Peer-graded Assignment: Capstone Project - Car accident severity in Seattle City

By Edward Scott

Table of Contents

Introduction

Car accidents are an inevitability of modern life, but wouldn't it be great if you could predict the severity of an accident on a given day? Factors such as weather, road condition, time of day and many others could have an impact on the severity of a collision. Knowing the likelihood of a severe accident happening could make individuals take a little extra care whilst driving, or they may even wish to seek alternative transport.

In this project I am going to try to predict the severity of a collision based on Seattle City collision data.

The data

The data I am using was collected between January 2004 and April 2020. There are 194,673 collision records with 1 row representing 1 collision. There are 38 columns included in the dataset, each representing a piece of information about the collision. I cannot verify the validity of the dataset as it was provided to me by Coursera.

I will create a Machine Learning model based on a training set of the cleaned data and will then will use the remaining data (test set) to evaluate the model. I will look at using a number of different Machine Learning models to ensure I choose the one that has the highest level of accuracy.

The column I will use as my dependent variable is SEVERITYCODE and the I will choose from the remaining columns whether or not to use them as independent variables. SEVERITYCODE has two possible outcomes within the dataset. We know that there are 2 other possible codes (2b and 3) but these have not been included. Therefore we can determine that this is a subset of data.

1 = Property Damage Only Collision
2 = Injury Collision

A link to the original dataset can be found here:
https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

A link to the MetaData form which describes the raw data file can be found here:
https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf

Methodology

Data Analysis and Preparation

Initially I did some high level analysis of the columns to understand whether I needed them or not. I used the MetaData form to determine if they were relevant. During this initial investigation of the data, I found a number of columns were not going to be useful to me as they were either identity columns, they were duplicate columns or the data was not useful for analysis. These columns are:

OBJECTID, INCKEY, COLDETKEY, REPORTNO, STATUS, INTKEY, LOCATION, EXCEPTRSNCODE, EXCEPTRSNDESC, SEVERITYCODE.1, SEVERITYDESC, INCDATE, SDOT_COLCODE, SDOT_COLDESC, SDOTCOLNUM, ST_COLCODE, ST_COLDESC, SEGLANEKEY, CROSSWALKKEY

After removing these columns, I did some analysis of the remaining columns to understand whether I needed to do any preparation of the data:

	SEVERITYCODE	X	Y	ADDRTYPE	COLLISIONTYPE	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	INCDTTM
count	194673.0	189339.0	189339.0	192747	189769	194673.0	194673.0	194673.0	194673.0	194673
unique	NaN	NaN	NaN	3	10	NaN	NaN	NaN	NaN	162058
top	NaN	NaN	NaN	Block	Parked Car	NaN	NaN	NaN	NaN	11/2/2006
freq	NaN	NaN	NaN	126926	47987	NaN	NaN	NaN	NaN	96
mean	1.3	-122.3	47.6	NaN	NaN	2.4	0.0	0.0	1.9	NaN
std	0.5	0.0	0.1	NaN	NaN	1.3	0.2	0.2	0.6	NaN
min	1.0	-122.4	47.5	NaN	NaN	0.0	0.0	0.0	0.0	NaN
25%	1.0	-122.3	47.6	NaN	NaN	2.0	0.0	0.0	2.0	NaN
50%	1.0	-122.3	47.6	NaN	NaN	2.0	0.0	0.0	2.0	NaN
75%	2.0	-122.3	47.7	NaN	NaN	3.0	0.0	0.0	2.0	NaN
max	2.0	-122.2	47.7	NaN	NaN	81.0	6.0	2.0	12.0	NaN

	JUNCTIONTYPE	INATTENTIONIND	UNDERINFL	WEATHER	ROADCOND	LIGHTCOND	PEDROWNOTGRNT	SPEEDING	HITPARKEDCAR
count	188344	29805	189789	189592	189661	189503	4667	9333	194673
unique	7	1	4	11	9	9	1	1	2
top	Mid-Block (not related to intersection)	Y	N	Clear	Dry	Daylight	Y	Y	N
freq	89800	29805	100274	111135	124510	116137	4667	9333	187457

From this analysis we can see that in the refined dataset there are 19 columns. SEVERITYCODE is the dependent variable that we will be using to model our data and the remaining columns I will analyse and make a decision whether to include them in my model or not.

Looking at the data we can see the following:

- Some of the columns contain NULL values
- Some columns do not match the MetaData form (e.g. UNDERINFL should only contain Y/N values)
- We have a mixture of datatypes (integers, floats and categorical columns)
- We have a datetime but we would prefer Year, Month, Day and Time as separate columns

Firstly I changed any Y/N columns to be 1/0 instead to make analysis of the data easier. Also to use these columns in a model, they will need to be numeric.

To deal with the NULL values within the dataset I did the following:

X - replace null values with mean

Y - replace null values with mean

ADDRTYPE - As there is no clear majority in terms of frequency and the number of records is below the threshold of 5%, I am going to drop the records with null values in this column

COLLISIONTYPE - As there is no clear majority in terms of frequency and the number of records is below the threshold of 5%, I am going to drop the records with null values in this column

JUNCTIONTYPE - As there is no clear majority in terms of frequency and the number of records is below the threshold of 5%, I am going to drop the records with null values in this column

INATTENTIONIND - replace nulls with “N” as there is only an indication of "Y" in the table

UNDERINFL - replace null values with “N” as this is the most frequently occurring value by far

WEATHER - changing null values to "Unknown" as this is an option within the dataset

ROADCOND - changing null values to "Unknown" as this is an option within the dataset

LIGHTCOND - changing null values to "Unknown" as this is an option within the dataset

PEDROWNOTGRNT - replace nulls with 0 as there is only an indication of "Y" in the table.

SPEEDING - replace nulls with 0 as there is only an indication of "Y" in the table

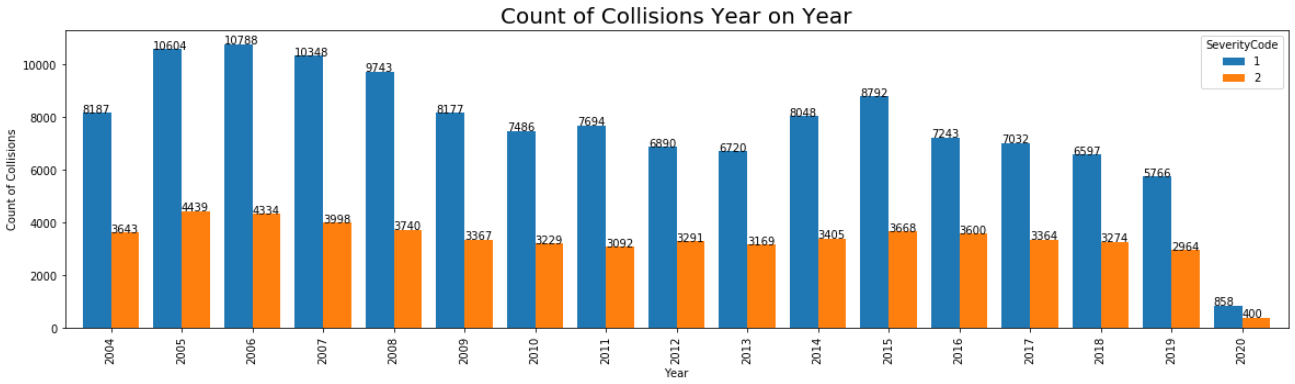
Once all of the values were filled I proceeded to change the column names to make the data easier and more recognisable to work with. The names were changed as follows:

SeverityCode, X, Y, LocationType, CollisionType, PersonCount, PedestrianCount, CyclistCount, VehicleCount, JunctionType, DriverInattention, DriverUnderInfluence, Weather, RoadCondition, LightCondition, PedROWNG, DriverSpeeding, HitParkedCar, Year, Month, Day, Hours, Minutes

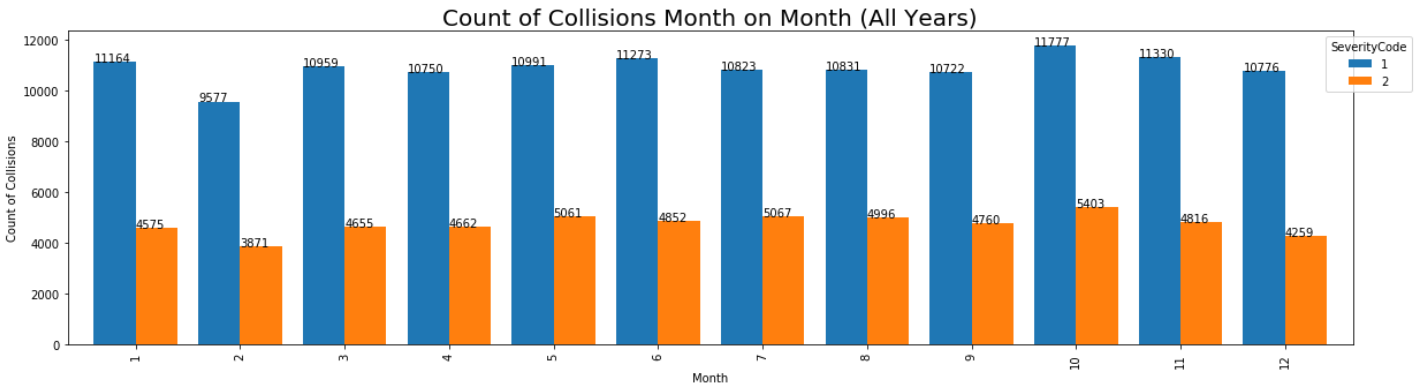
Data Visualisation

Once I had a cleaned dataset I was able to visualise the data to get a better grasp of what I was looking at.

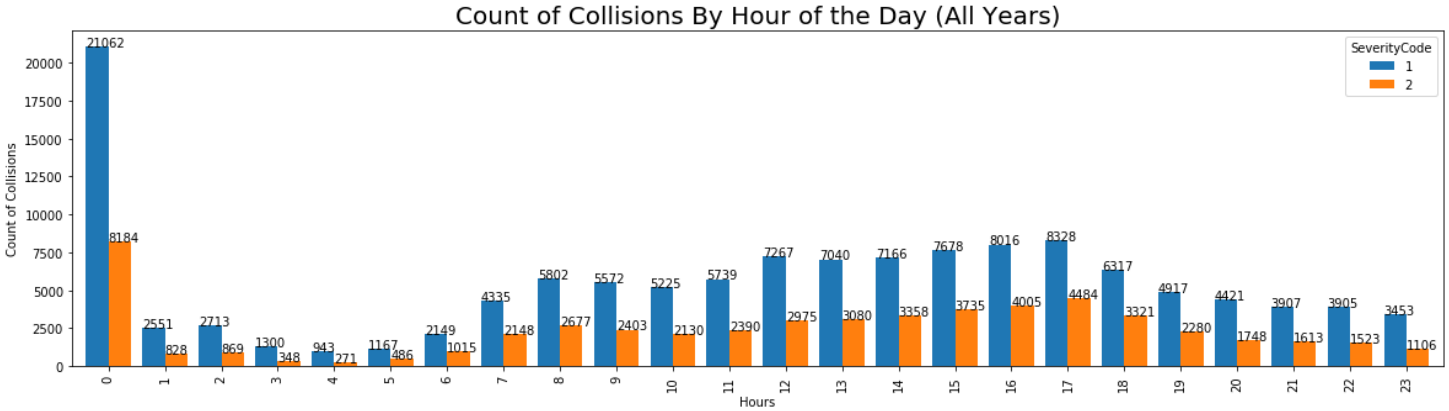
Time Periods



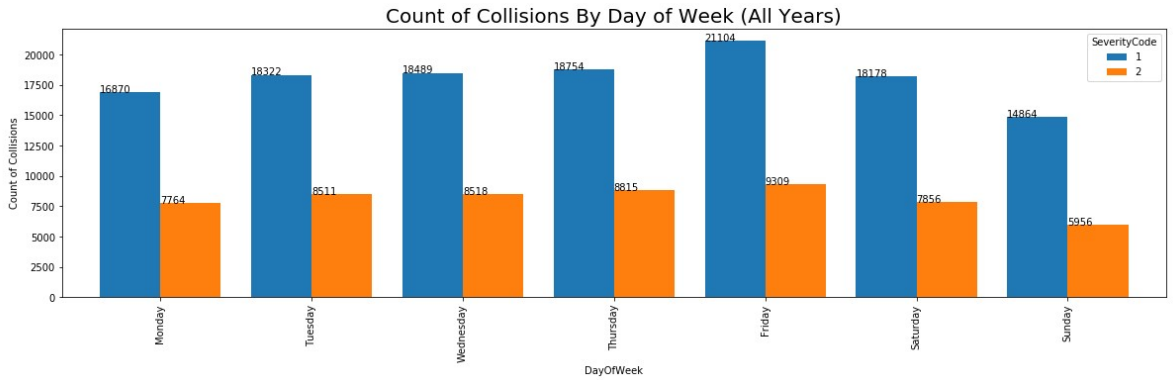
It appears that the volume of collisions was trending downwards year on year. This could potentially be put down to improvements in car technology or road safety improvements by Seattle City Council.



There doesn't appear to be significant differentiation between months in terms of collisions.



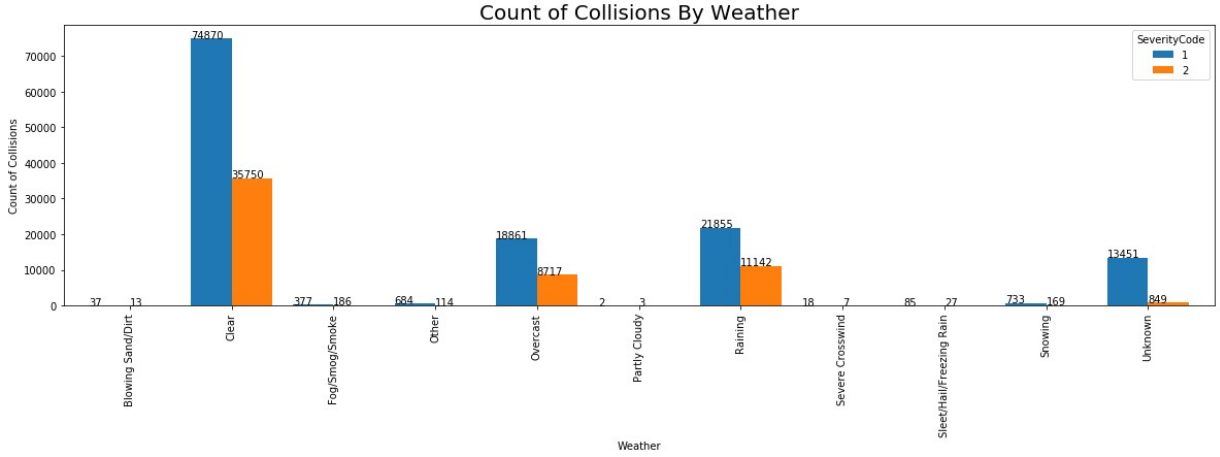
Looking at the collisions by hour we see a very large spike between 00:00 and 01:00. This may be a data anomaly within the DateTime column. If I end up using this field in the model I need to do some further analysis on this field to determine whether or not the data is correct. Other than this spike, we see more collisions between 17:00 and 18:00 which is expected as this is peak time.



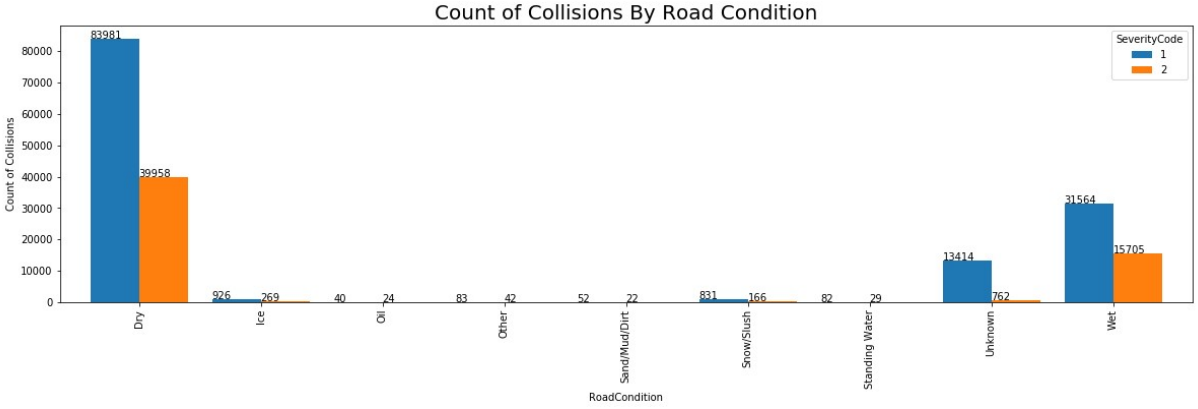
The last time period that I looked at was the day of the week. I hypothesized that collisions would be higher during the week (Monday to Friday) compared to the weekends. I thought this because traffic volumes during the week are higher (especially at peak drive times) than at the weekends, therefore more traffic leads to more collisions.

Looking at the chart above, it appears that the day with the highest frequency of collisions was Friday. This may be caused by individuals looking to get home quicker leading to more reckless driving, or potentially due to fatigue from the working week.

Weather, Road Condition and Light

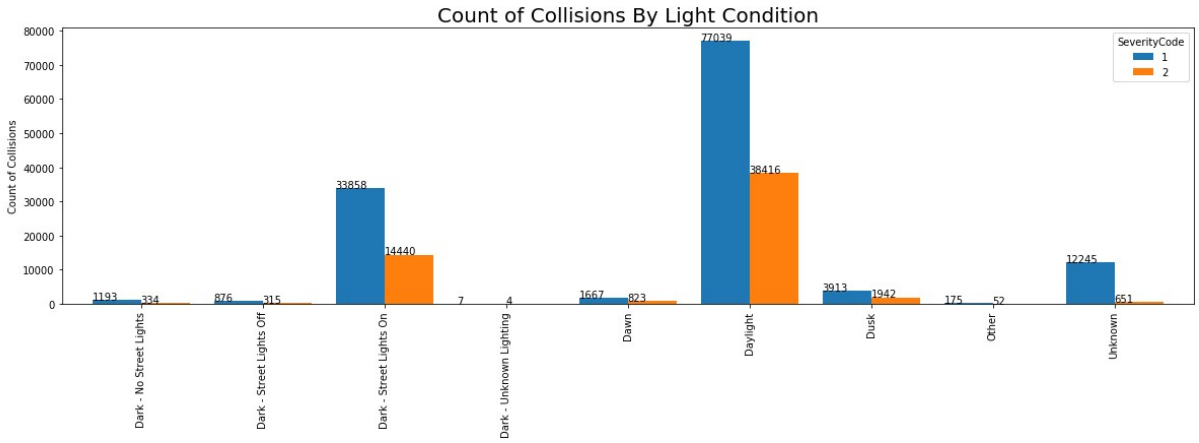


The majority of collisions occurred when the weather was clear. It is not indicative that more collisions occur during clear weather as we do not have weather data included in this dataset. Another thing to note is that we have “Unknown” weather conditions. If I am to use this column within my model I will need to exclude “Unknown” values.



Again, although the majority of collisions occur whilst the road is dry it is not indicative of more collisions occurring whilst the road is dry as it may be dry more of the time. To further analyse this we would need another source of data (e.g. rainfall).

Again, if I am to use this column within my model I will need to exclude “Unknown” values.



The majority of collisions occur during daylight. This is probably expected as more people are driving during the daytime. Again, if I am to use this column within my model I will need to exclude "Unknown" values.

Negligence

There are a number of collisions that where either an individual involved in the collision was negligent. Within the data, these scenarios are:

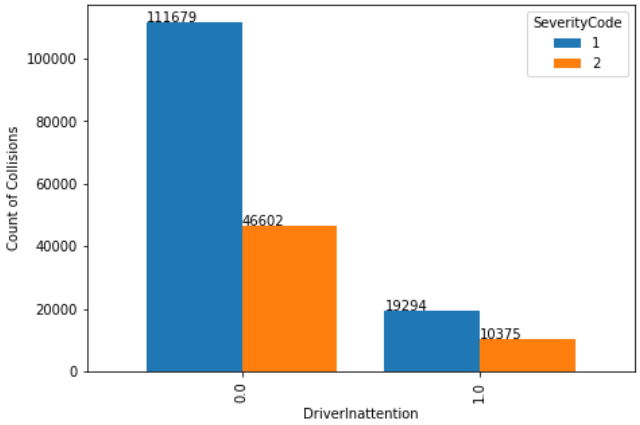
- A person involved in the collision was not paying full attention (column = INATTENTIONIND)
- A driver involved was under the influence of alcohol or drugs (column = UNDERINFL)
- A driver did not grant the pedestrian right of way (column = PEDROWNOTGRNT)
- A driver was caught speeding (column = SPEEDING)

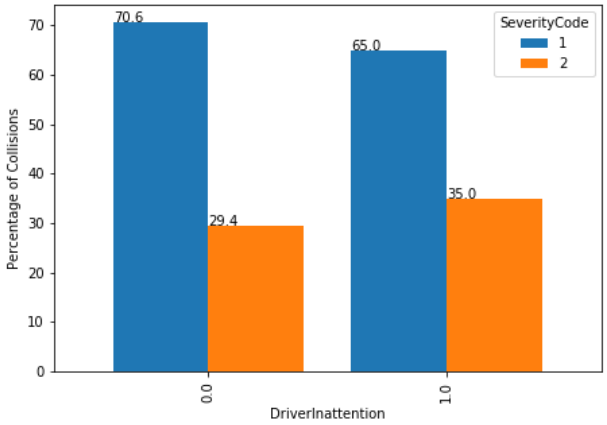
As you can see from the graphs on the next page, there is a slightly higher percentage of injury collisions where in all of these scenarios. Therefore we can speculate that it is likely that a collision will be more severe when an individual involved has been more negligent. However, because it is not possible to know this information before setting off for a journey, we will not be able to use these columns in our model.

If we wanted to be more thorough, we could look for the times of year where individuals are more negligent (e.g. more people driving under the influence of alcohol during Christmas/New Year). These are things to consider for future development of the model but we will not be considering them in this report.

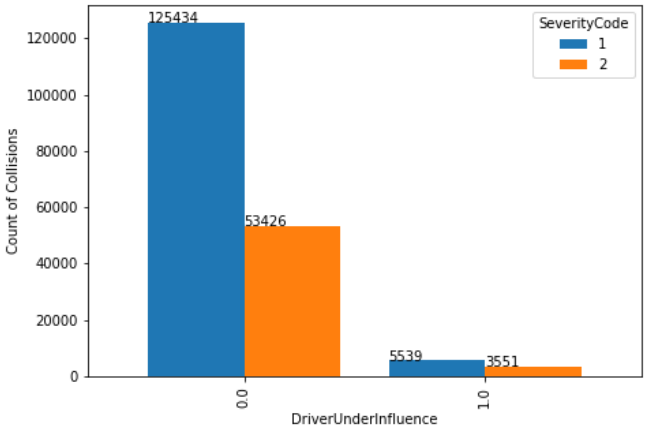
Negligence Charts

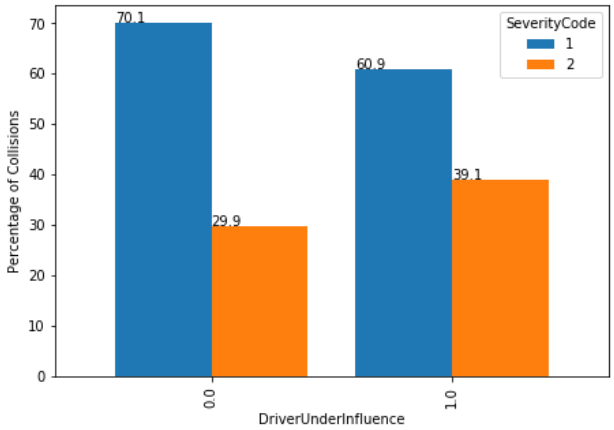
Count & Percentage of Collisions caused by an individual not paying full attention



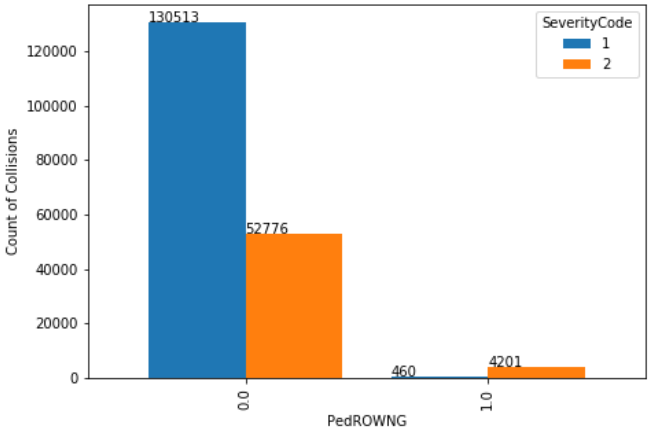


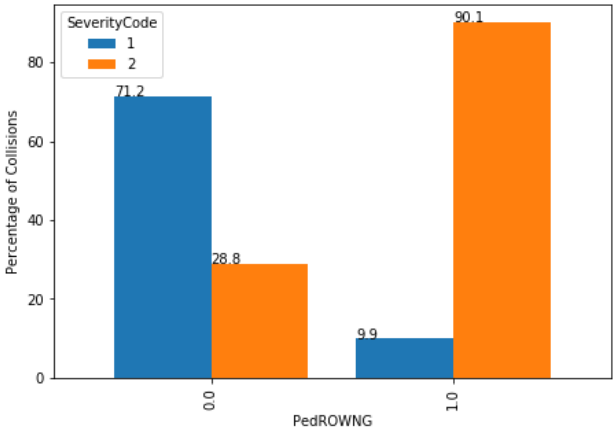
Count & Percentage of Collisions where a driver was under the influence of Drugs/Alcohol



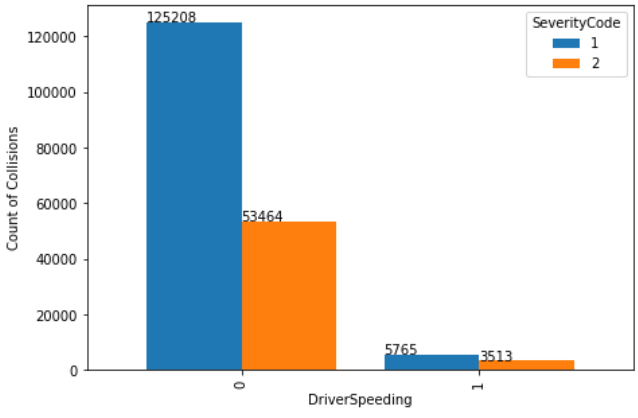


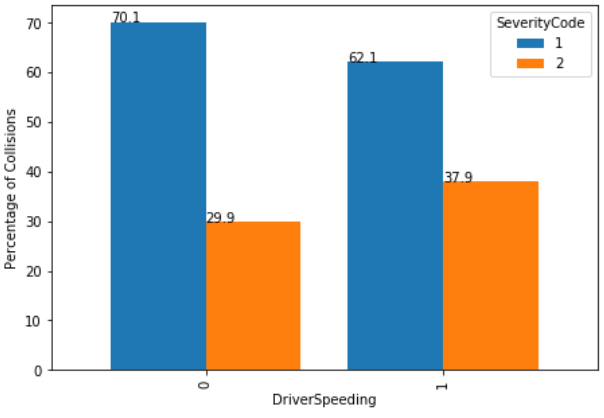
Count & Percentage of Collisions where a driver did not grant Pedestrian Right of Way





Count & Percentage of Collisions where a driver was Speeding





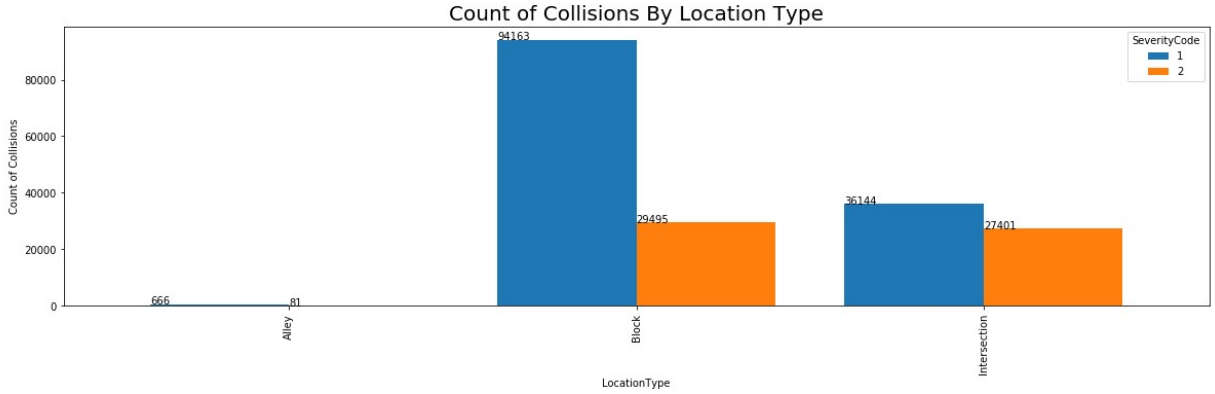
Collision Details

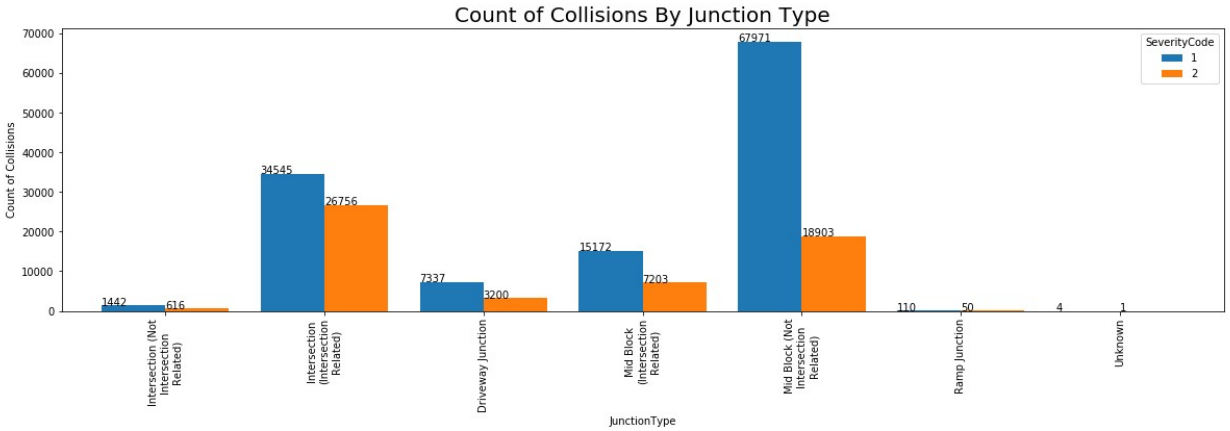
Within the dataset there are a number of columns that describe the location at which the collision occurred. These columns are:

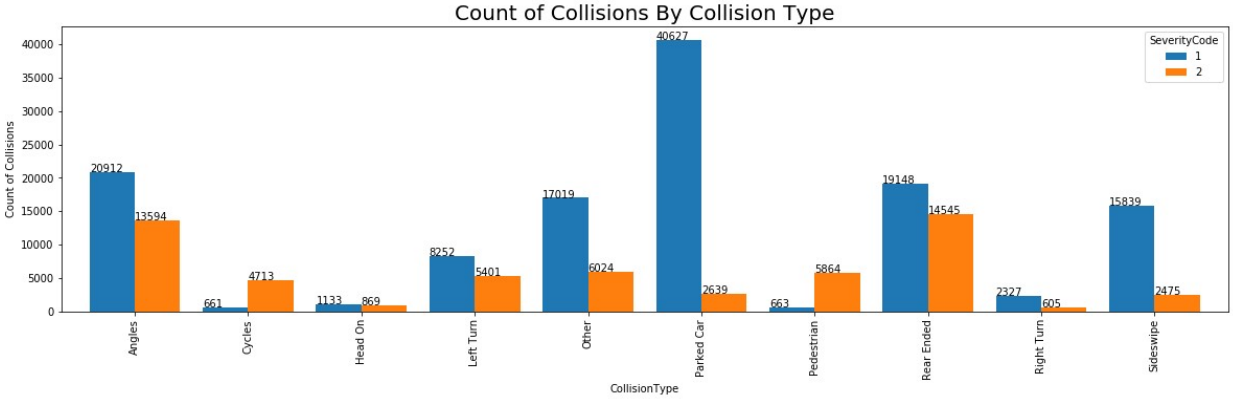
- Address Type (e.g. Alley, Block, Intersection) (column = ADDRTYPE)
- Collision Type (e.g. Head On, Sideswipe) (column = COLLISIONTYPE)
- Junction Type (e.g. Driveway Junction, Mid Block) (column = JUNCTIONTYPE)

As you can see from the graphs below, there appears to be a relationship between injury collisions and collisions that are related to intersections. We can also see a relationship between certain collision types and injury collisions, specifically Cycles, Head On, Pedestrian, and Rear End collisions.

Similar to Negligence, it is difficult for us to determine this information before the journey, therefore we would not be able to use this information in our current model. We may be able to use this in future iterations to link collision types with specific times or day, months of the year or days of the week in order to tweak the model and allow further variables to play a factor.







Parties Involved

Within the dataset there are a number of columns that describe the number of individuals/vehicles involved in the collision. These columns are:

- Number of people involved in the collision ranging from 0 to 81 (column = PERSONCOUNT)
- Number of pedestrians involved in the collision ranging from 0 to 6 (column = PEDCOUNT)
- Number of cyclists involved in the collision ranging from 0 to 2 (column = PEDCYLCOUNT)
- Number of vehicles involved in the collision ranging from 0 to 12 (column = VEHCOUNT)

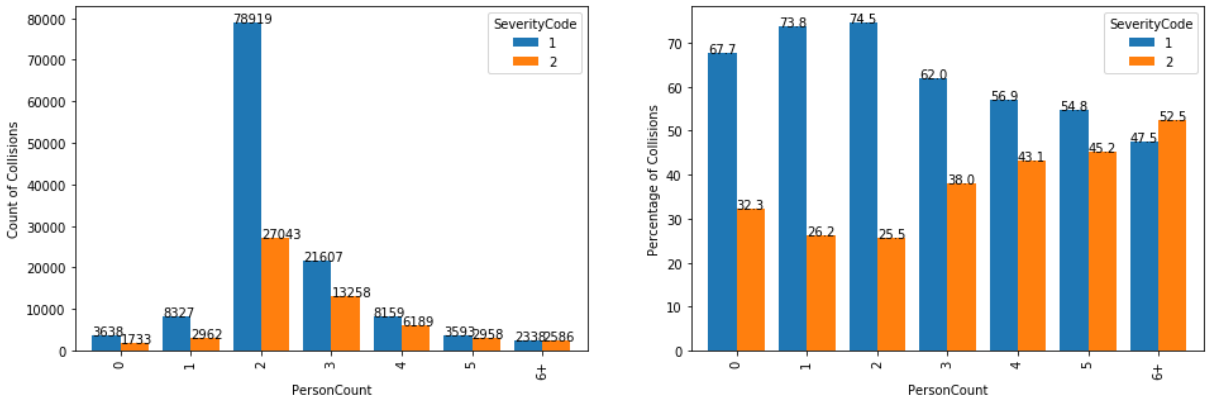
As you can see from the charts on the next page, I have grouped the data to make the visualisations easier to read. This also helps us draw better insights as some of the higher volume counts (e.g. 81) only had 1 or 2 collisions which would not be readable on a graph. I chose the groupings by first looking at the visuals without the grouping and making a judgement call as to where the best place to group is.

For persons involved, we can quite easily draw some good insights from the data. For example, as the number of Persons involved, Pedestrians involved or cyclists involved increases, the percentage of injury related collisions increases. This is understandable as it is more likely for a pedestrian to be injured in a car accident than a person driving a vehicle.

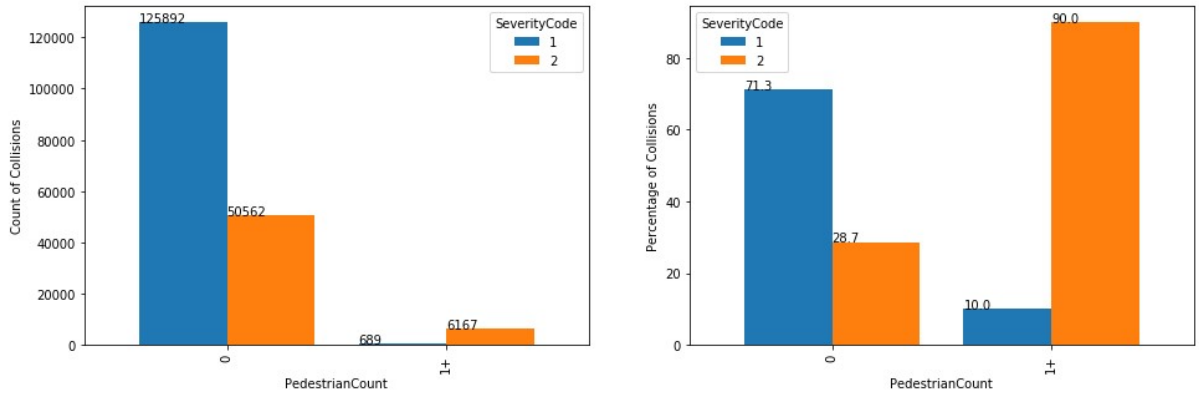
Another interesting insight is the number of vehicles involved. Where the vehicle count is 1, the percentage of injury related collisions is significantly higher. This supports our previous observation whereby if a pedestrian is involved (which is the likely circumstance if there is only one vehicle) then the severity of the collision will be higher.

Parties involved Charts

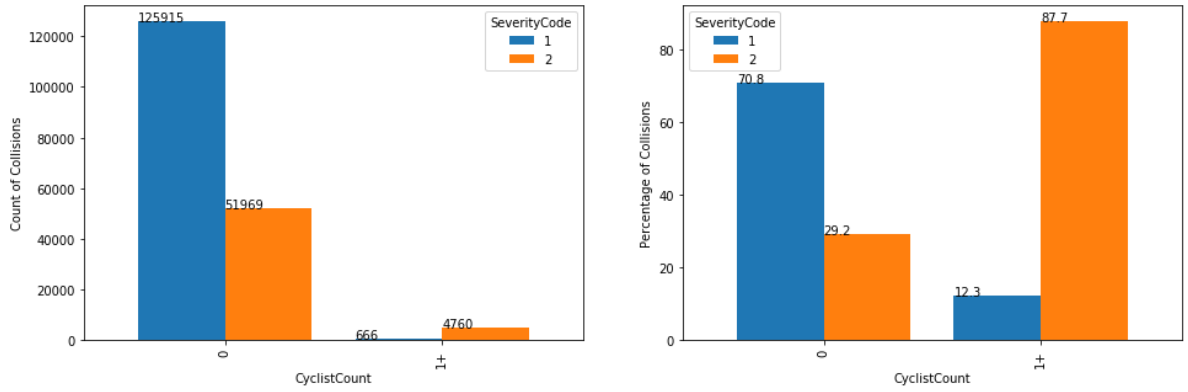
Count & Percentage of Collisions by Count of persons involved



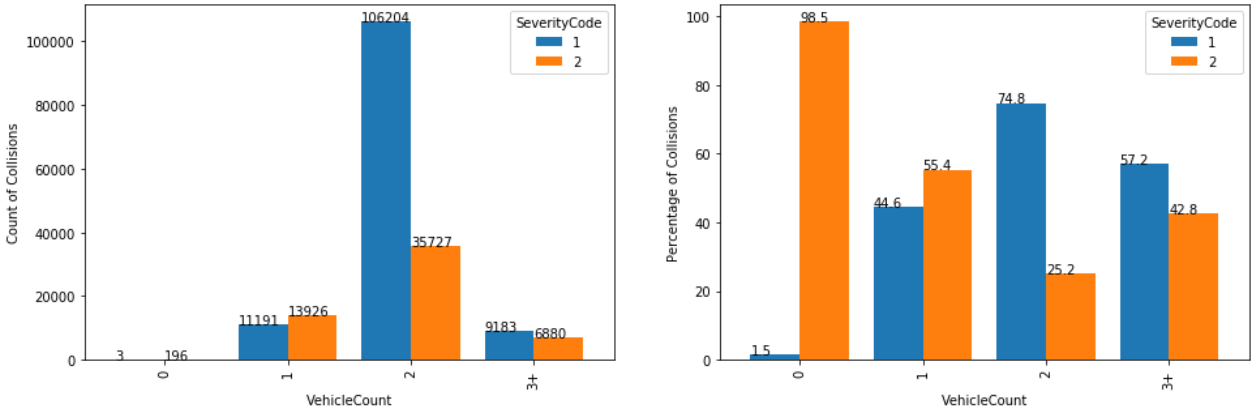
Count & Percentage of Collisions by Count of pedestrians involved



Count & Percentage of Collisions by Count of cyclists involved



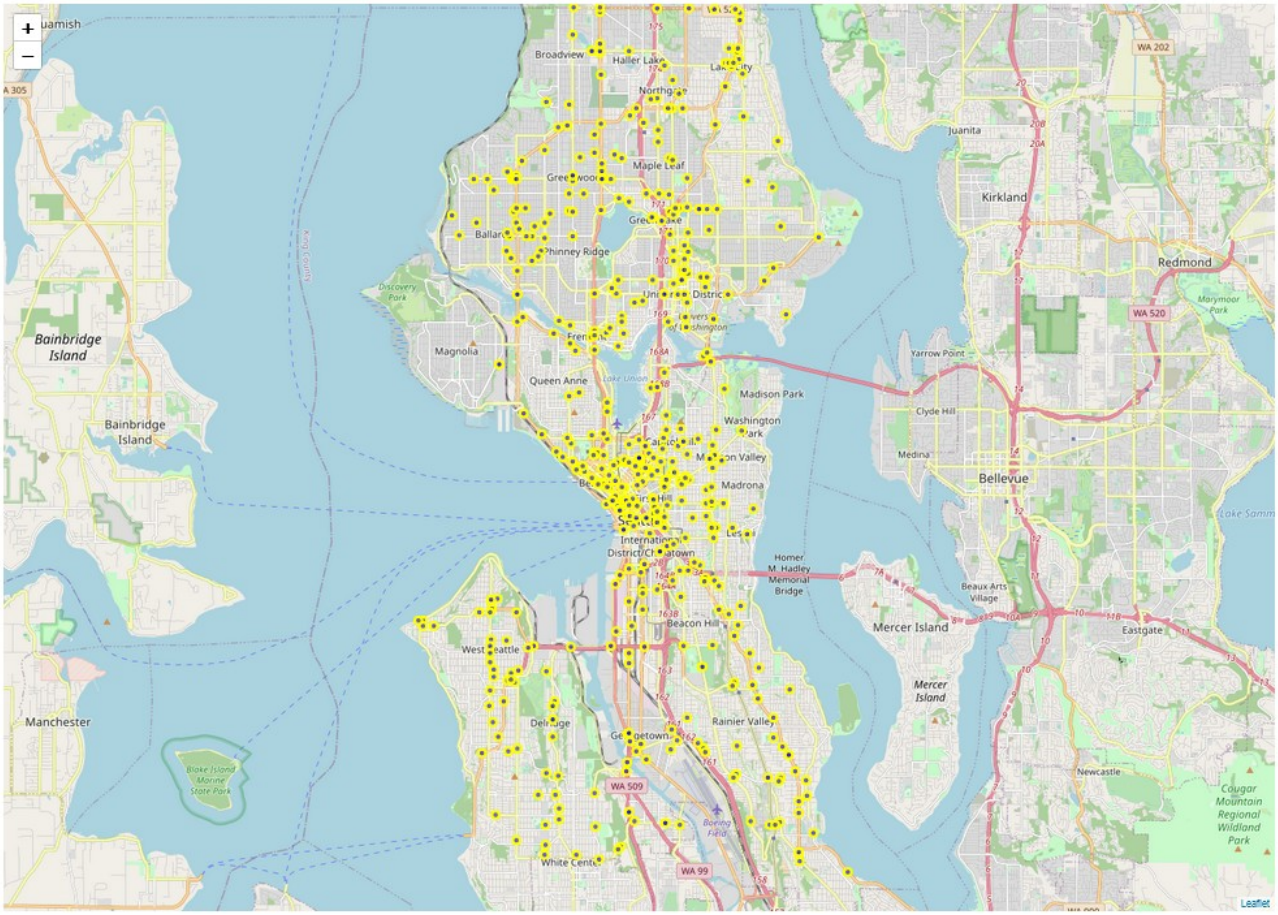
Count & Percentage of Collisions by Count of vehicles involved



Map of Collisions

To help understand the location of the accidents better, I decided it was best to plot the X Y coordinates onto a map. As you can see from the map below, the majority of collisions are located near the centre of Seattle City. This is an expected outcome as it is likely the area with the highest volume of traffic.

We may also say that this fits with the most frequent collision type being “Parked Car” as there is likely a lot more parked cars in the city centre than any other area of the city.



Modelling

Looking through all of the metrics and taking into account what information we would have to hand on a given day, I have decided to include the following fields within my model:

- Severity Code
- Month
- DayOfWeek
- Weather
- RoadCondition
- LightCondition

I will see how the model performs using a train and test model of 80% train, 20% test. I will use multiple model types including:

- Decision Tree
- KNN
- Logistic Regression
- Support Vector Machine

Once the models are all built I will evaluate their effectiveness using the following evaluation techniques:

- F1 Score
- Jaccard Similarity Score
- Pecision Score
- Recall Score
- Log Loss

Results

From the different machine learning algorithms I used, here are the results of the various evaluation techniques:

Model	Jaccard Score	F1 Score	Accuracy Score
KNN	0.667	0.545	0.668
Decision Tree	0.669	0.541	0.668
Logistic Regression	0.6715	0.539	0.672

Discussion

The models all performed slightly worse than expected. However before concluding the models are not fit for purpose we would need to take another real world dataset and confirm we see the same results.

Another point is that this dataset is limited to Seattle City; for the model to be robust it would need to be tested on other cities. It will likely not perform well, but it may benefit from further data being added from other cities to help the model make better predictions.

The dataset only included two severity levels, Property only and Injury. In the real world this is an unlikely scenario as over a 16 year period there is bound to have been at least one fatality. This shows that the data was manipulated beforehand – if we wanted to make the model complete we would need to include this data.

Conclusion

Overall, the models produced appeared to be ineffective. I would likely need to go back, re-evaluate the data, refine the model in order to get a better performing set of models.

Once happy, I would need to test the model in a real world scenario. After testing in the real world scenario we would likely find issues leading us to go back and adjust the model either by including data that we excluded (e.g. Time of Day) or potentially add more data.