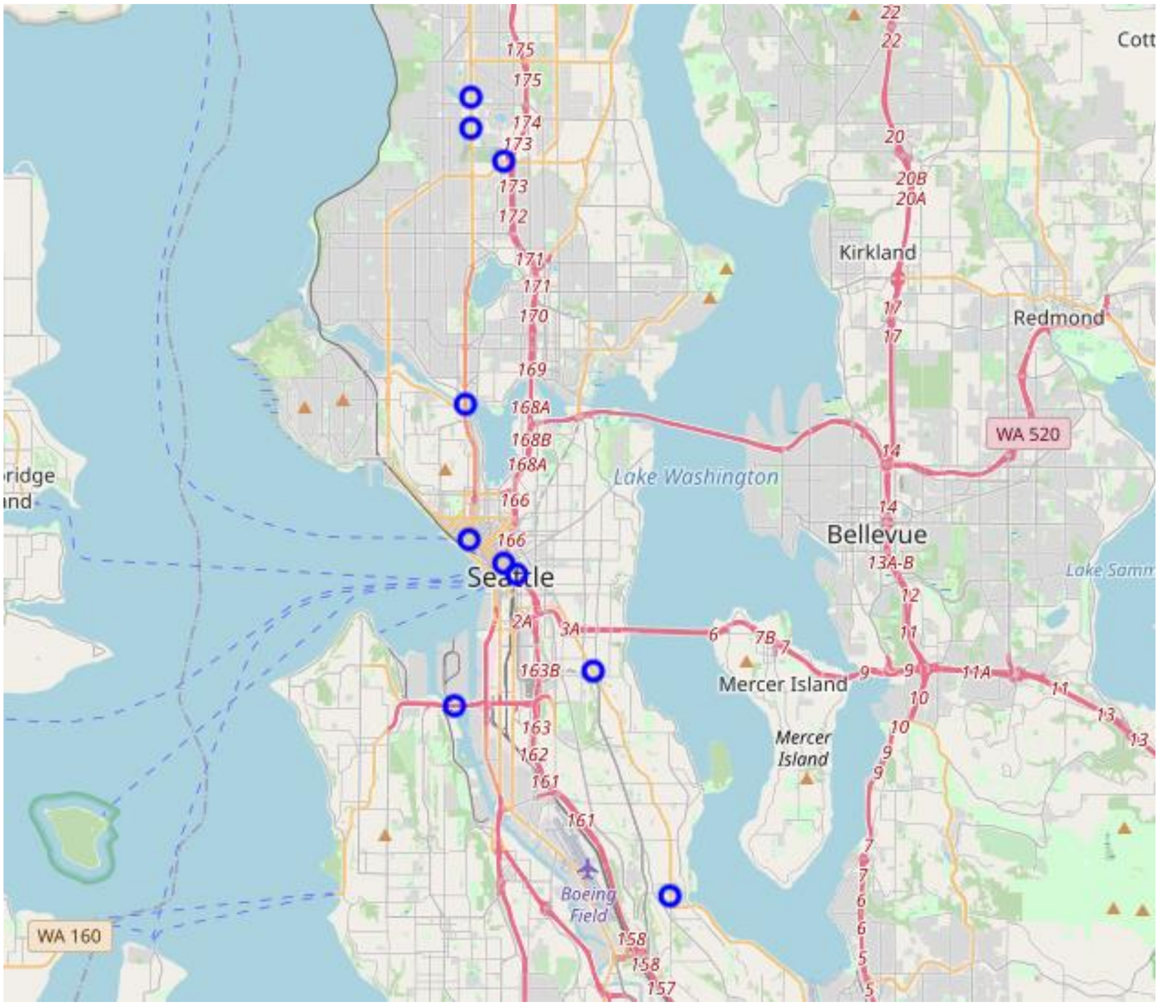






MAP OF 10 LOCATIONS WITH MOST TRAFFIC ACCIDENTS

The heavier the traffic is, the more likely there is to be an accident. Using coordinates from columns ‘x’ and ‘y’ it is possible to plot the top 10 locations with the higher rate of traffic accidents. Also, these locations may have the highest amounts of traffic jams.



Methodology

In order to process the data, this has been treated. As I mentioned in the Data Section, values with no hour in the incdttm column have been removed as well as unknown and null values from the selected features. Additionally, data types have been converted to integers in order to run the models correctly and data has been normalized.

I found several columns with a low amount of data. For example, 'speeding' could be a good feature for this analysis but only a small amount of the data has this column not null.

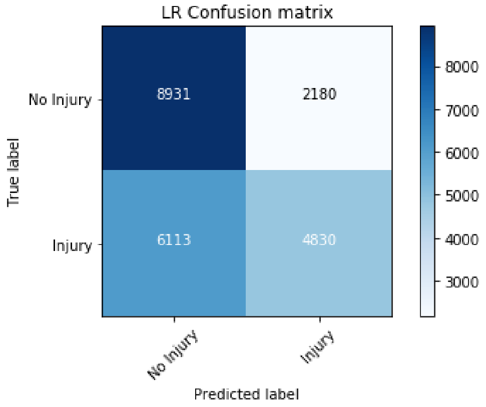
The next step is to balance the data. Standardizing data scales it to unit variance and is used so that all features' scales are equal. There are more collisions without injury than with injury so it is needed to undersample severitycode 1. Then, data has been split into training data (80%) and testing data (20%). In the final step, we standardize the data to feed into the classifiers.

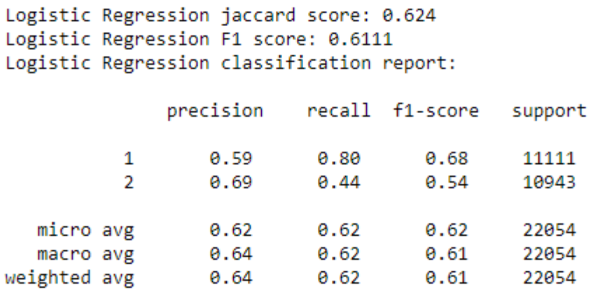
I chose the Logistic Regression, Decision Tree, K-Nearest Neighbors and Support Vector Machine classifiers to evaluate which create the most accurate model.

Results

Confusion matrix and report is given for all the classifiers studied.

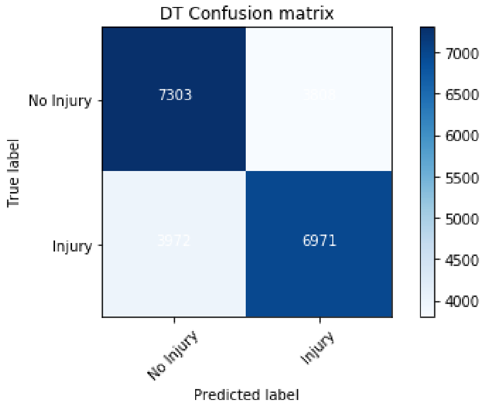
LOGISTIC REGRESSION CLASSIFIER





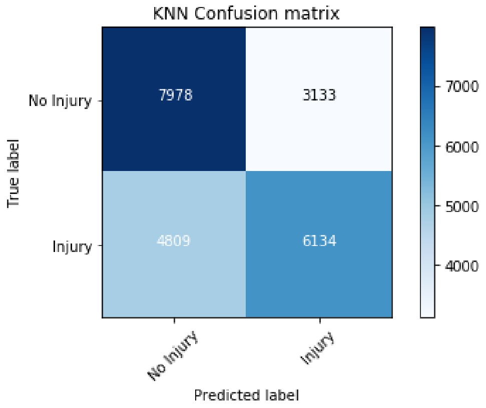
```
Logistic Regression jaccard score: 0.624
Logistic Regression F1 score: 0.6111
Logistic Regression classification report:

              precision    recall  f1-score   support

           1       0.59      0.80      0.68     11111
           2       0.69      0.44      0.54     10943

   micro avg       0.62      0.62      0.62     22054
   macro avg       0.64      0.62      0.61     22054
weighted avg       0.64      0.62      0.61     22054
```

DECISION TREE CLASSIFIER



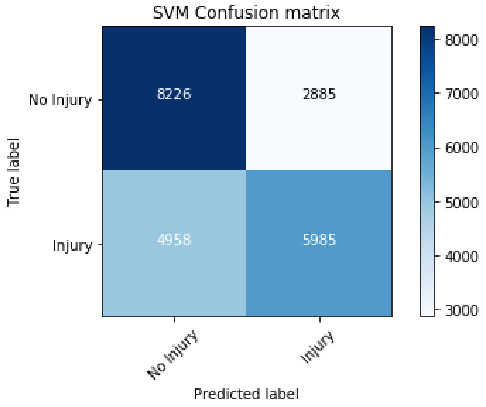
```
Decision Tree jaccard score: 0.6472
Decision Tree F1 score: 0.6472
Decision Tree classification report:

              precision    recall  f1-score   support

           1       0.65      0.66      0.65     11111
           2       0.65      0.64      0.64     10943

   micro avg       0.65      0.65      0.65     22054
   macro avg       0.65      0.65      0.65     22054
weighted avg       0.65      0.65      0.65     22054
```

K-NEAREST NEIGHBORS (KNN) CLASSIFIER



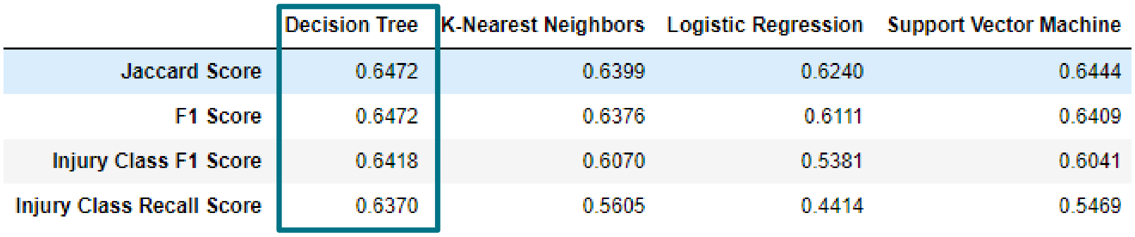
```
KNN jaccard score: 0.6399
KNN F1 score: 0.6376
KNN classification report:

              precision    recall  f1-score   support

           1       0.62      0.72      0.67     11111
           2       0.66      0.56      0.61     10943

   micro avg       0.64      0.64      0.64     22054
   macro avg       0.64      0.64      0.64     22054
weighted avg       0.64      0.64      0.64     22054
```

SUPPORT VECTOR MACHINE (SVM) CLASSIFIER



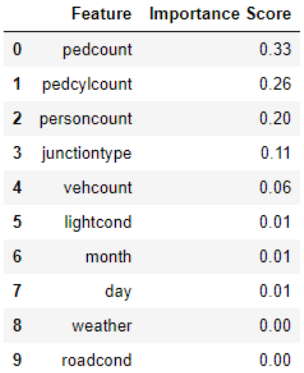
```
SVM jaccard score: 0.6444
SVM F1 score: 0.6409
SVM classification report:

              precision    recall  f1-score   support

           1       0.62      0.74      0.68     11111
           2       0.67      0.55      0.60     10943

   micro avg       0.64      0.64      0.64     22054
   macro avg       0.65      0.64      0.64     22054
weighted avg       0.65      0.64      0.64     22054
```

In the following table Jaccard, F1 scores are represented for all the classifiers. F1 and Recall scores have been ploted as well.

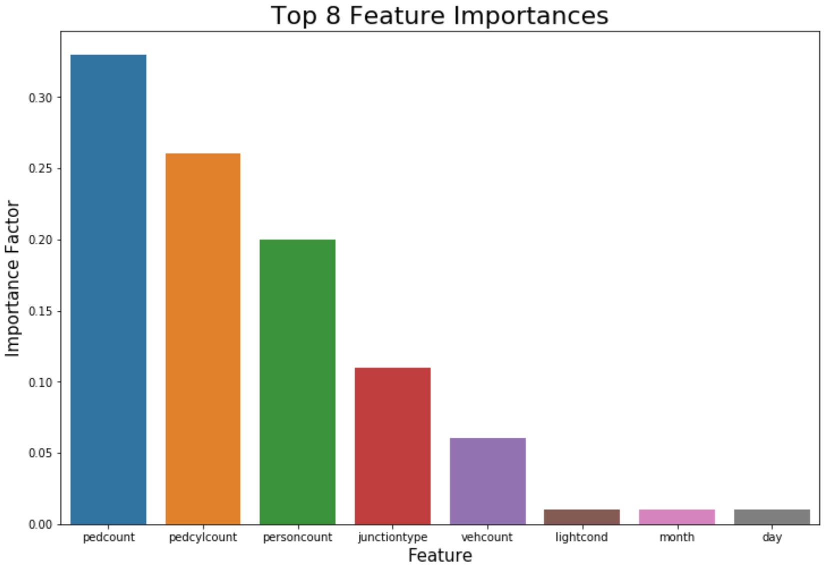


	Decision Tree	K-Nearest Neighbors	Logistic Regression	Support Vector Machine
Jaccard Score	0.6472	0.6399	0.6240	0.6444
F1 Score	0.6472	0.6376	0.6111	0.6409
Injury Class F1 Score	0.6418	0.6070	0.5381	0.6041
Injury Class Recall Score	0.6370	0.5605	0.4414	0.5469

Running the model for Decision Tree Classifier, it can be seen which features influence the most in the model:



	Feature	Importance Score
0	pedcount	0.33
1	pedcylcount	0.26
2	personcount	0.20
3	junctiontype	0.11
4	vehcount	0.06
5	lightcond	0.01
6	month	0.01
7	day	0.01
8	weather	0.00
9	roadcond	0.00



Discussion

Jaccard and F1 Score were used to analyses the results of the ML models. Each model had similar metrics but Decision Tree was the strongest. Choosing different hyperparameters values helped to improve the predictive power of the models.

Features with the biggest influence in the model were 'pedcount' and 'pedcylcount'. This seems logical, as pedestrians and cyclists receive the impact with low protection, resulting in possible injuries. Other variables with high influence were ‘pesoncount’, ‘junctiontype’ and ‘vehcount’.

It is remarkable that light, weather, road conditions have low influence in the model meaning they do not affect an accident to be mild or severe.

Influence of month and day have been taken into account. In a previous version of the model, ‘hour’ was taken into account as well. As it is depicted in the chart, during daytime appears to be more accidents than during nighttime, but as according to the model, it has no influence on the severity of the accident. Due to the great amount of time it required to run the model it was dropped in the final version.

Plotting top 10 locations with highest traffic accidents can help the shareholders to take actions to this specific points. Top 10 locations concentrate in the city center (3 locations), at the north of the city (3 locations), 2 bridges and 2 locations of the same avenue. As it was mentioned previously, these locations may suffer the highest amounts of traffic jams.

Conclusion

In this capstone the traffic accidents of Seattle since 2004 have been studied. Several models have been built helped by python and sklearn. In the process, data has been analyzed, plotted and processed to achieve the best model possible.

Some variables had to be dropped as the amount of data was low; in other cases, accidents were not taken into account because null or unknown values for a studied variable.

It must be highlighted that when analyzing higher number of accidents by ‘location’ there were several that had not coordinates associated, making difficult to plot it on a map. Coordinates has been studied instead.

It is interesting the influences of the variables studied in the outcome of an accident. From this analysis it can be concluded that a special attention has to be made in protect pedestrians and cyclists as they have contributed the most to the severity of the accident.

On the other hand, specific actions should be taken to areas with higher rates of accidents, and possible of traffic jams. Studying in more detail these areas and lowering their rates would help the city and people to be less stressed and to reduce the number of injured.