# Applied Data Science Capstone
# Car accident severity

By

Kevin Pursey

## Business Understanding

There is a worldwide competition underway! All major cities want to attract new businesses and workers that can grow their city into a model of prosperity. Seattle is in this competition.

One of the ways Seattle can compete better is to improve its livability score, specifically the transportation section of its livability score. With 49% of workers driving to work, any improvements in this section would have a positive impact and go a long way to improving daily life in Seattle.

Within the city of Seattle, transportation is a major factor when resident's rate their quality of life. The better the transportation the higher the quality of life. Within the category of transportation there is public and private with both experiencing accident related delays. A lot of factors contribute to delays, example: time of day, traffic volume, road conditions, weather conditions and traffic accidents. Currently, drivers are only aware of delays along their route when they encounter them. The more severe the incident the longer the delay as emergency response is generally proportional to the incident. The more severe the more emergency vehicles and first responders required. The city would like to help drivers avoid delays by giving them the ability to understand the current traffic situation, thus allowing them to adjust their driving plans. Access to such planning information is expected to improve a commute and thus improve daily life in Seattle.

As a key stakeholder, Seattle city, would like to be able to present a graphical depiction of traffic incidents on the city website. The collision dataset does not contain the amount of time required to clean an accident. In lieu of this data we will assume that incidents only impact traffic for a few hours and the graphic will limit its content to vehicle accidents reported within the previous two hours. This information will be accessible to everyone with the main user community expected to be the personal vehicle driver.

In addition, there will be the ability to predict the severity of an accident. The assumption in this model is that a prediction of severity will be based on the initial information provided by a caller. The caller, at minimum, should provide weather, road condition, light condition and include the number of vehicles involved in the collision.

## Data acquisition and preparation

### Data source

The data used will be a subset of the information provided in the Seattle *Data-Collisions.csv*. The site also provides a metadata pdf which describes each element within the Data-Collision file.

The Data-Collision file contains 194673 accident records spanning the years 2004 to 2020. These records cover both vehicle and non-vehicle accidents. I will be focusing on vehicle accidents and as such will limit the data to incidents involving vehicles.

### Data Preparation

Having access to this large volume of historical data provides a number of advantages in that it makes it easy to identify trends and increases the probability of being able to obtain a usable data-subset without having to manipulate the raw data by filling in missing values.

The following steps were taken to prepare the *Collision-Data.csv* source data:

- removal all incomplete accident records
- removal of accident records that did not involve at least one vehicle

The data preparation steps resulted in the initial 194673 accident records was reduced to 183950 accident records. I think this is sufficient to identify trends, model severity prediction.

I further sub-divided the data into three groups with different purposes. Group one was the complete 183950 records for uses in plotting accident trends. The second group contained accident records from 2019-01-01 and forward for use in the train test split when modeling and predicting severity, 9861 records. The final group is comprised of the most recent six accident records with the intended purpose of demonstrating how the accidents may appear as plots or icons on a Seattle city map.

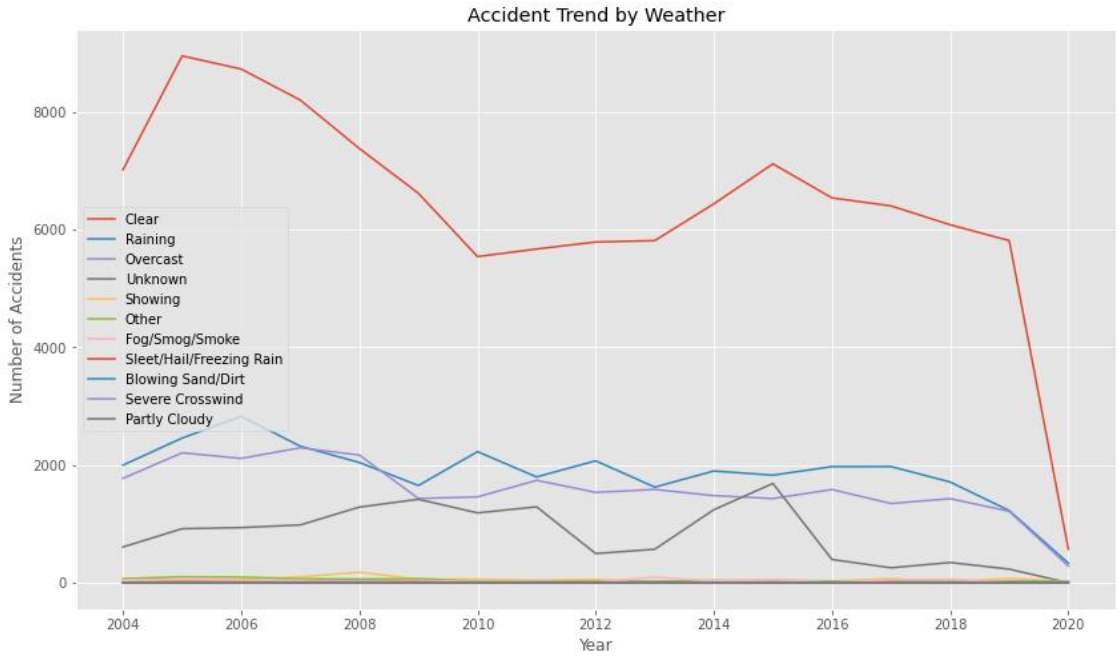
### Data observations

Each accident record includes a numeric severity code (SEVERITYCODE) with the following meanings: 0 = unknown, 1 = prop damage, 2 = injury, 2b = serious injury, 3 = fatality. I reviewed all the records from the Data-Collision file before removing any records and found that the severity code values were limited to either the value of 1 or 2. After I completed the data preparation steps, there were 128143 records with a severity code of 1 and 55807 records with a severity code of 2. This imbalance surfaces in the prediction accuracy and highlights that severity codes outside the dataset will not be predicted accurately if introduced in future accidents. This why any modeling should be viewed as iterative and revisited when new data becomes available.

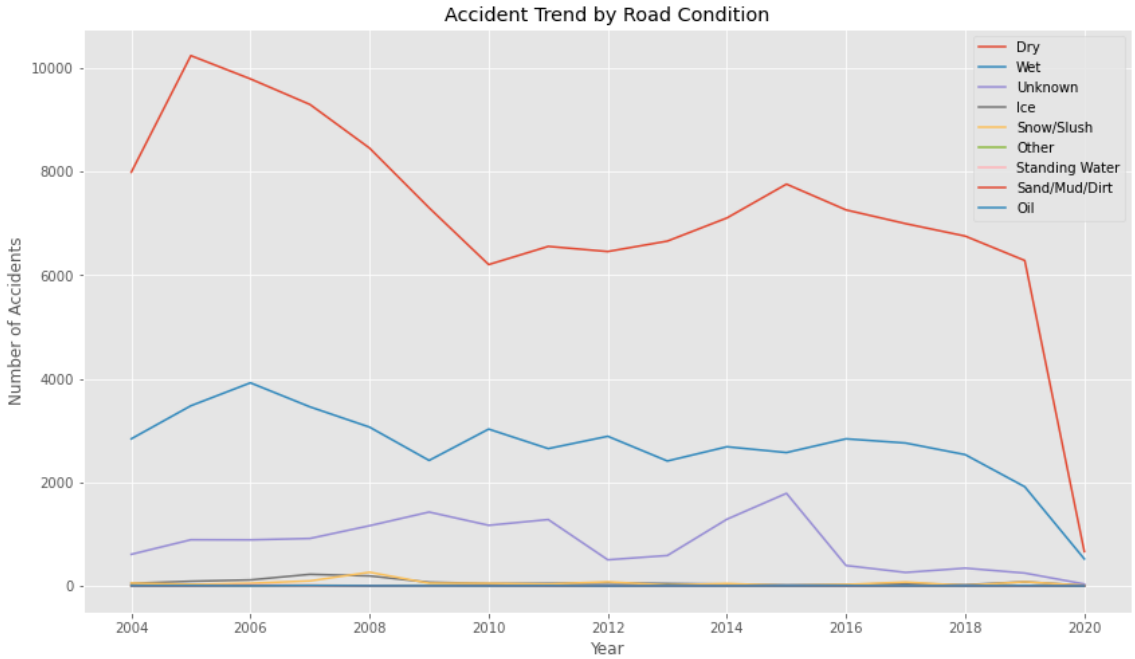
I was able to identify trends for the weather, road conditions, light conditions as well as the daily timing of accidents.

The weather trend, shown on the line graph, indicates that Clear, Raining and Overcast days have the most accidents and that they have always been the top three categories. Looking at the trend it is

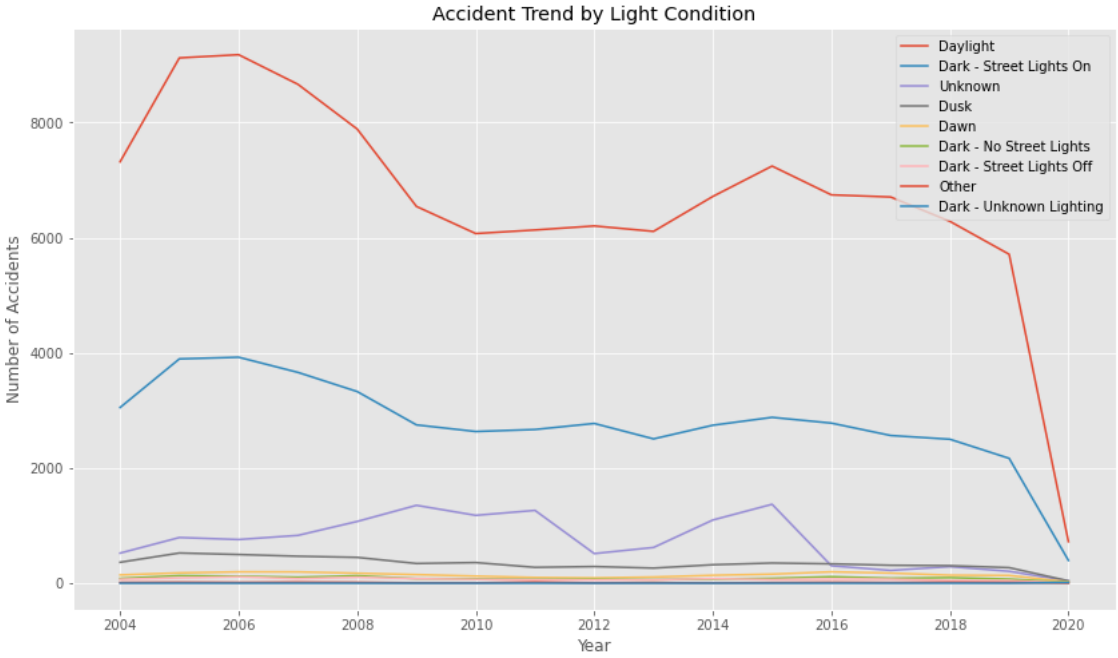
obvious that Clear days have always had more accidents, by multiples, and that this is not an emerging or changing trend.



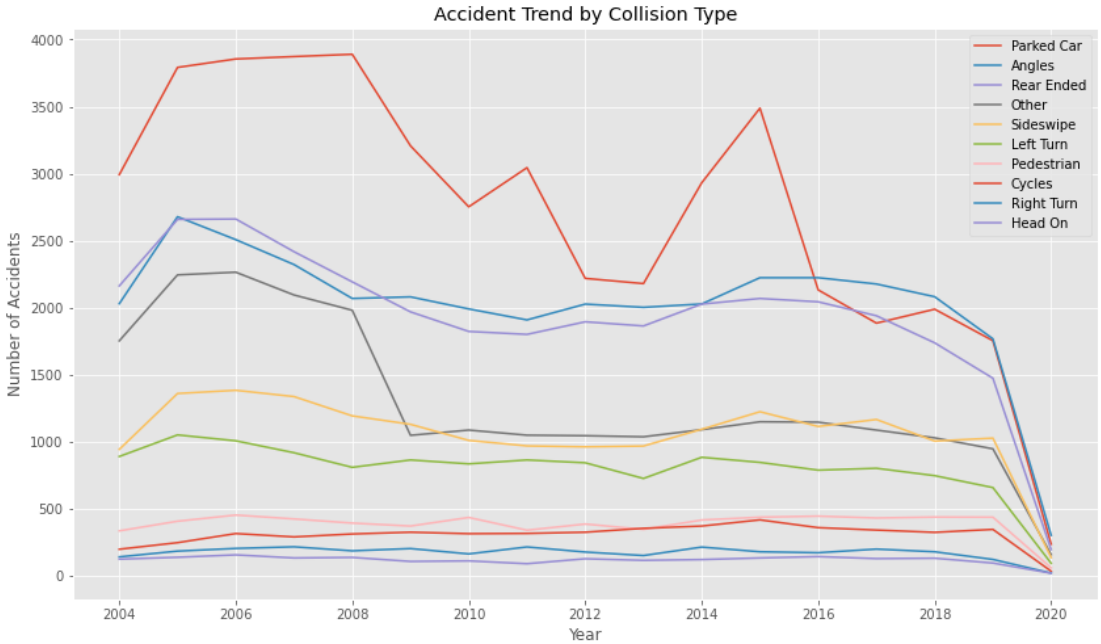

With the dataset spanning 2004 to 2020 the recorded road conditions at the time of accidents really do span every season. The prominent trend is that top three road conditions have remained so since 2004. Most accidents happen on Dry roads with the second being Wet roads. The interesting think I see with the Unknown road condition is that the trend appears to be decreasing which leads me to believe that historically this option my of been a ‘go to’ or ‘default’ selection in recording an accident by with the decrease it appears the more effort is being made in more accurately recording accidents.



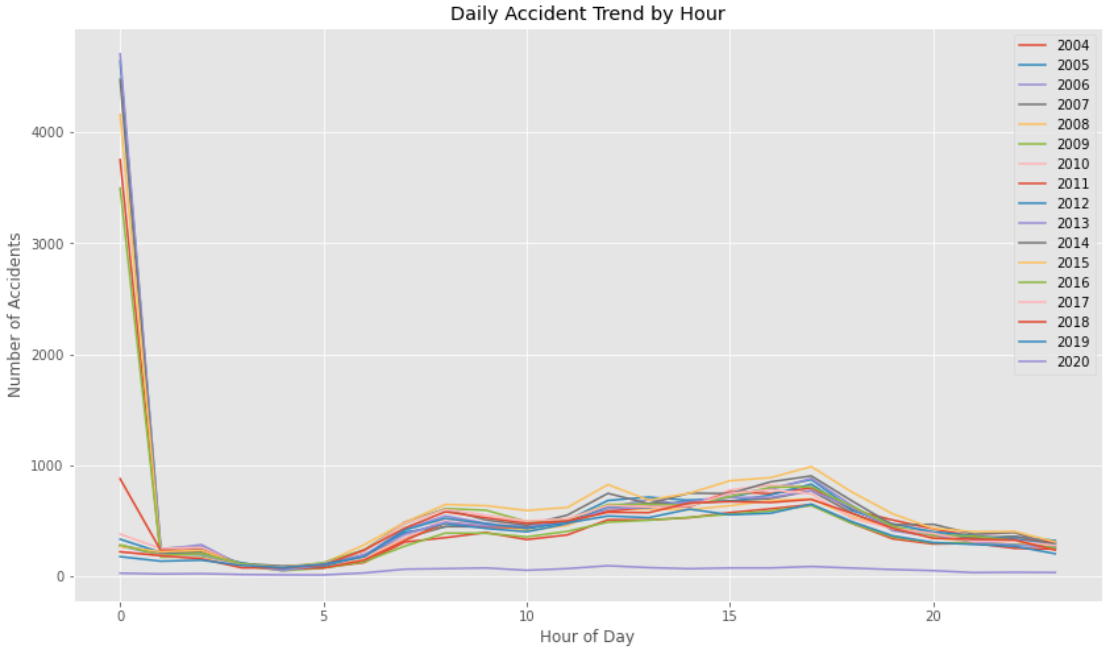
The light condition trend shows the most accidents happening in Daylight conditions. This trend has been unchanged since 2004. Like the trends in road conditions, the Unknown light condition trend is decreasing, and I suspect it is for the same reason, better processes in recording accidents.




The collision type trend shows the top three types Parked Car, Angles and Read Ended are approaching an equal distribution.



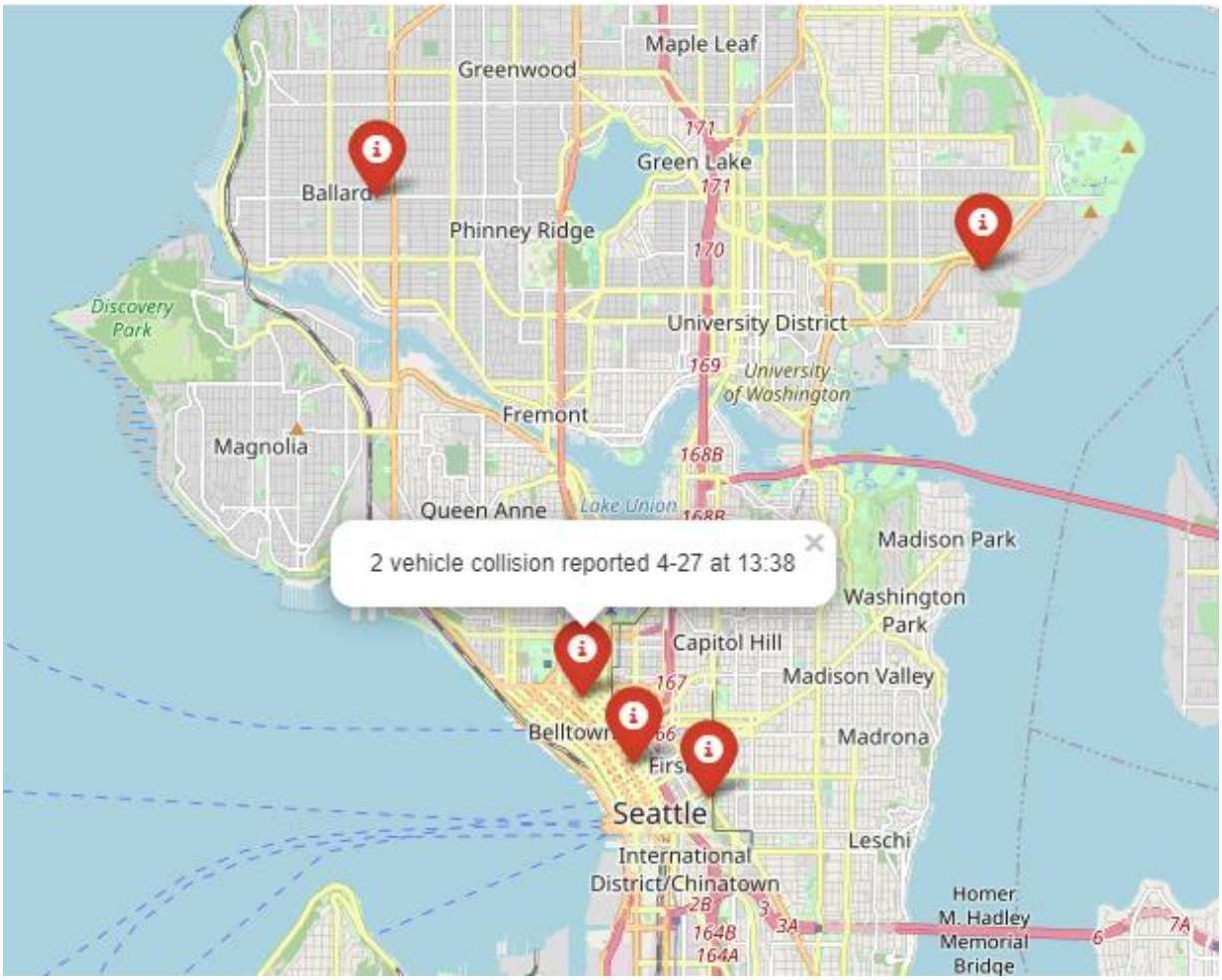
When we look at the full range of accidents, I can show the daily hour accident rates for each year, comparing year over year. We can see that the midnight to 1am hour has and remains a high accident period. However, and more clearly seen are the three spikes during the day. The hours of 8am, noon and 5pm. This daily trend is not surprising and confirms what most might believe; high traffic volumes means more accidents. The other trend highlighted here is that our individual driving habits may not be improving thus we can probably expect a continuing trend in the coming years.



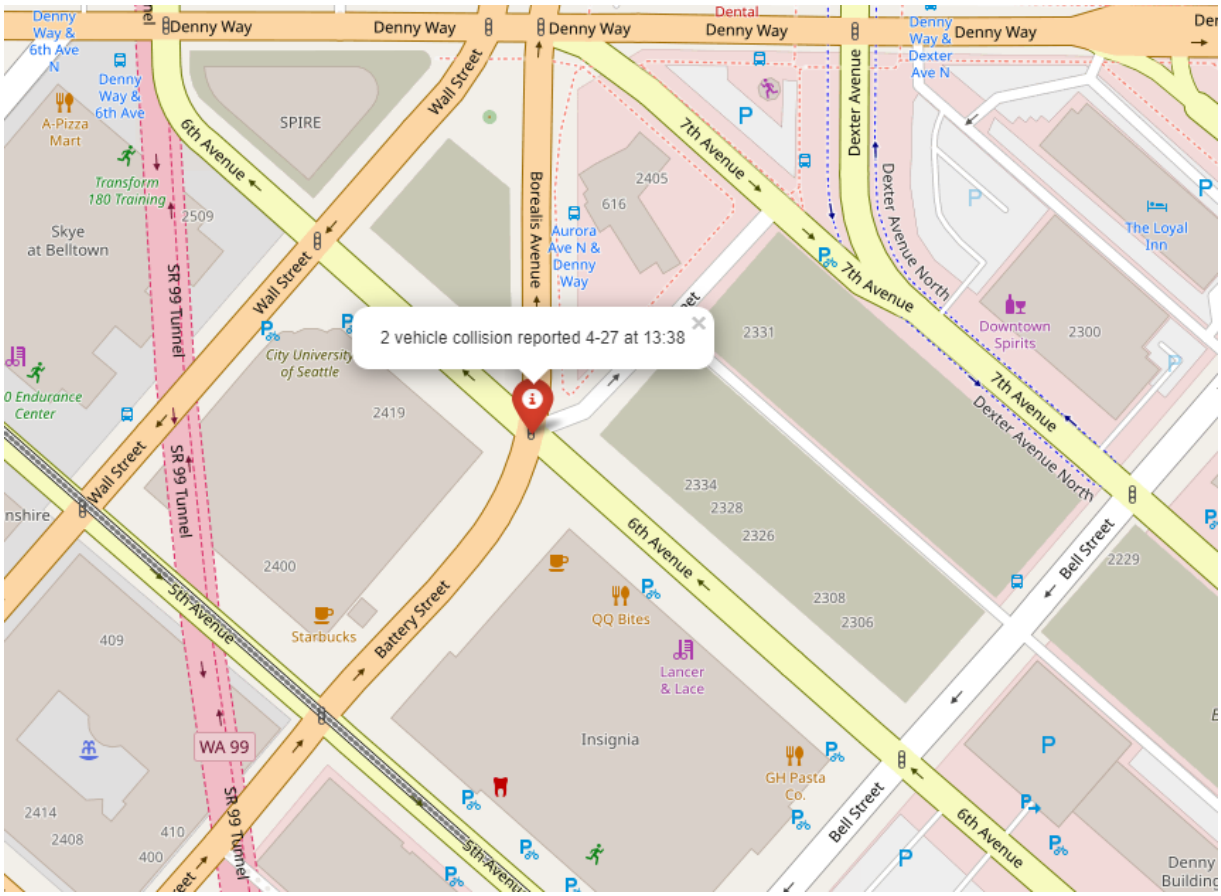
## Modeling

### Seattle City Map

A goal of the capstone was to provide a way for drivers to avoid delays. To address this, I have chosen to use folium for presenting a Seattle city map that can indicate where accidents are located. The accident location comes from the X and Y coordinates associated with each accident record. As shown below, I have harvested the accident information which is most likely to provide drivers with an easy to understand picture of the traffic conditions. Now drivers can plan accordingly and avoid delays.



The following image demonstrates the level of detail available when a user zooms on a specific accident area of the map.




## Predicting Severity

In addressing the goal of predict accident severity I took the approach that a determination of severity would be the output of an accident being reported and would inform the type of response required by the first responders. I selected a feature set that I believe corresponds to the information a witness would be able to easily provide when reporting an accident.

The selected feature set:

- number of vehicles involved in accident (VEHCOUNT)
- collision type (COLLISIONTYPE)
- weather (WEATHER)
- road condition (ROADCOND)
- light condition (LIGHTCOND)

## Train Test Split

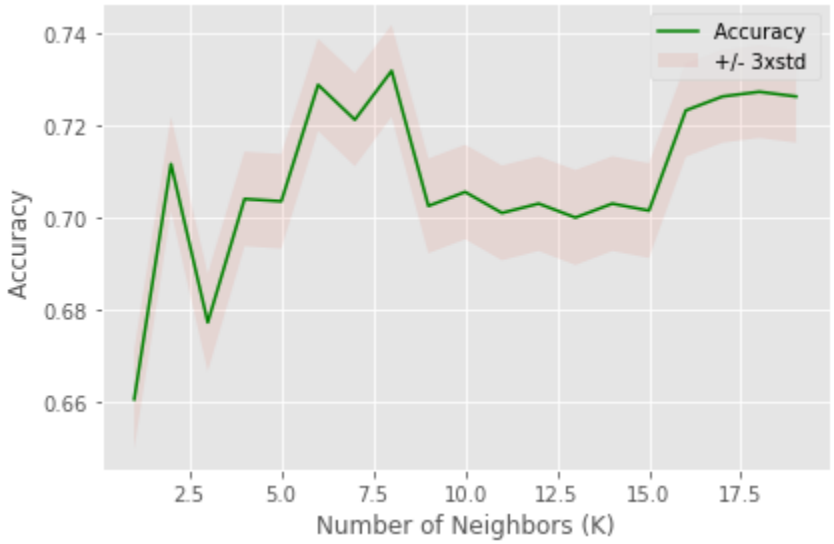
In order to train and test each model the dataset of 9861 accidents was split so that 80% could be used for training and the remaining 20% used to test.

- Train set = 7888 records
- Test set = 1973 records

I created four models to determine which would have the highest accuracy, KNN, Decision Tree, Support Vector Machine and a Log Regression.

## KNN model

In running numerous iterations of the KNN model I found that the best accuracy was achieved when using 8 neighbors. As shown in the Model Evaluation grid, the KNN model has the best accuracy with the selected feature set.




## Model Evaluation

The following table provides a view of the accuracy achieved using each model. The KNN model is the most accurate in determining the accident severity based on the chosen feature set.

| ***Model \ Accuracy*** | Jaccard | F1-score | Log |
|---|---|---|---|
| KNN (neighbor=8) | 73.19 | 70.09 | |
| Decision Tree | 73.14 | 67.60 | |
| SVM | 73.19 | 68.73 | |
| Log Regression | 72.98 | 68.04 | 52 |

## Conclusion

Given that only two of the five severity codes have been used in recording accidents it can be expected that the use of additional severity codes in the data should act as a trigger to re-model the prediction and either change or confirm existing models.

The data and models in this report are limited to vehicle accidents but could easily be expanded to include non-vehicle accidents if it is determined that the stakeholders should be expanded beyond drivers.

With regards to the Seattle city map accident plots, the refresh interval would need to be refined to ensure the most recent information is available at peak use time. This should include communication with the source file owner once it is determined what pear times are. To begin, I would suggest doing a refresh of the map every 10 minutes leading up to the peak periods of 8am, noon, 5pm and midnight. Outside of peak periods, a map refresh could be reduced to every 15 minutes.