Capstone Project - Collision Severity Analysis

Hai Dang Nguyen

October 18th , 2020

Table of Contents

I. Introduction

A traffic collision, or car accident, occurs when a vehicle collides with another vehicle(s), road user(s), animal(s), and object(s) such as road debris, tree, or building. Traffic collisions often result in injuries, disabilities, deaths, and property damages. The other problems are traffic congestion and delay, leading to lost time, a reduction in productivity and an increased response time by police, fire, and emergency medical services. Consequently, the collisions are associated with substantial costs to the economy, the society and the individuals involved. According to the World Health Organization fact sheet in February 2020, every year approximately 1.35 million lives are taken away, and 20 to 50 million more people suffer from non-fatal injuries, with many incurring a disability because of their injury. The collisions cost most countries around 3% of their gross domestic product. Moreover, the number of traffic crashes and their victims has been a rising trend globally due to increases in population and motorization.

Seattle is the largest city in both the state of Washington and the Pacific Northwest region of North America, with a total population of 3.4 million. In 2018, it was reported that over 80% of Seattle households owned at least one vehicle, which means about 457,000 vehicles crammed into the city's 84 square miles of land area. In 2019, there were over 11,000 collisions which resulted in around 4,400 injuries, 26 fatalities and numerous property damages. Road safety is thus a major public health issue throughout the world, and it is crucial for many government sectors to work in partnership in order to improve it. The severity is undoubtedly a fundamental aspect of a collision event. Accurate prediction of accident severity and identification of the key factors can provide vital information for an effective management of traffic collisions. This is important to reduce accident frequency and severity in near future, restore the traffic capacity quickly and enhance traffic safety and transportation system efficiency, thus saving many lives and wealth.

This study aims at developing a model system using machine learning algorithms to predict the collision severity. Indicators for accident severity will be set, which represents number of fatalities, number of injuries, and property damage, respectively. In addition, it intends to give a good insight into the factors that could be contributing to collisions, for example, crash location, car speeding, weather conditions, lightning condition, and road condition. The severity of future collisions will be based on the similarity of their initial conditions to those of historical collision records.

II. Data

The collision data is available City of Seattle Open Data Portal, collected and recorded by the Seattle Department of Transportation's (SDOT). It includes all types of collisions that happened in Seattle city from 2004 to Sep 2020. This dataset is updated weekly, labelled and contains 221,525 accident records with 40 attributes for each accident as follows:

Field	Description
OBJECTID	ObjectID ESRI unique identifier.
INCKEY	A unique key for the incident.
COLDETKEY	Secondary key for the incident.
ADDRTYPE	Collision address type.
INTKEY	Key that corresponds to the intersection associated with a collision.
LOCATION	Description of the general location of the collision.
SEVERITYCODE	A code that corresponds to the severity of the collision.
COLLISIONTYPE	Collision type.

PERSONCOUNT	The total number of people involved in the collision.
PEDCOUNT	The number of pedestrians involved in the collision. This is entered by the state.
PEDCYLCOUNT	The number of bicycles involved in the collision. This is entered by the state.
VEHCOUNT	The number of vehicles involved in the collision. This is entered by the state.
INJURIES	The number of total injuries in the collision. This is entered by the state.
SERIOUSINJURIES	The number of serious injuries in the collision. This is entered by the state.
FATALITIES	The number of fatalities in the collision. This is entered by the state.
INCDATE	The date of the incident.
INCDTTM	The date and time of the incident.
JUNCTIONTYPE	Category of junction at which collision took place.
SDOT_COLCODE	A code given to the collision by SDOT.
INATTENTIONIND	Whether or not collision was due to inattention. (Y/N)
UNDERINFL	Whether or not a driver involved was under the influence of drugs or alcohol.
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.
PEDROWNOTGRNT	Whether or not the pedestrian right of way was not granted. (Y/N)
SDOTCOLNUM	A number given to the collision by SDOT.
SPEEDING	Whether or not speeding was a factor in the collision. (Y/N)
ST_COLCODE	A code provided by the state that describes the collision.
SEGLANEKEY	A key for the lane segment in which the collision occurred.
CROSSWALKKEY	A key for the crosswalk at which the collision occurred.
HITPARKEDCAR	Whether or not the collision involved hitting a parked car. (Y/N)

1. Data Understanding

Since we would like to identify the key factors that cause a collision and predict the level of collision severity, we will use SEVERITYCODE as our dependent variable Y, and try different combinations of independent variables X to get the model result.

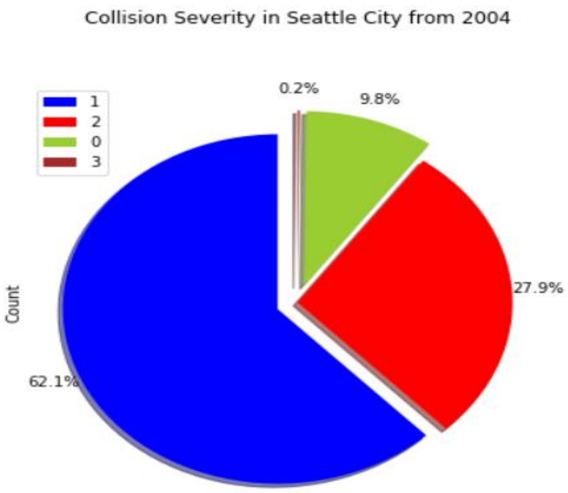
Link for dataset: Seattle Collision Data

Level of collision severity

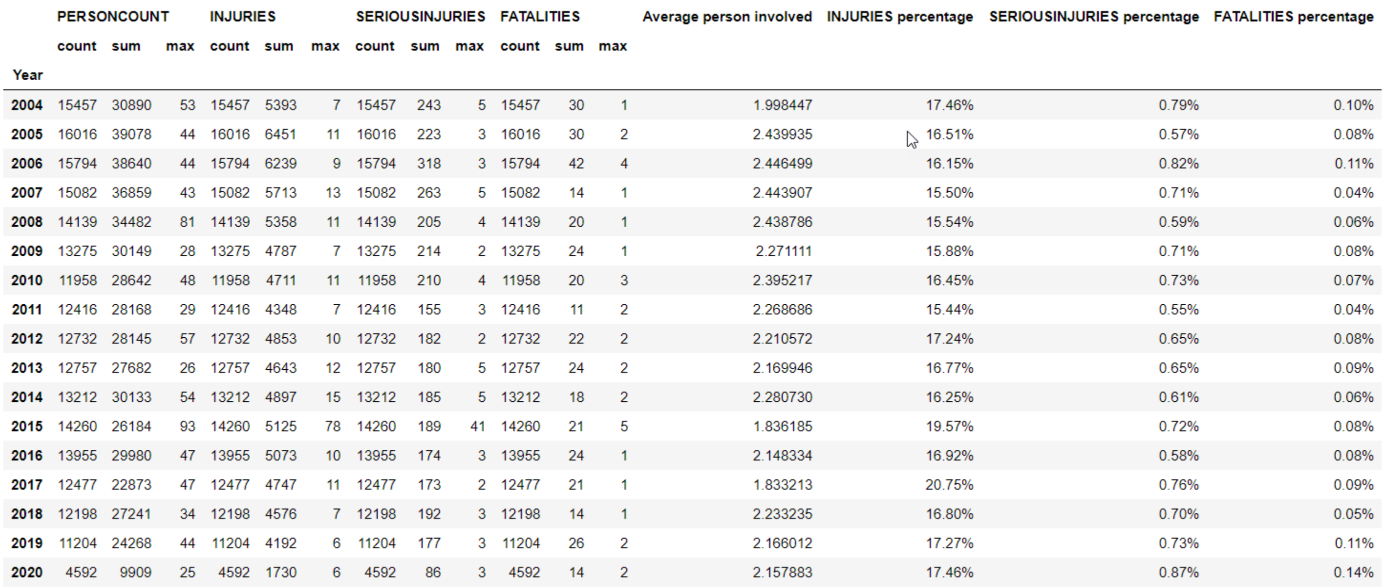
Column SEVERITYCODE encodes the Seattle Department of Transport (SDOT) accident severity metric, according to the following schema:

Code	Description
0	Unknown/no data
1	Property damage only
2	Minor injury collision
2b	Major injury collision
3	Fatality collision

To get a clearer display of the data, we combine the group 2b of major injury collision into the group 2 of minor injury collision, so that we can compare the three levels of severity which are Property damage, Injury and Fatality. From the pie chart, we can see that approximately two-thirds of the collisions resulted in no apparent injury with around 10% being unknown severity and 62% being property damage. Out of the remaining third, around 28% resulted in some kind of injury even if not severe or fatal, and the number of fatal injury-involved collisions represents a small fraction of the total number of collisions in Seattle (0.2%).



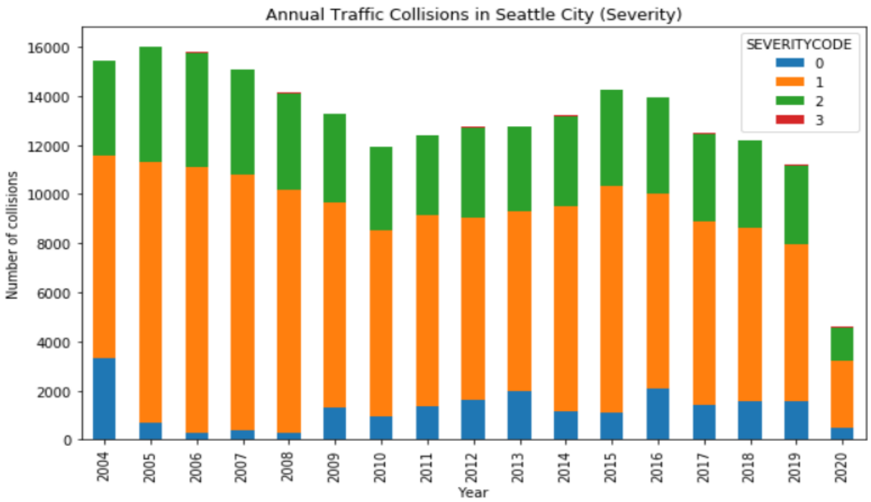
On average, there were 2 people involved in a collision. The number of injuries and fatalities seemed to drop over time, from 6,000 to 4,000 for injuries, from 300 to 170 for serious injuries and from 40 to 20 for fatalities. However, the percentage of people involved resulting in injuries and fatalities seemed to have no drastic change, with around 17% resulting in injuries, 0.7% resulting in serious injuries and 0.1% resulting in fatalities. In 2015, there was a dramatic collision where the total number of people engaged in a collision peaked at 93, leading to 78 injuries and 5 fatalities. But these numbers have dropped in recent years, still Seattle City needs effective strategies to be mitigate future collision severity.



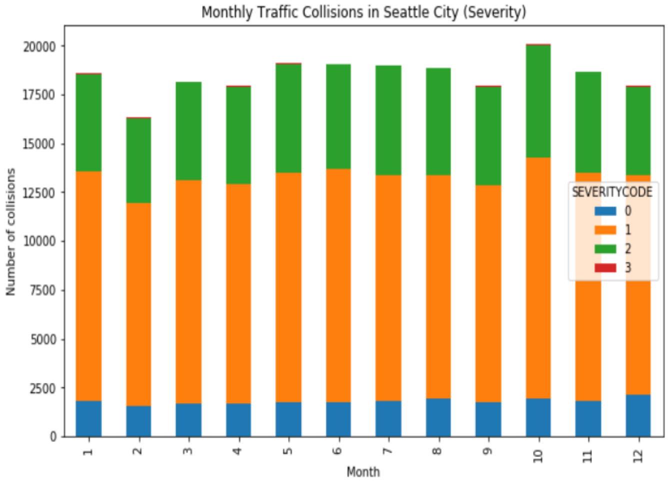
	PERSONCOUNT			INJURIES			SERIOUSINJURIES			FATALITIES			Average person involved	INJURIES percentage	SERIOUSINJURIES percentage	FATALITIES percentage
	count	sum	max	count	sum	max	count	sum	max	count	sum	max				
Year																
2004	15457	30890	53	15457	5393	7	15457	243	5	15457	30	1	1.998447	17.46%	0.79%	0.10%
2005	16016	39078	44	16016	6451	11	16016	223	3	16016	30	2	2.439935	16.51%	0.57%	0.08%
2006	15794	38640	44	15794	6239	9	15794	318	3	15794	42	4	2.446499	16.15%	0.82%	0.11%
2007	15082	36859	43	15082	5713	13	15082	263	5	15082	14	1	2.443907	15.50%	0.71%	0.04%
2008	14139	34482	81	14139	5358	11	14139	205	4	14139	20	1	2.438786	15.54%	0.59%	0.06%
2009	13275	30149	28	13275	4787	7	13275	214	2	13275	24	1	2.271111	15.88%	0.71%	0.08%
2010	11958	28642	48	11958	4711	11	11958	210	4	11958	20	3	2.395217	16.45%	0.73%	0.07%
2011	12416	28168	29	12416	4348	7	12416	155	3	12416	11	2	2.268686	15.44%	0.55%	0.04%
2012	12732	28145	57	12732	4853	10	12732	182	2	12732	22	2	2.210572	17.24%	0.65%	0.08%
2013	12757	27682	26	12757	4643	12	12757	180	5	12757	24	2	2.169946	16.77%	0.65%	0.09%
2014	13212	30133	54	13212	4897	15	13212	185	5	13212	18	2	2.280730	16.25%	0.61%	0.06%
2015	14260	26184	93	14260	5125	78	14260	189	41	14260	21	5	1.836185	19.57%	0.72%	0.08%
2016	13955	29980	47	13955	5073	10	13955	174	3	13955	24	1	2.148334	16.92%	0.58%	0.08%
2017	12477	22873	47	12477	4747	11	12477	173	2	12477	21	1	1.833213	20.75%	0.76%	0.09%
2018	12198	27241	34	12198	4576	7	12198	192	3	12198	14	1	2.233235	16.80%	0.70%	0.05%
2019	11204	24268	44	11204	4192	6	11204	177	3	11204	26	2	2.166012	17.27%	0.73%	0.11%
2020	4592	9909	25	4592	1730	6	4592	86	3	4592	14	2	2.157883	17.46%	0.87%	0.14%

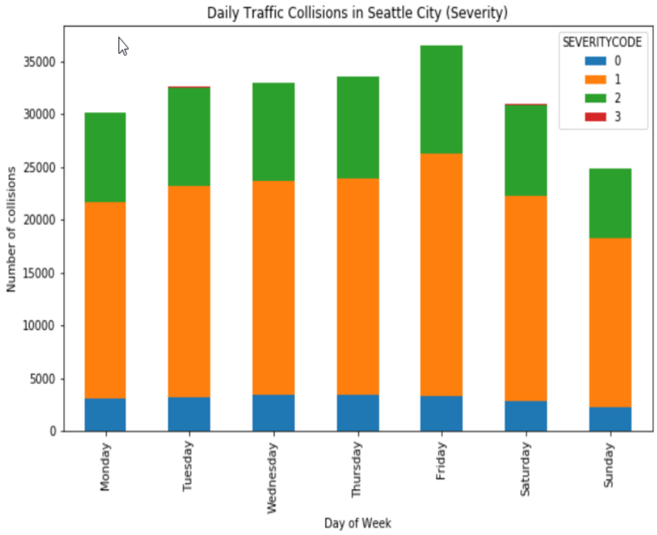
Number of collisions over time

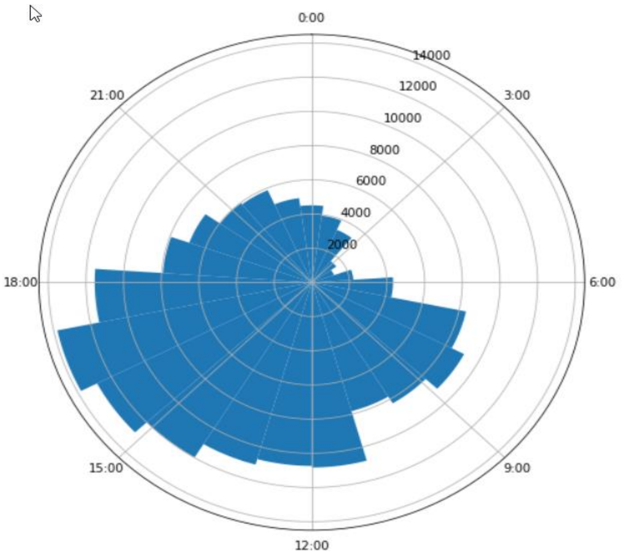
Since 2004, the number of collisions in Seattle has remained at a high level, around 11,000 annually and peaked at 16,000 collisions in 2005. It declined steadily from 2006 to 2013, only to rebound slightly in the four subsequent years but remained well below the 2005 numbers. In the recent years from 2015, the number of collisions has decreased again, but the incident scale has no significant change in this period. In 2020 during the COVID-19 outbreak, on top of the closure of all nonessential businesses and schools until April, residents are being urged and pleaded with to stay home. And with that, there has been less congestion and traffic, hence resulting in a large reduction in accidents compared to 2019. The number of injury-involved collisions stayed roughly constant while the amount of unknown severity data increased from 2008, leading to a drop on property damage-involved collisions. Combining with the insight above about the level of collision severity, this suggests an enhancement in data recording regarding the severity.


Annual Traffic Collisions in Seattle City (Severity)

With similar approach for monthly data, October was the worst month with more car accidents than the average month, in contrast to February with fewer accidents than the average month. In a week, the number of collisions grew through the working week, peaking on Friday. Sunday was the quietest day, with much fewer collisions than the average day. However, fatalities peaked on the weekends with Saturday having more deaths than the average day. In addition, there seemed to be more collision occurring during the morning rush between 8am and 9am, during lunch time between 11am and 12pm, but most during the evening jam from 2pm to 6pm experiencing 40% of the day's collisions.


Monthly Traffic Collisions in Seattle City (Severity)

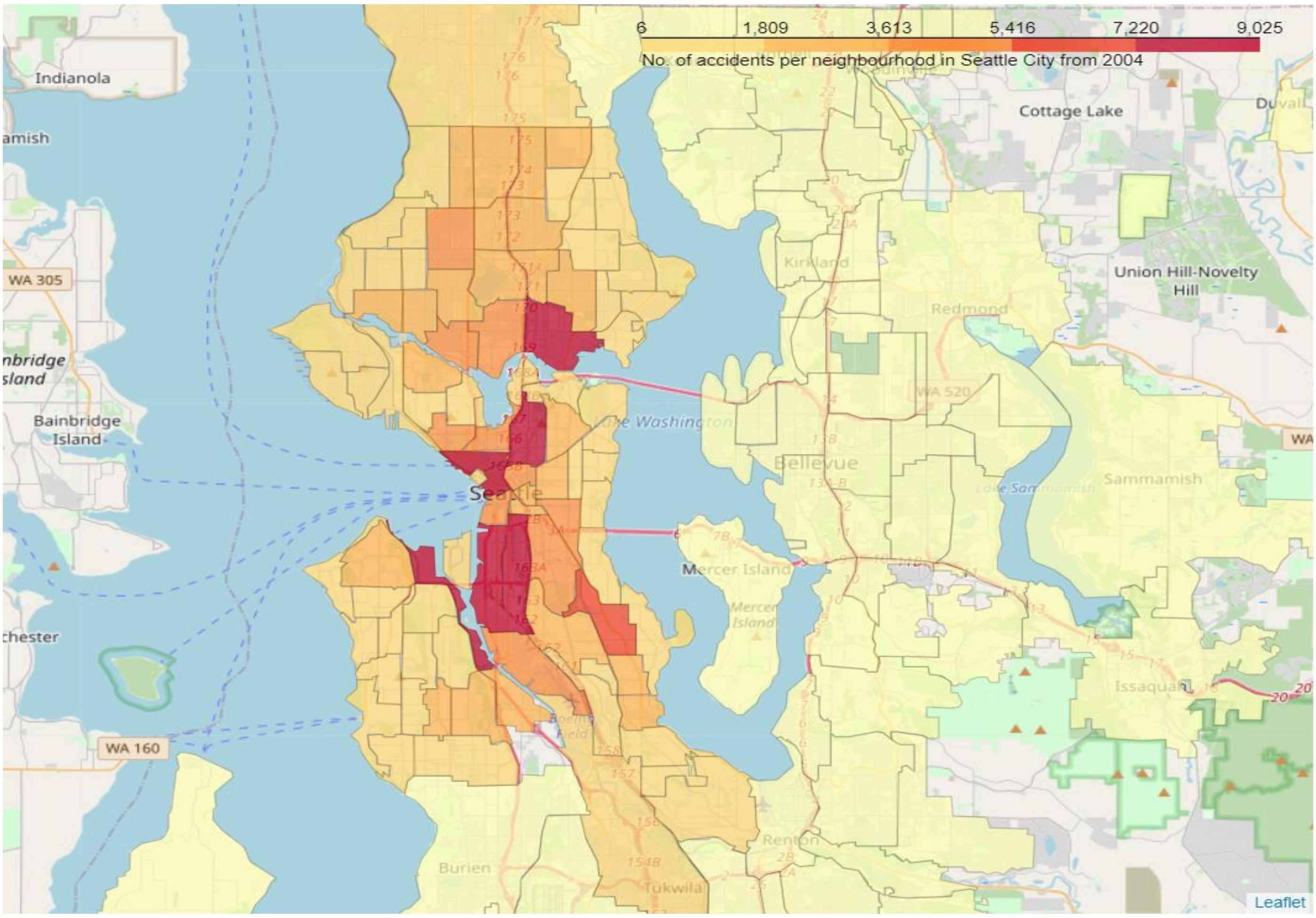

Daily Traffic Collisions in Seattle City (Severity)

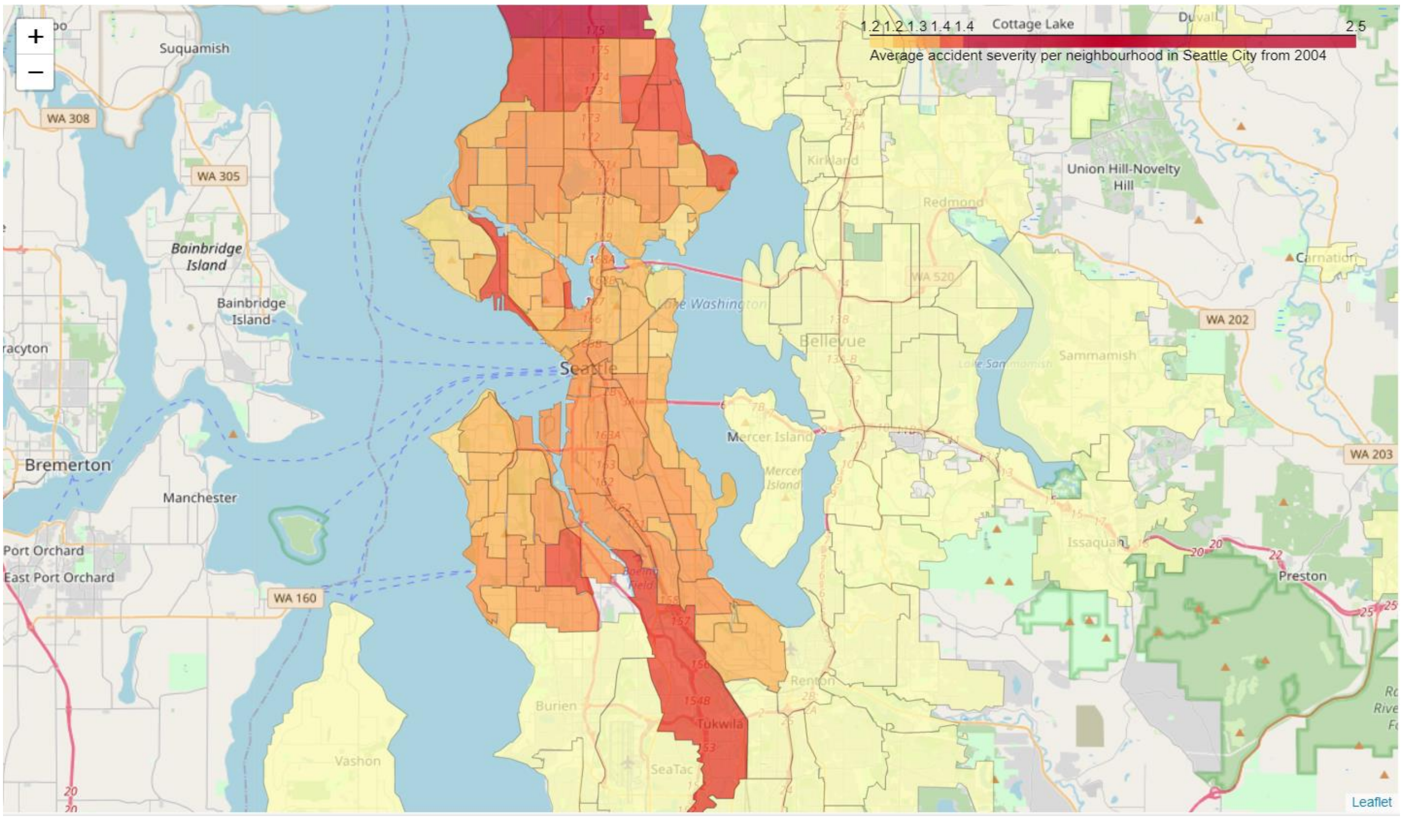


Location of Collision

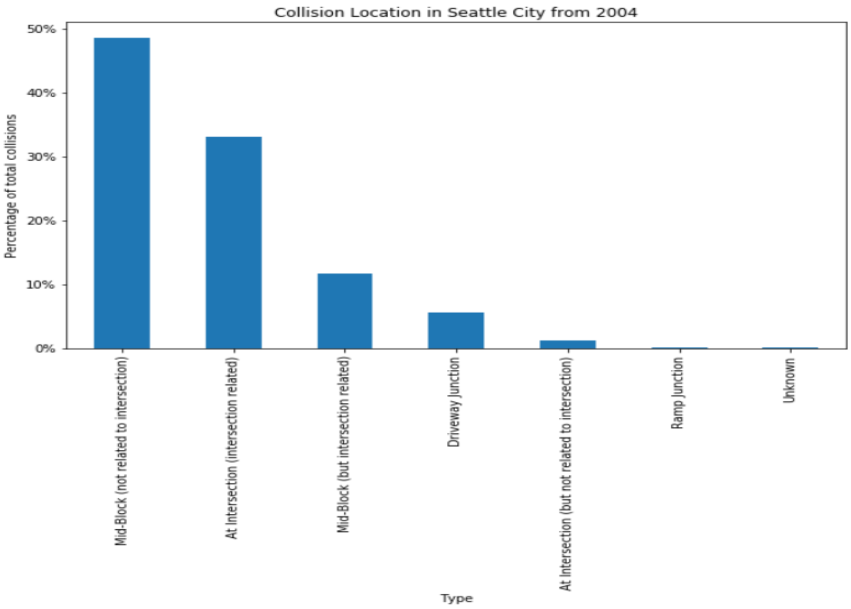
By using the columns X and Y which are the coordinates of the collisions, Choropleth maps reveal that accidents occur more frequently towards the centre of the city, and in neighbourhoods at either end of road bridges which straddle Seattle's major waterways. For instance, the table below shows the top ten areas regarding to the number of accidents. They include some of the most dangerous intersections in Seattle such as James Street and Sixth Avenue, Boren Avenue and Pike Street, Lake City Way NE and NE 130th Street, and Dexter Avenue North and Thomas Street. Also, they are one of the busiest arterials, and the corresponding higher pedestrian volumes and/or higher vehicle volumes would increase the opportunities for pedestrian-vehicle and vehicle-vehicle conflicts. We remove the severity code 0 for Unknown data, and these top areas reported an average severity code higher than 1 i.e. more accidents involved injuries and fatalities than others.

Neighborhood	Count	Mean_Severity
Belltown	9024	1.286348
Industrial District	8919	1.330530
University District	7811	1.298809
Central Business District	7487	1.309203
Broadway	7281	1.279907
Columbia City	5618	1.330367
Greenwood	5045	1.349653
South Lake Union	4773	1.309030
North Beacon Hill	4650	1.316344
Wallingford	4406	1.304585



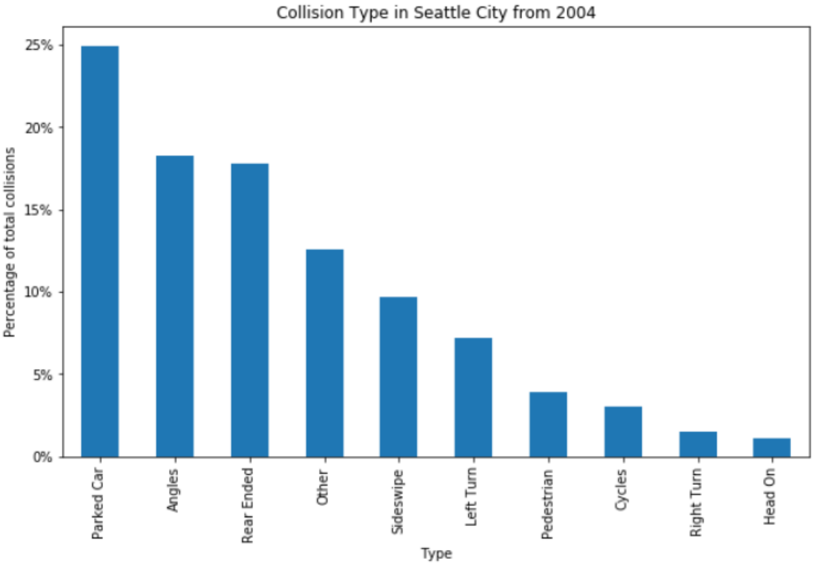


It seemed that collision occurred mostly in the mid-block of a segment (around 60% of total collision), with majority not related to an intersection (around 50%). But many collisions also took place at an intersection (33% of total collision), which is not surprising given intersections have the highest potential for conflicts—they have more users interacting and more movements. Around 5.5% of total collision was at the driveway junctions. And perhaps crashes were more likely to be severe at locations without a traffic signal.

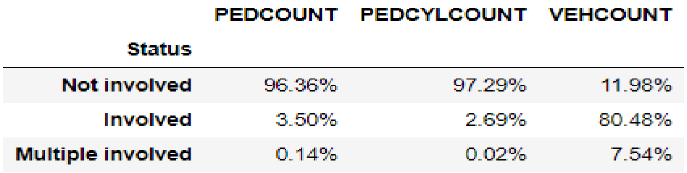


Type of Collision

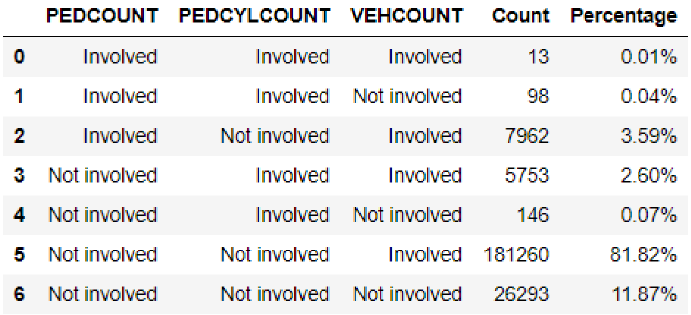
Perhaps surprisingly, the most common type of collision was collisions between a moving vehicle and a parked vehicle (around 25% of total collision). The following most common collisions were vehicles entering the flow of traffic at an angle (typically at intersections) and vehicles traveling in the same direction (the vast majority of which were likely rear-end collisions), each contributed around 20% of total collision in Seattle. Other common types of collisions are sideswipes, left turns, and right turns. Only 1% resulted in head-on collisions where the drivers might cross into another lane of traffic or go the wrong way on an exit ramp or street. There were some other types such as broadside collisions, vehicles hitting fixed objects (6% of total collisions involved hitting a parked car), etc., and they are grouped as 'Other' and accounted for 12% of total collisions.



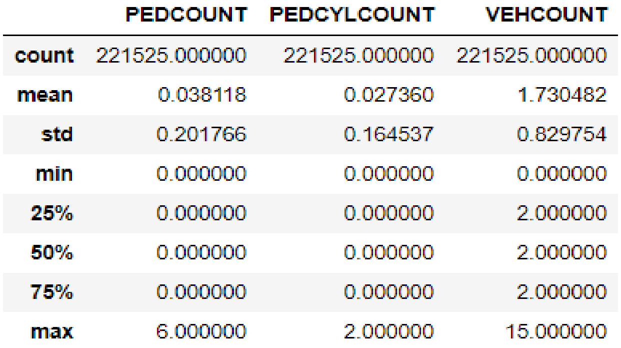
The bicyclists and pedestrians were involved in only 7% of all crashes, while around 90% involved at least one vehicle. Around 4% were pedestrian-vehicle collision and approximately 3% were bicyclist-vehicle collision. There were several cases where multiple people and vehicles are involved, nearly 8% of total collisions involved more than 2 vehicles. The maximum number of engagements in a collision was 6 for pedestrians, 2 for bicyclists and 15 for vehicles.



	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT
Status			
Not involved	96.36%	97.29%	11.98%
Involved	3.50%	2.69%	80.48%
Multiple involved	0.14%	0.02%	7.54%



	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	Count	Percentage
0	Involved	Involved	Involved	13	0.01%
1	Involved	Involved	Not involved	98	0.04%
2	Involved	Not involved	Involved	7962	3.59%
3	Not involved	Involved	Involved	5753	2.60%
4	Not involved	Involved	Not involved	146	0.07%
5	Not involved	Not involved	Involved	181260	81.82%
6	Not involved	Not involved	Not involved	26293	11.87%



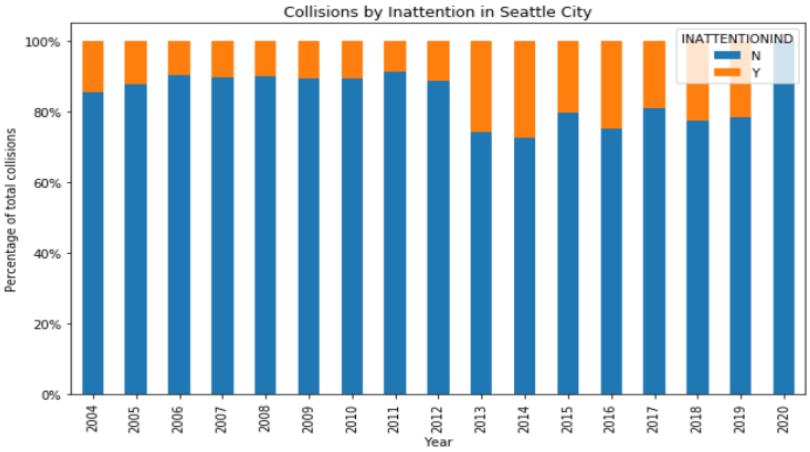
	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT
count	221525.000000	221525.000000	221525.000000
mean	0.038118	0.027360	1.730482
std	0.201766	0.164537	0.829754
min	0.000000	0.000000	0.000000
25%	0.000000	0.000000	2.000000
50%	0.000000	0.000000	2.000000
75%	0.000000	0.000000	2.000000
max	6.000000	2.000000	15.000000

Common Contributing Factors

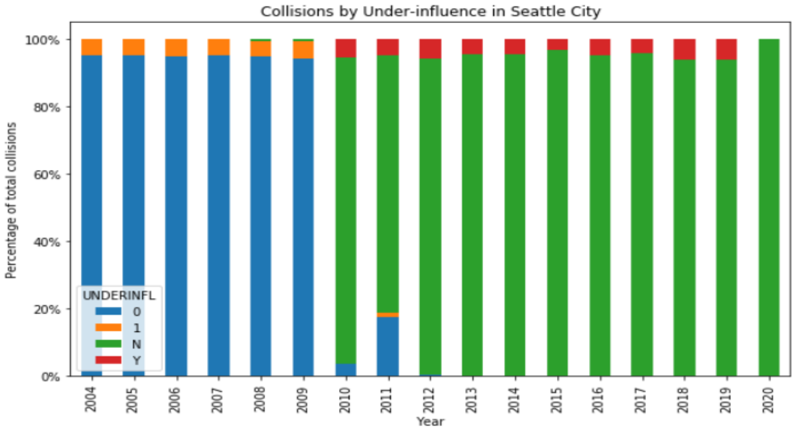
Car accidents happen for a host of reasons, including behavioural, environmental, and situational. A small number of car accidents are inevitable and cannot be prevented. Most of them, however, could at least be prevented, and many result from poor decisions by drivers who should have done better. The most common causes of car accidents are:

1. **Driver distraction**: distracted driving is a prevalent risky behaviour and hence a leading contributing cause of collisions in Seattle. Many circumstances are considered inattention like eating, grooming, using mobile phone, adjusting an audio system, other distractions inside the vehicle and distractions outside the vehicle. Every year around 1,000 to more than 2,000 cases were distraction-involved and the number has been increasing after 2011.

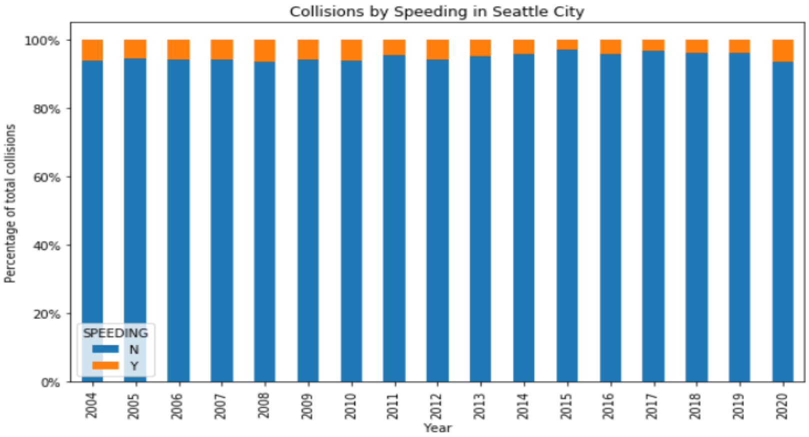
This accounted for around 20% of total collisions starting from 2013. In 2020, luckily there has been no such case occurring yet.



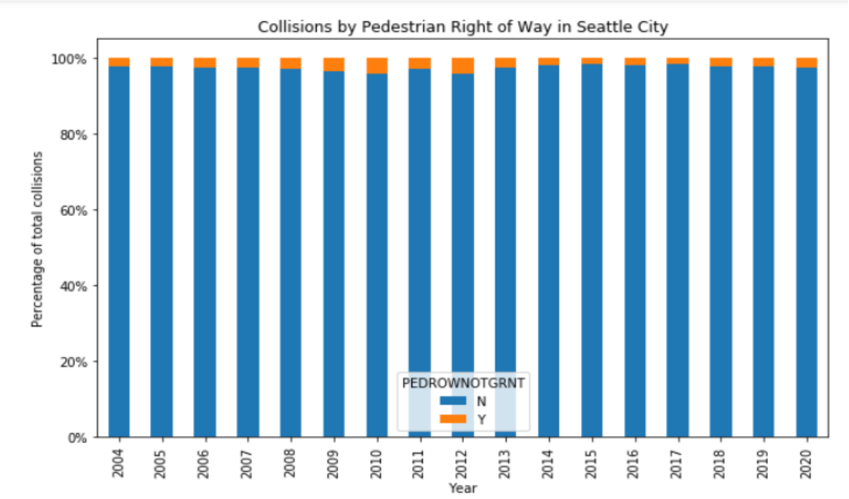
2. **Alcohol and drug impairment**: Driver impairment due to alcohol and/or drugs is one of the most contributing factors in fatal crashes and is involved in over 600 collisions per year, and this number has just slightly reduced over time without oblivious improvement. The dataset entries had a complete change starting 2012, switching from 0 - 1 to N - Y.



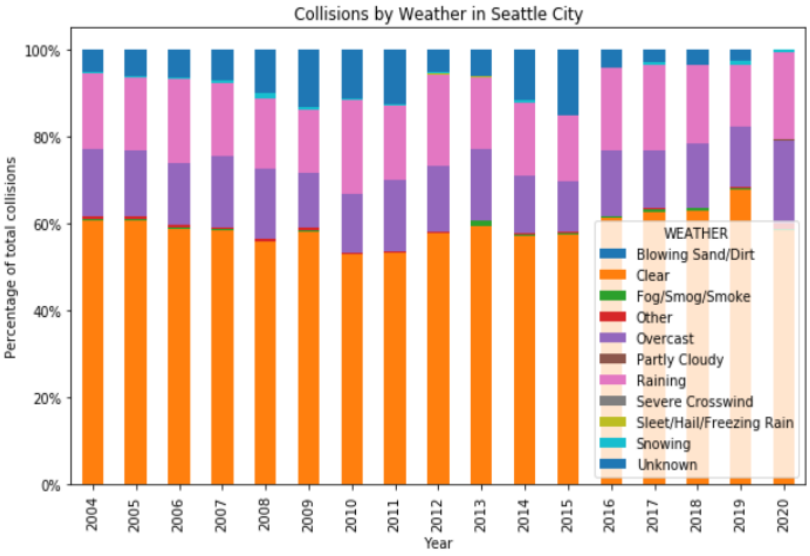
3. **Speeding**: Speeding remains a significant cause of car accidents in Seattle, noted as a contributing factor in over 400 collisions each year under the categories “exceeding reasonable and safe speed” and “exceeding the speed limit”. Conditions for pedestrians and bikers being hit by a vehicle also deteriorate with speed. But with continually improving road safety strategies, this number reduced from 900 collisions annually during the peak period 2005-2008.



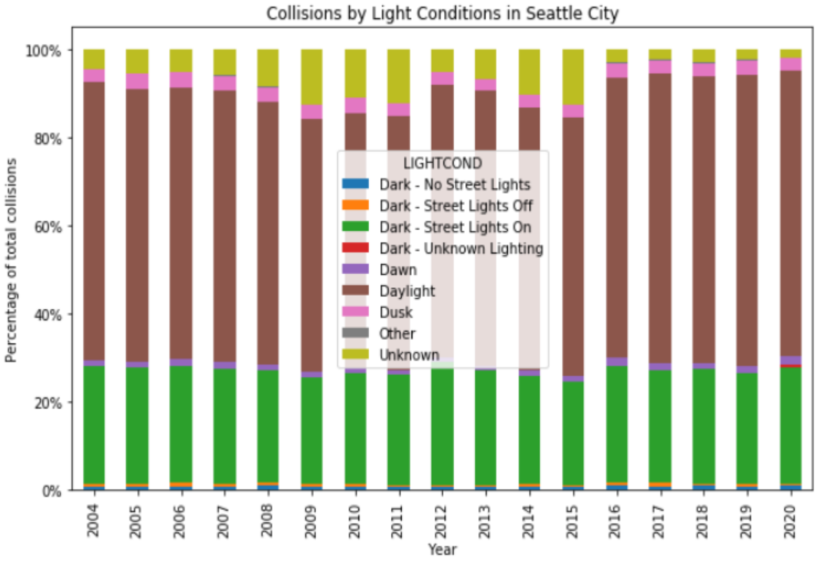
4. **Whether the pedestrian right of way was not granted**: This is a commonly cited factor for pedestrian and bicyclist-involved collisions in Seattle, causing over 200 cases annually but this number has decreased in recent years. This contributing factor generally indicates that a driver, pedestrian, or bicyclists stopped a fellow traveler from continuing their legal path. Examples of a “did not grant right-of-way” collision are unsafe crossing practices (such as crossing against a signal or failing to use a crosswalk), and disregarding traffic signals or stop signs.



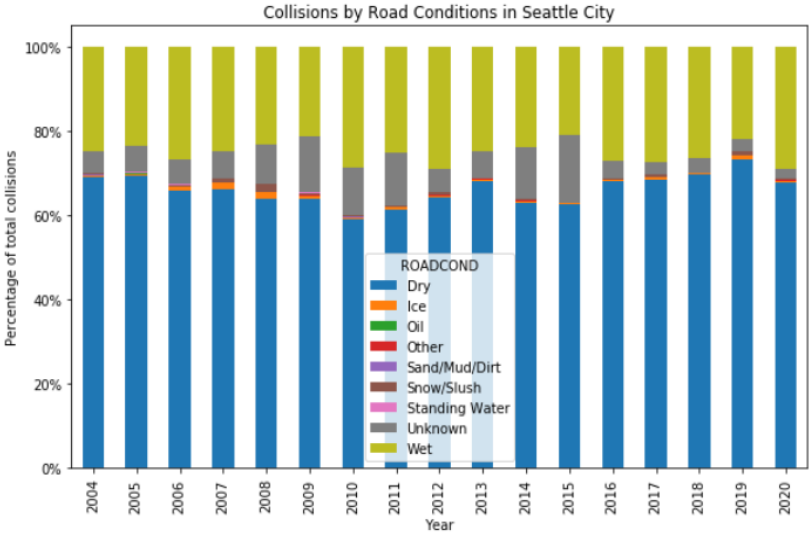
5. **Weather**: Weather acts through visibility impairments, slick pavement, high winds, and temperature extremes to affect driver capabilities and vehicle performance, leading to crash risk. Surprisingly, over 60% of total collisions occurred in a clear weather, whereas the vast majority of most weather-related crashes happened during rainfall and overcast weather, around 20% of all crashes respectively. A much smaller percentage of weather-related crashes occur during high winds blowing sand/dirt. The remaining cases took place in the presence of fog and snow, on icy pavement or under severe crosswinds.



6. **Light conditions**: Poor lighting conditions can limit the ability of a road user to see other users or road signals. In Seattle, again surprisingly, over 60% of the accidents occurred during daylight hours. Over 30% of these collisions occurred when it was dark out but there were streetlights on. There were still over 100 cases each year that were caused by the unavailability of streetlights. The remainder of the collisions occurred at dawn, dusk, or there was no record about the light conditions.



7. **Road conditions**: Drivers have an obligation to operate their vehicles in a manner appropriate for road conditions, but that is not always an easy task. Wet roads, in particular, contributed to nearly 30% of Seattle car accidents. The fact that over 60% of the accidents occurred with dry roads, combining with the weather and light conditions data, means that these collisions were due to human-related factors. The remainder of the collisions occurred with poor road conditions such as ice and snow not properly cleared, standing water, oil built up on the roads, and more.

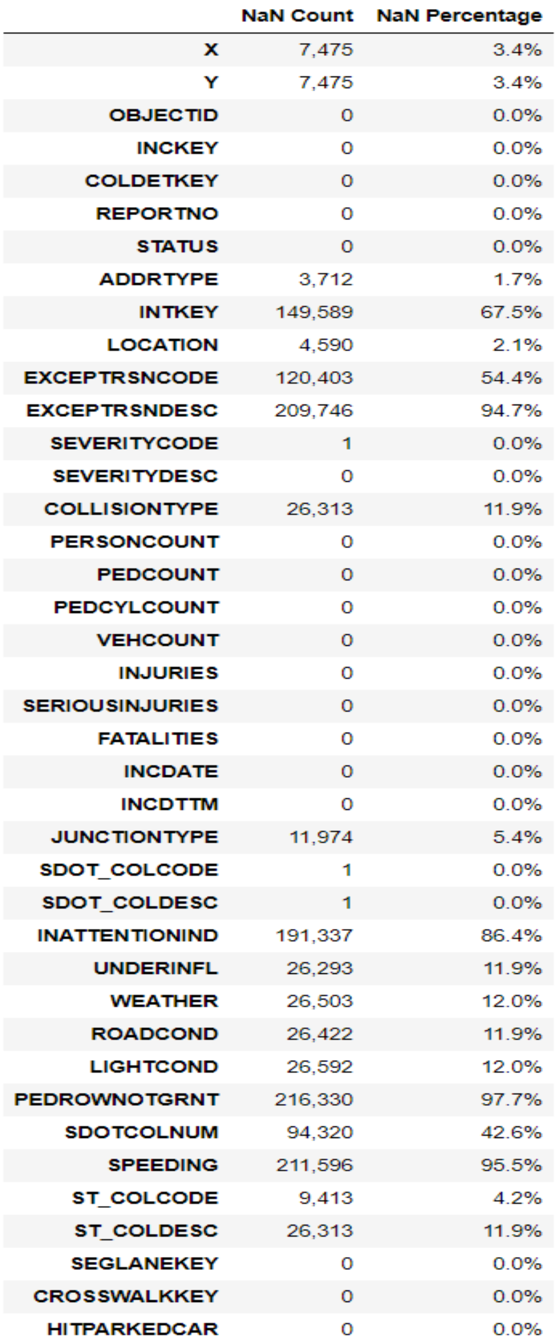


2. Data Preparation

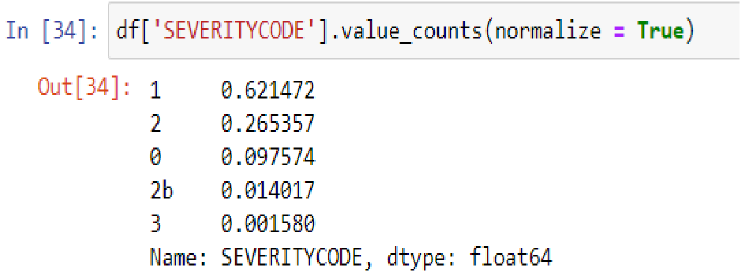
In their original form, the Seattle collision data is not suitable for quantitative data analysis and so likely to create noise in the modeling result. The main reasons for this are as follows:

Missing Values

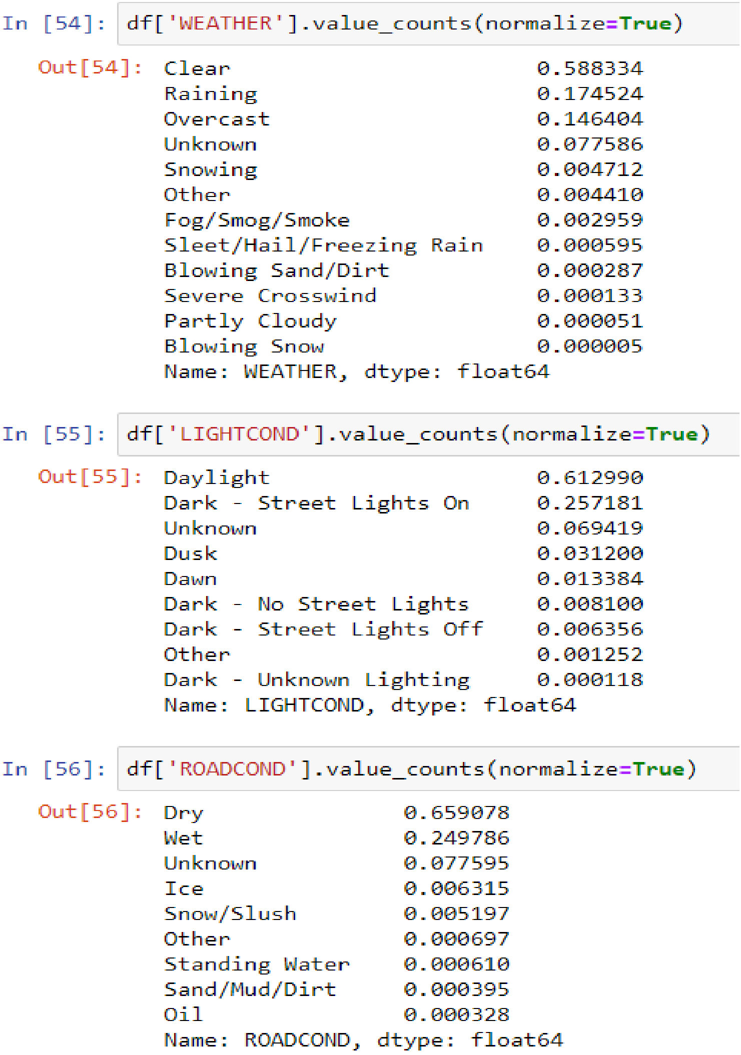
There are missing values on part of the database, around 10% of the target variable SEVERITYCODE is unknown and some attributes have over 40% of missing data (such as INTKEY, EXCEPTRSNCODE). Common attributing factors (such as JUNCTIONTYPE, WEATHER, LIGHTCOND, ROADCOND) include unknown entries which have same meaning as missing data.



	NaN Count	NaN Percentage
X	7,475	3.4%
Y	7,475	3.4%
OBJECTID	0	0.0%
INCKEY	0	0.0%
COLDETKEY	0	0.0%
REPORTNO	0	0.0%
STATUS	0	0.0%
ADDRTYPE	3,712	1.7%
INTKEY	149,589	67.5%
LOCATION	4,590	2.1%
EXCEPTRSNCODE	120,403	54.4%
EXCEPTRSNDESC	209,746	94.7%
SEVERITYCODE	1	0.0%
SEVERITYDESC	0	0.0%
COLLISIONTYPE	26,313	11.9%
PERSONCOUNT	0	0.0%
PEDCOUNT	0	0.0%
PEDCYLCOUNT	0	0.0%
VEHCOUNT	0	0.0%
INJURIES	0	0.0%
SERIOUSINJURIES	0	0.0%
FATALITIES	0	0.0%
INCDATE	0	0.0%
INCDTTM	0	0.0%
JUNCTIONTYPE	11,974	5.4%
SDOT_COLCODE	1	0.0%
SDOT_COLDESC	1	0.0%
INATTENTIONIND	191,337	86.4%
UNDERINFL	26,293	11.9%
WEATHER	26,503	12.0%
ROADCOND	26,422	11.9%
LIGHTCOND	26,592	12.0%
PEDROWNOTGRNT	216,330	97.7%
SDOTCOLNUM	94,320	42.6%
SPEEDING	211,596	95.5%
ST_COLCODE	9,413	4.2%
ST_COLDESC	26,313	11.9%
SEGLANEKEY	0	0.0%
CROSSWALKKEY	0	0.0%
HITPARKEDCAR	0	0.0%



```
In [34]: df['SEVERITYCODE'].value_counts(normalize = True)

Out[34]: 1     0.621472
         2     0.265357
         0     0.097574
         2b    0.014017
         3     0.001580
         Name: SEVERITYCODE, dtype: float64
```



```
In [54]: df['WEATHER'].value_counts(normalize=True)

Out[54]: Clear                       0.588334
         Raining                     0.174524
         Overcast                    0.146404
         Unknown                     0.077586
         Snowing                     0.004712
         Other                       0.004410
         Fog/Smog/Smoke              0.002959
         Sleet/Hail/Freezing Rain    0.000595
         Blowing Sand/Dirt           0.000287
         Severe Crosswind            0.000133
         Partly Cloudy               0.000051
         Blowing Snow                0.000005
         Name: WEATHER, dtype: float64

In [55]: df['LIGHTCOND'].value_counts(normalize=True)

Out[55]: Daylight                    0.612990
         Dark - Street Lights On     0.257181
         Unknown                     0.069419
         Dusk                        0.031200
         Dawn                        0.013384
         Dark - No Street Lights     0.008100
         Dark - Street Lights Off    0.006356
         Other                       0.001252
         Dark - Unknown Lighting     0.000118
         Name: LIGHTCOND, dtype: float64

In [56]: df['ROADCOND'].value_counts(normalize=True)

Out[56]: Dry                0.659078
         Wet                0.249786
         Unknown            0.077595
         Ice                0.006315
         Snow/Slush         0.005197
         Other              0.000697
         Standing Water     0.000610
         Sand/Mud/Dirt      0.000395
         Oil                0.000328
         Name: ROADCOND, dtype: float64
```

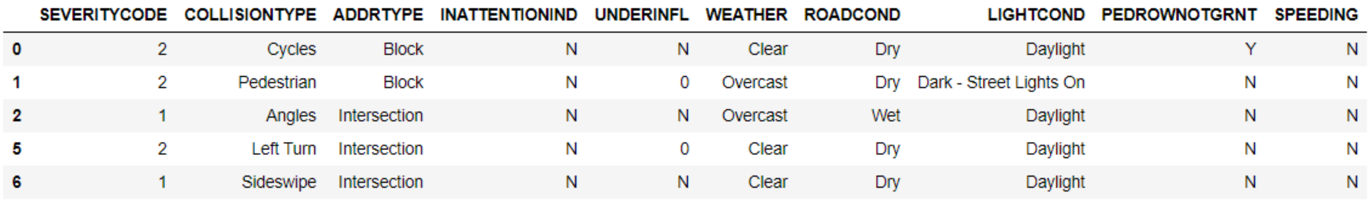
Unnecessary Attributes

1. The dataset includes columns that provide unnecessary information for the machine learning algorithms, for instance, columns of metadata (such as OBJECTID, INCKEY, COLDETKEY, REPORTNO, STATUS, SDOT_COLCODE, SDOTCOLNUM, ST_COLCODE, SEGLANEKEY, CROSSWALKKEY) and description columns (such as EXCEPTRSNDESC, SEVERITYDESC, SDOT_COLDESC, ST_COLDESC).
2. As the purpose of building the model is to see how various factors interact and influence the collision severity, the columns of location, date and involved numbers (such as X, Y, Location, INCDATE, INCDTTM, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, INJURIES, SERIOUSINJURIES, FATALITIES) would give redundant information.

3. Some columns give duplicated information, for example, the column HITPARKEDCAR indicates whether a collision involves hitting parked car which is also given in column COLLISIONTYPE, the column JUNCTIONTYPE and column ADDRTYPE both gives the type of collision location.

Feature Selection

Overall, such missing data entries and irrelevant data attributes will be removed from the dataset. Except for INATTENTIONIND, PEDROWNOTGRNT and SPEEDING columns where there are two types of entry which are 'Yes' and NaN, so we assume that NaN means 'No' entry. We have an updated dataset collision_df with 175,068 records and 10 attributes as follows:



	SEVERITYCODE	COLLISIONTYPE	ADDRTYPE	INATTENTIONIND	UNDERINFL	WEATHER	ROADCOND	LIGHTCOND	PEDROWNOTGRNT	SPEEDING
0	2	Cycles	Block	N	N	Clear	Dry	Daylight	Y	N
1	2	Pedestrian	Block	N	0	Overcast	Dry	Dark - Street Lights On	N	N
2	1	Angles	Intersection	N	N	Overcast	Wet	Daylight	N	N
5	2	Left Turn	Intersection	N	0	Clear	Dry	Daylight	N	N
6	1	Sideswipe	Intersection	N	N	Clear	Dry	Daylight	N	N

Encoding categorical variables

For target variable SEVERITYCODE, to focus on the collision severity of property damage - injury - fatality, we combine the group 2b of serious injury with group 2 of injury. As we removed the severity code 0 for unknown data, we re-code the collision severity to include code 0 for property damage, code 1 for injury and code 2 for fatality. The remaining attributes in the dataset, which are the independent X variables, are all categorical variables and this is not applicable for Machine Learning algorithms.

- For INATTENTIONIND, UNDERINFL, PEDROWNOTGRNT and SPEEDING that have two types of record ('N' and 'Y'), we use label encoding to recast each of these categorical variables as a series of numerical values where 'N' → 0, 'Y' → 1.
- For COLLISIONTYPE, ADDRTYPE, WEATHER, LIGHTCOND and ROADCOND, we recast these variables as numeric data using the one-hot encoding. This involves the pandas function get_dummies to create indicator variables (or dummy variables) for each unique category, and to assign value 1 if the data falls into that category and value 0 otherwise. Afterwards, the original columns of categorical variables are dropped from the dataset. This step increases the total attributes in the dataset from 10 attributes to 44 attributes.

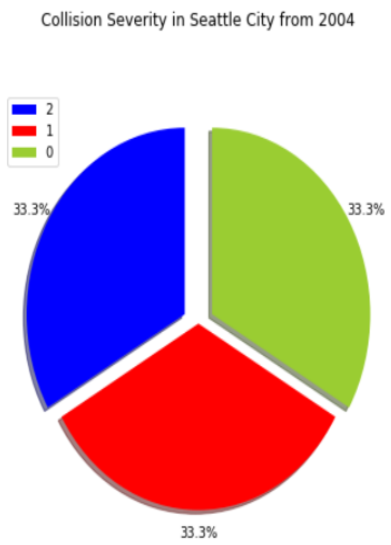
Balancing data

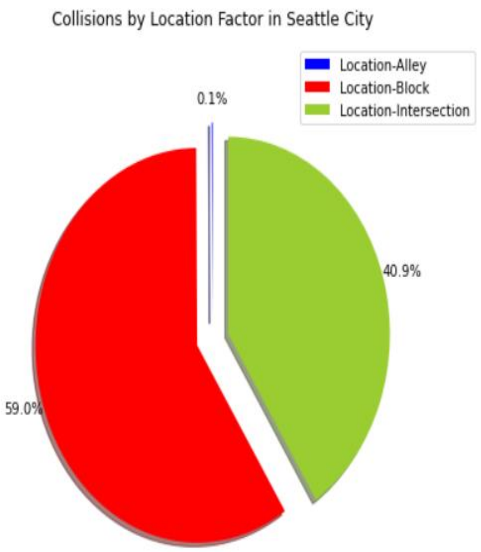
As the requirement of the Machine Learning algorithms, the clean dataset need to be balanced before being used. Here if we train a model to predict the collision severity using a dataset in which over 60% of the collisions have one particular outcome (property damage), it is likely that the model will produce some biased results. Hence, we apply the data balancing by randomly selecting N_2 samples from the larger data groups (severity code 0 for property damage and code 1 for injury) and re-group these values into smaller groups, where N_2 is the number of accidents with severity code 2 for fatality.

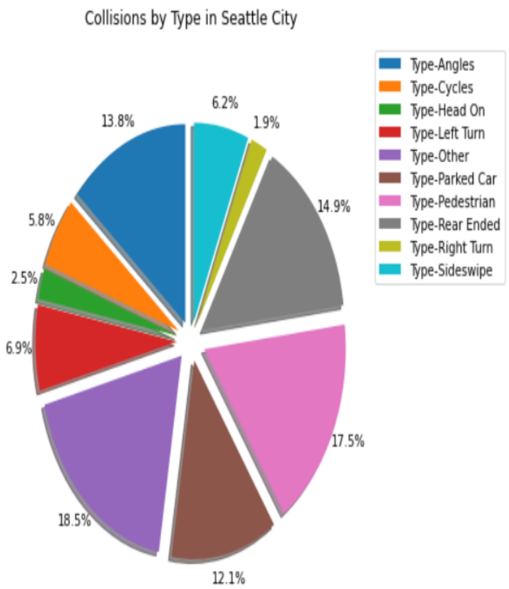
We ensure that no attribute is left with no variation in values, i.e. that attribute only has one value available for modeling; thus, Weather-Blowing Sand/Dirt, Road-Oil, Road-Other, Light-Dark - Unknown Lighting, Light-Other columns are removed. As a result, we have a resampled clean dataset resampled_df with 1,008 records and the 39 attributes:

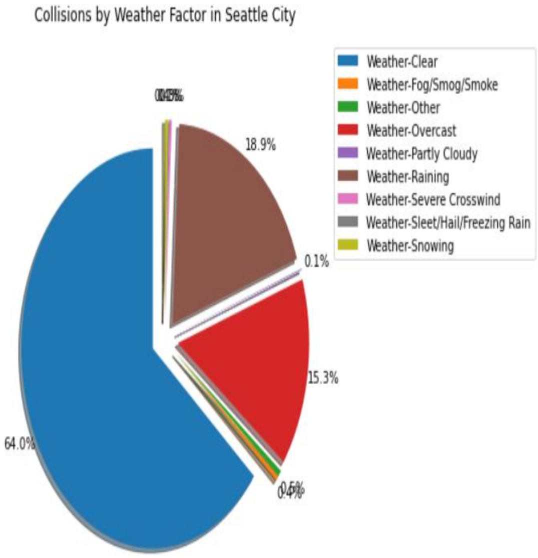
	No. of collisions
SEVERITYCODE	1008
INATTENTIONIND	128
UNDERINFL	132
PEDROWNOTGRNT	81
SPEEDING	137
Type-Angles	139
Type-Cycles	58
Type-Head On	25
Type-Left Turn	70
Type-Other	186
Type-Parked Car	122
Type-Pedestrian	176
Type-Rear Ended	150
Type-Right Turn	19
Type-Sideswipe	63
Location-Alley	1
Location-Block	595
Location-Intersection	412
Weather-Clear	645
Weather-Fog/Smog/Smoke	4
Weather-Other	5
Weather-Overcast	154
Weather-Partly Cloudy	1
Weather-Raining	191
Weather-Severe Crosswind	3
Weather-Sleet/Hail/Freezing Rain	1
Weather-Snowing	4
Road-Dry	734
Road-Ice	5
Road-Sand/Mud/Dirt	1
Road-Snow/Slush	4
Road-Standing Water	2
Road-Wet	262
Light-Dark - No Street Lights	6
Light-Dark - Street Lights Off	9
Light-Dark - Street Lights On	339
Light-Dawn	16
Light-Daylight	602
Light-Dusk	36

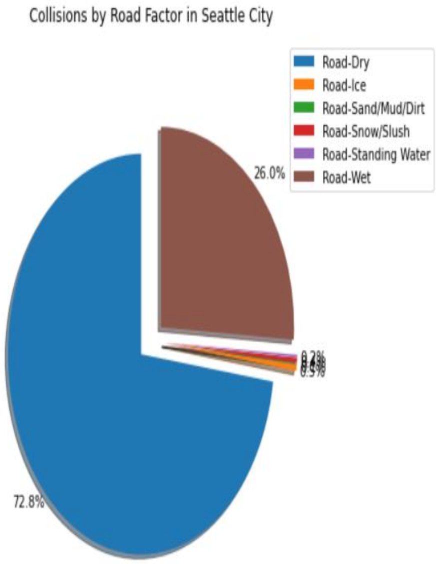
With the resampled dataset, we have an equal amount of data between three levels of severity, namely property damage - injury - fatality. The distribution of each remaining attributes is like that before resampling.

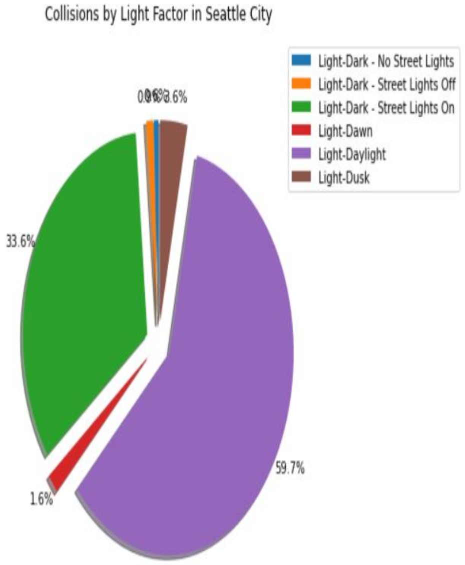


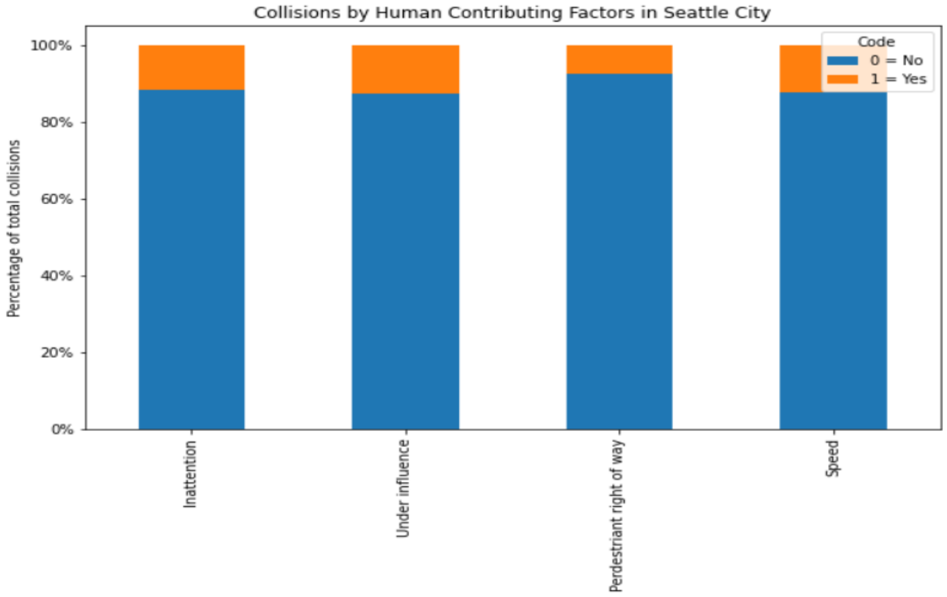






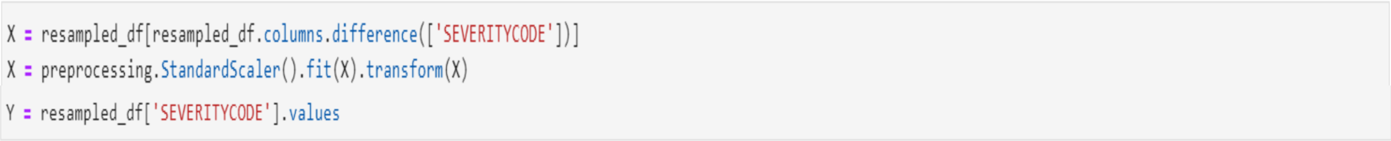






Normalizing Data

Having a cleaned and balanced dataset, we separate the independent variables to feature set X and dependent variable SEVERITYCOD to set Y. We use StandardScaler package from Scikit Learn to standardise the feature set. This package recasts each feature in the feature set having distribution with zero mean and unit variance. It helps to ensure that all features are on a similar scale and the model will not be biased towards or away from certain features due to the range numerical values assigned to those features.



```
X = resampled_df[resampled_df.columns.difference(['SEVERITYCODE'])]
X = preprocessing.StandardScaler().fit(X).transform(X)
Y = resampled_df['SEVERITYCODE'].values
```

Creating the Training and Testing subsets

After the data wrangling, we split the dataset into training and testing subsets, which are mutually exclusive, as training and testing on the same dataset will most likely have low out-of-sample accuracy due to the likelihood of the model being over-fit. It is important for the model to have a high out-of-sample accuracy, because the purpose is to make correct predictions on new unknown data. The split is performed using the following ratio:

- 70% for training to build an accurate model, and
- 30% for testing to report the accuracy of the model.

III. Collision Severity Model

1. Machine Learning Algorithm Selection

The following algorithms are used for model development:

- **K Nearest Neighbor (KNN)**: This is a classification algorithm that takes a bunch of labelled data points and uses them to learn how to label other points, based on their similarity to other cases. Similar cases with the same class labels are near each other, and the distance between two data points can be calculated using Euclidean distance. Once a new point is to be predicted, it takes into account the 'K' nearest points to it to determine its classification.
- **Decision Tree**: This is another classification algorithm that breaks down a dataset into smaller subsets to incrementally develop a decision tree. It tests an attribute and branching the data points based on the result of the test. So, a decision tree includes internal nodes each corresponding to a test, each branch in an internal node corresponding to a result of the test, and each leaf node in a branch assigning a data point to a class.
- **Support Vector Machine**: This works by mapping data to a high-dimensional feature space so that data points can be categorized. A separator between the categories is estimated, then the data is transformed in such a way that the separator could be drawn as a hyperplane. Following this, characteristics of new unknown data can be used to predict the group to which it should belong.
- **Logistic Regression**: This is a statistical and machine learning technique that, in its basic form, produces a formula to predict the probability of the class label regarding the dependent variable Y as a logistic function of the independent variables Xs. It is useful when variable Y is categorical, as Linear regression is more suited for estimating continuous values.

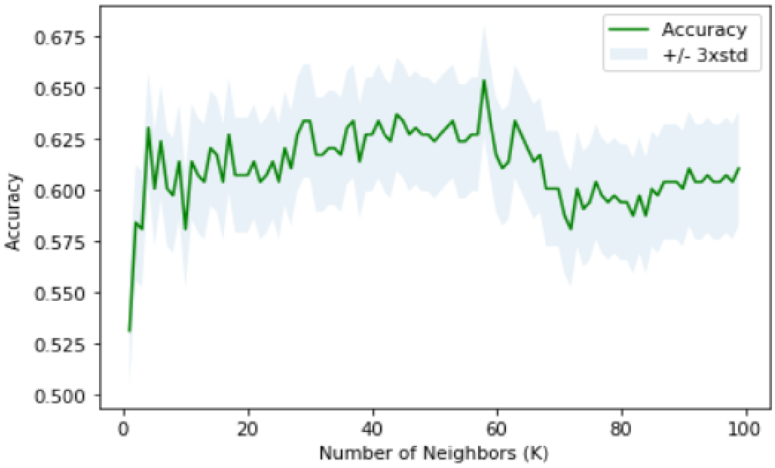
The model performance is evaluated using the accuracy measurements, such as:

- **Jaccard index**: This is calculated using jaccard_similarity_score function (accuracy classification score function would provide the same result), to show how closely the actual labels and predicted labels are matched in the test set. It is the size of the intersection divided by the size of the union of two label sets Y and $\widehat{Y}$, where Y is the true label based on the testing data and $\widehat{Y}$ is the predicted label using the model.
- **F1-score**: This is the harmonic average of the precision and recall of the model, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0 (if either the precision or the recall is 0). The precision is the number of correctly identified positive results divided by the number of all positive results, including those not identified correctly. The recall is the number of correctly identified positive results divided by the number of all samples that should have been identified as positive.
- **Log loss**: This is accuracy measurement for Logistic regression, where the predicted output is probability value between 0 and 1 of a class label, instead of the label. A model with a lower log loss would have a better accuracy.
- **Confusion matrix**: This shows the corrected and wrong predictions, in comparison with the actual labels. Each row shows the Actual/True labels in testing set, and columns show predicted labels by the model.

2. Model Development

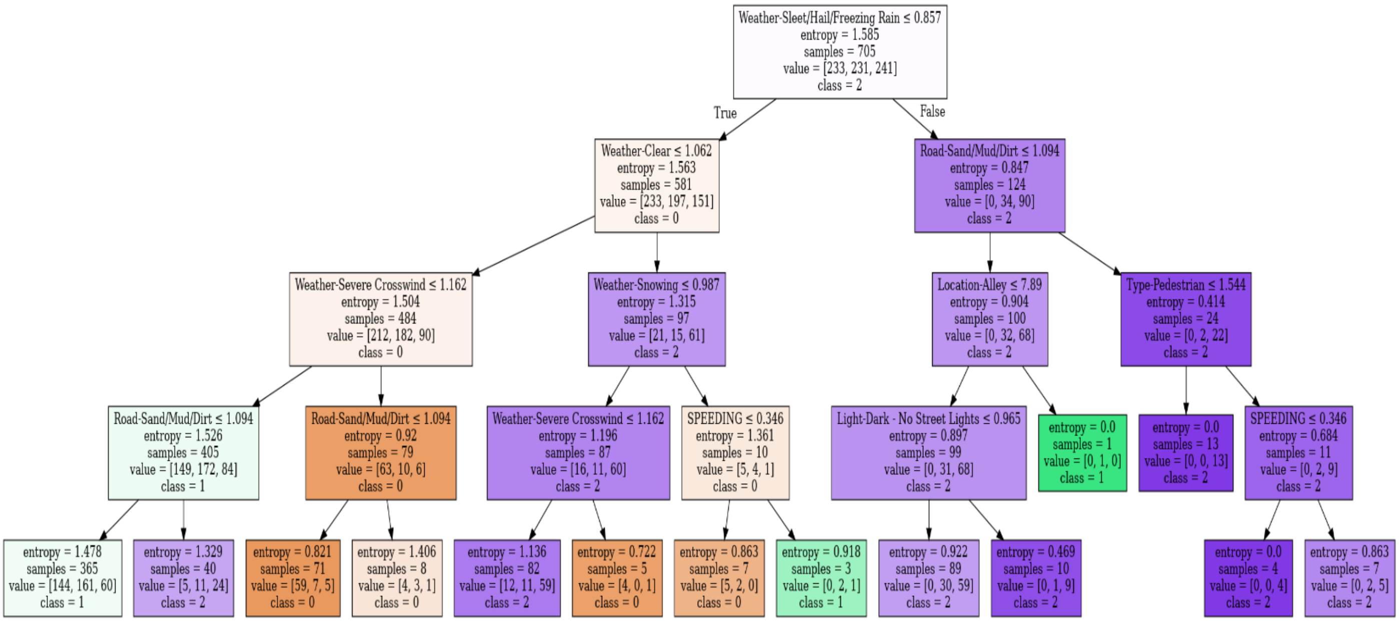
K Nearest Neighbor (KNN)

K in KNN is the number of nearest neighbors to examine. The optimal choice of K is highly dependent on the dataset and is important for the model performance. a low value of K can cause an overly complex model, which might result in overfitting of model i.e. the prediction process is not generalized enough to be used for out-of-sample cases. On the other hand, a high value of K can cause an overly generalized model. To choose the right value for K, we perform several iterations with different values of K. Each iteration involves using the training set for modeling and calculating the accuracy of prediction using the testing set. The KNN model is built using the KNeighborsClassifier package in Scikit Learn with parameter n_neighbors = K. Aside from updating K after each iteration, all other parameters in the model are kept at default. It turns out the best K = 58 which provides the best accuracy of 0.65. So, K = 58 will be used to train our model for collision severity that later predicts the testing set.



Decision Tree

The objective is to first find an attribute from the dataset that is most predictive in splitting data within the training set, based on the values of that attribute. Then in each branch, the process is repeated using other attributes to split the subtrees, until all the leaf nodes are pure. The impurity of the nodes is calculated by entropy which is the amount of information disorder. A node is considered pure if in all cases they fall into a specific category of target variable, the severity code. The decision tree model is built using the DecisionTreeClassifier package in Scikit Learn, with criterion 'entropy'. The tree requires 19 layers of branching depth to achieve leaf node purity, and WEATHER is the most predictive attribute as being the initial branching level. To get a sample review of the decision tree, we set the branching depth to 4, as shown below:

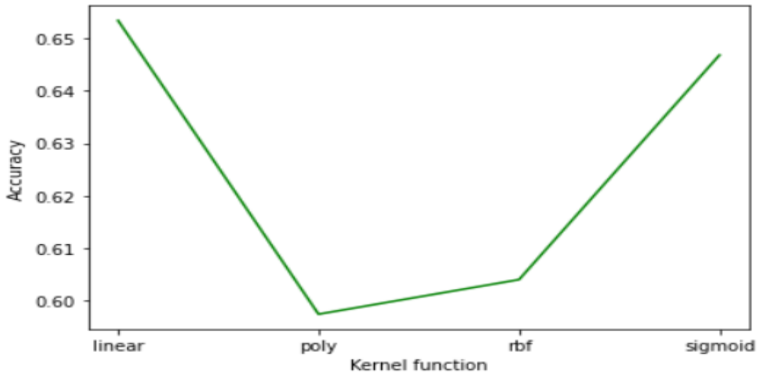

Weather-Sleet/Hail/Freezing Rain ≤ 0.857 entropy = 1.585 samples = 705 value = [233, 231, 241] class = 2
True
False
Weather-Clear ≤ 1.062 entropy = 1.563 samples = 581 value = [233, 197, 151] class = 0
Road-Sand/Mud/Dirt ≤ 1.094 entropy = 0.847 samples = 124 value = [0, 34, 90] class = 2
Weather-Severe Crosswind ≤ 1.162 entropy = 1.504 samples = 484 value = [212, 182, 90] class = 0
Weather-Snowing ≤ 0.987 entropy = 1.315 samples = 97 value = [21, 15, 61] class = 2
Location-Alley ≤ 7.89 entropy = 0.904 samples = 100 value = [0, 32, 68] class = 2
Type-Pedestrian ≤ 1.544 entropy = 0.414 samples = 24 value = [0, 2, 22] class = 2
Road-Sand/Mud/Dirt ≤ 1.094 entropy = 1.526 samples = 405 value = [149, 172, 84] class = 1
Road-Sand/Mud/Dirt ≤ 1.094 entropy = 0.92 samples = 79 value = [63, 10, 6] class = 0
Weather-Severe Crosswind ≤ 1.162 entropy = 1.196 samples = 87 value = [16, 11, 60] class = 2
SPEEDING ≤ 0.346 entropy = 1.361 samples = 10 value = [5, 4, 1] class = 0
Light-Dark - No Street Lights ≤ 0.965 entropy = 0.897 samples = 99 value = [0, 31, 68] class = 2
entropy = 0.0 samples = 1 value = [0, 1, 0] class = 1
entropy = 0.0 samples = 13 value = [0, 0, 13] class = 2
SPEEDING ≤ 0.346 entropy = 0.684 samples = 11 value = [0, 2, 9] class = 2
entropy = 1.478 samples = 365 value = [144, 161, 60] class = 1
entropy = 1.329 samples = 40 value = [5, 11, 24] class = 2
entropy = 0.821 samples = 71 value = [59, 7, 5] class = 0
entropy = 1.406 samples = 8 value = [4, 3, 1] class = 0
entropy = 1.136 samples = 82 value = [12, 11, 59] class = 2
entropy = 0.722 samples = 5 value = [4, 0, 1] class = 0
entropy = 0.863 samples = 7 value = [5, 2, 0] class = 0
entropy = 0.918 samples = 3 value = [0, 2, 1] class = 1
entropy = 0.922 samples = 89 value = [0, 30, 59] class = 2
entropy = 0.469 samples = 10 value = [0, 1, 9] class = 2
entropy = 0.0 samples = 4 value = [0, 0, 4] class = 2
entropy = 0.863 samples = 7 value = [0, 2, 5] class = 2

Support Vector Machine (SVM)

The kernel function used for the transformation can be of different types, such as:

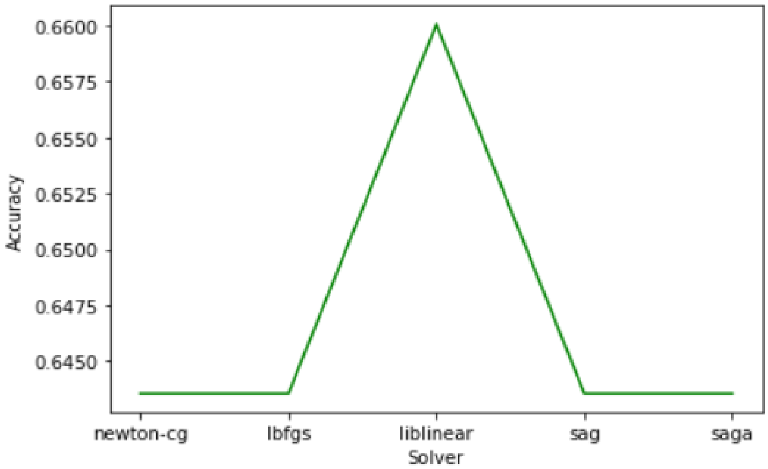
- Linear
- Polynomial
- Radial basis function (RBF)
- Sigmoid

To choose the kernel function for SVM model, we perform several iterations with different function. Each iteration involves using the training set for modeling and calculating the accuracy of prediction using the testing set. The SVM model is built using the SVM package in Scikit Learn. Aside from updating parameter kernel after each iteration, all other parameters in the model are kept at default. It turns out the best kernel function is linear function which provides the best accuracy of 0.65. So, kernel = 'linear' will be used to train our model for collision severity that later predicts the testing set.



Logistic Regression

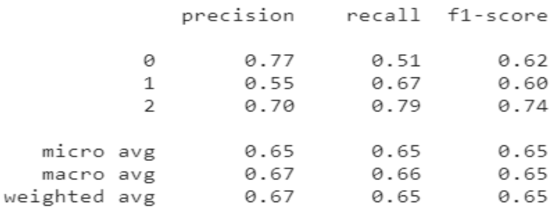
The SVM model is built using the LogisticRegression package in Scikit Learn. This function implements logistic regression and can use different numerical optimizers to find parameters, including ‘newton-cg’, ‘lbfgs’, ‘liblinear’, ‘sag’, ‘saga’ solvers. To choose the solver for SVM model, we perform several iterations with different function. Each iteration involves using the training set for modeling and calculating the accuracy of prediction using the testing set. Aside from updating parameter solver after each iteration, all other parameters in the model are kept at default. It turns out the best solver is liblinear which provides the best accuracy of 0.66. So, solver = 'liblinear' will be used to train our model for collision severity that later predicts the testing set.



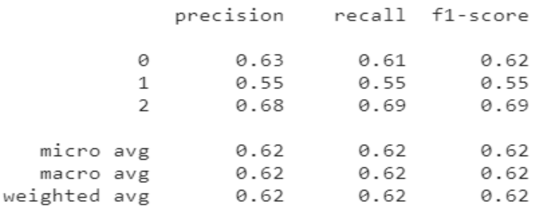
3. Model Evaluation

The comparison of the accuracy produced by the models is as follows:

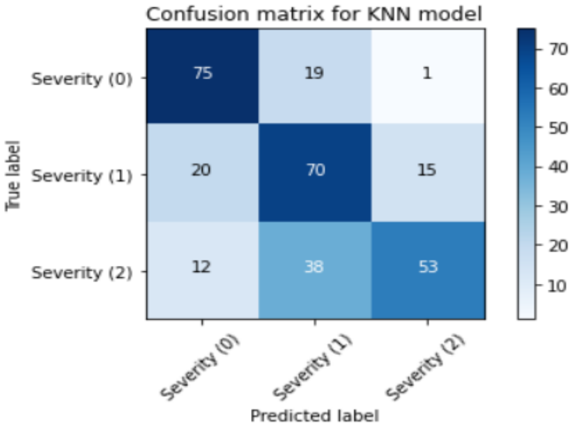
Algorithm	Jaccard	F1-score	LogLoss
KNN	0.65	0.65	NA
Decision Tree	0.62	0.62	NA
SVM	0.65	0.65	NA
Logistic Regression	0.66	0.65	0.87

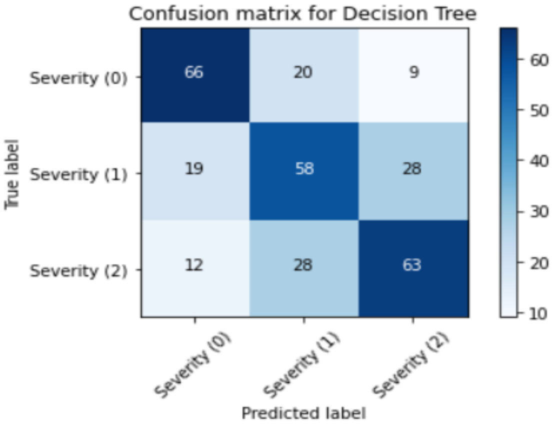


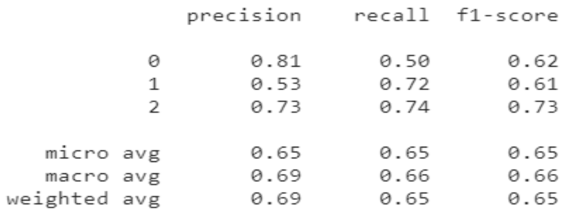
```
              precision    recall  f1-score

           0       0.77      0.51      0.62
           1       0.55      0.67      0.60
           2       0.70      0.79      0.74

   micro avg       0.65      0.65      0.65
   macro avg       0.67      0.66      0.65
weighted avg       0.67      0.65      0.65
```



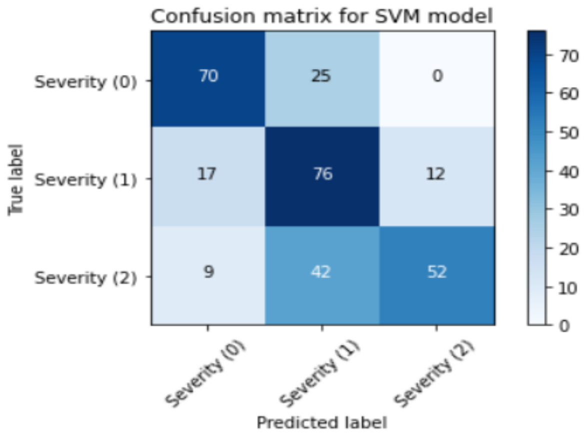
```
              precision    recall  f1-score

           0       0.63      0.61      0.62
           1       0.55      0.55      0.55
           2       0.68      0.69      0.69

   micro avg       0.62      0.62      0.62
   macro avg       0.62      0.62      0.62
weighted avg       0.62      0.62      0.62
```

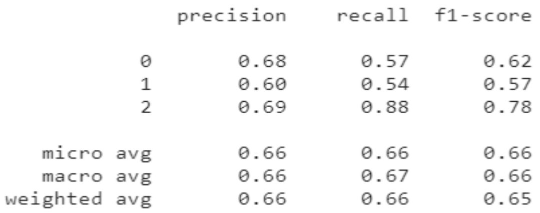




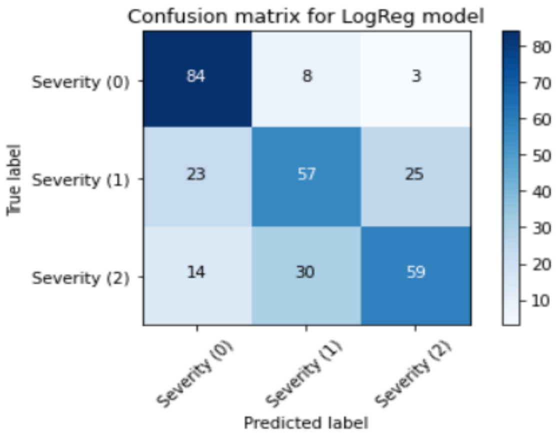


```
              precision    recall  f1-score

           0       0.81      0.50      0.62
           1       0.53      0.72      0.61
           2       0.73      0.74      0.73

   micro avg       0.65      0.65      0.65
   macro avg       0.69      0.66      0.66
weighted avg       0.69      0.65      0.65
```





```
              precision    recall  f1-score

           0       0.68      0.57      0.62
           1       0.60      0.54      0.57
           2       0.69      0.88      0.78

   micro avg       0.66      0.66      0.66
   macro avg       0.66      0.67      0.66
weighted avg       0.66      0.66      0.65
```



On examining the scores, the confusion matrix, and the classification report, we can see that:

- The accuracy values fluctuated from 60% to 80%, which means the models could offer the prediction of the severity of future collisions with 60% ~ 80% accuracy. All four models are good predictor of severity code 2, but the accuracy drastically drops for severity code 0 and code 1.
- The severity code 2 should be given more preference than the severity code 0 and code 1 due to a bigger impact. In that case, the Logistic Regression Model is the best model for a better prediction for the collision severity, with the highest overall Jaccard index (66%) and highest F1-score for severity code 2.
- Decision Tree has the lowest accuracy scores among the four preferred models. The accuracy scores of KNN Model and SVM models are slightly lower (65%) than those of Logistic Regression Model but shows a small higher F1-score for severity code 1, if that is the main aim for modeling.

IV. Conclusion

In this study, the collision data from 2004 of SDOT was used to train and test the models which predict the severity of a collision. The main purpose is to help the transportation and health departments in Seattle City to efficiently allocate their resources to mitigate the occurrence of collisions and minimise the severity when they do occur. Based on the output and accuracy scores, it concludes that Logistic Regression Model is the preferred model to predict the collision severity in Seattle City, with an accuracy score of ~66%.

The collision type, address type, human-related factors (inattention, under influence, speeding, and right of way) and external condition (weather, road, and light) constructively affect the collision severity. It was found that over 60% of the collisions happened in Clear weather, Daylight on Dry Road, which could mean that the human-related factors and road design play a major role in the occurrence and severity of collisions.

Further Suggestions

The SDOT should enhance data recording procedure to mitigate the missing values in the collision dataset. This could increase the accuracy of machine learning algorithms. In addition, the number of accidents in Seattle City has declined considerably since 2006, and there might have been a gradual change in the contributing factors as a result of different road strategies and enforcements being applied. Thus, the models may be biased if including the old data that is no longer relevant. In future, we need to review the models for performance and impact, and refine the models if there are changes in e.g. data recording, behaviour of road users, etc. This helps the models to improve their performance and remain on track to generate the intended solution.

Seattle City can use the insights gained from the data to prevent crashes, for example:

- Extra precautions on the road condition and light condition that were not ideal for a specific area,
- Reduction in speed limits on residential streets and review of arterial speed limits,
- Installation of signage and pavement markings, signal timing and lane allocation improvements,
- Clearing obstructions for intersections to improve visibility of pedestrians in crosswalks,
- Events and contests to distribute safety information and promote road safety awareness.