IBM Data Science Capstone Project Report

Chan Ching Kwan

August 2020

1 Introduction

1.1 Background

The advance of traffic technology have brought a huge benefit to both individuals and society by saving time and linking up different places. Countries are building more roads, and the industry are manufacturing more cars. Expansion of traffic usage is expected. However, the increasing number of car usage have also brought the increasing amount of accidents which caused many death and economic loss. Research should be done on car accidents to reduce the accidents ratio, otherwise the loss from accidents would only be more serious with the rise of traffic. From 2004 to half of 2020, there are more than 12000 collision every year in Seattle, which means in total around 200000 have happened in just on city. In a compact city Hong Kong, according to the Road Safety Council of HKSAR, there are on average 41 accidents per day resulted around 3 man killed every week in 2015. And this number stay almost the same. Collision is a serious and very common problem.

Data science can come into play on the traffic accident problem. By investigating the data collected from the previous traffic accidents, data science can diagnose the cause of traffic accidents. From uncontrollable factors, like weather and animals, to controllable factors, like roads design and improper use of car, data science can find out the impact of those factors

1.2 People of interest

The huge number of traffic accidents is serious public problem which may involve city design, education of transport and advertisement on safe driving. The first party would like to solve this problem is surely the government. As government is the authority that can make big changes to reduce this loss.

Insurance company would also be interested in this research on traffic accidents. As all cars should have insurance. When there is accidents, insurance

company have to compensate. If there are predictive model of accidents, insurance company then can set the price accordingly and even reduce the compensation rate by reducing the accidents rate.

With the knowledge of different factors causing the traffic accidents, then effective solution can be found to deal with this problem. Thus reducing the loss from accidents. The data analysis on car accidents is beneficial to government because of above reason. It may also be a good basis for traffic technology to predict and prevent accidents by analysing the traffic conditions and warming the driver with high risk.

2 Data

2.1 Data information

In this capstone project, the example data of ArcGIS collision data is used. This set of data recorded the conditions of in total 194673 collisions from 2004/01/01 to 2020/05/19 at City of Seattle by SDOT Traffic Management Division, Traffic Records Group. The following informations are highlighted as expected to have determining factor on accidents:

1. Geographic location
2. Light condition
3. Driver condition on drugs and alcohol
4. Road condition
5. Weather
6. Speeding
7. Parked car
8. Seriousness of the collision

These information are useful for finding the most impactful way to reduce the collision. For example, one can correlate the light condition and the frequency of the collsion to determine which light condition is the most dangerous for driver. It may also be possible to reveal the location of imporper light design by locating collision with different light conditions, then we may be able to find some region or road has bad light design which will cause higher density of collision.

With different combination of information, this data set can reveal how we can improve road design to reduce collision effectively by the proper use of data analysis method.

2.2 Data prepossessing

The data set contains in total 194673 records of accidents with 37 information, including time, geographic location and conditions. But some of the information may not be useful for analysis and some are redundant. For example, 'SEVERITYDESC' is column contains detailed description of the column 'SEVERITYCODE' which is a number describe the type of collision given by the police department. The following table shows the deleted columns and the reason of the removal.

Column Name	Delete Reason
INCKEY, COLDETKEY, OBJECTID, REPORTNO, INTKEY, SDOTCOLNUM,	Number or ID given by the state without useful information
SEVERITYDESC, LOCATION, SDOT_COLDESC, COLLISIONTYPE, ADDRTYPE, ST_COLDESC	Detail description of other data
INCDATE, SEVERITYCODE.1, CROSSWALKKEY, SEGLANEKEY	Repeated data
EXCEPTRSNCODE, EXCEPTRSNDESC	Unknown data without explaination in meta data
PEDROWNOTGRNT, JUNCTIONTYPE	Too specific data which may complicate the analysis

Table 1: The table shows the deleted columns and the reason of removal. 20 columns are dropped to simplify the data set for better analysis.

After obtaining the useful data column, we may consider the problem of different rows, which is the missing data. To prevent bias and simplify the analysis, all missing data would be removed from the data set. If any row contain 'unknown', 'missing' or empty space, the row will be removed. Therefore, after the columns and rows cleaning, in total 184146 rows with 18 columns are remained as out data set.

Finally, some of the data type is no int or float, which are needed to be converted to adopt the library requirement. All non-numeric data are converted into integer or floating point number. The detail is written on the Juypter notebook.

3 Methodology

3.1 Visualizing data

To begin with, it is useful to understand data through different graphs. The following figures show the distribution of different data.

To start with, we would like to find out which types of collision is the most common.

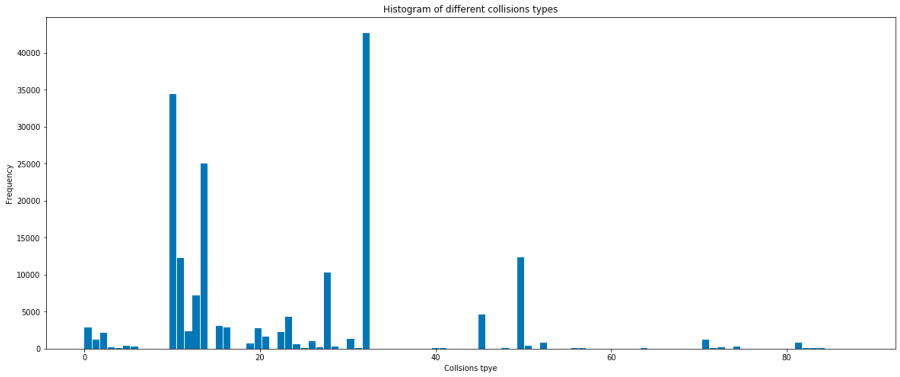

Figure 1: The figure shows the distribution of different types of collisions. The highest three types are 32, 10 ,14 which are obviously higher than other types. According to the meta data provided by the Seattle Police Department, 32 is One Parked - One Moving, 10 is 'Entering At Angle' and 14 is From Same Direction - Both Going Straight - One Stopped - Rear End. These 3 types are already 57.8% of the total cases.

From the data, the most frequently occur collisions are 32 which is 'One Parked - One Moving' case, with 42711 cases which is 23.2% of the total case. This has already revealed the parking problem of the Seattle. If government provides a safe parking space or abandoned the illegal parking, this may hugely reduce the problem of parking car collisions which is a large portion of the collisions.

The second common case is 10, 'Entering At Angle' with 34407 cases (18.6%). Further analysis can be in the following section to find out where has the most case 10 collisions, which may reveal the bad design of the road.

Then we may consider other factors that may cause collisions, including weather, road, light condition. Pie charts below demonstrate the distribution of different columns. As shown in the figure, most of the collisions happens at the normal conditions with clean dry weather with day light. However, this does not provide much because this may due to the traffic conditions and ratio of

good weather. This depends on the number of raining days. According to the Weather and Climate website, there are average about 50% measurable raining days in Seattle[1]. This is surprising that most of the collisions were happened in the clean and dry days even though the numbers of shiny and raining days are the same.

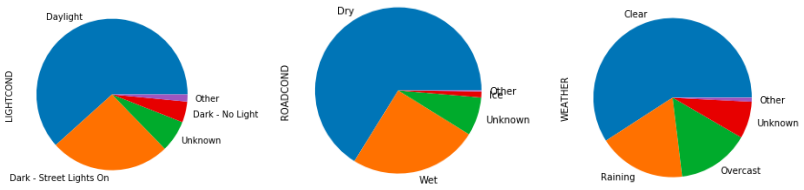

Figure 2: These figures show the distribution of different road conditions. The first one is the light condition. From the pie chart, most collisions happened with daylight, and second majority with street lights. The second pie chart is the about the road, where dry is the majority. The last pie chart is the weather. Over 50% of the collisions happened with clear sky, which is not normal as Seattle has relatively high precipitation.

Another factors that may cause collision are the status of the driver, include speeding, inattention and effect of medicine. The following tables shows the ratio of the positive-negative cases of each conditions. It is clear that most of the cases are not caused by any of those factors, the total number of positive cases are 46661, 25.3%. This number is over-estimated as the union is double counted. This shows driver's conditions would play a role but not major role.

The final part of the this part discuss about the seasonal effect. As shown in the following figure, the collisions happened in every month are roughly the same, which reveals there are no seasonal effect on the traffic accidents. On the other hand, the trend of number of collisions can be found by plotting the histogram of number of collisions happened in each year. From the histogram, we can see a decreasing trend from 2005 to 2019. The data of 2020 was recorded until May, and 2004 is the first year of record, thereby the data from these two years are excluded from this plot. Using the linear regression technique, the trend can be calculated to predict the number of collisions in the future and being a indicator of effectiveness of improving.

3.2 Exploratory data analysis

In this section we would explore the relation of information using the several data analysis technique.

[1] `https://weather-and-climate.com/average-monthly-Rainy-days,seattle,United-States-of-America`

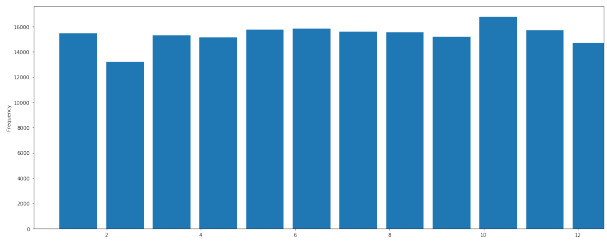

Figure 3: The figure shows the histogram of the distribution of collisions happened in different month. The bars have almost the same height which means the seasonal difference has a very minor effect on the number of collisions.

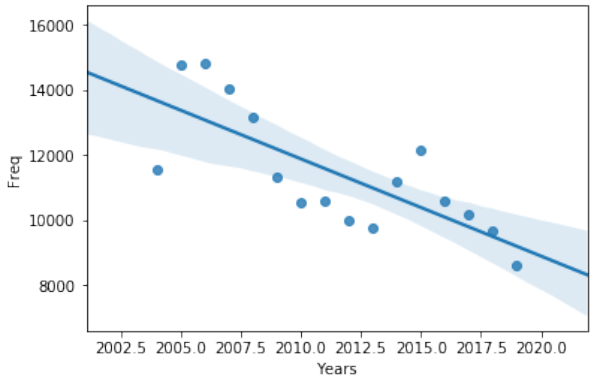

Figure 4: The figure shows the linear regression result of the yearly distribution of collisions. The scatter points are the number of collisions, the blue line is the fitting of regression and the shaped region is the $\sigma = 1$ boundary. The number of collisions shows a decreasing trend. The coefficient of determination of the linear regression is 0.567.

The simplest thing to start is the accident black-spot. Using the locations from the data, density-based clustering method can be applied. The density-based method draw a small circle around the location, and then count the number of nearing points. If the nearing points exceed certain threshold, that data and connected data within the circle is cluster into a group. Then there will be groups of data and non-grouped data. In this case, those grouped data are the accident black-spot. Figure below show the location of black-spot on the Seattle map.

Another things that we may do is calculating the correlation between different information. Pearson correlation coefficient is a great indicator which calculates the overlapping of different variables. However, this coefficient can or be calculated when the data is numeric data. In this case, the data are not

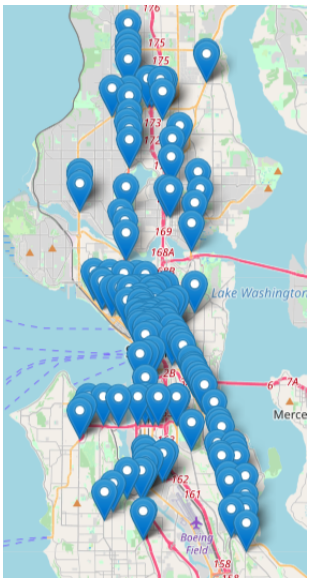

Figure 5: The figure shows the result of density-based clustering technique. Using this method, the spotted location represent a group of dense collisions occurred in the past. The criteria of grouping is 100 cases within the 200m radius. In total 287 groups were found based on this criteria. One thing to be notice is the black spots in the middle to lower part of the map formed a dense network. In fact, if one changes the criteria with higher radius, those dense spots would group together and form a grant group.

numeric but categorical. Therefore, transformation from categorical to numeric data should be done. The first thing we can do is to classify the type of data. If the data is bi-category, 0 and 1 are good enough replacements. When the data is multi-category, we can give rank to this number according to the seriousness. For example, the weather data (excluding the other and unknown data) can be judged by how 'bad' the weather is . Clean is 0, overcast is 1, raining is 2 and so on. This method is also applied to other multi-category data. Finally, with the built in correlation function of pandas, the Pearson correlation coefficient can be found as follow.

3.3 Predictive model

In the predictive model section, two methods can be used practically to prevent traffic accidents.

Combining the accident black-spot analysis in the previous section, dangerous level of the location can be provided and reminds the driver. When the

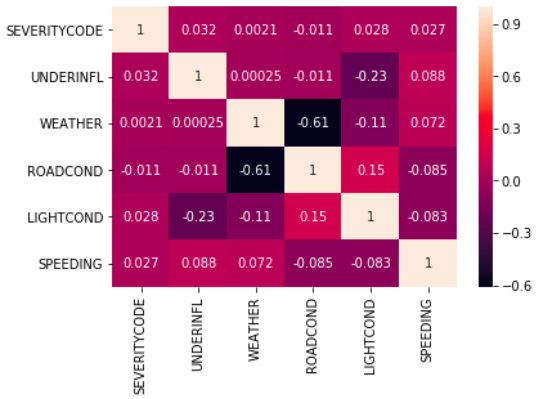

Figure 6: The heat map is the Pearson-correlation coefficient matrix of selected factors from the data. The data is transformed into numerical data. From the matrix, the first row shows the correlation of seriousness with other conditions, but all the elements are small. In fact, the only large element in the matrix is the correlation of weather and road condition. This is very normal result. Therefore, we can conclude that the properties of each collisions are not related.

driver is currently driving on the location with many previous collisions, the application show different colours to indicate the dangerous level. The location can be applied by the Google map API which is out of the scope of this report. When the location is input, the program will calculate the number of collisions occurred within 50 meter radius. If there are 3 different levels: Safe, Careful, Attention which all based on the number of collisions. The program will also show the most common type of collisions to further remind the driver.

```
Your lattitude should be within the location of Seattle
What is your latitude:-122.3
What is your longitude:47.5
Total number of neatby collisions: 36
Dangerous level is HIGH and the most common collsions are type: 50

Your lattitude should be within the location of Seattle
What is your latitude:-122.25
What is your longitude:47.6
Total number of neatby collisions: 0
Dangerous level is LOW and the most common collsions are type: none
```

Figure 7: The above figures show the output of the dangerous level prediction. The threshold was set to be 30 within 100meter radius. The first test ($x = -122.3, y = 47.5$) gave a high dangerous level with total case=36. The most common type of collisions are 50. The second test is ($x = -122.25, y = 47.6$). In this case, the location is safe with no collisions happened before. Therefore the common type is none.

From the correlation of different data with seriousness of the collisions, logistic regression can be used to determine which collisions would cause serious

injury or massive injury. Using the built in function from skilearn and numerical data, the seriousness of the collisions can be predicted through this model. The following table shows the accuracy score of this logistic regression model. However, the logistic regression is actually invalid. Because the number of injurious accidents are relatively small, the model would not be able to capture the information difference. So the accuracy can be reasonably large even if the model predicting all positive. The logistic regression is failed in this case.

	precision	recall	f1-score	support
1	0.67	1.00	0.80	22327
2	0.00	0.00	0.00	10912
micro avg	0.67	0.67	0.67	33239
macro avg	0.34	0.50	0.40	33239
weighted avg	0.45	0.67	0.54	33239

Figure 8: The above figure is the result statistic of the logistic regression. The model have trained to give all result '1' which is no injury collision. This is because the low number of injury collisions. The accuracy of the logistic regression is not bad even if it give all '1' result.

4 Results

In section 2, non-intuitive result are found. In Seattle, where has average 50% raining day, the collisions were most likely happened in the good weather day. Most of the collisions are not affected by the weather conditions and driver conditions. On the other hand, some types of collisions were common. 'One Parked - One Moving' and 'Entering At Angle' in 84 types of collisions had occupied 40% of the total cases.

Using the Pearson correlation coefficient, the seriousness of the collisions are independent from the other conditions, including weather and driver conditions. Also there are no seasonal effect as the monthly distribution is almost flat. This shows that either the reason of most collisions are hidden behind the data or collisions and seriousness are just random.

The most successful out-come is the study of geographical information of the collisions. Using the clustering technique, the accidents black-spots can be easily located. In total more than 250 locations had been spotted. Furthermore, using the same concept, we can determine the dangerous level of certain location and provide simple analysis to driver. The output can remind the driver what is the most dangerous and common accidents happened in the past, so the driver can pay more attention to avoid collisions.

5 Conclusion

After the analysis of the data, we found out that most of the conditions are not major factors to cause serious accidents, and in fact, most of the collisions happened in the normal conditions which is a bit non-intuitive. The major factor of collisions are the type and location of the collisions. That means of government would like to lower the number of collisions, it would be more efficient to put more effort on the parking car collisions problem than improving other conditions. The Seattle government can build more safe parking place and ban illegal parking have a higher chance to get accidents. Also, the government can take track the accidents black-spots. The reason of high number of collisions in those locations are not discovered in this data set, this may due to the improper design of the roads.

For future research, similar analysis can be preformed for other cities to understand the accidents problem in US. Also more data can be collected about collisions with different seriousness as the current data set has small portions of serious injury collisions. Thereby, the logistic regression can not be applied to investigate the relation of seriousness and other factors.

6 Acknowledgement

The author would like take the chance to thank IBM and Coursera for offering high quality data science education.