Capstone Project - Collision Severity Analysis

Hai Dang Nguyen

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 accidents will be based on the similarity of their initial conditions to those of other accidents in the historical 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:

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.

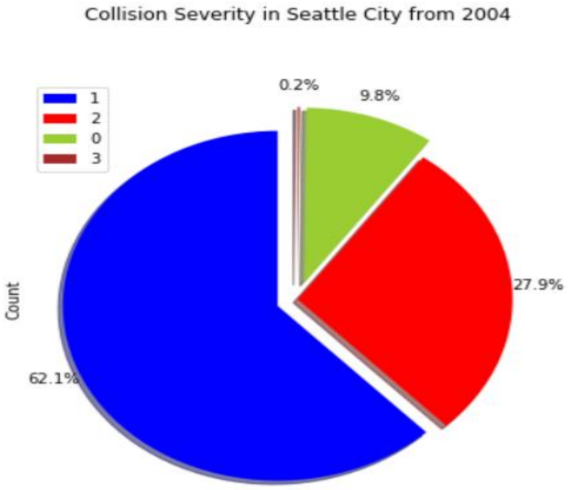
Link for dataset: Seattle Collision Data

Level of collision severity

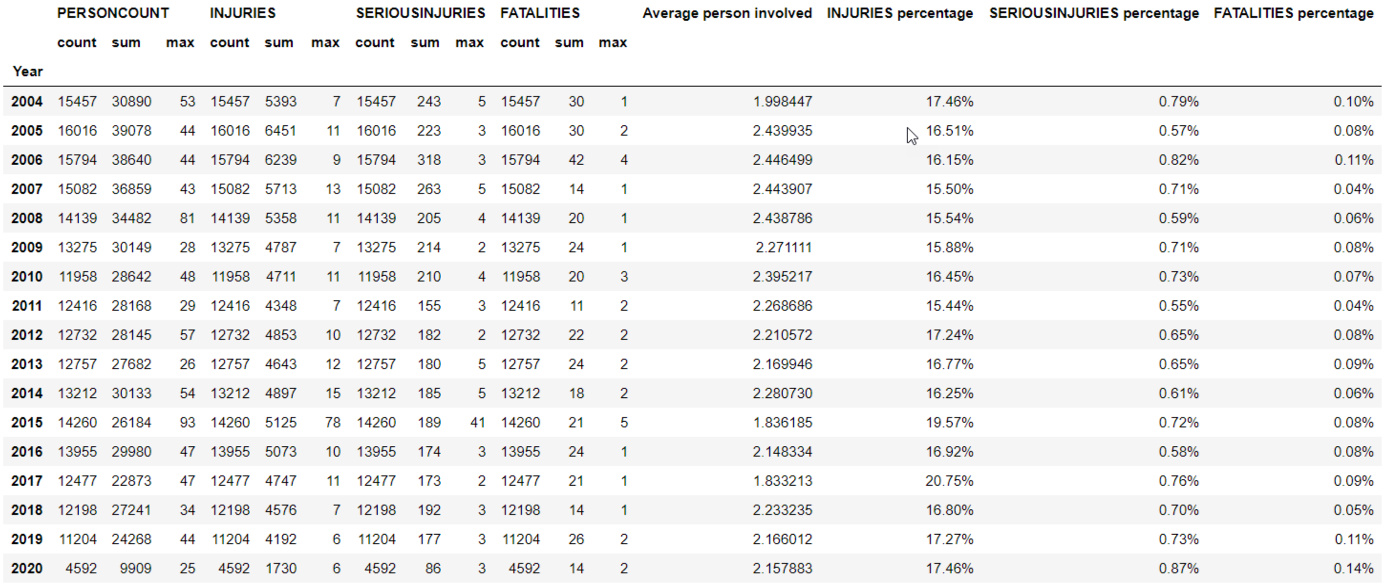
Column SEVERITYCODE encodes the Seattle Department of Transport (SDOT) accident severity metric, according to the following schema:

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%).



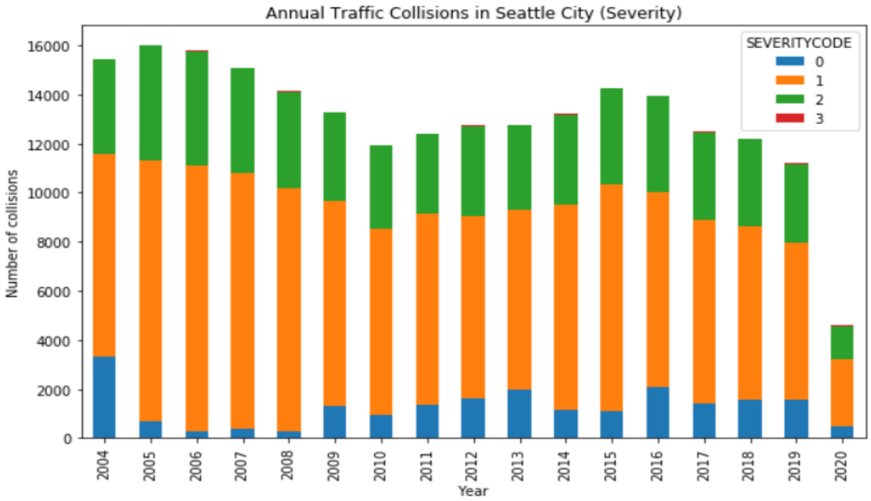
On average, there were 2 people involved in a collision. The number of injuries and fatalities seemed to drop over time, from 6,000 to 4,000 for injuries, from 300 to 170 for serious injuries and from 40 to 20 for fatalities. However, the percentage of people involved resulting in injuries and fatalities 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.



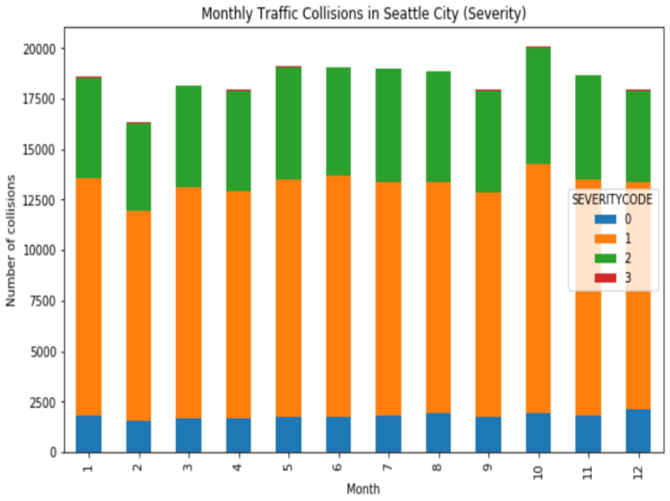
	PERSONCOUNT			INJURIES			SERIOUSINJURIES			FATALITIES			Average person involved	INJURIES percentage	SERIOUSINJURIES percentage	FATALITIES percentage
	count	sum	max	count	sum	max	count	sum	max	count	sum	max				
Year																
2004	15457	30890	53	15457	5393	7	15457	243	5	15457	30	1	1.998447	17.46%	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%

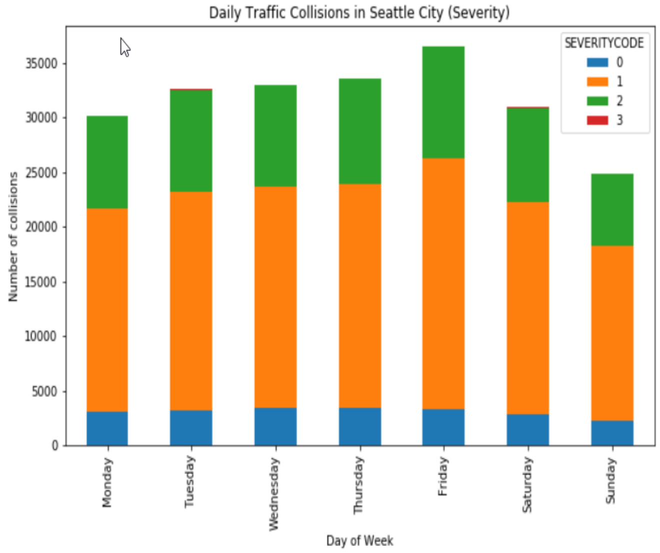
Number of collisions over time

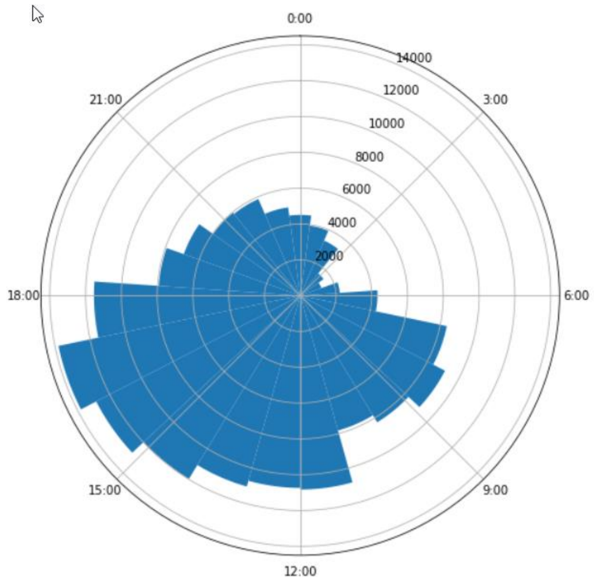
Since 2004, the number of collisions in Seattle has remained at a high level, around 11,000 annually and peaked at 16,000 collisions in 2005. It declined steadily from 2006 to 2013, only to rebound slightly in the four subsequent years but remained well below the 2005 numbers. In the recent years from 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 regading the 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.



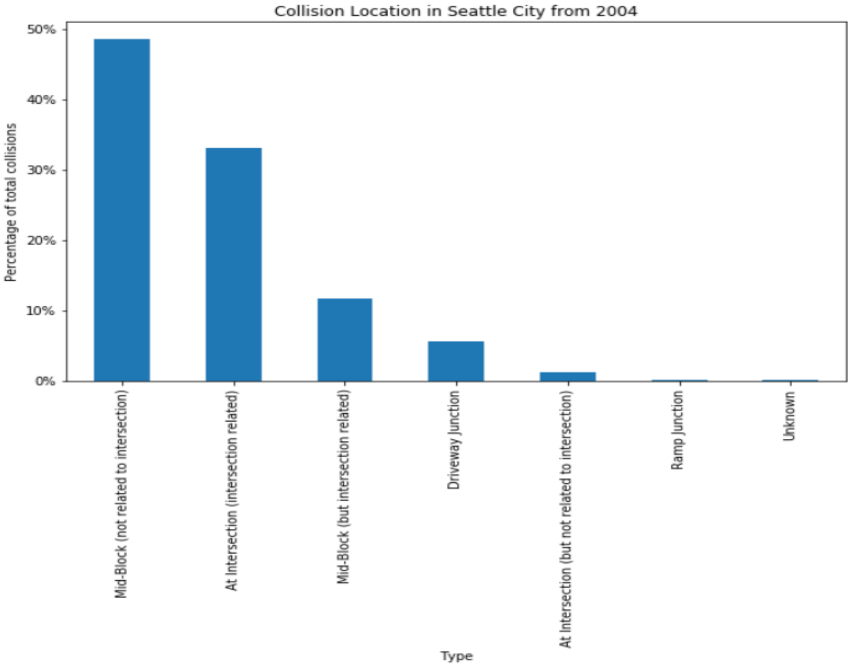




Location of Collision

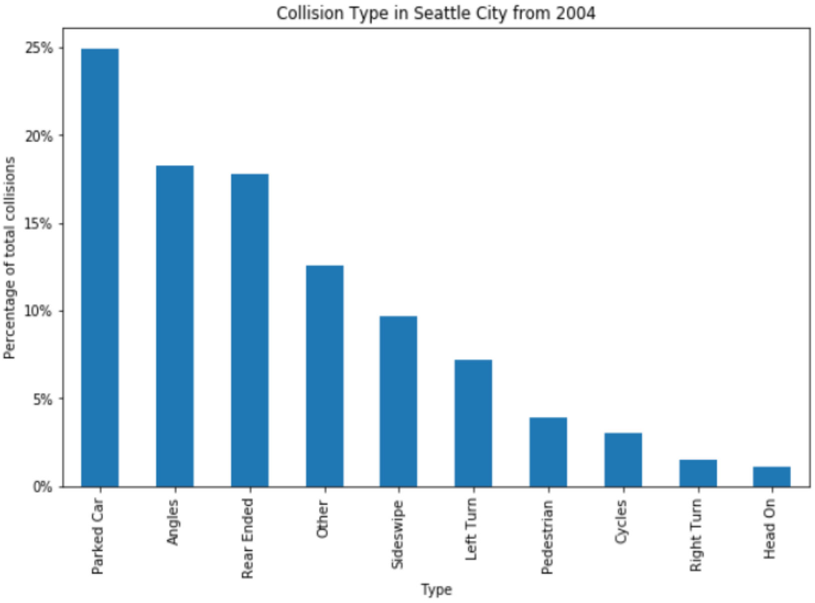
By using the columns X and Y which are the coordinates of the collisions, a Folium map reveals that accidents occur more frequently towards the centre of the city, and in neighbourhoods at either end of road bridges which straddle Seattle’s major waterways. For instance, James Street and Sixth Avenue, Boren Avenue and Pike Street, Lake City Way NE and NE 130th Street, and Dexter Avenue North and Thomas Street have continuously topped the list of Seattle’s the most dangerous intersections. The West Seattle Bridge, the East Marginal Way South, the Ballard Bridge, the Montlake Bridge, the Fremont Bridge, 1st Ave S Bridge, and S Michigan St made the list for the busiest arterials and reported a higher incidence of accidents than average. The corresponding higher pedestrian volumes and/or higher vehicle volumes would increase the opportunities for pedestrian-vehicle and vehicle-vehicle conflicts.

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.

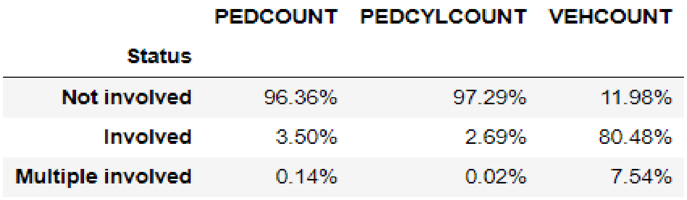


Type of Collision

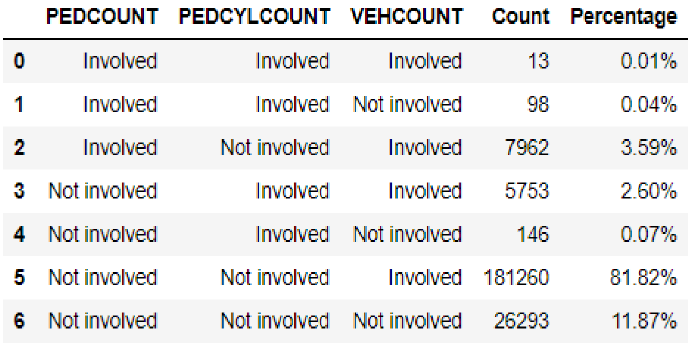
Perhaps surprisingly, the most common type of collision was collisions between a moving vehicle and a parked vehicle (around 25% of total collision). The following most common collisions were vehicles entering the flow of traffic at an angle (typically at intersections) and vehicles traveling in the 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.



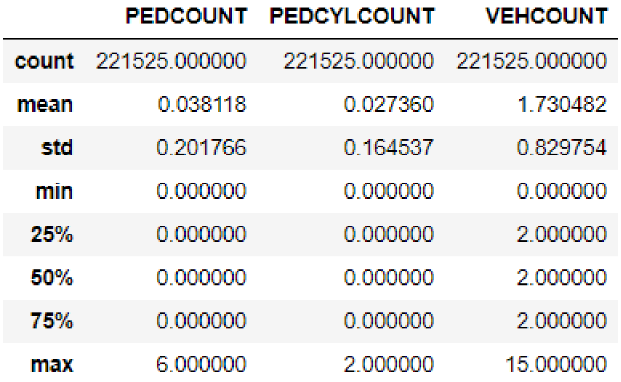
The bicyclists and pedestrians were involved in only 7% of all crashes, while around 90% involved at least one vehicle. Around 4% were pedestrian-vehicle collision and approximately 3% were bicyclist-vehicle collision. There were several cases where multiple people and vehicles are involved, nearly 8% 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%

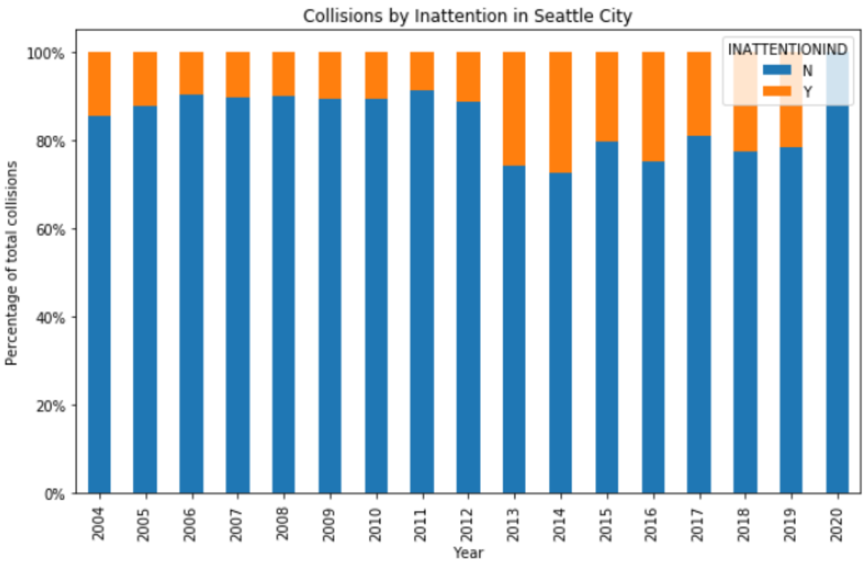


	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT
count	221525.000000	221525.000000	221525.000000
mean	0.038118	0.027360	1.730482
std	0.201766	0.164537	0.829754
min	0.000000	0.000000	0.000000
25%	0.000000	0.000000	2.000000
50%	0.000000	0.000000	2.000000
75%	0.000000	0.000000	2.000000
max	6.000000	2.000000	15.000000

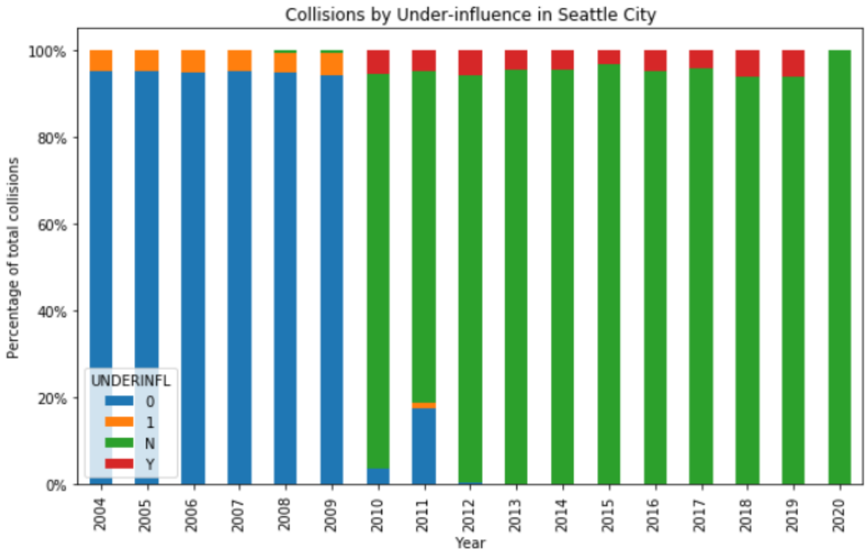
Common Contributing Factors

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:

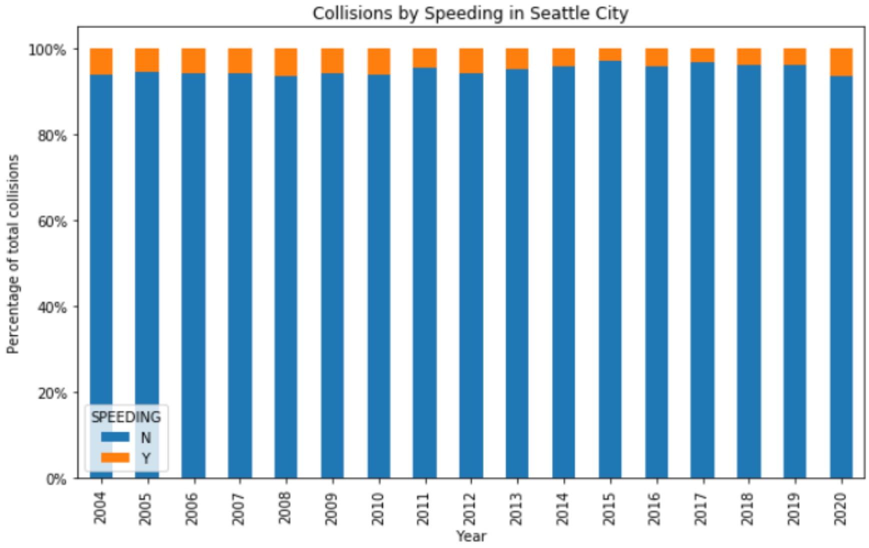
1. **Driver distraction**: distracted driving is a prevalent risky behaviour and hence a leading contributing cause of collisions in Seattle. Many circumstances are considered inattention like eating, grooming, using mobile phone, adjusting an audio system, other distractions inside the vehicle and distractions 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.



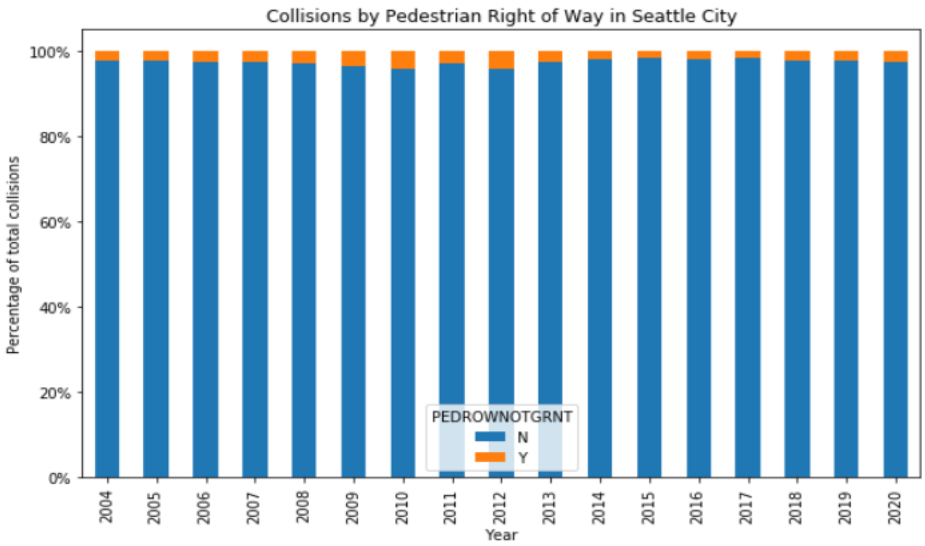
2. **Alcohol and drug impairment**: Driver impairment due to alcohol and/or drugs is one of the most contributing factors in fatal crashes and is involved in over 600 collisions per year, and this number has just slightly reduced over time without oblivious improvement. The dataset entries had a complete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



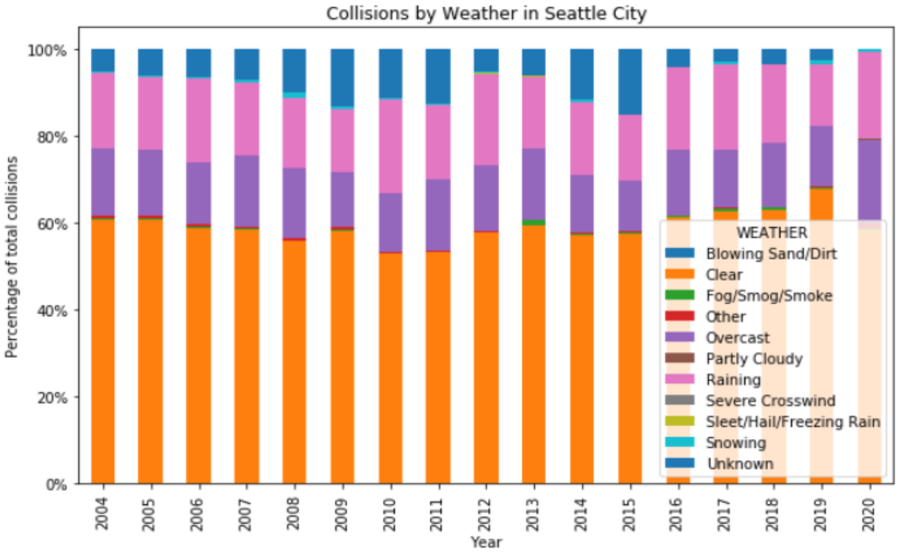
3. **Speeding**: Speeding remains a significant cause of car accidents in Seattle, noted as a contributing factor in over 400 collisions each year under the categories "exceeding reasonable and safe speed" and "exceeding the speed limit". Conditions for pedestrians and bikers being hit by a vehicle also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



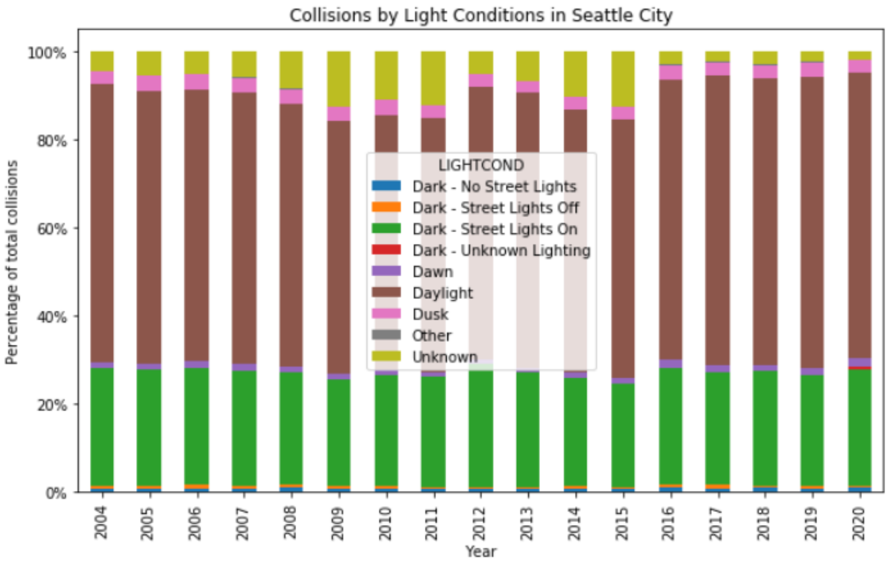
4. **Whether the pedestrian right of way was not granted**: This is a commonly cited factor for pedestrian and bicyclist-involved collisions in Seattle, causing over 200 cases annually but this number has decreased in recent years. This contributing factor generally indicates that a driver, pedestrian, 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.



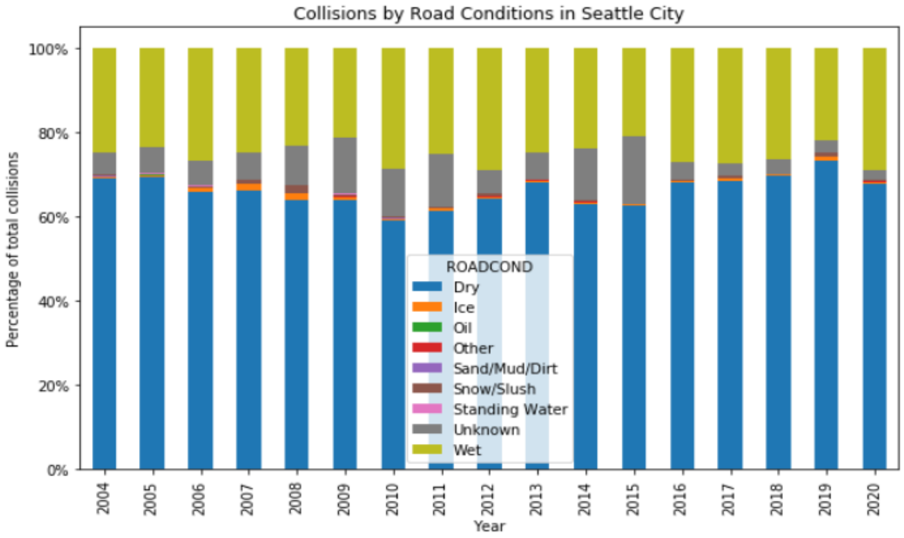
5. **Weather**: Weather acts through visibility impairments, slick pavement, high winds, and temperature extremes to affect driver capabilities and vehicle performance, leading to crash risk. Surprisingly, over 60% of total collisions occurred in a clear weather, whereas the vast majority of most weather-related 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.

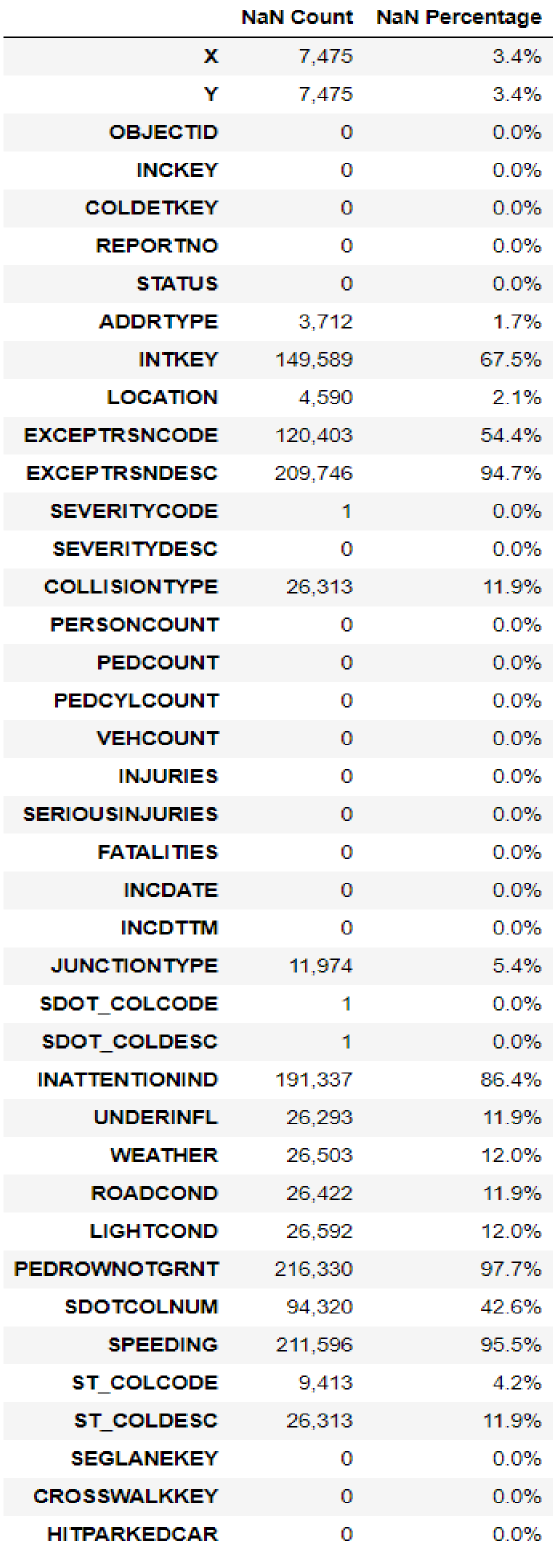


2. Data Preparation

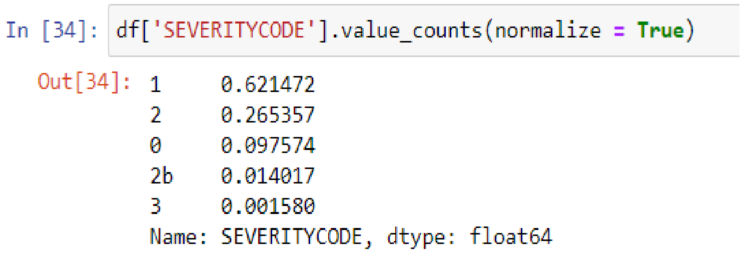
In their original form, the Seattle collision data is not suitable for quantitative data analysis and so likely to create noise in the modeling result. The main reasons for this are as follows:

Missing Values

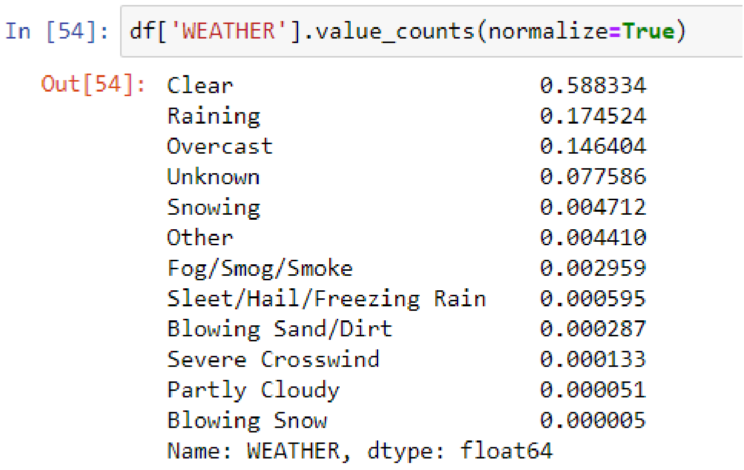
There are missing values on part of the database, around 10% of the target variable SEVERITYCODE is unknown and some attributes have over 40% of missing data (such as INTKEY, EXCEPTRSNCODE). Common attributing factors (such as JUNCTIONTYPE, WEATHER, LIGHTCOND, ROADCOND) include unknown entries which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



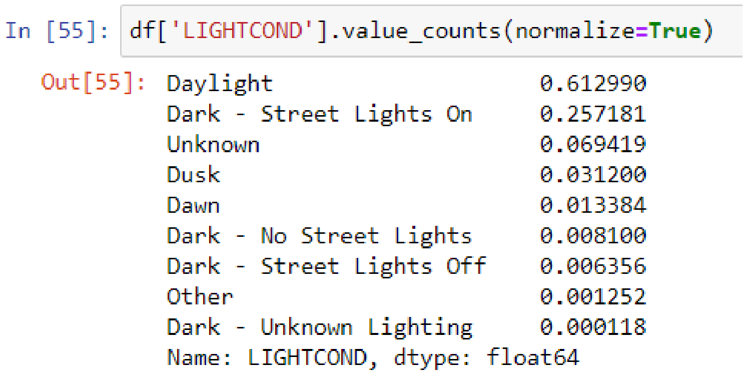
	NaN Count	NaN Percentage
X	7,475	3.4%
Y	7,475	3.4%
OBJECTID	0	0.0%
INCKEY	0	0.0%
COLDETKEY	0	0.0%
REPORTNO	0	0.0%
STATUS	0	0.0%
ADDRTYPE	3,712	1.7%
INTKEY	149,589	67.5%
LOCATION	4,590	2.1%
EXCEPTRSNCODE	120,403	54.4%
EXCEPTRSNDESC	209,746	94.7%
SEVERITYCODE	1	0.0%
SEVERITYDESC	0	0.0%
COLLISIONTYPE	26,313	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%



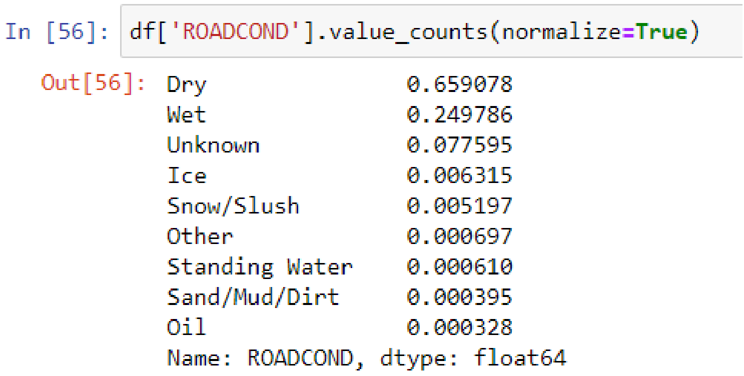
```
In [34]: df['SEVERITYCODE'].value_counts(normalize = True)

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

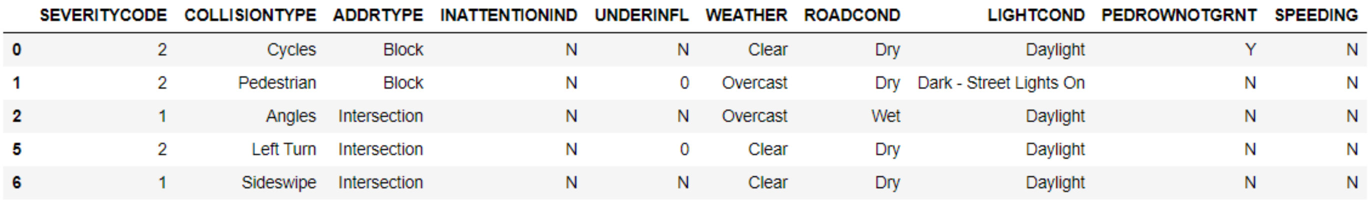
1. The dataset includes columns that provide unnecessary information for the machine learning algorithms, for instance, columns of metadata (such as OBJECTID, INCKEY, COLDETKEY, REPORTNO, STATUS, SDOT_COLCODE, SDOTCOLNUM, ST_COLCODE, SEGLANEKEY, CROSSWALKKEY) and description columns (such as EXCEPTRSNDESC, 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

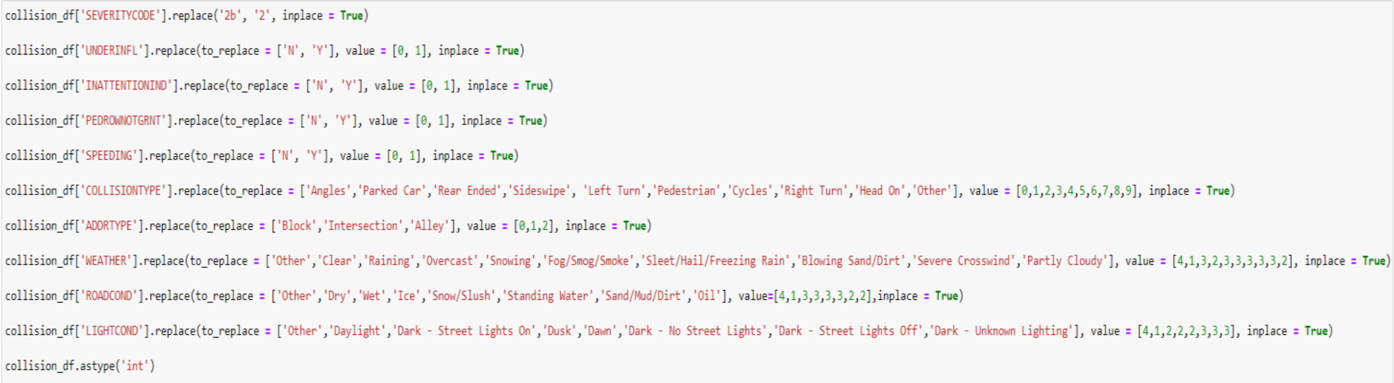
Overall, such missing data entries and irrelevant data attributes will be removed from the dataset. Except for INATTENTIONIND, PEDROWNOTGRNT and SPEEDING columns where there are two types of entry which are 'Yes' and NaN, so we assume that NaN means 'No' entry. We have an updated dataset collision_df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



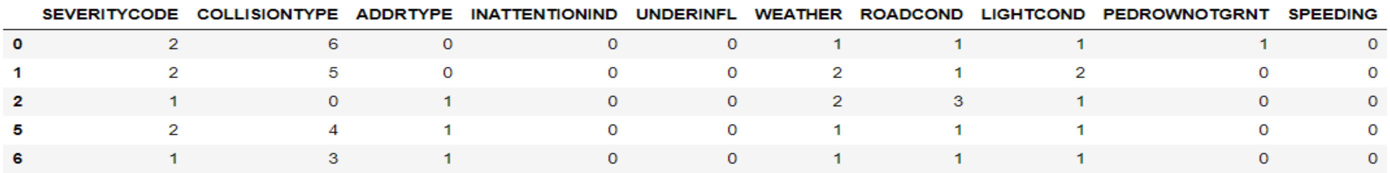
	SEVERITYCODE	COLLISIONTYPE	ADDRTYPE	INATTENTIONIND	UNDERINFL	WEATHER	ROADCOND	LIGHTCOND	PEDROWNOTGRNT	SPEEDING
0	2	Cycles	Block	N	N	Clear	Dry	Daylight	Y	N
1	2	Pedestrian	Block	N	0	Overcast	Dry	Dark - Street Lights On	N	N
2	1	Angles	Intersection	N	N	Overcast	Wet	Daylight	N	N
5	2	Left Turn	Intersection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

Encoding categorical variables

To focus on the collision severity of property damage - injury - fatality, we will combine the group 2b of serious injury with group 2 of injury. 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. We will use label encoding to recast each of these categorical variables as a series of numerical data. For COLLISIONTYPE and JUNCTIONTYPE, since code '1' defining property damage for dependent variable SEVERITYCODE has the largest proportion, we assume the code '0' as a element of an independent variable to depict the most probable cause of a collision. For WEATHER, LIGHTCOND and ROADCOND, we imply three codes for conditions 'Normal', 'Unfavourable' and 'Adverse' based on the description of data. Any group 'Other' of a variable will be encoded highest. We ensure that all the columns are numerical (integer).



```
collision_df['SEVERITYCODE'].replace('2b', '2', inplace = True)

collision_df['UNDERINFL'].replace(to_replace = ['N', 'Y'], value = [0, 1], inplace = True)

collision_df['INATTENTIONIND'].replace(to_replace = ['N', 'Y'], value = [0, 1], inplace = True)

collision_df['PEDROWNOTGRNT'].replace(to_replace = ['N', 'Y'], value = [0, 1], inplace = True)

collision_df['SPEEDING'].replace(to_replace = ['N', 'Y'], value = [0, 1], inplace = True)

collision_df['COLLISIONTYPE'].replace(to_replace = ['Angles','Parked Car','Rear Ended','Sideswipe', 'Left Turn','Pedestrian','Cycles','Right Turn','Head On','Other'], value = [0,1,2,3,4,5,6,7,8,9], inplace = True)

collision_df['ADDRTYPE'].replace(to_replace = ['Block','Intersection','Alley'], value = [0,1,2], inplace = True)

collision_df['WEATHER'].replace(to_replace = ['Other','Clear','Raining','Overcast','Snowing','Fog/Smog/Smoke','Sleet/Hail/Freezing Rain','Blowing Sand/Dirt','Severe Crosswind','Partly Cloudy'], value = [4,1,3,2,3,3,3,3,3,2], inplace = True)

collision_df['ROADCOND'].replace(to_replace = ['Other','Dry','Wet','Ice','Snow/Slush','Standing Water','Sand/Mud/Dirt','Oil'], value=[4,1,3,3,3,3,2,2],inplace = True)

collision_df['LIGHTCOND'].replace(to_replace = ['Other','Daylight','Dark - Street Lights On','Dusk','Dawn','Dark - No Street Lights','Dark - Street Lights Off','Dark - Unknown Lighting'], value = [4,1,2,2,2,3,3,3], inplace = True)

collision_df.astype('int')
```



	SEVERITYCODE	COLLISIONTYPE	ADDRTYPE	INATTENTIONIND	UNDERINFL	WEATHER	ROADCOND	LIGHTCOND	PEDROWNOTGRNT	SPEEDING
0	2	6	0	0	0	1	1	1	1	0
1	2	5	0	0	0	2	1	2	0	0
2	1	0	1	0	0	2	3	1	0	0
5	2	4	1	0	0	1	1	1	0	0
6	1	3	1	0	0	1	1	1	0	0

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_3 samples from the larger data groups (severity code 1 and 2) and re-group these values into smaller groups, where is the number of accidents with severity code of 3 (Fatality).