Analysing severity levels of collisions in Seattle

Capstone Project

Predicting severity level of collision in Seattle is valuable for citizens and local government

- Predictive model which will provide reasonable prediction if under certain conditions is higher probability of collision with a given severity
- Citizens of Seattle should be interested in this problem because knowing the relationships between conditions and likelihood of collision can save their money and life
- Goal is to provide them information about current situation on a roads and possible dangers – so it will be possible to avoid them

The problem being analyzed and target audience

- What is the severity magnitude for a collision to occur in Seattle?
- Audience consists of people who lives in Seattle or travel through it and goal is to provide them information about current situation on a roads and possible dangers.
- Audience should be interested in this problem because knowing the relationships between conditions and likelihood of collision can save their money and life.

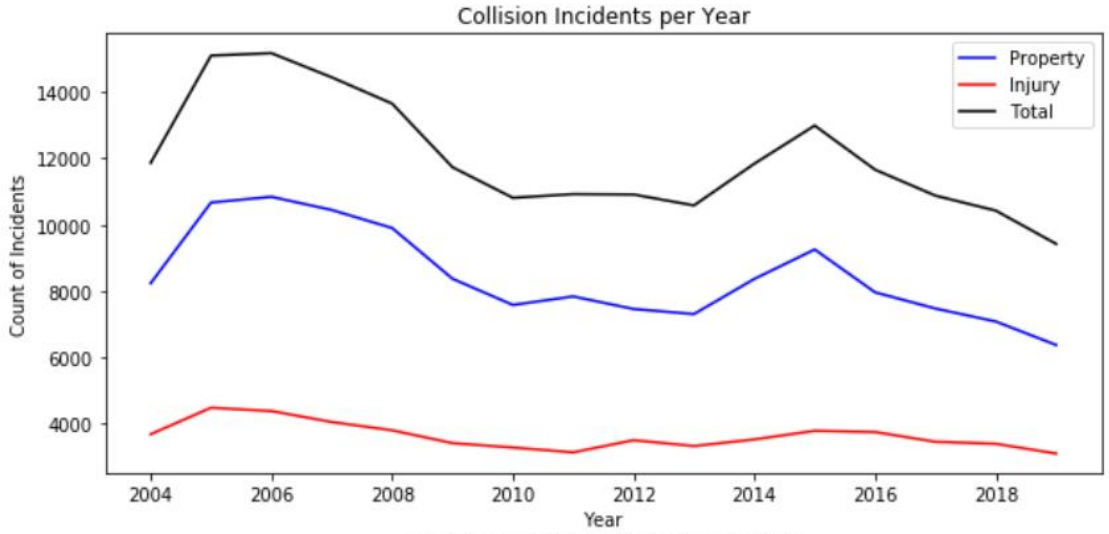
Data description

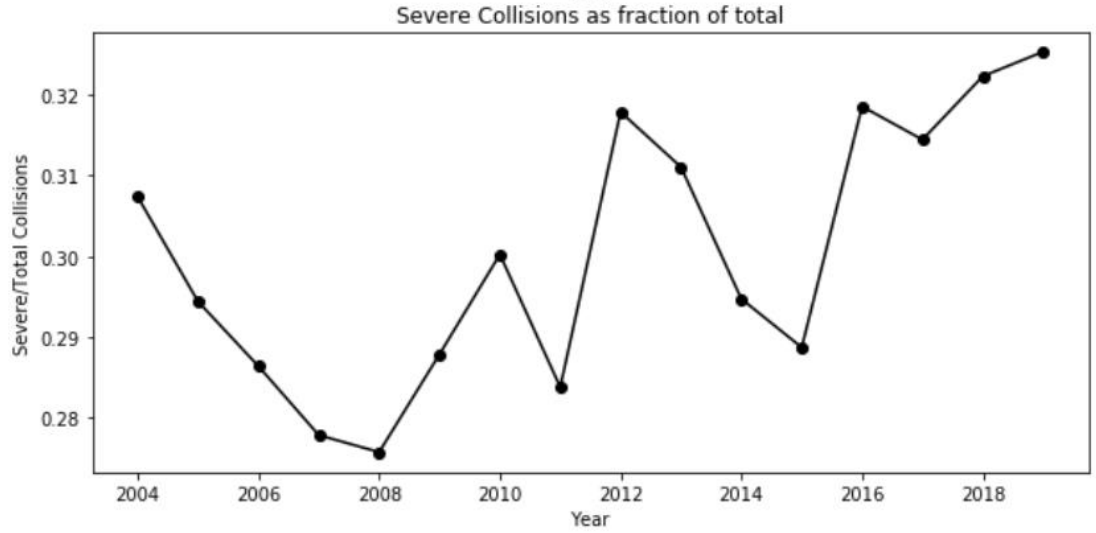
- This dataset contains about 6.5 mln car collisions records. The data include all types of collisions. Collisions will display at the intersection or mid-block of a segment. The data are collected from a timeframe of 2004 to Present. The data is updated on a weekly basis and is provided by SPD and recorded by Traffic Records.
- In the raw dataset were 194 673 rows and 38 columns
- Duplicate, highly similar or highly correlated features were dropped.

Data focus

The analysis focuses on the following columns because they are relevant to the analysis:

- COLLISIONTYPE: A keyword describing the collision, eg 'head-on', 'angled', 'cycles', etc.
- PERSONCOUNT: Total number of people involved in the collision.
- PEDCOUNT: The number of pedestrians involved in the collision.
- PEDCYLCOUNT: The number of bicycles involved.
- VEHCOUNT: The number of vehicles involved in the collision.
- INCDATE/INCDTTM: Date and time recordings for the incident records.
- INATTENTIONIND: If collision was due to inattention.
- UNDERINFL: If driver was under influence of drugs/alchohol.
- WEATHER: Weather at the time of incident.
- ROADCOND: Condition of road at the time of incident.
- LIGHTCOND: Light conditions at the time of incident.
- SPEEDING: If speeding was a factor in the collision.

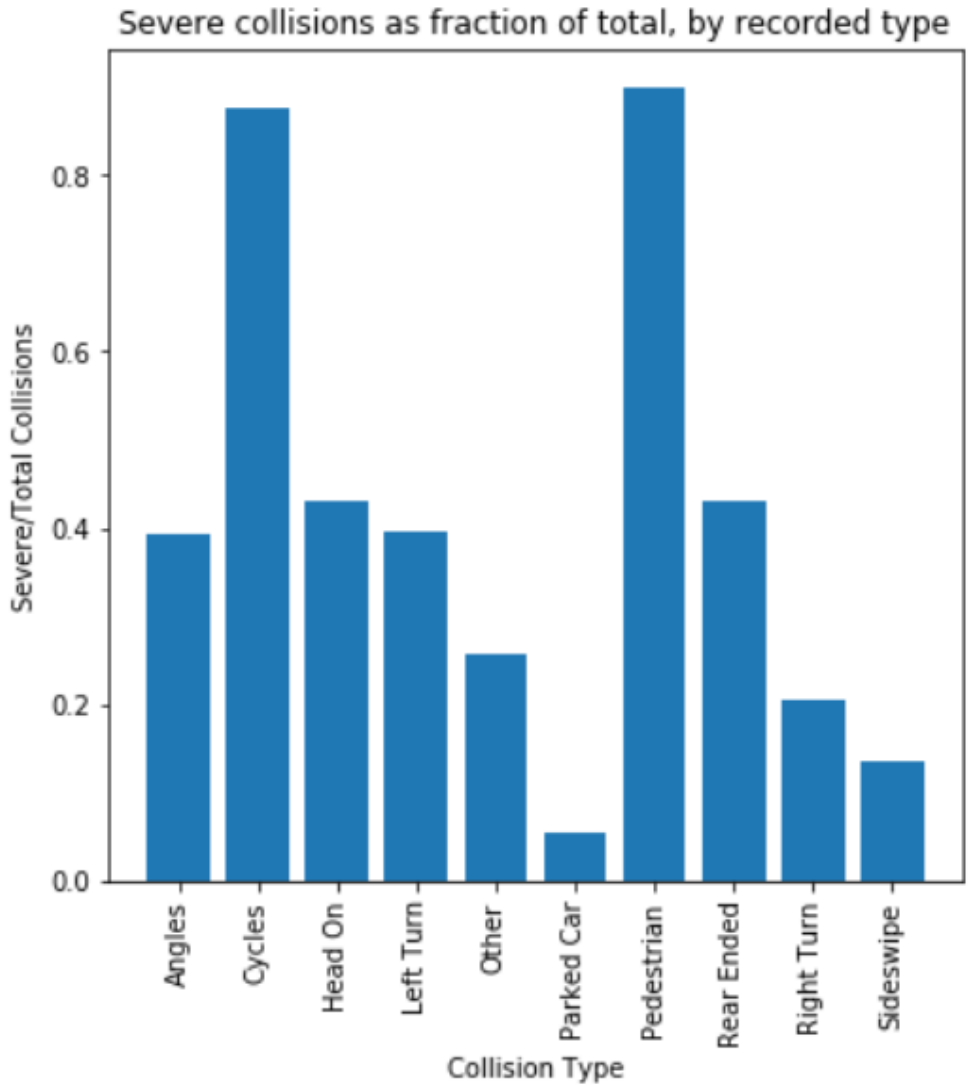




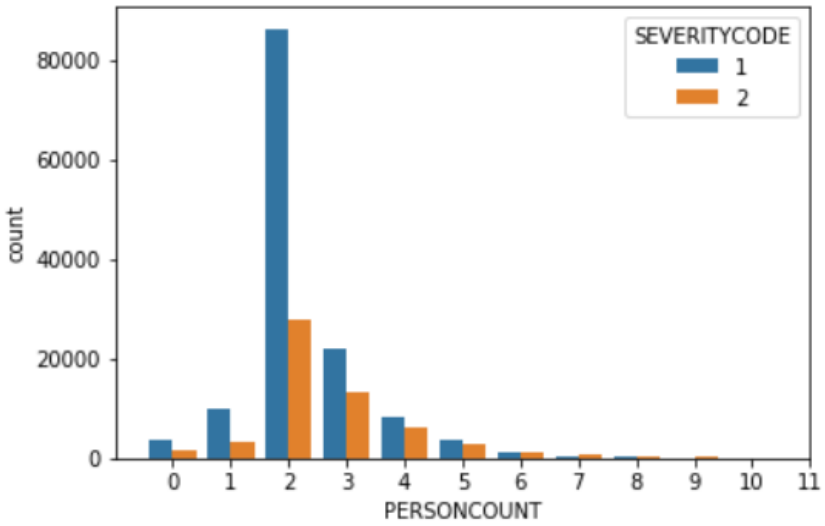
From the data, we can clearly see that the major amount of collisions are property incidents and the amount have dropped between 2004 and present. Despite the decrease in the overall amount of collisions, the severity in general have fluctuated reaching its lowest in 2008 and its highest in 2020.

```
Total collisions in number:  194673
Collisions on weekdays in number:  145329
Collisions on weekends in number:  49344
```

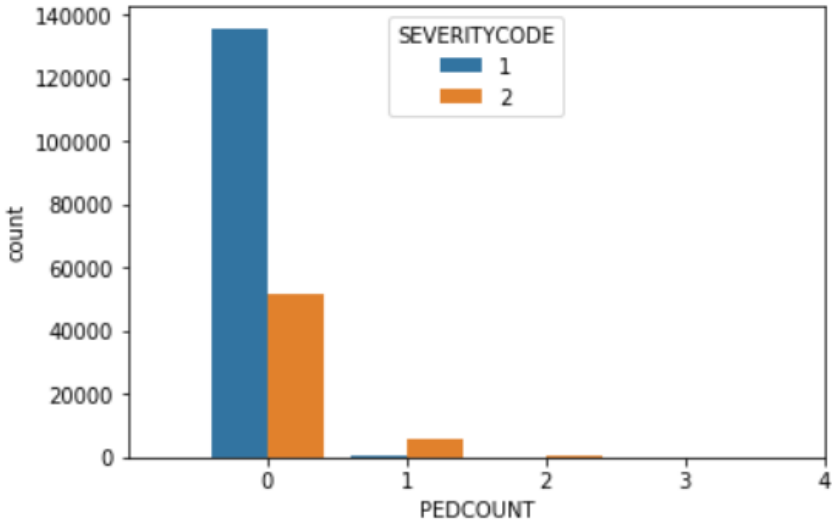
As expected the total volume of collisions is higher for weekdays than weekends. The volume gradually increased to a peak on Fridays before dropping on Saturday and Sunday. Separately, The severity and amount of collisions is at its highest during the middle of the day.



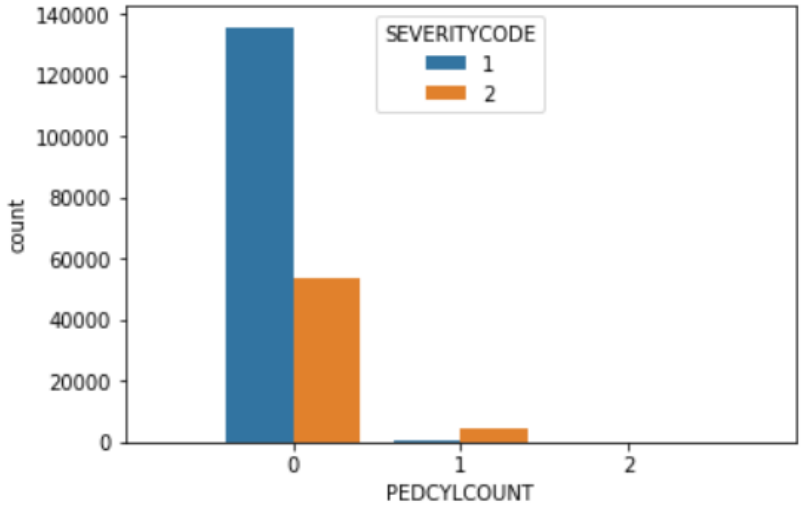
From the bar chart we can see that the highest number of collisions and severity occurred as a result of cycles and pedestrians.



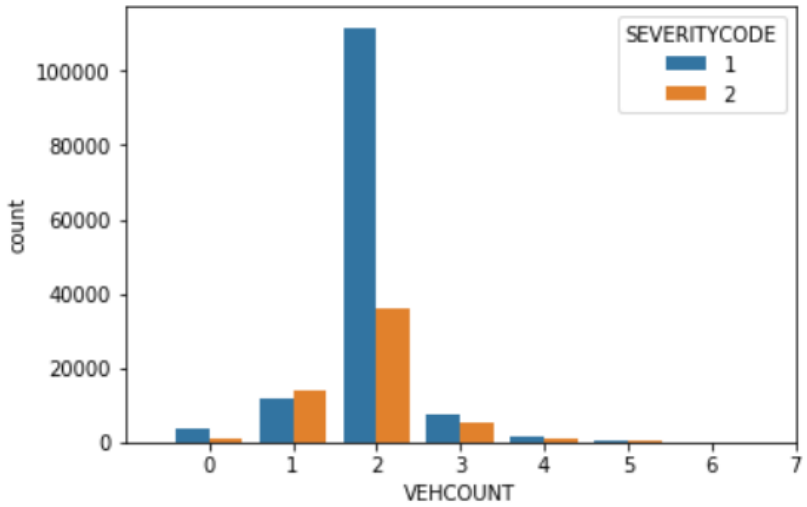
From the bar chart we can see that the highest number of collisions involved 2 people, and the biggest range involved 1-5 people. This range should be more inspected.



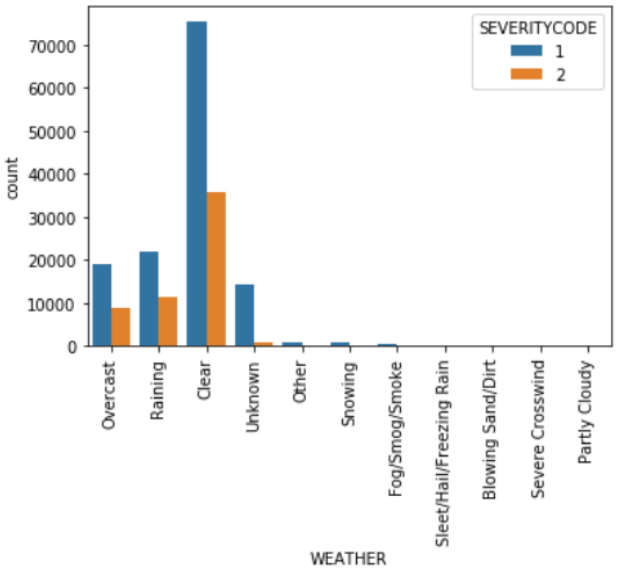
From the bar chart we can see that pedestrians struck by vehicles are more likely to withstand injuries than not.

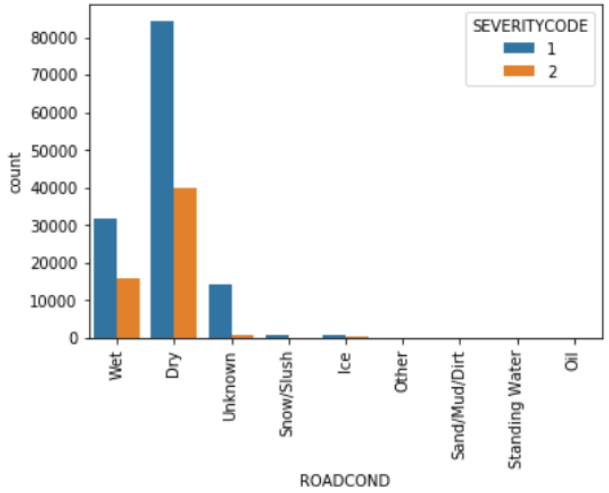


From the bar chart we can see that collisions didn't really involve in most cases bicycles.

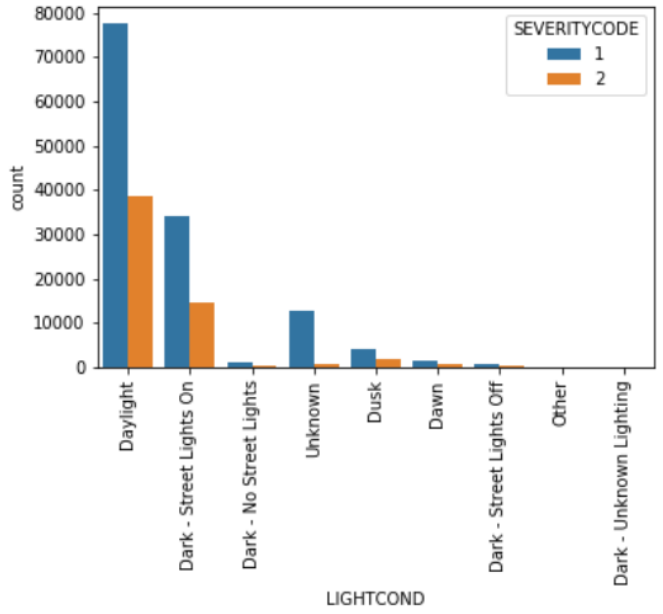


From the bar chart we can see that collisions involving one vehicle have resulted in more incidents with injury than property damage.

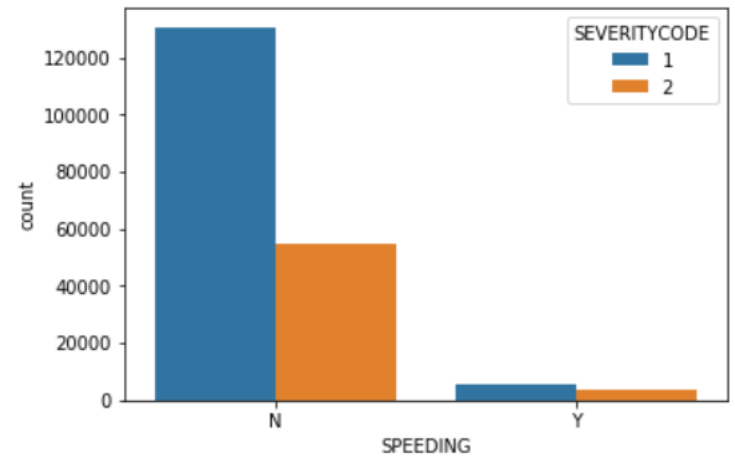




From the data we can conclude that rainy weather does increase the rate of injury accidents, but only slightly compared to clear conditions, noting that injury collision rates are not remarkably different from when conditions are dry.



We can conclude from the data that there is nothing remarkable about dark or poor lighting conditions versus daylight. It is more likely however, that the total volume and severity outcomes are lower during darker conditions because of traffic during those times.



From the data we can conclude that the proportion of collisions resulting in injury is 1.28 times higher when speeding is involved compared to when it's not involved.

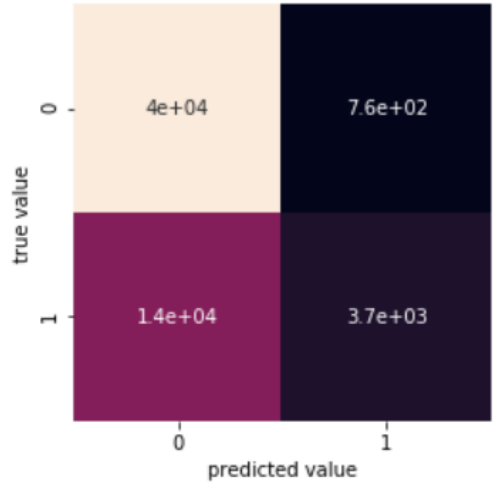
Discussions & Conclusions

- According to data provided we can predict with models what value will severity code have based on the model. It will be valuable to have complete dataset and minimise NaN values and also in some cases there is only one binary classifying value and it's impossible to deduct what the missing values are. These features will be useful for analysis of Severity of collision.
- Logistics regression has a score of 0.75.
- During the analysis we have observed several interesting relationships in data and this kind of classification problem is good to predict with a given accuracy because of minimal differences in Severity code 1 and Severity code 2 in reality. Generally, there is higher probability of SEVERITYCODE 1 collision, and they are mainly happening during daylight on a dry road, followed by dark - street lights on a wet road.

```
In [100]: model.score(test_x, test_y)
Out[100]: 0.7500941748570255

In [101]: model_y = model.predict(test_x)

          mat = confusion_matrix(test_y,model_y)
          sns.heatmap(mat, square=True, annot=True, cbar=False)
          plt.xlabel('predicted value')
          plt.ylabel('true value')
Out[101]: Text(91.68, 0.5, 'true value')
```



Conclusion and future directions

- Built useful models to predict severity code according to a given conditions
- Accuracy of the models has room for improvement
- Capture more data about binary conditions, traffic (number of vehicles) and hour of an accident