Capstone Assignment Report

Yunya Gao， 2020-09-30

Content:

1 Introduction and Problem statement

A traffic collision, also called a motor vehicle collision, car accident, or car crash, occurs when a vehicle collides with another vehicle, pedestrian, animal, road debris, or other stationary obstruction, such as a tree, pole or building. Traffic collisions often result in injury, disability, death, and property damage as well as financial costs to both society and the individuals involved.[1]

Every year the lives of approximately 1.35 million people are cut short as a result of a road traffic crash. Between 20 and 50 million more people suffer non-fatal injuries, with many incurring a disability as a result of their injury.[2]

Road traffic injuries cause considerable economic losses to individuals, their families, and to nations as a whole. These losses arise from the cost of treatment as well as lost productivity for those killed or disabled by their injuries, and for family members who need to take time off work or school to care for the injured. Road traffic crashes cost most countries 3% of their gross domestic product.[2]

More than 90% of road traffic deaths occur in low- and middle-income countries. Road traffic injury death rates are highest in the African region. Even within high-income countries, people from lower socioeconomic backgrounds are more likely to be involved in road traffic crashes. Road traffic injuries are the leading cause of death for children and young adults aged 5-29 years. From a young age, males are more likely to be involved in road traffic crashes than females. About three quarters (73%) of all road traffic deaths occur among young males under the age of 25 years who are almost 3 times as likely to be killed in a road traffic crash as young females.[2]

Due to huge losses caused by traffic collision, it is quite important to know the impacts of different factors which result in these collisions and make some effective recommendations to prevent these tragedies.

Usually, a number of factors contribute to the risk of collisions, including vehicle design, speed of operation, road design, road environment, driving skills, impairment due to alcohol or drugs, and behaviours, notably distracted driving, speeding and street racing.[1][3]

Generally, there are two types of factors. One is related to the personal behaviors like over-confidence, speeding, driving skills, drinking alcohol, etc. The other one is related to environmental conditions such as road design, road environment, weather and light.

Given that the importance of personal behaviors has been recognized by the public from long time ago, there are a lot of educational methods and legal regulations to reduce traffic collisions. This project aims to analyze the influences of environmental conditions on traffic collisions and then provide more specific recommendations to urban planners or traffic department to improve the conditions to reduce traffic collisions.

2 Data description

The dataset used in this project is "Data-Collisions.csv" from week 1 in this course. You can get data by clicking HERE.

There are 37 attributes in the given table, which include location (X and Y), timestamps, severity types, weather, light, road conditions, etc.

Among these attributes:

- SEVERITYCODE, PERSONCOUNT and COLLISIONTYPE can be used as dependent variables to indicate the severity of the traffic collisions;
- WEATHER, ROADCOND, LIGHTCOND, ADDRTYPE can be used as independent variables to show how environmental conditions influence the severity of the collisions;
- INCDATE, X and Y can be used as temporal and spatial variables to show how the severity of collisions change along temporal and spatial dimensions.

With the data, it is possible to realize the goal of this project, which is to analyze the influences of environmental conditions on traffic collisions and then provide more specific recommendations to urban planners or traffic department to improve the conditions to reduce traffic collisions.

3 Methodology

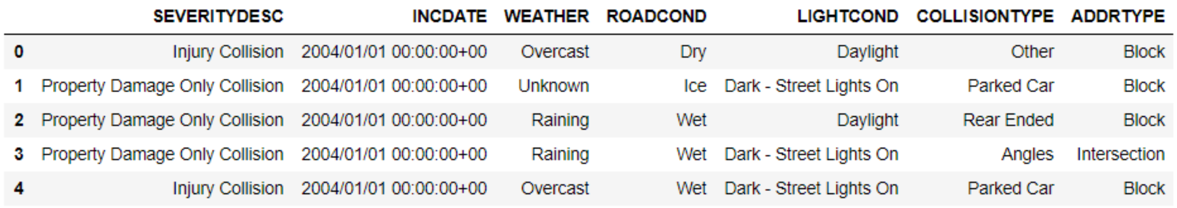
3.1 Data Preparation

From original dataset provided in the course, there are many steps done to prepare data for this project.

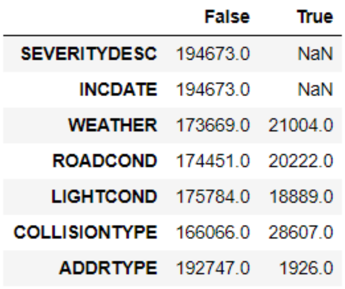
First, environmental conditions related variables were selected as independent variables. Second, recognize missing data and deal with them by deleting them from table due to large amount of existing data. Third, apply one-hot encoding to all variables because they are categorical data. Forth, calculate the sum of these independent variables and delete the variables whose total number is less than 500. Because the amount is too low given the large amount of data.

After all of the above steps, there are two dependent variables: Injury Collision, Property Damage Only Collision. And there are 25 independent variables related to environmental conditions. The total number of data is 146876.

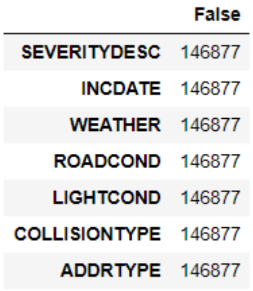
The following several screenshots show part of the results when dealing with missing data and one-hot encoding.



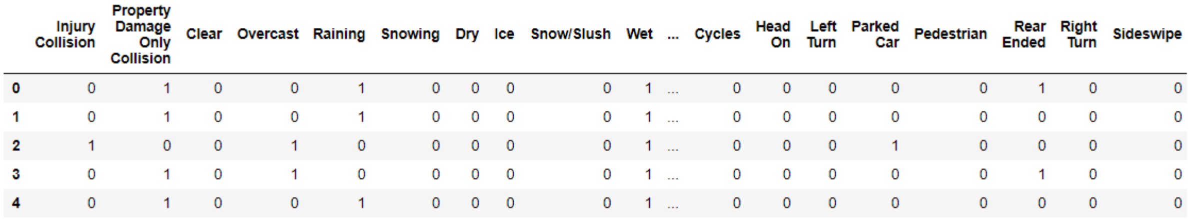
	SEVERITYDESC	INCDATE	WEATHER	ROADCOND	LIGHTCOND	COLLISIONTYPE	ADDRTYPE
0	Injury Collision	2004/01/01 00:00:00+00	Overcast	Dry	Daylight	Other	Block
1	Property Damage Only Collision	2004/01/01 00:00:00+00	Unknown	Ice	Dark - Street Lights On	Parked Car	Block
2	Property Damage Only Collision	2004/01/01 00:00:00+00	Raining	Wet	Daylight	Rear Ended	Block
3	Property Damage Only Collision	2004/01/01 00:00:00+00	Raining	Wet	Dark - Street Lights On	Angles	Intersection
4	Injury Collision	2004/01/01 00:00:00+00	Overcast	Wet	Dark - Street Lights On	Parked Car	Block



	False	True
SEVERITYDESC	194673.0	NaN
INCDATE	194673.0	NaN
WEATHER	173669.0	21004.0
ROADCOND	174451.0	20222.0
LIGHTCOND	175784.0	18889.0
COLLISIONTYPE	166066.0	28607.0
ADDRTYPE	192747.0	1926.0



	False
SEVERITYDESC	146877
INCDATE	146877
WEATHER	146877
ROADCOND	146877
LIGHTCOND	146877
COLLISIONTYPE	146877
ADDRTYPE	146877



	Injury Collision	Property Damage Only Collision	Clear	Overcast	Raining	Snowing	Dry	Ice	Snow/Slush	Wet	...	Cycles	Head On	Left Turn	Parked Car	Pedestrian	Rear Ended	Right Turn	Sideswipe
0	0	1	0	0	1	0	0	0	0	1	...	0	0	0	0	0	1	0	0
1	0	1	0	0	1	0	0	0	0	1	...	0	0	0	0	0	0	0	0
2	1	0	0	1	0	0	0	0	0	1	...	0	0	0	1	0	0	0	0
3	0	1	0	1	0	0	0	0	0	1	...	0	0	0	0	0	1	0	0
4	0	1	0	0	1	0	0	0	0	1	...	0	0	0	0	0	0	0	0

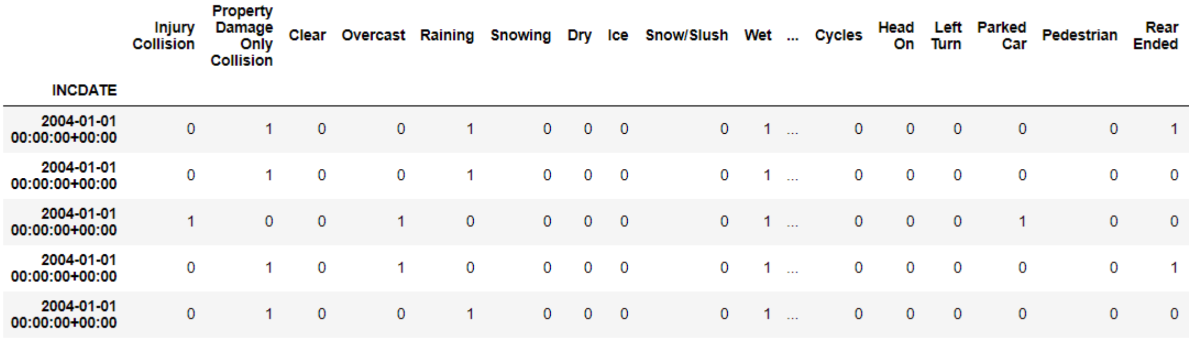
3.2 Analysis

To understand the data, firstly, the project made some exploratory analysis including time series analysis and regression analysis. To provide stakeholders a tool to help them check what the possibility is under a certain set of environmental conditions or predict how traffic collisions can change when climate changes or other conditions change. By this way, they can see the potential consequences for different scenarios they made to improve the situations. There are two types of models made, KNN for modelling categorical data and multiple linear regression for monthly aggregated number of traffic collisions.

4 Results and Discussion

4.1 Time series analysis

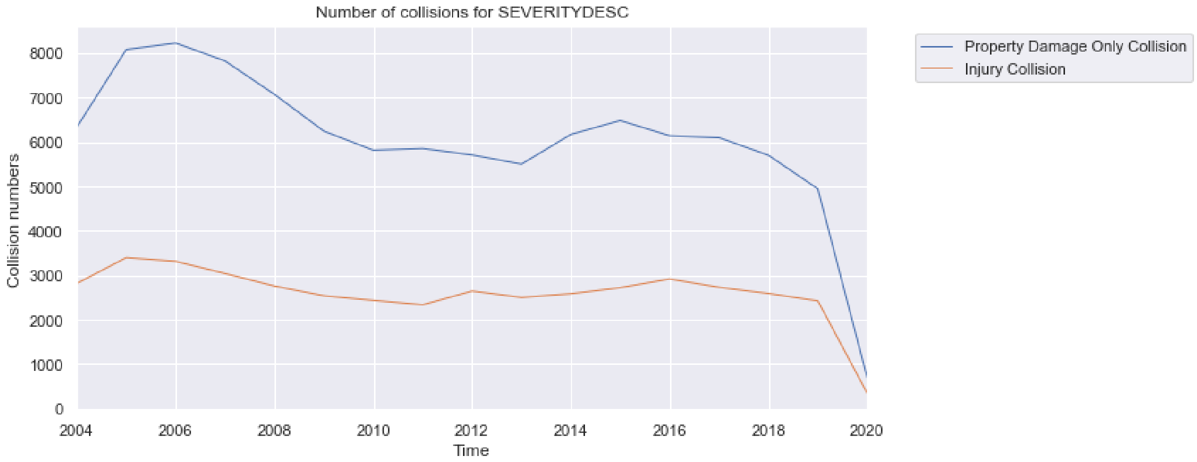
Before analysing how variables change along time, it is necessary to set the timestamps in the data as index, as the following screenshot shows.



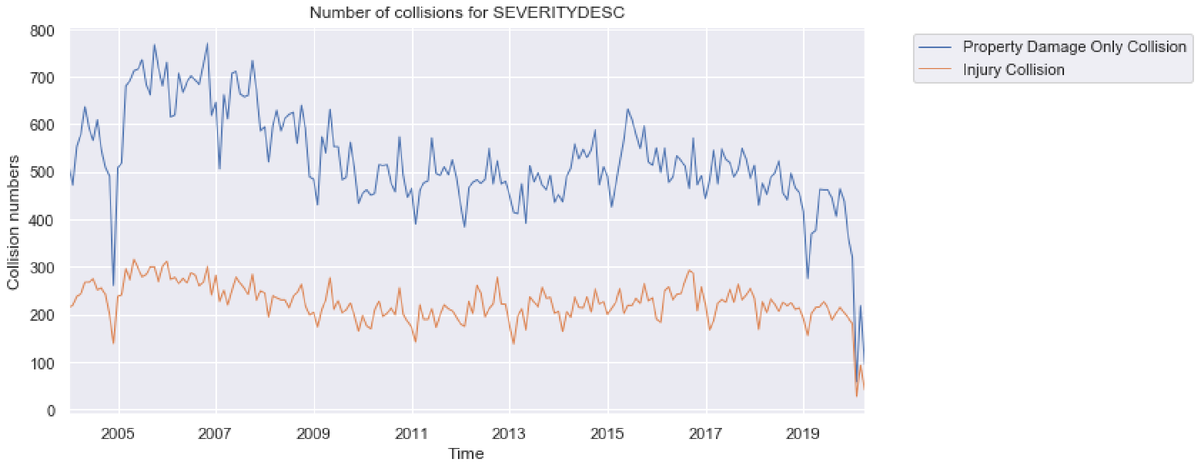
	Injury Collision	Property Damage Only Collision	Clear	Overcast	Raining	Snowing	Dry	Ice	Snow/Slush	Wet	...	Cycles	Head On	Left Turn	Parked Car	Pedestrian	Rear Ended
INCDATE																	
2004-01-01 00:00:00+00:00	0	1	0	0	1	0	0	0	0	1	...	0	0	0	0	0	1
2004-01-01 00:00:00+00:00	0	1	0	0	1	0	0	0	0	1	...	0	0	0	0	0	0
2004-01-01 00:00:00+00:00	1	0	0	1	0	0	0	0	0	1	...	0	0	0	1	0	0
2004-01-01 00:00:00+00:00	0	1	0	1	0	0	0	0	0	1	...	0	0	0	0	0	1
2004-01-01 00:00:00+00:00	0	1	0	0	1	0	0	0	0	1	...	0	0	0	0	0	0

4.1.1 Severity types

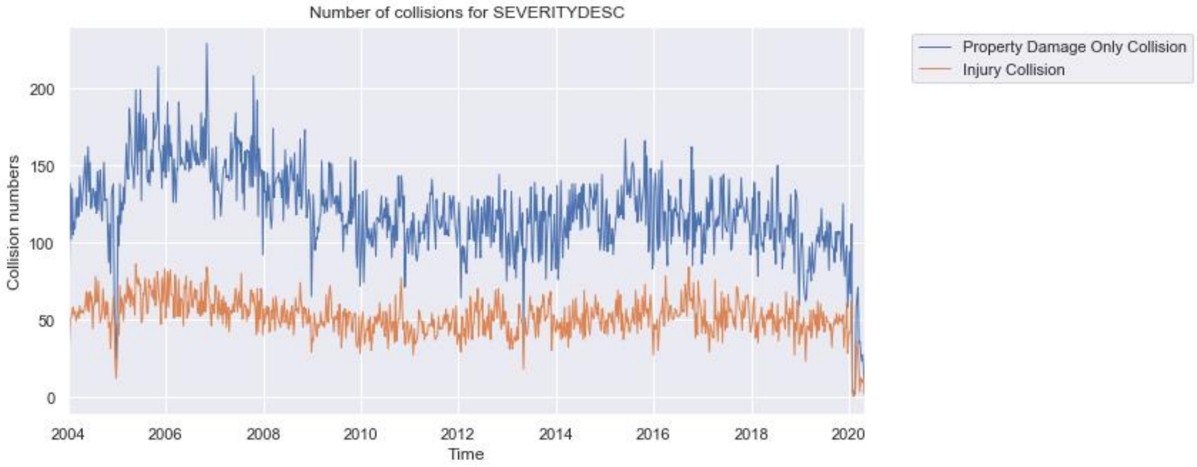
From annual aggregation, overall, it can be found that the annual numbers of both two types of collisions have decreased from 2004 to 2019. (Data for 2020 is not completed.) The annual numbers of property damage only collision (property collision) are all much higher than that of injury collision. The fluctuation of property collision is bigger than that of injury collision.



From monthly aggregation, it can be seen some "periodic cycles" like sin or cos curves on both two types. It indicates that the numbers of collisions are related to periodic factors. At the same time, we can see that the peaks and valleys are mostly matched in terms of time.

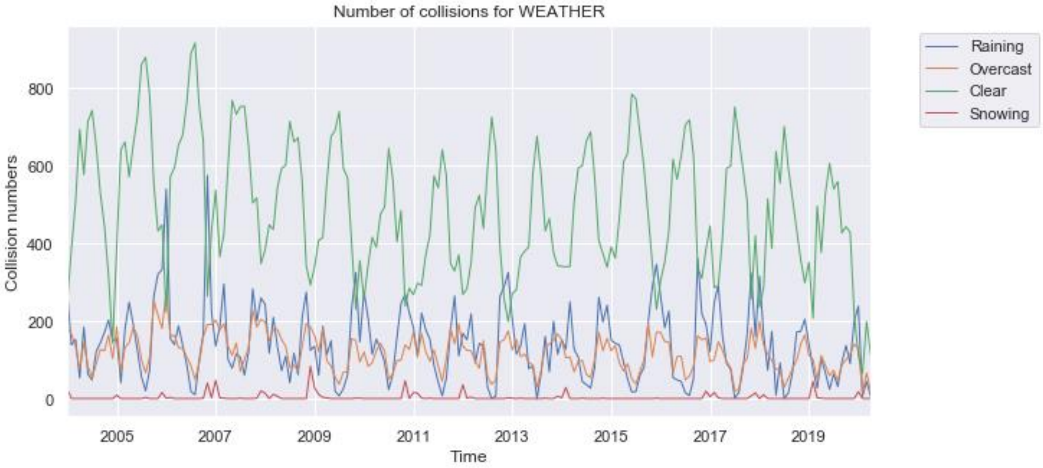


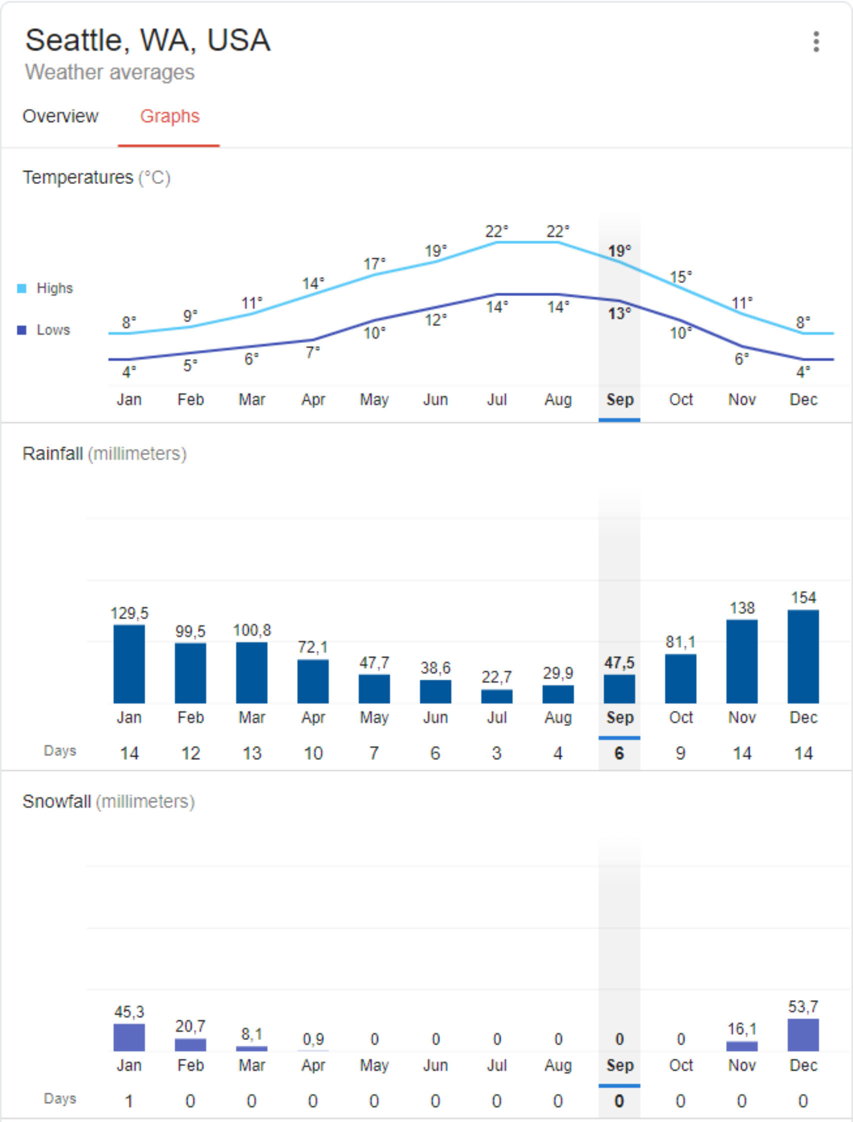
From weekly aggregation, "periodic cycles" are still there but more noisy than monthly aggregation. Therefore, the following analysis for other variables will use monthly aggregated data.



4.1.2 Weather

It can be observed that the occurrence of "Raining" is totally different from other weather conditions. This patterns are near the patterns of climate in Seattle.[Seattle, WA, USA, NOAA]. So we can see that the peak of "clear" usually occur in summer and autumn while the peak of other variables usually occur in spring and winter. The number of "snowing" is quite low.





Seattle, WA, USA

Weather averages

Overview Graphs

Temperatures (°C)

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Highs	8°	9°	11°	14°	17°	19°	22°	22°	19°	15°	11°	8°
Lows	4°	5°	6°	7°	10°	12°	14°	14°	13°	10°	6°	4°

Rainfall (millimeters)

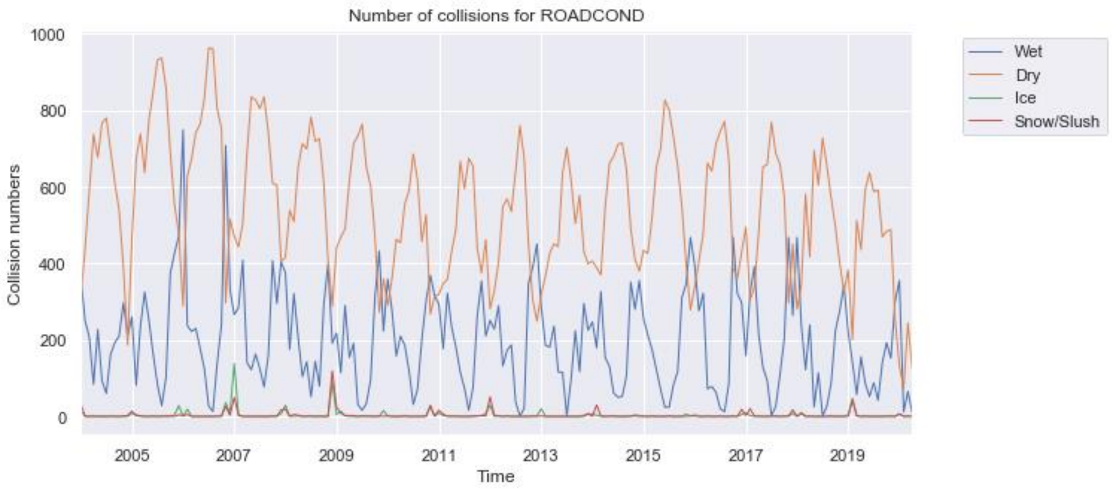
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
	129,5	99,5	100,8	72,1	47,7	38,6	22,7	29,9	47,5	81,1	138	154
Days	14	12	13	10	7	6	3	4	6	9	14	14

Snowfall (millimeters)

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
	45,3	20,7	8,1	0,9	0	0	0	0	0	0	16,1	53,7
Days	1	0	0	0	0	0	0	0	0	0	0	0

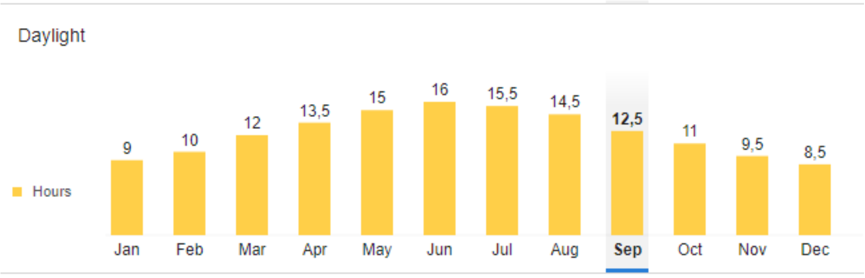
4.1.3 Road

It can be observed that the patterns of road conditions are quite similar to that of weather conditions. In most cases, the weather conditions decide the road conditions.

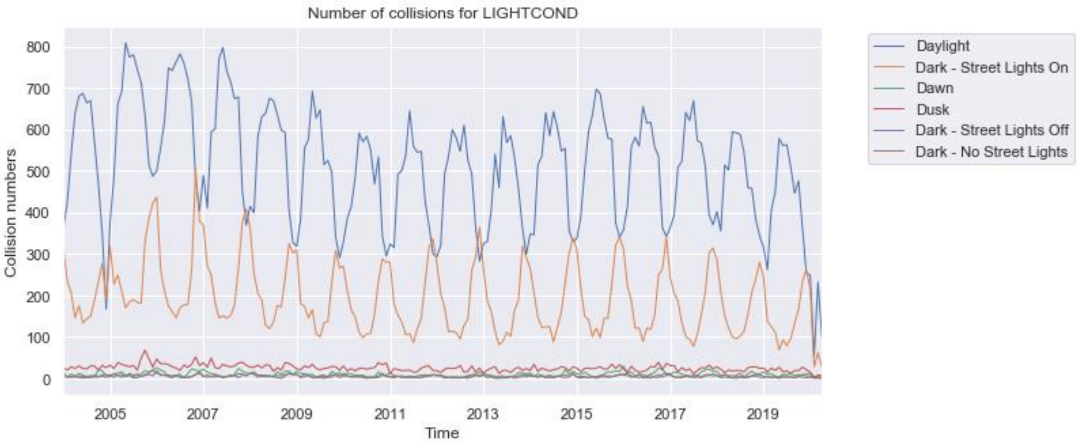


4.1.4 Light

It can be observed that the patterns of light are also periodic. This is mainly related to the location of Seattle, which decides the daily light time. Besides, it can be seen that the numbers of "Dark-Street Lights Off" and "Dark-No Street Lights" are quite low compared to other conditions. This indicates that the possibility of collisions caused by no street light is quite low. And thus increasing street light may not help reduce collisions a lot.

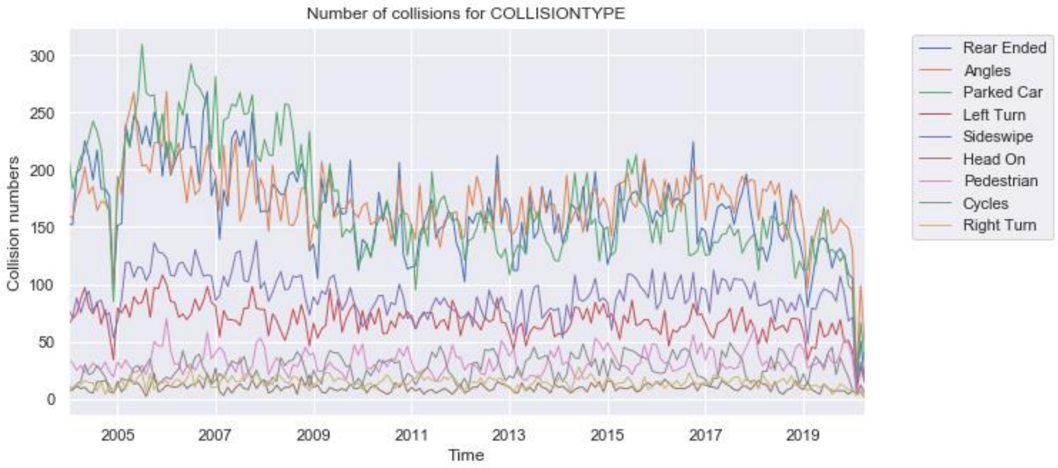


Source: NOAA

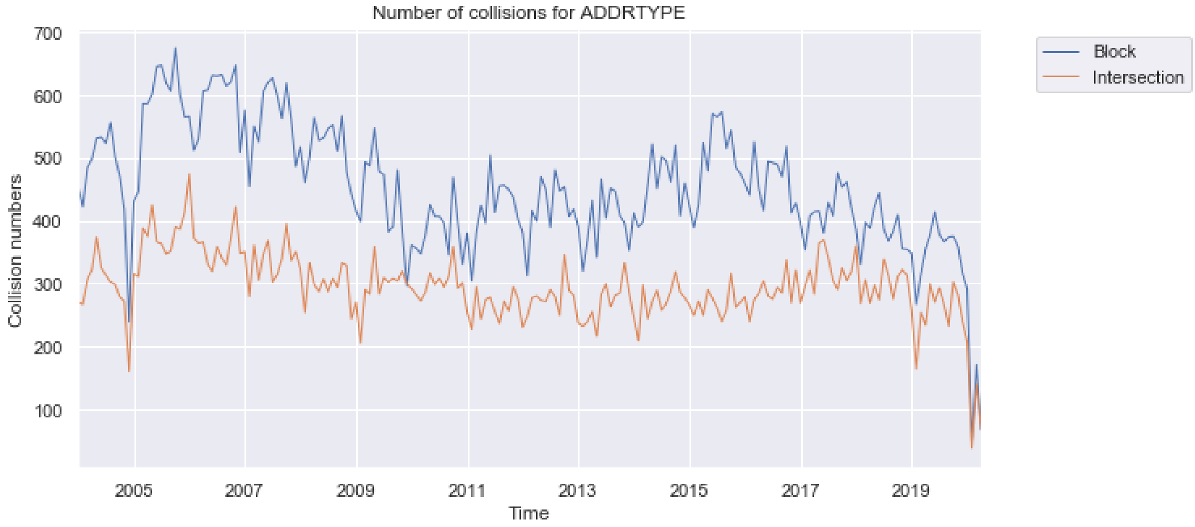


4.1.5 Collision types

It can be observed that there are mainly three categories for these types. The curve of category one (Rear Ended, Angles, Parked Car) is quite similar to property collision while the curve category two (left turn and sideswipe) is quite similar to injury collision. These similarities are interesting to be analysed deeper.



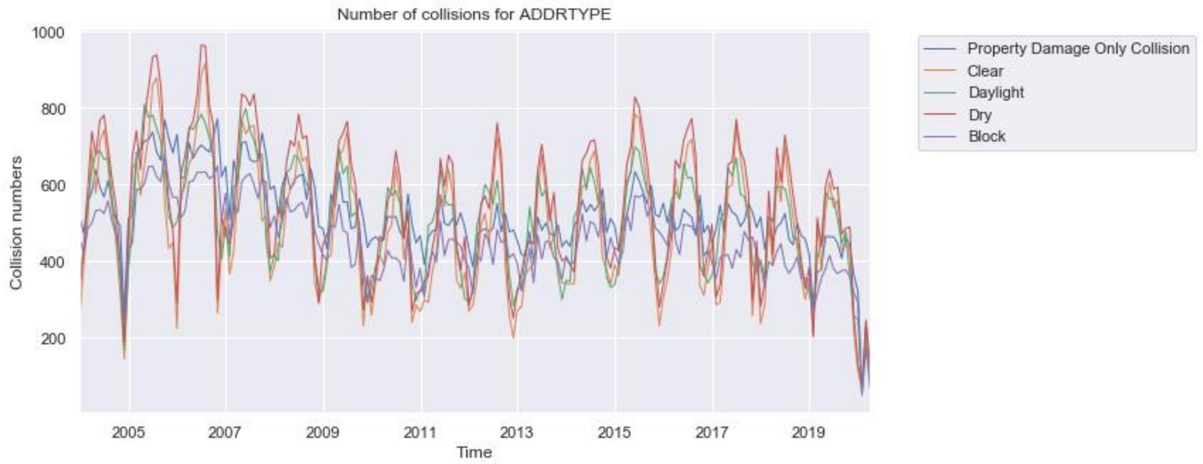
4.1.6 Address types

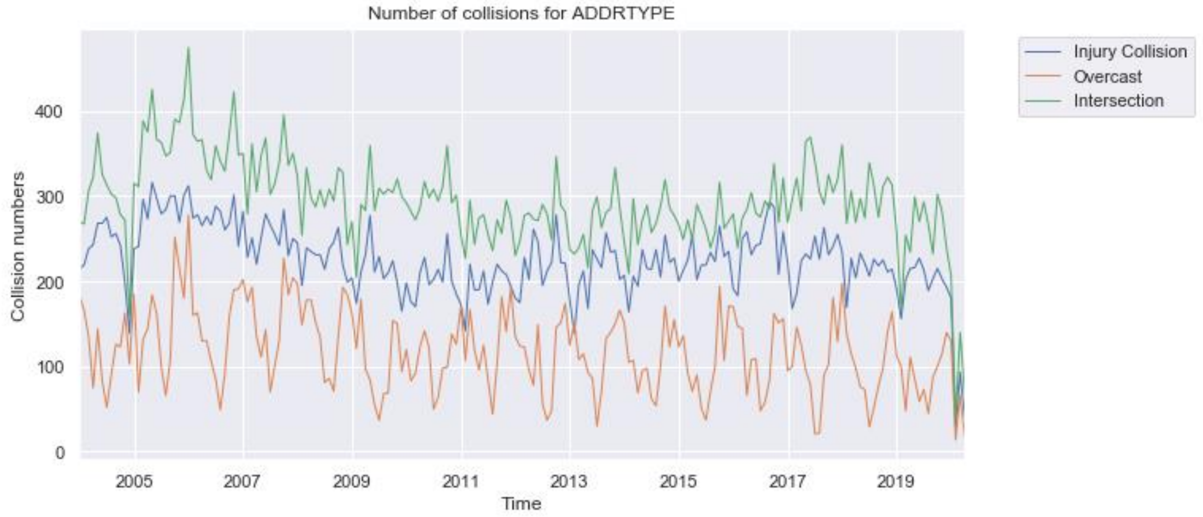


4.1.7 Findings

- The similarities in shapes of curves of two types of collisions and environmental conditions

After trials, the curve shape of property collision is similar to that of "clear", "daylight", "dry" and "block". The curve shape of injury collision is similar to that of "intersection". These similarities may indicate the relations between these variables. More researches are needed to dig out them. But if assuming there are some causal relationships, then it may reveal that property collisions tend to happen when environmental conditions are "good" for driving. The potential reason may be that people don't pay enough attention to driving due to over-confidence. If assuming property collisions are less severe than injury collisions, then it may be concluded that good environmental conditions may let people be over-confident on driving. And the collisions caused by the over-confidence are not as severe as the collisions caused by bad environmental conditions. Because weather plays an important role in deciding environmental conditions, so the curves of property collisions look more periodic and the numbers are much more than injury collisions.

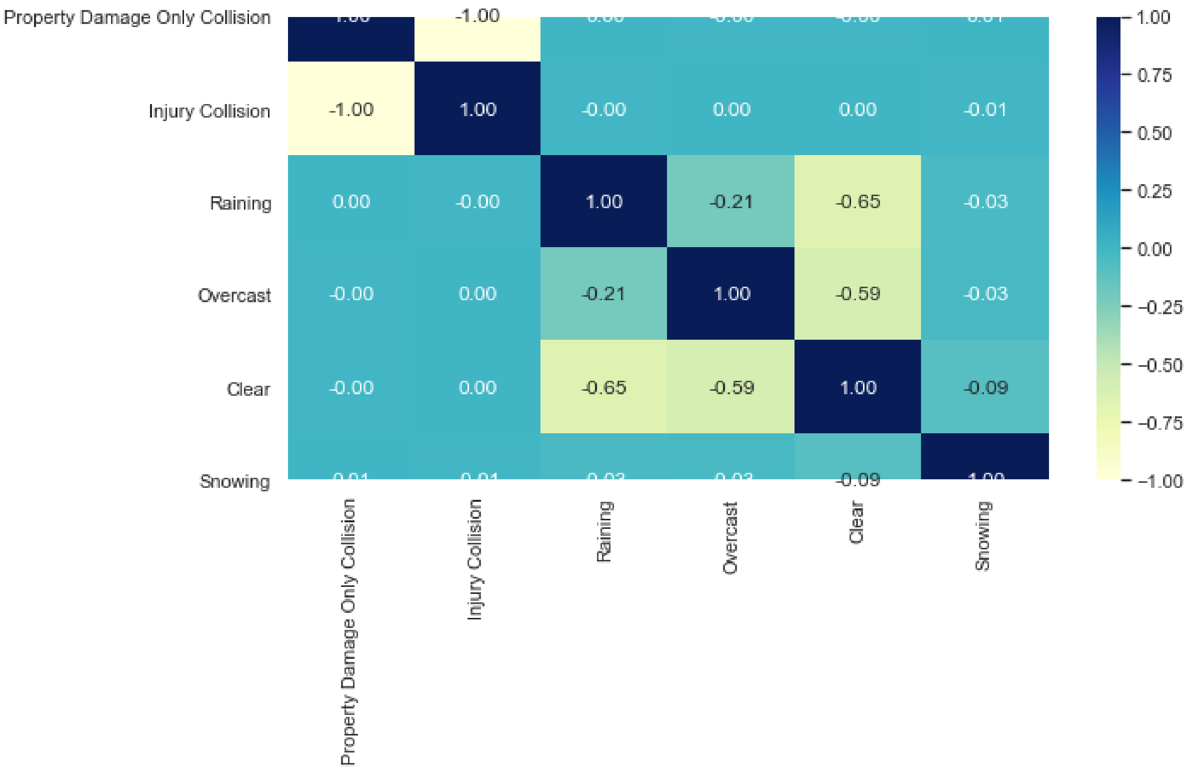




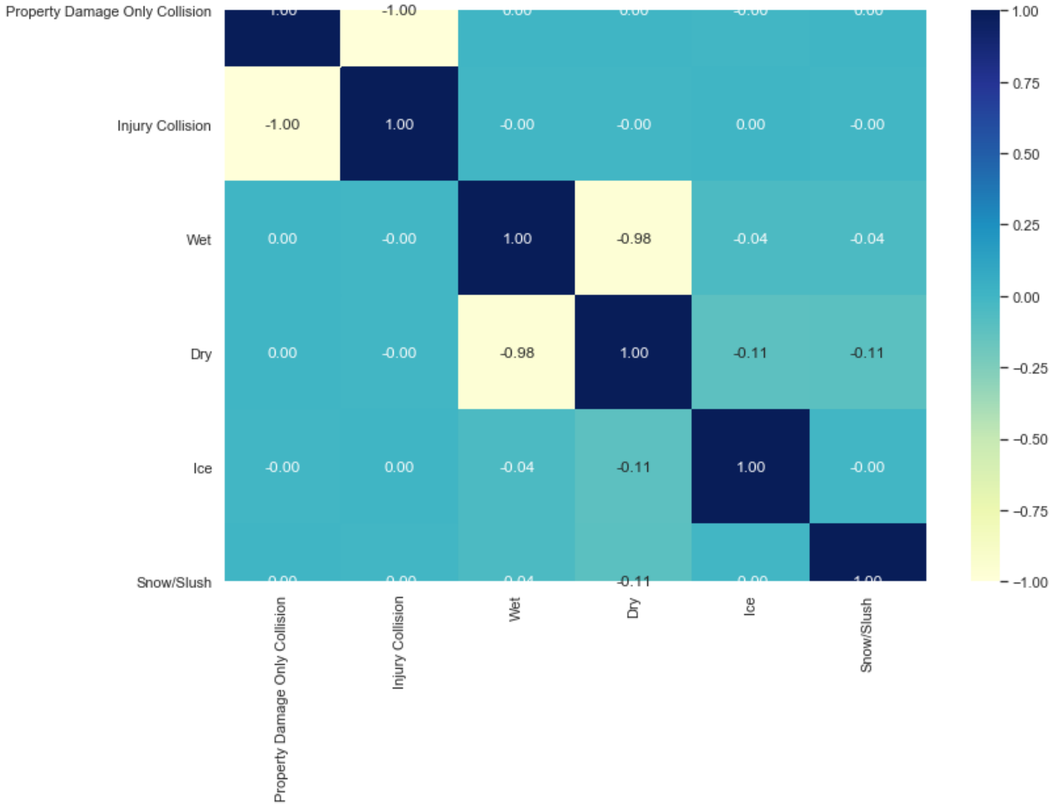
4.2 Regression analysis

This section analyzed the linear correlations between different variables. But overall, there are no obvious findings from these results except some common knowledge like findings. For example, weather "raining" and road "dry" have a coefficient -0.65. These results may reveal that for data with time series, simple linear regression may not work well.

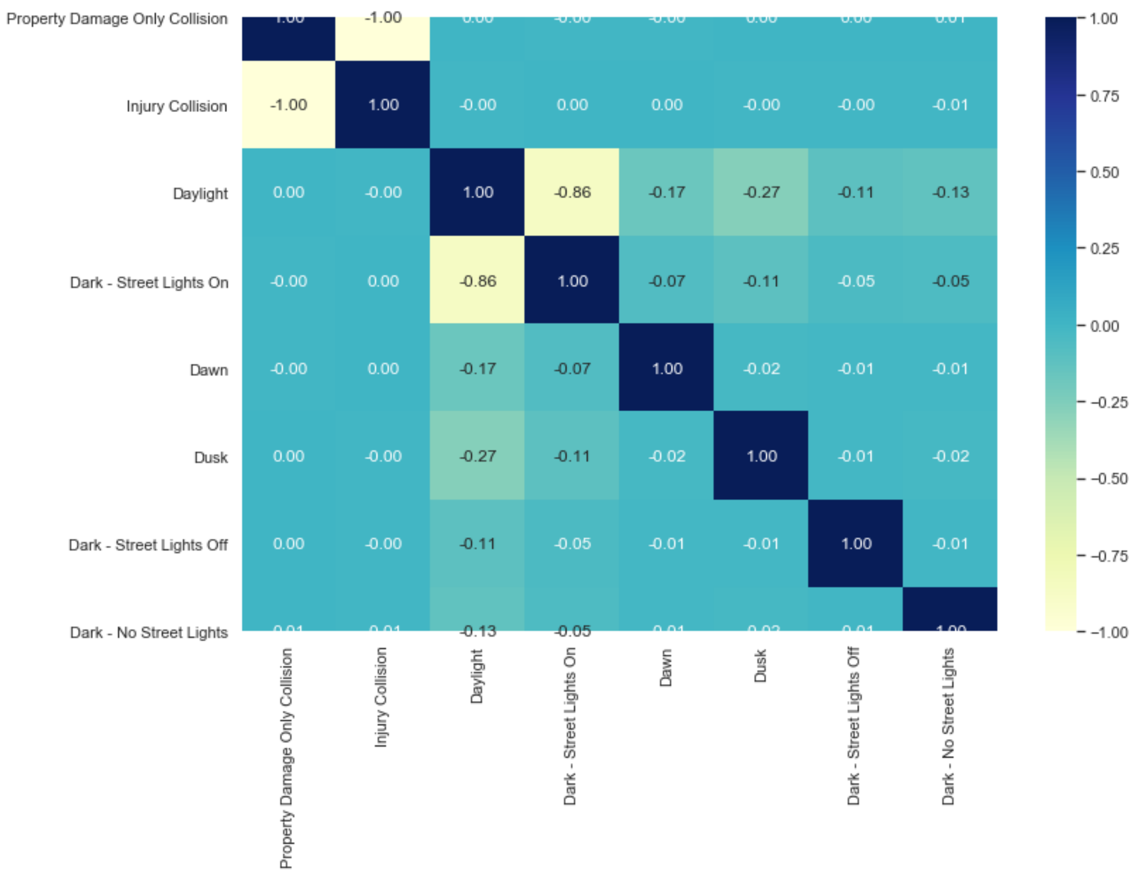
Correlations between severity types and weather conditions:



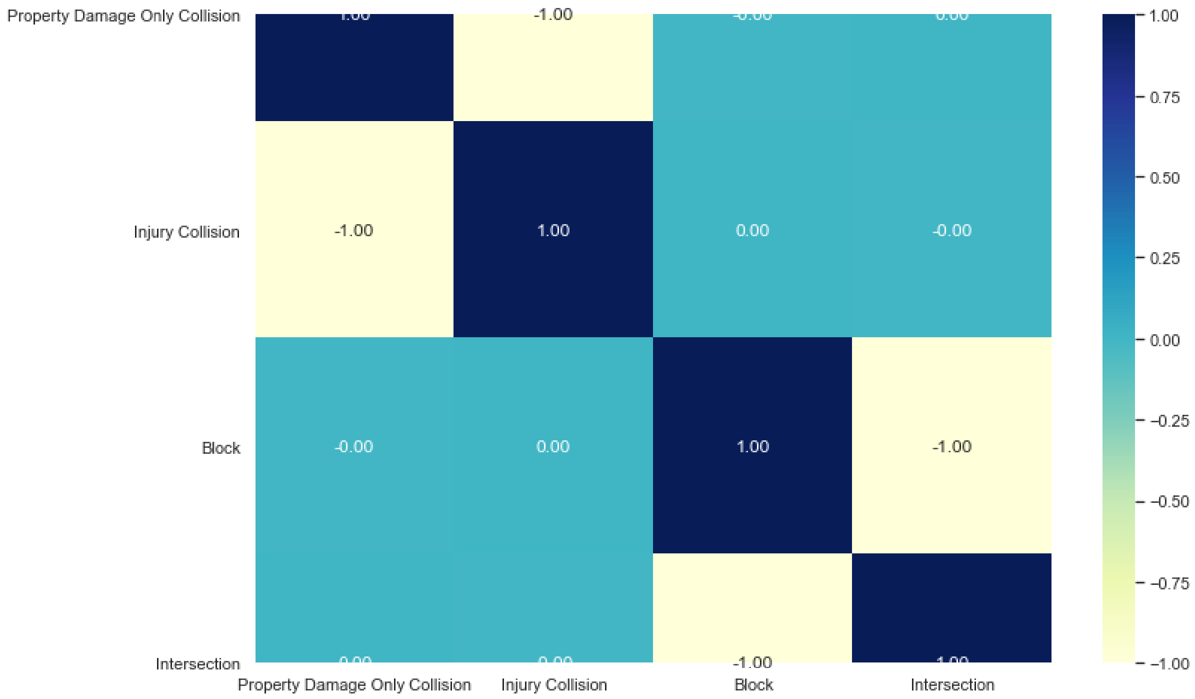
Correlations between severity types and road conditions:



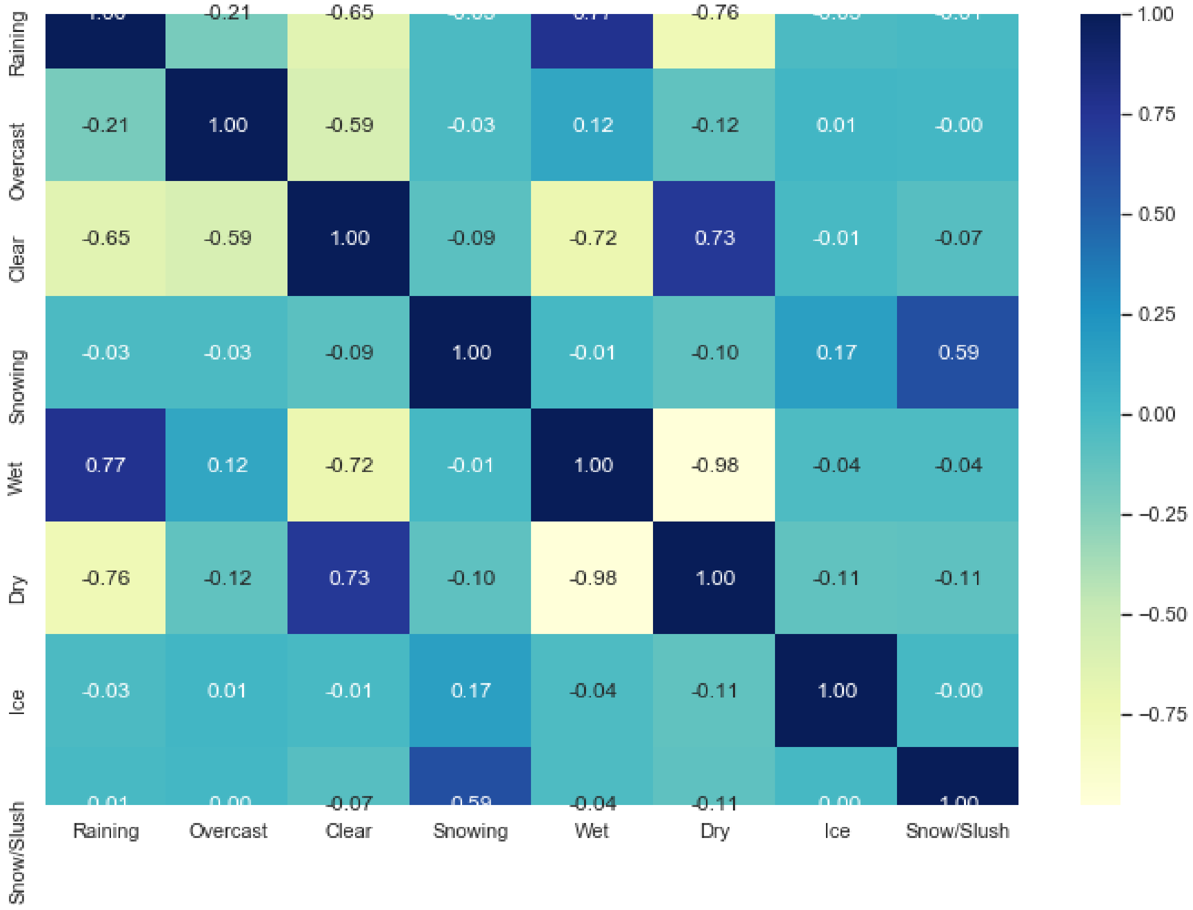
Correlations between severity types and light conditions:



Correlations between severity types and address conditions:



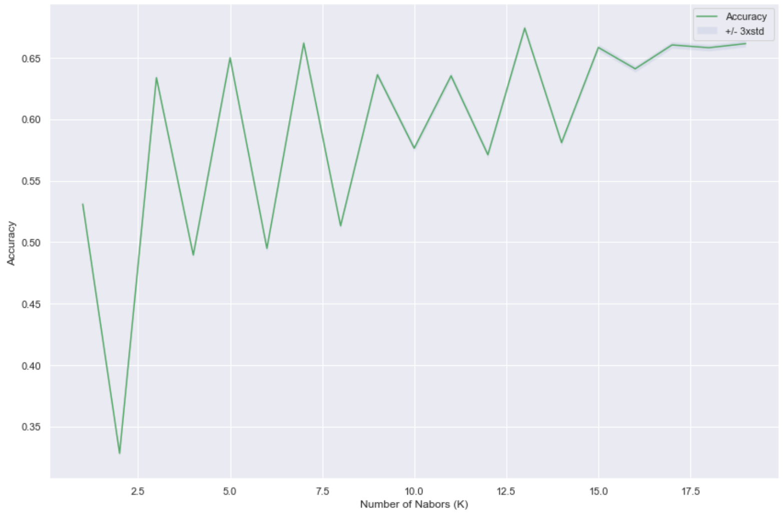
Correlations between weather and road conditions:



4.3 Build KNN model for classification of original categorical data

Building classification models for categorical data (without monthly aggregated) is for urban planners or traffic departments to analyse the possibility of the occurrence of traffic collisions, such as predicting how traffic collisions can change under the big background of climate change.

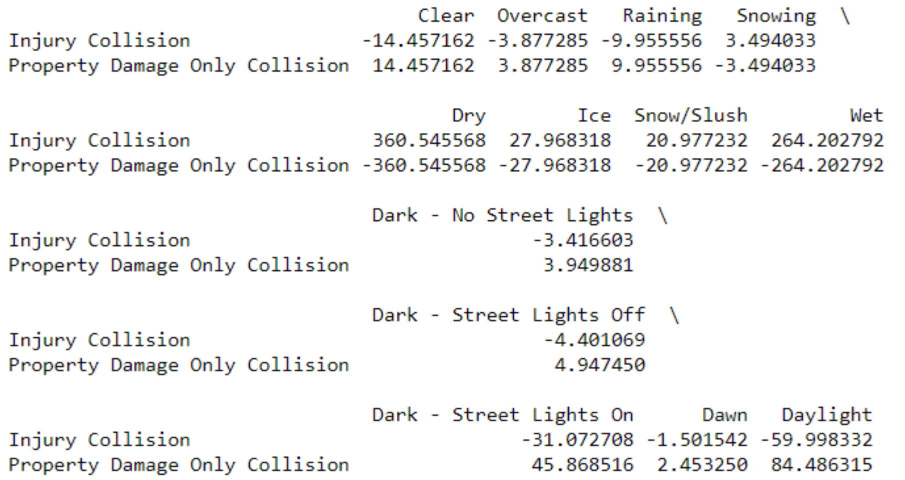
In this project, 80% of data were selected for training while 20% for testing. To select the k value which can produce the highest accuracy, k values from 1 to 20 have been tested. The results show that k=13 performs best with an accuracy score 0.67. At the end, all data were used for training KNN mode with k = 13. The trained model can be helpful for those urban planners or traffic departments.



4.4 Build multiple linear regression models for monthly aggregated data

Initially, building multiple regression models for monthly aggregated data is for urban planners to see which factors play an important role in resulting traffic collisions. The X variables have been normalized before being applied to the model. So, the weights shown in the below table can reveal the importance of a certain variables on the occurrence of traffic collisions. But as mentioned above, the linear regression may not be suitable for this case. So, the values of the weights cannot provide a clear insight though the accuracy is not low (R2 score: 0.9607).

The following screenshots show the weights of the trained model.



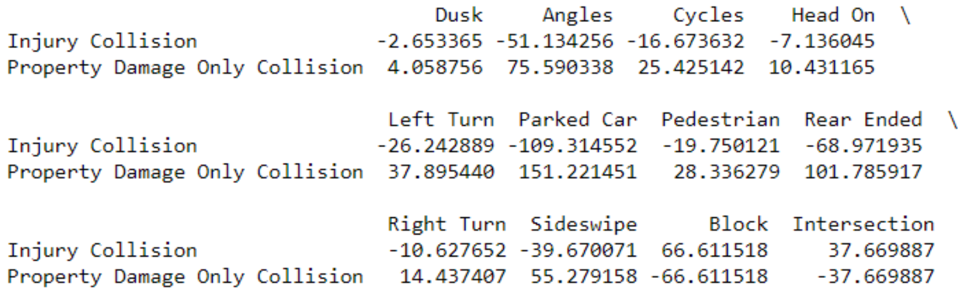
```
                                  Clear  Overcast   Raining   Snowing  \
Injury Collision             -14.457162 -3.877285 -9.955556  3.494033
Property Damage Only Collision  14.457162  3.877285  9.955556 -3.494033

                                     Dry        Ice  Snow/Slush        Wet  \
Injury Collision              360.545568  27.968318   20.977232  264.202792
Property Damage Only Collision -360.545568 -27.968318  -20.977232 -264.202792

                               Dark - No Street Lights  \
Injury Collision                             -3.416603
Property Damage Only Collision                3.949881

                               Dark - Street Lights Off  \
Injury Collision                              -4.401069
Property Damage Only Collision                 4.947450

                               Dark - Street Lights On      Dawn   Daylight
Injury Collision                            -31.072708 -1.501542 -59.998332
Property Damage Only Collision               45.868516  2.453250  84.486315
```



```
                                  Dusk     Angles     Cycles    Head On  \
Injury Collision              -2.653365 -51.134256 -16.673632  -7.136045
Property Damage Only Collision  4.058756  75.590338  25.425142  10.431165

                               Left Turn  Parked Car  Pedestrian  Rear Ended  \
Injury Collision              -26.242889 -109.314552  -19.750121  -68.971935
Property Damage Only Collision  37.895440  151.221451   28.336279  101.785917

                               Right Turn  Sideswipe      Block  Intersection
Injury Collision               -10.627652 -39.670071  66.611518     37.669887
Property Damage Only Collision  14.437407  55.279158 -66.611518    -37.669887
```

5 Conclusions

This project aims to analyse how environmental conditions can influence the occurrences of traffic collisions. The following parts concluded the major content of each section and the findings from analysis.

- Time series analysis

Most of the curves of selected variables related to environmental conditions are periodic. The basic reason for is related to the location and climate of Seattle.

From the shapes of these curves of the variables, it can be found that property collisions tend to happen when environmental conditions are "good" for driving. Good environmental conditions may let people be over-confident on driving. And the collisions caused by the over-confidence are not as severe as the collisions caused by bad environmental conditions. Because weather plays an important role in deciding environmental conditions, so the curves of property collisions look more periodic and the numbers are much more than injury collisions. The above conclusion was deferred from data only, more researches should be done to further explore the mechanism.

- KNN model

To select the most suitable K value for this project, values from 1 to 20 have been tested. The number of training data is 117501; the number of testing data is 29376. The k which produced highest accuracy (accuracy score=0.67) is 13. Then both training and testing data merged together trained the KNN model with k=13. The final trained model can be used for urban planners or traffic department to check the possibility of occurrence of traffic collisions for different environmental conditions so as to make some preparations in advance.

- Multiple linear model

Linear regression models may not work well for this case. The weights of trained models cannot help explain the influences of the variables on the occurrence of collisions.

Reference

[1]https://en.wikipedia.org/wiki/Traffic_collision
[2]https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries
[3]https://www.sciencedirect.com/science/article/pii/S0001457518300873#:~:text=What%20are%20the%20main%20contributing,the%20collisions%20of%20older%20drivers