Analyzing Seattle Traffic Accidents Data

Baron Purwa Hartono

August 21, 2020

1. Introduction and Problem Understanding

Everywhere in the world can have a traffic accident, and there are many different things that cause the accident to happen, whether its weather condition, road condition (maybe there is a hole in the road, or the road slippery, etc), light condition (Daylight, Dark - Street Lights On, Dark - Street Lights Off, etc), drunk driving, and other things.

So the goal is we want to reduce traffic accidents as much as possible by understanding what cause the accident to happen, and to know if we can notif the drivers/people to be careful through mobile application like maybe google navigation, etc, if the condition at that time likely cause an accident to happen.

Or we can educate the population/residents what cause a traffic accident to happen to make them to be more careful.

Hope this project can help many Government to reduce traffic accidents that happen in their area.

2. Data acquisition and cleaning

2.1 Data sources

The data that we have right now to help solve the problem came from the IBM Data Capstone Course. This data give us information about all accidents/collisions that happens in Seattle. So we will analyze the data to understand what caused an accident to happen, and what makes the accidents more severe. And if it possible, we can make machine learning model to predict whether the accident will likely to happen or not based on the condition at that time.

2.2 Data Set Summary

Title : Collisions—All Years

- **Abstract :** All collisions provided by SPD and recorded by Traffic Records.
- **Descriptions :** This includes all types of collisions. Collisions will display at the intersection or mid-block of a segment. Timeframe: 2004 to Present.
- **Update Frequency :** Weekly
- **Keyword(s) :** SDOT, Seattle, Transportation, Accidents, Bicycle, Car, Collisions, Pedestrian, Traffic, Vehicle
- **Contact Organization :** SDOT Traffic Management Division, Traffic Records Group
- **Contact Person :** SDOT GIS Analyst
- **Contact Email :** DOT_IT_GIS@seattle.gov

2.3 Attribute Information

Attribute	Data Type, Length	Description
OBJECTID	Long	ESRI unique identifier
SHAPE	Geometry	ESRI geometry field
INCKEY	Long	A unique key for the incident
ADDRTYPE	Text, 12	Collision address type: Alley, Block, Intersection
INTKEY	Double	Key that corresponds to the intersection associated with a collision
LOCATION	Text, 255	Description of the general location of the collision
EXCEPTRSNCODE	Text, 10	
EXCEPTRSNDESC	Text, 300	
SEVERITYCODE	Text, 100	A code that corresponds to the severity of the collision: 3—fatality, 2b—serious injury, 2—injury, 1—prop damage, 0—unknown
SEVERITYDESC	Text	A detailed description of the severity of the collision
COLLISIONTYPE	Text, 300	Collision Type
PERSONCOUNT	Double	The total number of people involved in the collision
PEDCOUNT	Double	The number of pedestrians involved in the collision. This is entered by the state.
PEDCYLCOUNT	Double	The number of bicycles involved in the collision. This is entered by the state.
VEHCOUNT	Double	The number of vehicles involved in the collision. This is entered by the state.
INJURIES	Double	The number of total injuries in the collision. This is entered by the state.
SERIOUSINJURIES	Double	The number of serious injuries in the collision. This is entered by the state.
FATALITIES	Double	The number of fatalities in the collision. This is entered by the state.
INCDATE	Date	The date of the incident.
INCDTTM	Text, 30	The date and time of the incident.
JUNCTIONTYPE	Text, 300	Category of junction at which collision took place
SDOT_COLCODE	Text, 10	A code given to the collision by SDOT.
SDOT_COLDESC	Text, 300	A description of the collision corresponding to the collision code.
INATTENTIONIND	Text, 1	Whether or not collision was due to inattention. (Y/N)
UNDERINFL	Text, 10	Whether or not a driver involved was under the influence of drugs or alcohol.
WEATHER	Text, 300	A description of the weather conditions during the time of the collision.
ROADCOND	Text, 300	The condition of the road during the collision.
LIGHTCOND	Text, 300	The light conditions during the collision.
PEDROWNOTGRNT	Text, 1	Whether or not the pedestrian right of way was not granted. (Y/N)
SDOTCOLNUM	Text, 10	A number given to the collision by SDOT.
SPEEDING	Text, 1	Whether or not speeding was a factor in the collision. (Y/N)
ST_COLCODE	Text, 10	A code provided by the state that describes the collision. For more information about these codes, please see the State Collision Code Dictionary.
ST_COLDESC	Text, 300	A description that corresponds to the state's coding designation.
SEGLANEKEY	Long	A key for the lane segment in which the collision occurred.
CROSSWALKKEY	Long	A key for the crosswalk at which the collision occurred.
HITPARKEDCAR	Text, 1	Whether or not the collision involved hitting a parked car. (Y/N)

2.4 Data Cleaning

there are missing values and incorrect data types, useless features/columns for analysis so first i drop some columns that not needed in my analysis, columns that have too many missing values, and columns that only have 1 unique value (all values is the same), and i will use the desc column not the code column for easier to read the data, this is the list of columns that I drop/delete :

```
df.drop(['SEVERITYCODE','X','Y','OBJECTID','INCKEY','COLDETKEY',
         'REPORTNO','STATUS','INTKEY','LOCATION','EXCEPTRSNCODE','EXCEPTRSNDESC',
         'SEVERITYCODE.1','SDOT_COLCODE','INATTENTIONIND','PEDROWNOTGRNT',
         'SDOTCOLNUM','SPEEDING','SEGLANEKEY','CROSSWALKKEY'],axis=1, inplace=True)
```

And then I fix ST_COLCODE data, because the data type isnt int but object(string), and the sum of the null values not the same as ST_COLDESC, after I checked it, it appears that there are duplicate values and an empty space data like this.

```
array(['10', '11', '32', '23', '5', '22', '14', '30', ' ', '28', '51',
       '13', '50', '12', '45', '0', '20', '21', '1', '52', '16', '15',
       '74', '81', '26', '19', '2', '66', '71', '3', '24', '40', '57',
       '6', '83', '25', '27', '4', '72', '29', '56', '73', '41', '17',
       '65', '82', '67', '49', '84', '31', '43', '42', '48', '64', '53',
       32, 50, 15, 10, 14, 20, 13, 22, 51, 11, 28, 12, 52, 21, 0, 19, 30,
       16, 40, 26, 27, 83, 2, 45, 65, 23, 24, 71, 1, 29, 81, 25, 4, 73,
       74, 72, 3, 84, 64, 57, 42, 41, 48, 66, 56, 31, 82, 67, '54', '60',
       53, 43, 87, 54, '87', nan, '7', '8', '85', '88', '18'],
      dtype=object)
```

After I fix it, the data looks like this

```
array([10., 11., 32., 23.,  5., 22., 14., 30., nan, 28., 51., 13., 50.,
       12., 45.,  0., 20., 21.,  1., 52., 16., 15., 74., 81., 26., 19.,
        2., 66., 71.,  3., 24., 40., 57.,  6., 83., 25., 27.,  4., 72.,
       29., 56., 73., 41., 17., 65., 82., 67., 49., 84., 31., 43., 42.,
       48., 64., 53., 54., 60., 87.,  7.,  8., 85., 88., 18.])
```

And then I change the INCDATE, INCDTTM columns data type to datetime type from string type. And then I drop all rows that have missing values, because i need the data as correct as possible. and the missing values not that much compare to the entire dataset.

The data in UNDERINFL column isnt consistent, there is N & 0, and 1 & Y, they explained the same thing. so i chose one format, the one i chose is the 0 and 1 format.

And then I check outlier of the numerical columns, it turns out there are some outliers but the outliers data still have a make sense value, so I didn’t do anything about it.

3. Exploratory Data Analysis

First I check the distribution of numerical columns using df.describe().

	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	UNDERINFL	ST_COLCODE
count	182895.000000	182895.000000	182895.000000	182895.000000	182895.000000	182895.000000
mean	2.476268	0.038995	0.029831	1.971984	0.049192	22.583411
std	1.370912	0.202960	0.171435	0.563237	0.216270	14.575891
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	2.000000	0.000000	0.000000	2.000000	0.000000	11.000000
50%	2.000000	0.000000	0.000000	2.000000	0.000000	15.000000
75%	3.000000	0.000000	0.000000	2.000000	0.000000	32.000000
max	81.000000	6.000000	2.000000	12.000000	1.000000	88.000000

i can see the distribution of the numerical column quite clear with this. and i see there is a quite big accident have happened that includes 81 people in it. Then i check the details of that accident.

Turns out it happened in Intersection, it caused Injury, it includes only 2 vehicles, it happened in 2008-12-19 at noon, the driver not under influenced of alcohol or drugs, the weather is clear, but the road is filled with snow, and the light condition is daylight because it happened at noon.

this is the accident description MOTOR VEHICLE RAN OFF ROAD - HIT FIXED OBJECT.

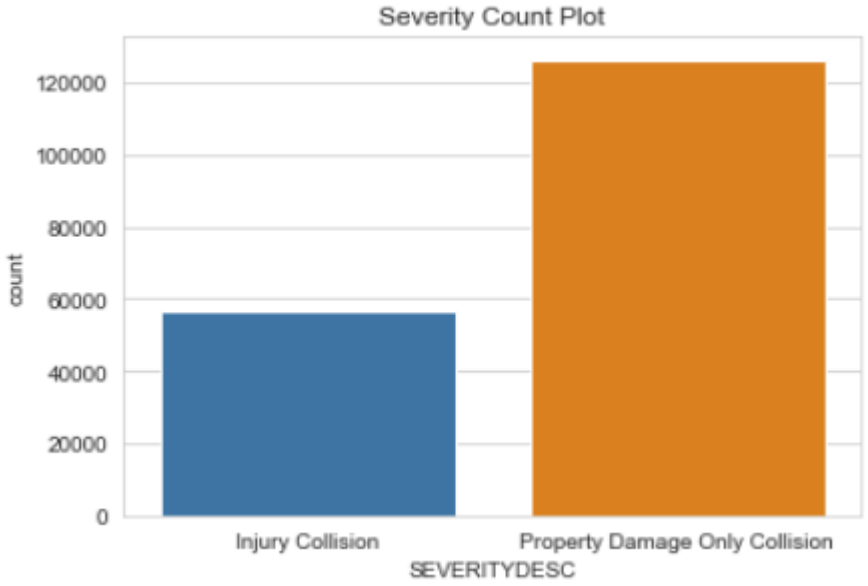
And then I check the accident that includes 12 vehicles too. And i see that the accident gladly didnt caused any injuries, and it happend at night time with street lights on. the driver must be careless, because he/she not under influence of alcohol and drugs, maybe he/she driving while playing with his/her phone.

And then I check the distribution of the categorical columns too.

	ADDRTYPE	SEVERITYDESC	COLLISIONTYPE	JUNCTIONTYPE	SDOT_COLDESC	WEATHER	ROADCOND	LIGHTCOND	ST_COLDESC	HITPARKEDCAR
count	182895	182895	182895	182895	182895	182895	182895	182895	182895	182895
unique	3	2	10	7	39	11	9	9	62	2
top	Block	Property Damage Only Collision	Parked Car	Mid-Block (not related to intersection)	MOTOR VEHICLE STRUCK MOTOR VEHICLE, FRONT END ...	Clear	Dry	Daylight	One parked--one moving	N
freq	119362	126270	43119	86609	83024	109059	122153	113837	39619	177205

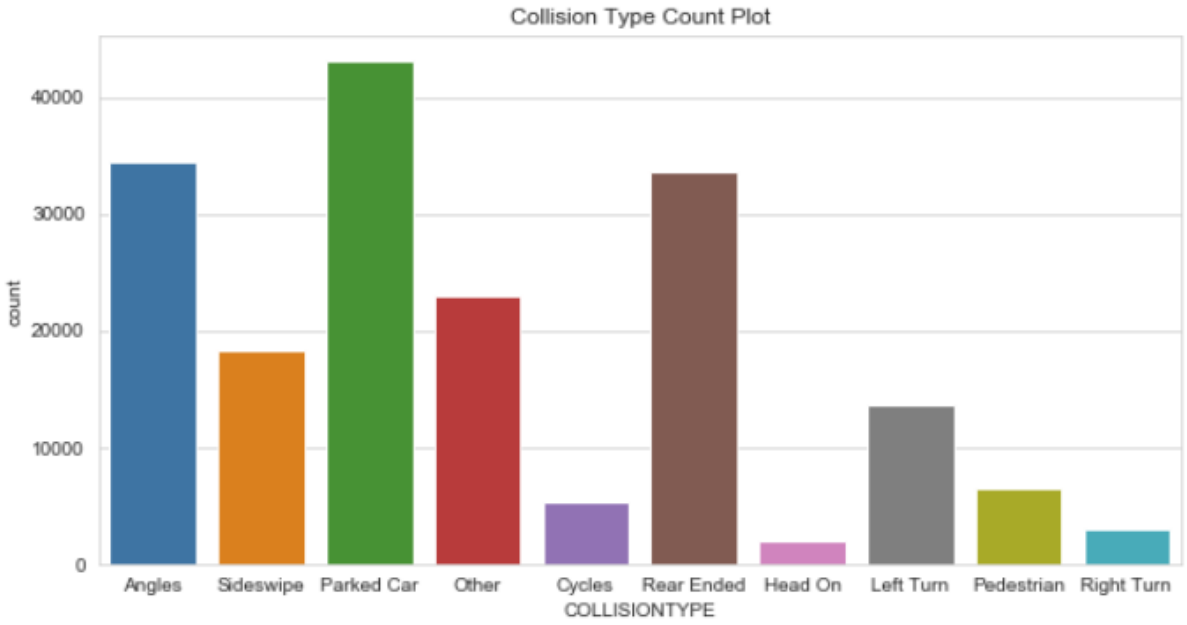
Count = how many rows in each columns, unique = how many unique values in each columns, top = the most frequent category data (mode), freq = how many data/rows of the mode.

And then I check the distribution of the SEVERITYDESC column



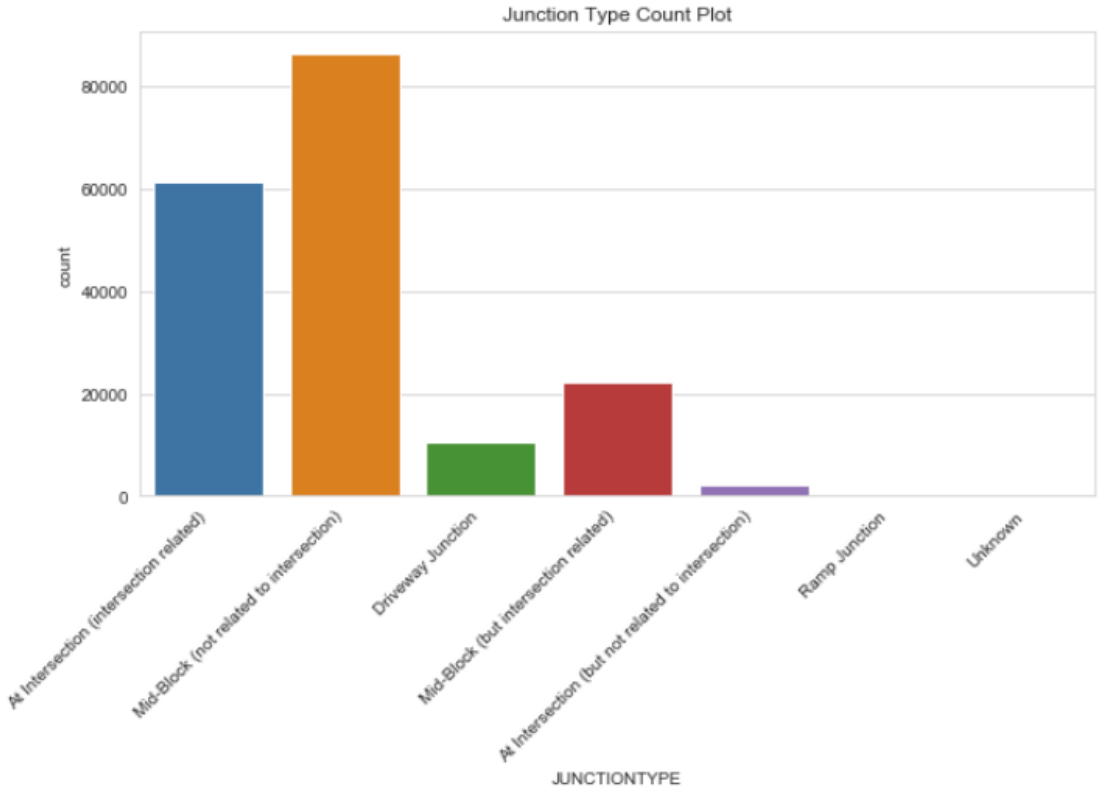
gladly i see that there are more accidents that cause only property damage, but the accidents that cause injuries quite a lot too.

Then I see the distribution of COLLISIONTYPE data



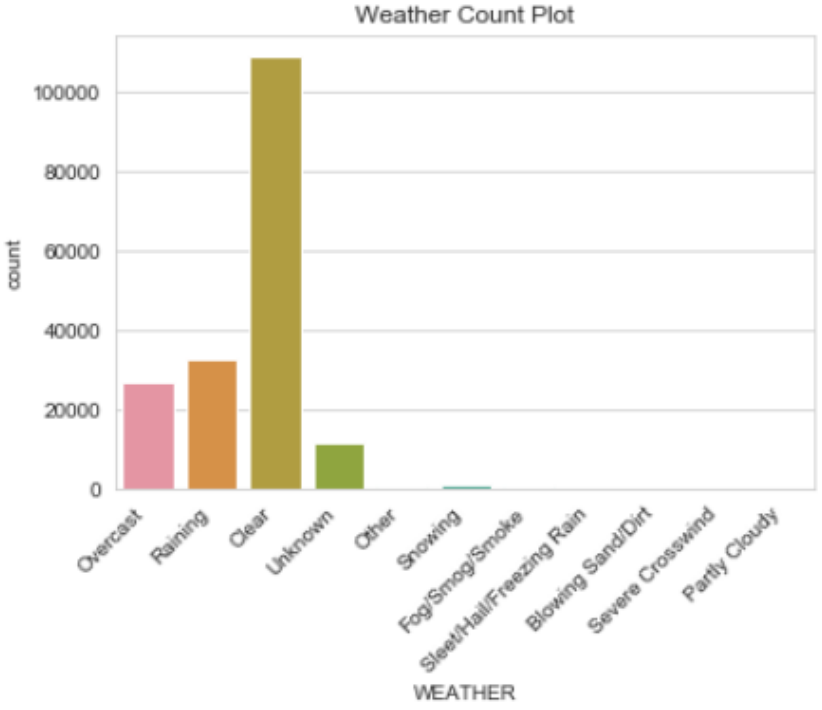
the 3 most often happen collision types is Angles, Parked Car, and Rear Ended. so the accidents mostly involves 2 vehicles.

Then i take a look at the distributions of JUNCTIONTYPE column

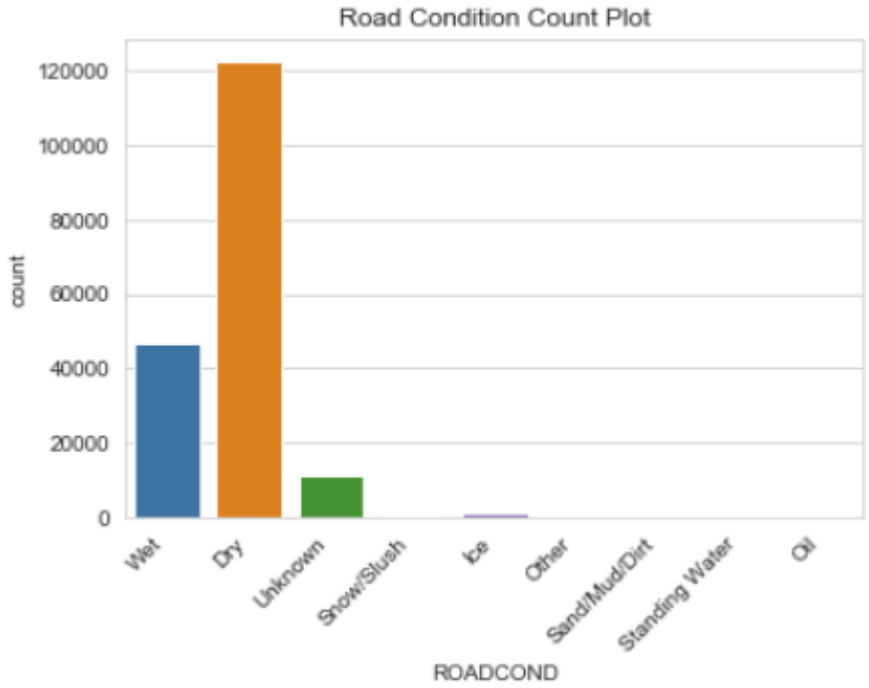


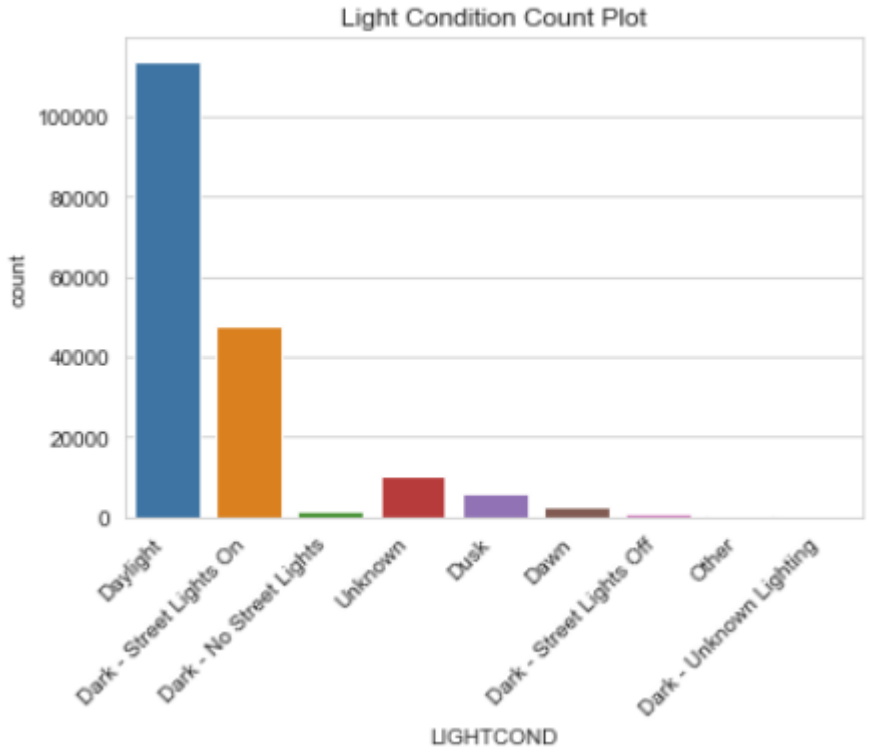
the accidents/collisions most often happened at Mid-Block (not related to intersection) Junction type and then followed by At Intersection (intersection related). so we need to be more careful when we are at these 2 Junction Types.

Then I went to check the count distributions of other categorical columns.



In this data, most accidents happened in Clear Weather condition.

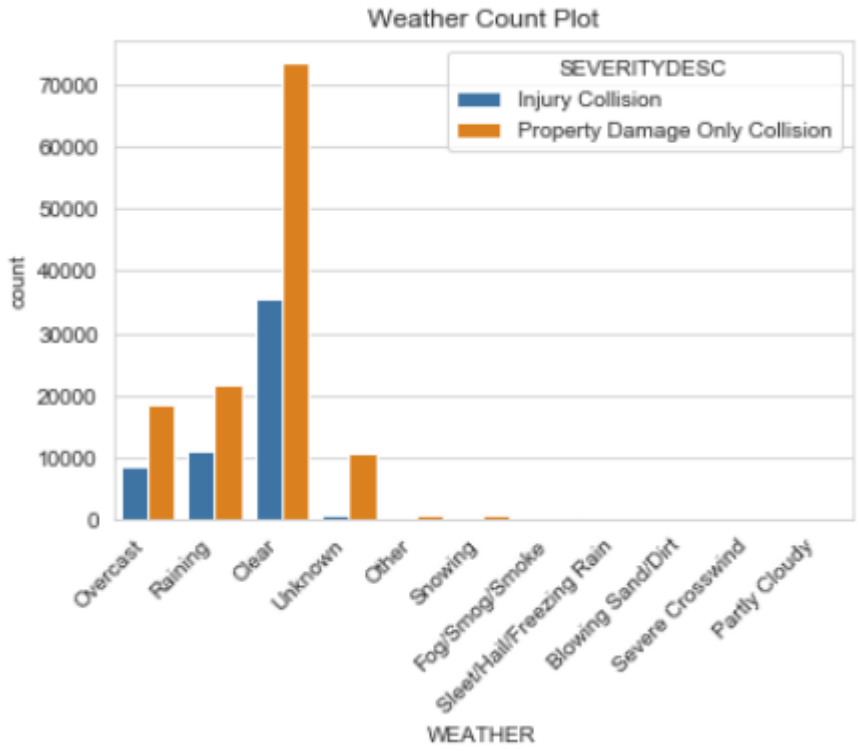




actually most accidents happened in normal condition when there is still Daylight, dry road, clear weather. this is make sense to me because based on my experience in my city where I live, traffic is tend to be more crowded when the condition is like the above.

and actually if we want to see/know what condition that more likely to cause an accident, we need to see the ratio comparison between accident data and no accident data, but for this case it didnt have the no accident data, thats make sense because if it keep track of every travel/trip that happened, the database will need more space to keep it, and it will be harder too to process the data, because it will be much larger in size.

so then i just compare the ratio between SEVERITYDESC Injury Collision and Property Damage Only Collision in each condition

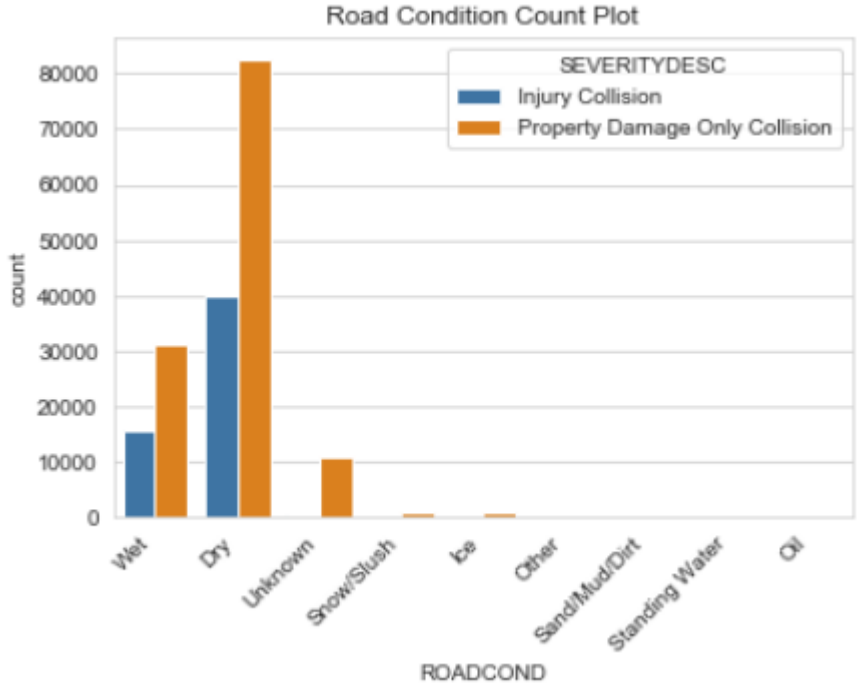


because the Weather data so imbalance, we cant see the bar clearly for some weather, so tried another method to see it more clearly.

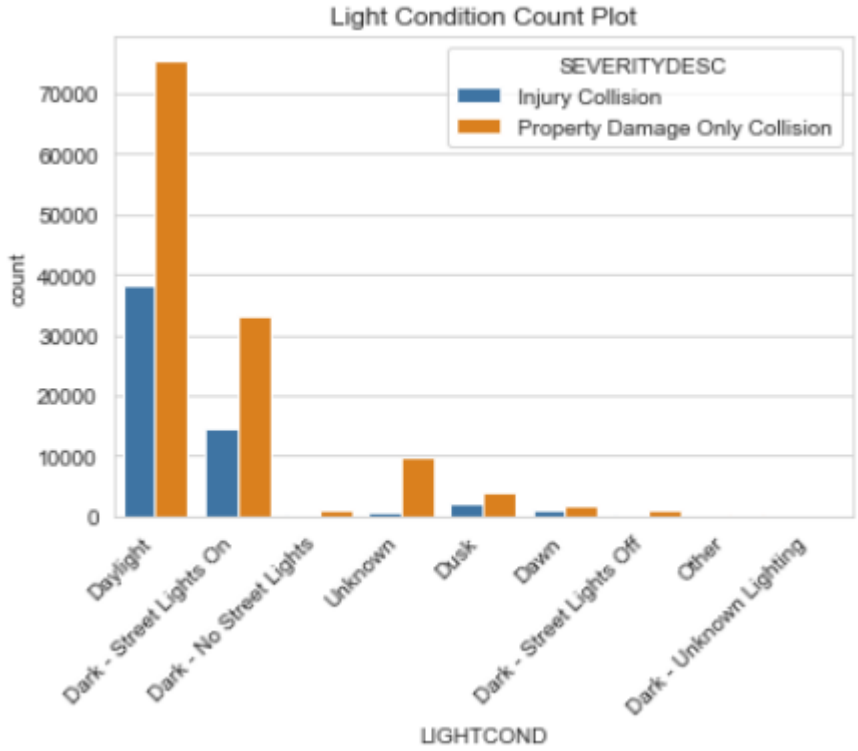
SEVERITYDESC WEATHER	Injury Collision	Property Damage Only Collision
Partly Cloudy	0.600000	0.400000
Raining	0.339532	0.660468
Fog/Smog/Smoke	0.334532	0.665468
Clear	0.326273	0.673727
Overcast	0.318986	0.681014
Severe Crosswind	0.280000	0.720000
Blowing Sand/Dirt	0.265306	0.734694
Sleet/Hail/Freezing Rain	0.241071	0.758929
Snowing	0.189557	0.810443
Other	0.152815	0.847185
Unknown	0.066254	0.933746

the highest ratio of Injury Collision is when the Weather is Partly Cloudy, but the Partly Cloudy data is not many. and Unknown is the lowest ratio of Injury Collision, based on my assumption, what caused the Unknown data to be mostly Property Damage Only Collision is because people tend to more pay attention to Injury Collision than the Property Damage Only Collision.

Then I see the other condition columns.



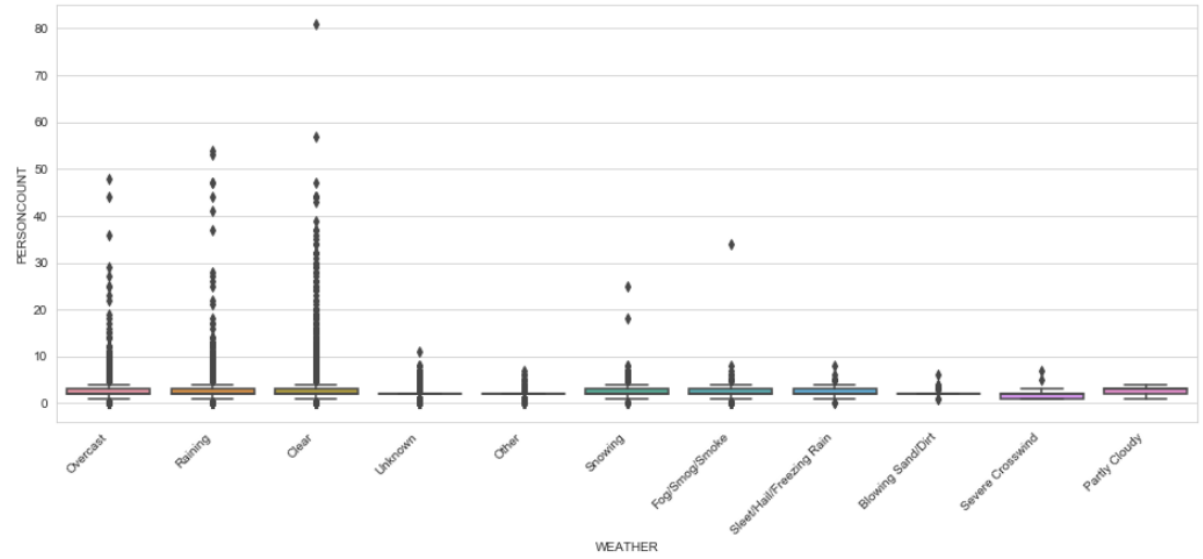
SEVERITYDESC ROADCOND	Injury Collision	Property Damage Only Collision
Oil	0.400000	0.600000
Other	0.341463	0.658537
Wet	0.334618	0.665382
Sand/Mud/Dirt	0.328358	0.671642
Dry	0.325322	0.674678
Standing Water	0.268519	0.731481
Ice	0.226848	0.773152
Snow/Slush	0.168712	0.831288
Unknown	0.061377	0.938623

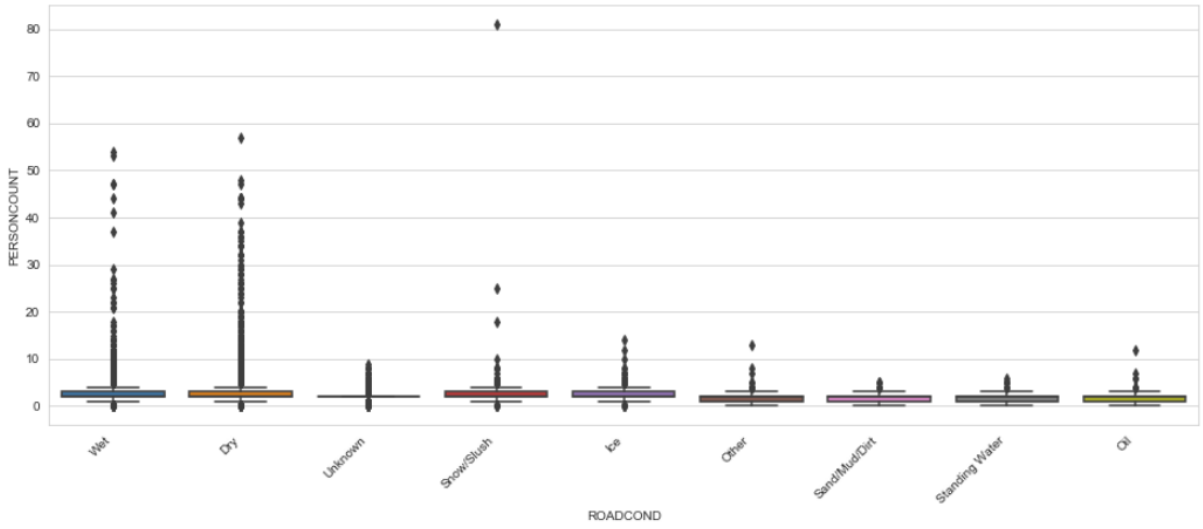


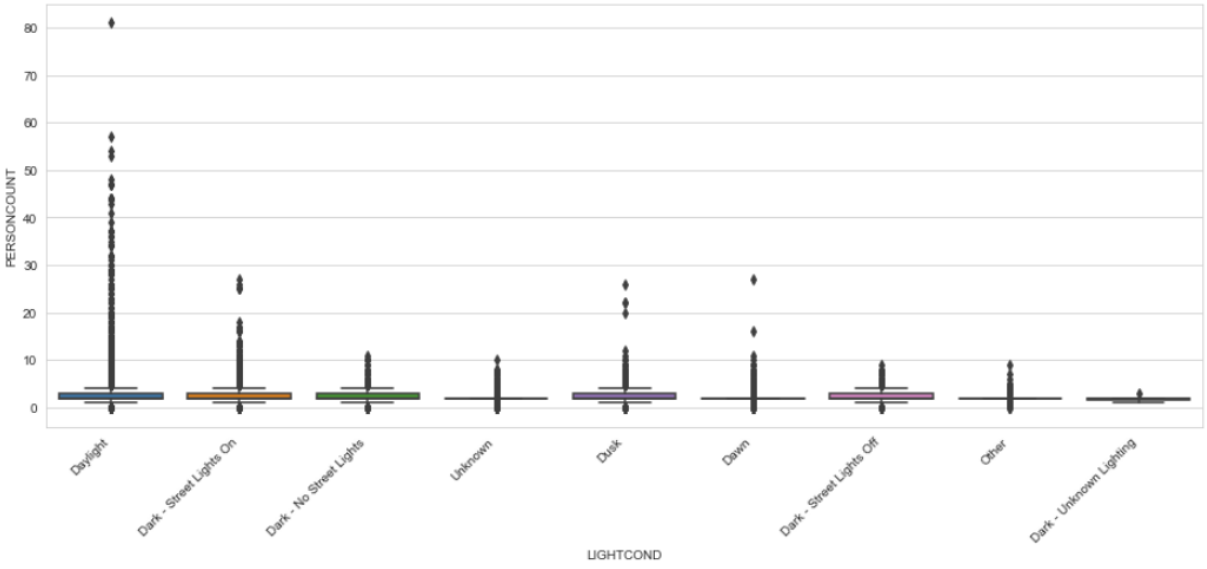
SEVERITYDESC LIGHTCOND	Injury Collision	Property Damage Only Collision
Dark - Unknown Lighting	0.363636	0.636364
Daylight	0.336059	0.663941
Dawn	0.333877	0.666123
Dusk	0.333738	0.666262
Dark - Street Lights On	0.301828	0.698172
Dark - Street Lights Off	0.270527	0.729473
Other	0.247619	0.752381
Dark - No Street Lights	0.224504	0.775496
Unknown	0.055130	0.944870

I see the same pattern for Unknown data in each condition columns. maybe my assumption is right.

now i check if weather, road, and light condition affect The total number of people involved in the collision.



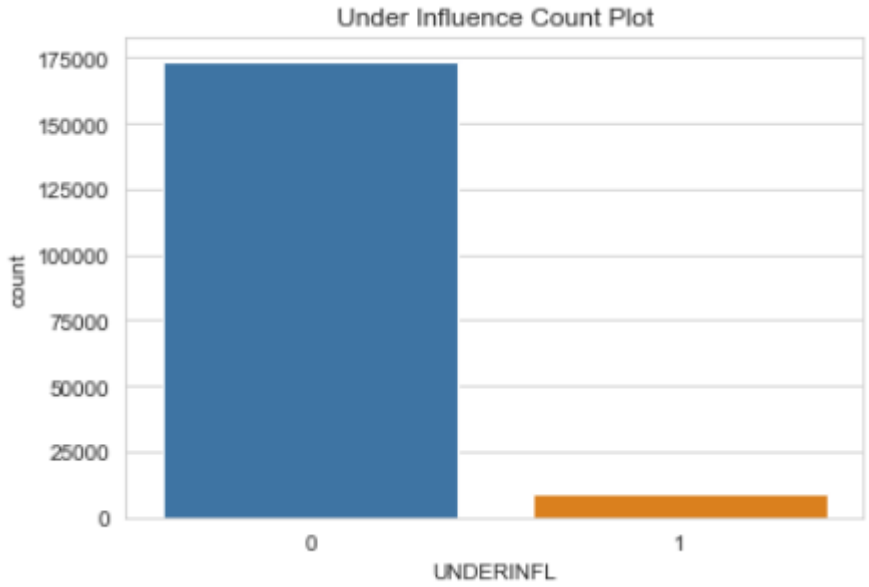




The distribution is almost the same in all conditions, but there are more outliers data in some conditions than the others.

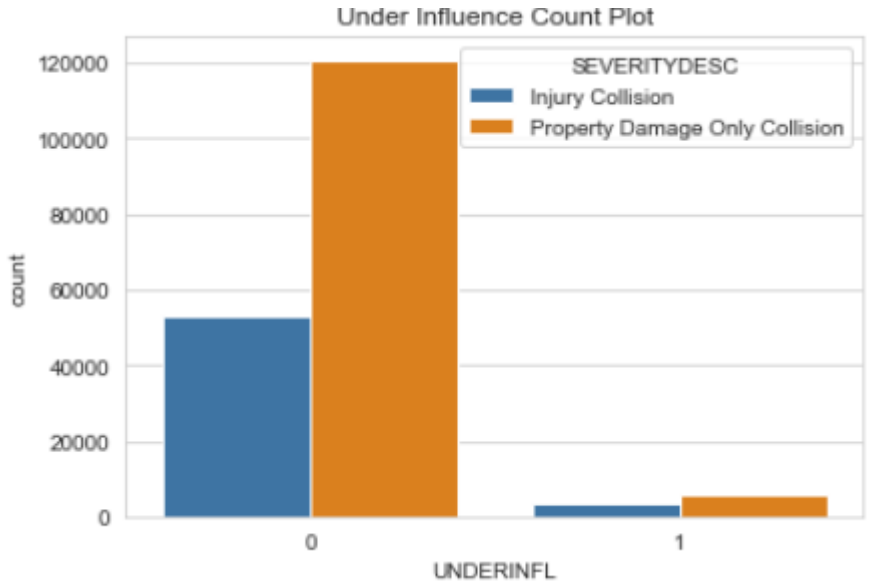
based on this analysis, we can see that each condition give a little impact of what type of Collision/Accident that will happened and how many people involved in the accident. but later i tried to analyze this even further by trying to make a machine learning model to predict the type of Collision based on the Weather, Road, and Light Condition.

next i check if the accidents that happen often more happened when the driver under influence of alcohol/drugs or not.



there are a lot more accidents that happend with the driver not under influenced of alcohol/drugs, my assumption is because most people obey the law to not drive when under influenced of alcohol or drugs.

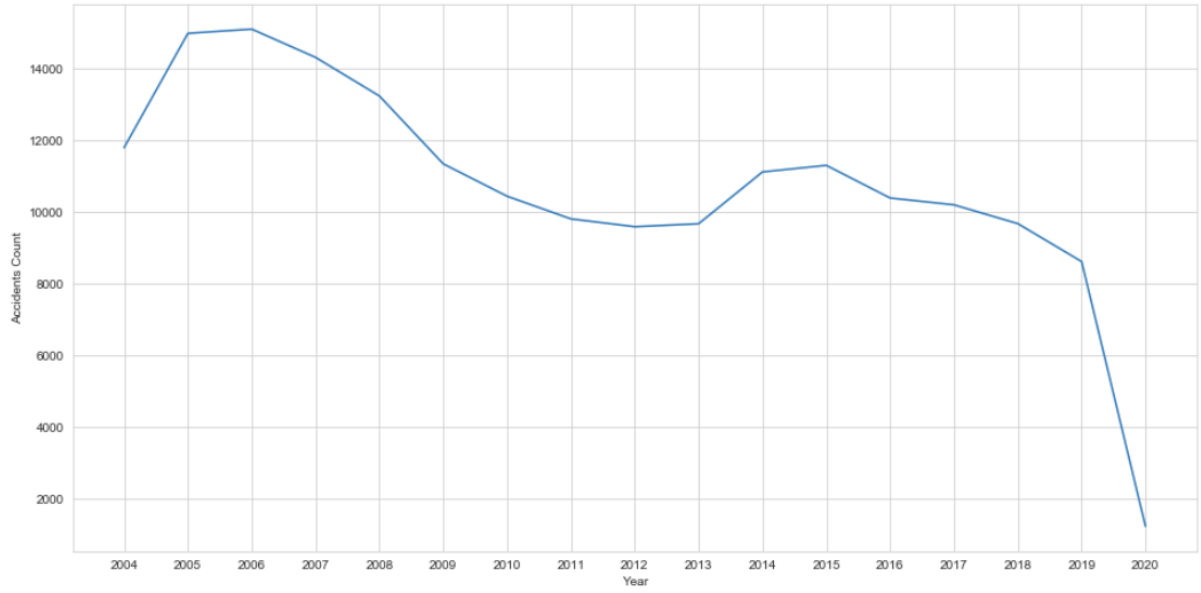
Then I check if Under influence of alcohol/drugs affect the type of the collision that happened.



SEVERITYDESC	Injury Collision	Property Damage Only Collision
UNDERINFL		
1	0.392131	0.607869
0	0.305334	0.694666

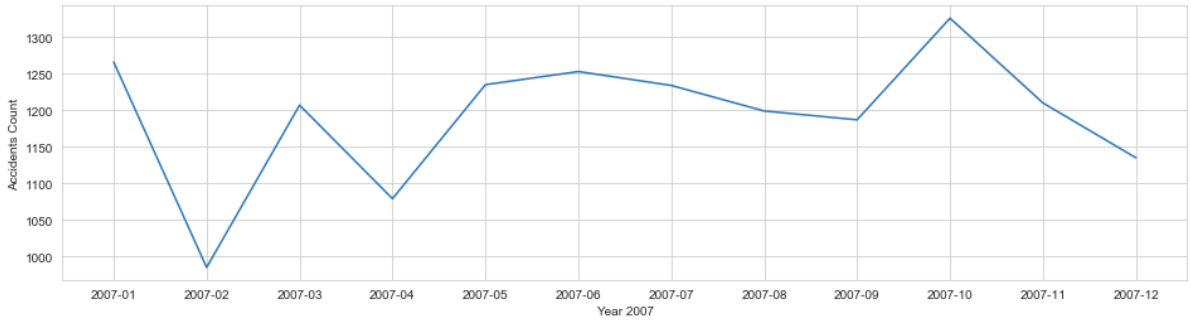
under influence of alcohol/drugs not really make a difference of the type of the collisions/accidents that happened, but under influence still higher than not under influence in Injury Collision, so its still safer to not be under influence of alcohol/drugs when you drive.

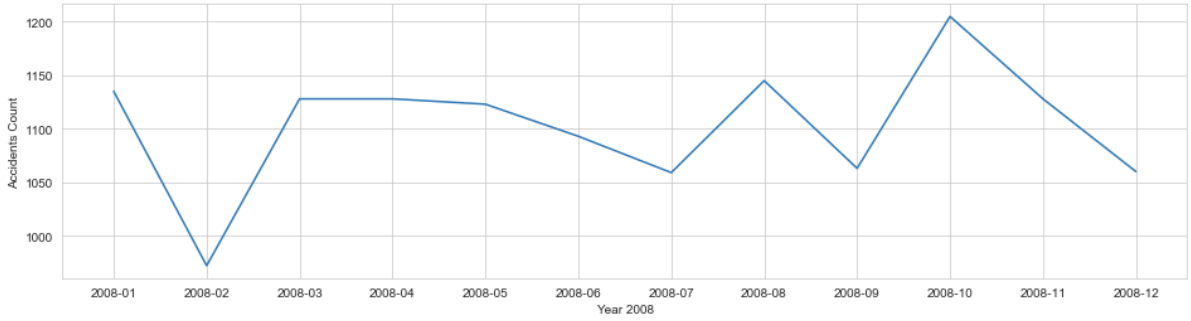
Then I went to see the accident trend with the help of accident date data, first i focused on per year accident trend from 2004 to 2020.

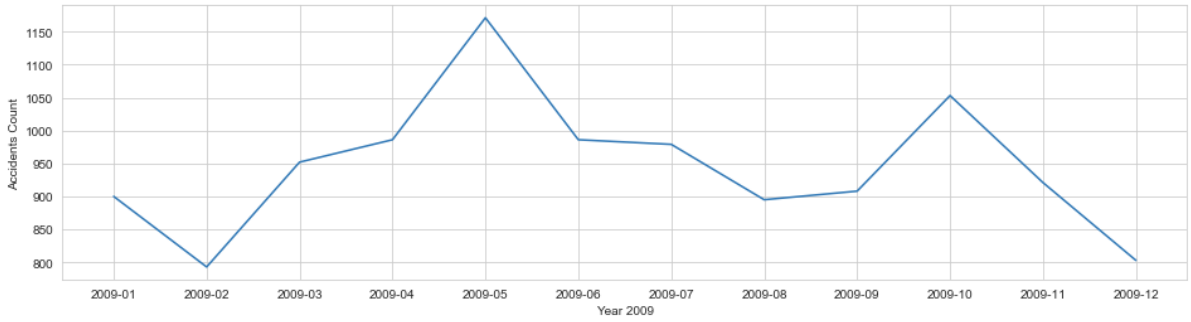


the accidents that happen in Seattle each year have trend to get lower starting from year 2006 to 2019, i cant used the year 2020 because the year isnt over yet and the data only until 2020 April. but my prediction is in year 2020 the traffic accidents that will happen in the world not only in Seattle will decrease, because of the covid-19 that happening right now, it makes people to stay at home more than when there is no covid-19.

Then i check the trend of the accidents each year per month this is some the years I checked, (you can see the full plots in the notebook):







the traffic accidents trend each year in Seattle quite similar, the lowest accidents almost always happened in February each year. I dont know why the trend is like this, but this is only based on the data that i have and use.

4. Predictive Modeling

i tried to make a classification machine learning model to predict the types of the collision/accident based on the Weather, Road, and Light Condition. the purpose is just to know if we can actually know what type of collision that will happened based on the 3 conditions above.

i used Random Forest Classifier Algorithm for easier target class balancing.

first i droped the unknown data from each 3 conditions columns because we need to know these 3 specific conditions to try predict the outcome. Then i did One Hot Encoding to the 3 conditions categorical columns, because its often give more accuracy to the model

then I did a train test split, and train the model using the train data. This is the hyperparameter I chose to balance the target class

```
rfc = RandomForestClassifier(n_estimators=100, random_state=101,
                             max_depth=10, class_weight={'Injury Collision':2, 'Property Damage Only Collision': 1 })
rfc.fit(X_train, y_train)
```

And then I evaluate the model using the test data, and here is the evaluation.

	precision	recall	f1-score	support
Injury Collision	0.34	0.79	0.47	16597
Property Damage Only Collision	0.70	0.25	0.37	33757
accuracy			0.43	50354
macro avg	0.52	0.52	0.42	50354
weighted avg	0.58	0.43	0.40	50354

ROC AUC Score = 0.523438199180945

5. Result

The model performance isnt great like i expected. its make sense because of course its hard to know what type of accidents/collisions that will happen only by that 3 conditions and it will be almost impossible if we want to try predict when will the accident happen, thats why its called accident right? because there are so many uncontrollable variables that caused the accident to happen.

6. Conclusion and Future Directions

first this data only represent accidents that happened in Seattle, and there is no record of travel that not have accident. as you see in the data analysis, first the condition of the weather,road, and light doesnt really affect what type of accident that will happen and how many people involved in the accident. Most of the time accidents that happened only involve 2 person.

second, its safer to drive without influence by alcohol/drugs, but it is not really significant based on the data that we have.

third, the trend of the accidents in Seattle from 2004 to 2020 is going down so its a good sign. and the trend per year is quite similar from 2004-2020 when we sum the accidents per month (Month February almost always have the lowest accidents).

so actually for this problem and this data, simple machine learning model wont solve the problem. if we really want to solve the problem and reduce traffic accidents that happen, i suggest we can use AI Reinforcement Learning to build an auto driving car like Tesla did. Because with machine there is no human error, but its only if we trained the AI well, the sensor isnt broken and the program isnt hackable.