Data analysis on Seattle traffic collisions from 2004 to 2020

Chan Ching Kwan

14 Sep, 2020

Traffic collisions as society loss

- Starting from 2004 to May 2020, total collisions in Seattle: 194673 cases

  which means more than 32 collisions per day
- Collisions
  1. Losing life (Human capital loss)
  2. Damage on property (Economic loss)
  3. Traffic jam (Society time loss)
  4. Citizen felt insecure (lost in quality of life)
- Loses covered by individual or mainly insurance company and government

Conclusion: Need to reduce traffic collisions effectively

Objectives

Investigation the property of different collisions

1. Trend
2. Types
3. Causes
4. Correlation

Build predict model for future development

Suggestion to reduce collisions

Data

- Data from Seattle Police Department, recorded from 01-01-2004 to 30-05-2020
- Total 194673 rows with 38 types of information
- Meta data file can be found:

https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf

Data Removal

- Rows with empty data

  184146 rows remain
- Repeated or unexplained information

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

After removal: the size of data

184146 rows x 17 columns

Data preprocessing

1

Changing string data into numeric data

- For easier data analysis
- Include weather, road condition or other bi-category data

2

Grouping small data into other category

Visualization

Summary :

Most of the collisions happened in **Normal** conditions

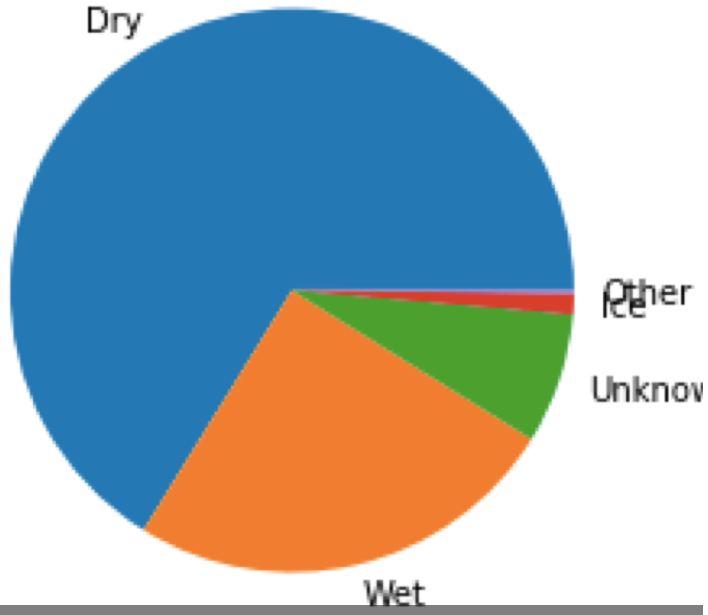


Figure 2: Pie chart for road

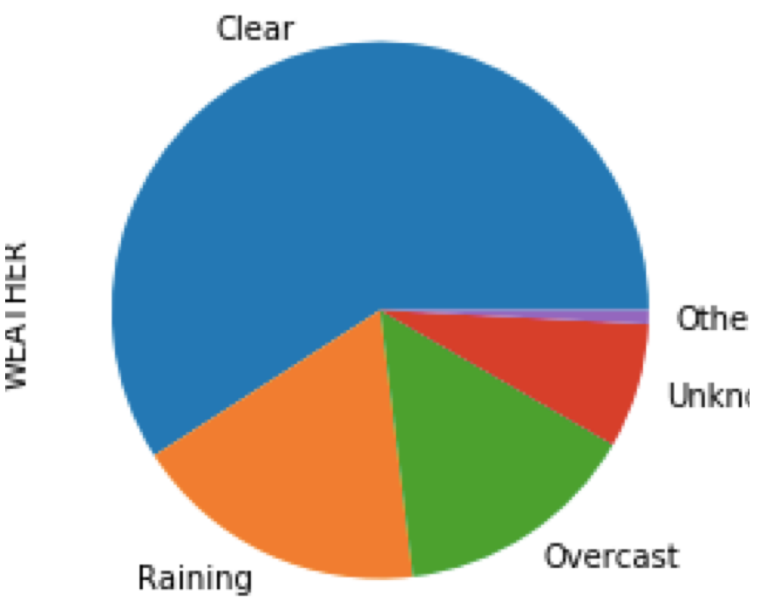


Figure 1: Pie chart for weather

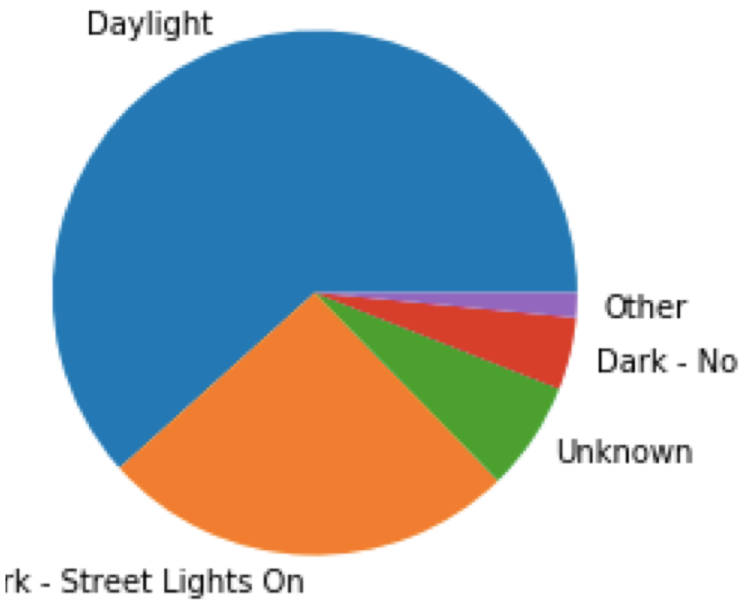


Figure 1: Pie chart for light

Monthly distribution visualization

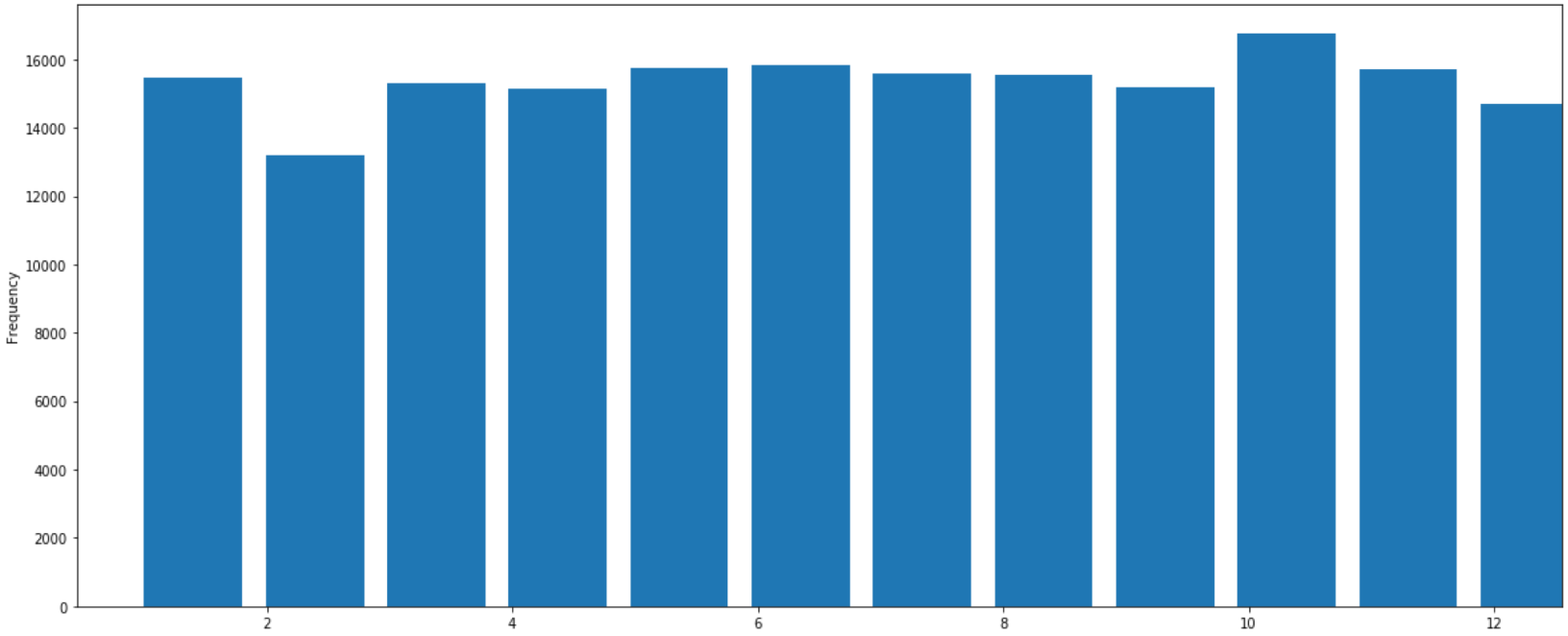


Figure 4: monthly distribution of collisions.

The distribution is very even which means the seasonal effect is weak on collisions.

Year trend

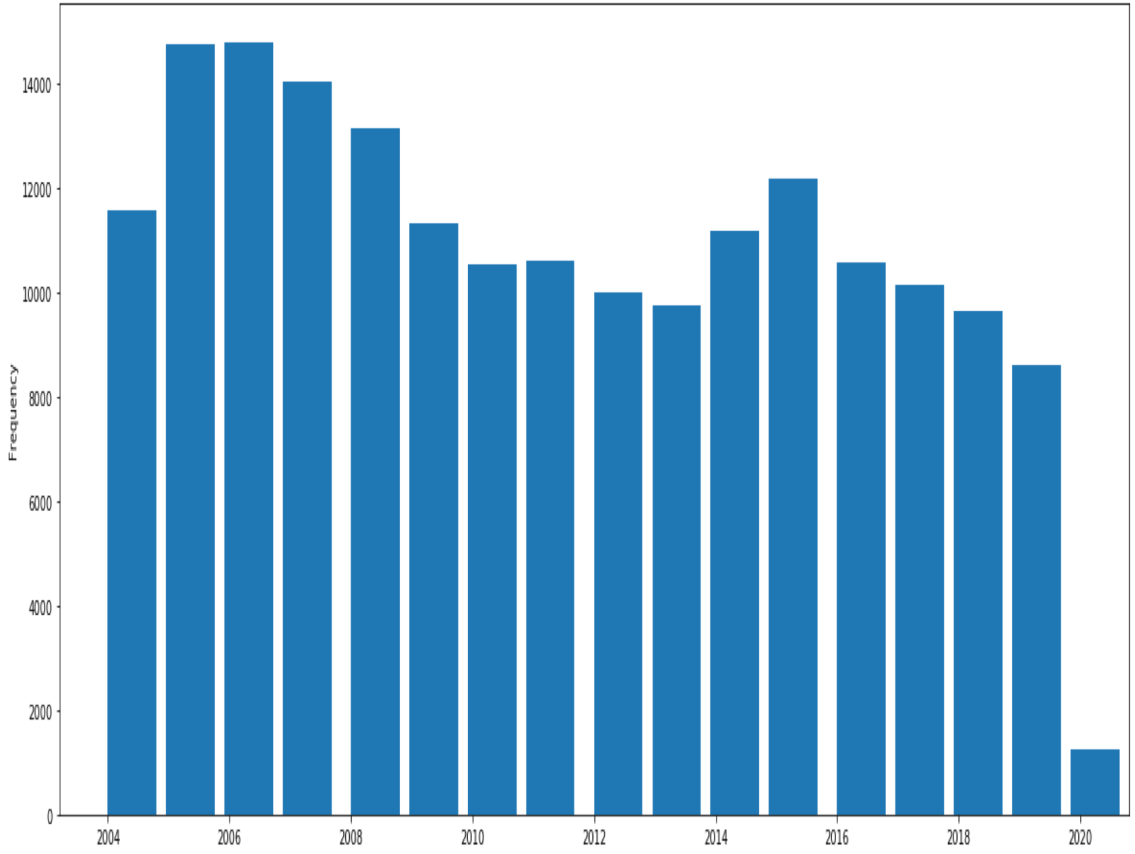


Figure 5: The histogram of yearly distribution.
As shown as above, the number of 2020 is exceptionally small because of the incomplete record. Therefore it is remove from further analysis

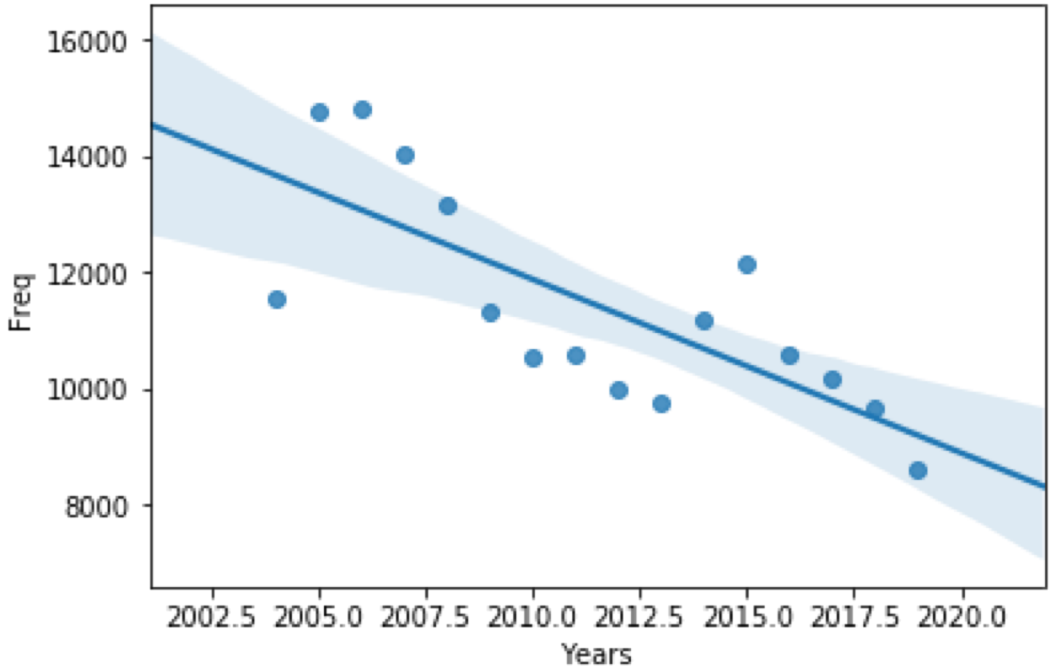


Figure 6: The linear regression result of year distribution. The blue linear show the regression line and the shaped area represent 68.3% confident zone. The number of collisions are decaying. The R^2 coefficient is 0.568.

Type of collisions

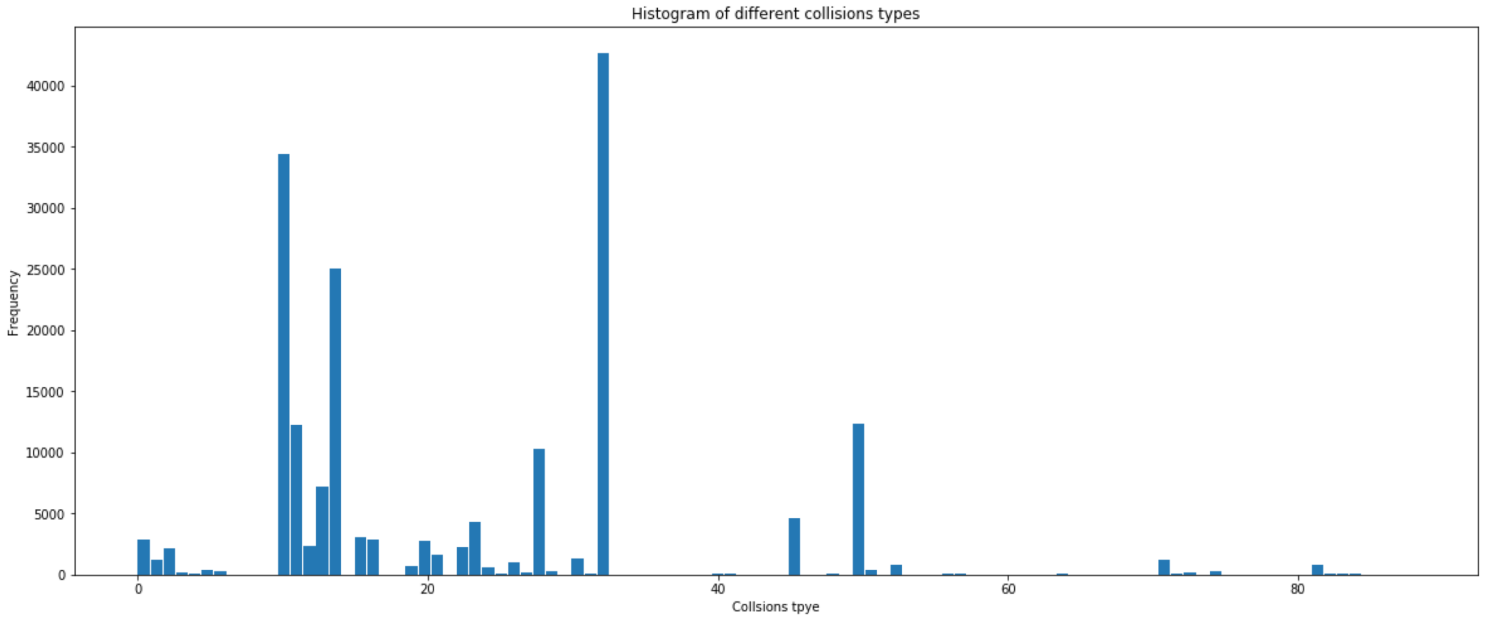


Figure 7: The histogram of the type of collisions
The top 3 types of collisions dominate over 50% of total number.

The top 5 types of collisions:

32: One Parked - One Moving(23.2%)

10: Entering At Angle(18.6%)

14: From Same Direction - Both Going Straight - One Stopped - Rear End (13.6%)

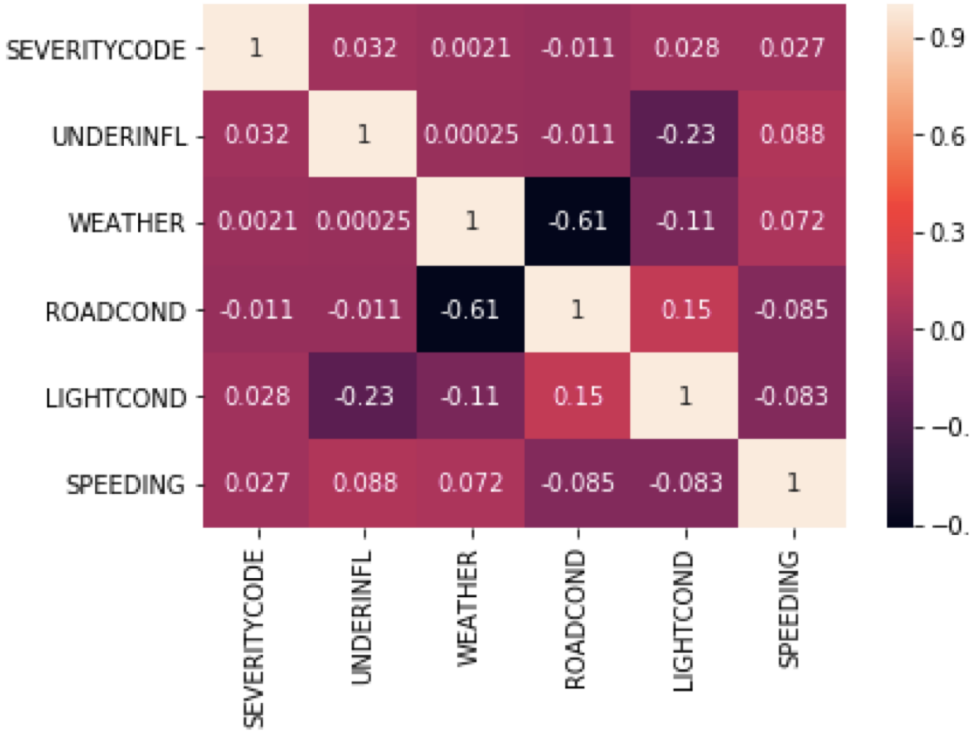
50: Struck Fixed Object(6.7%)

11: From Same Direction -Both Going Straight-Both Moving- Sideswipe(5.6%)

Summary: Collisions are dominated by only 3 types, with over 50%

Correlation matrix

Figure 8: The heat map is the correlation matrix of different factors. The first row is seriousness against other factors but none of them show high correlation value, which means they are not main factors

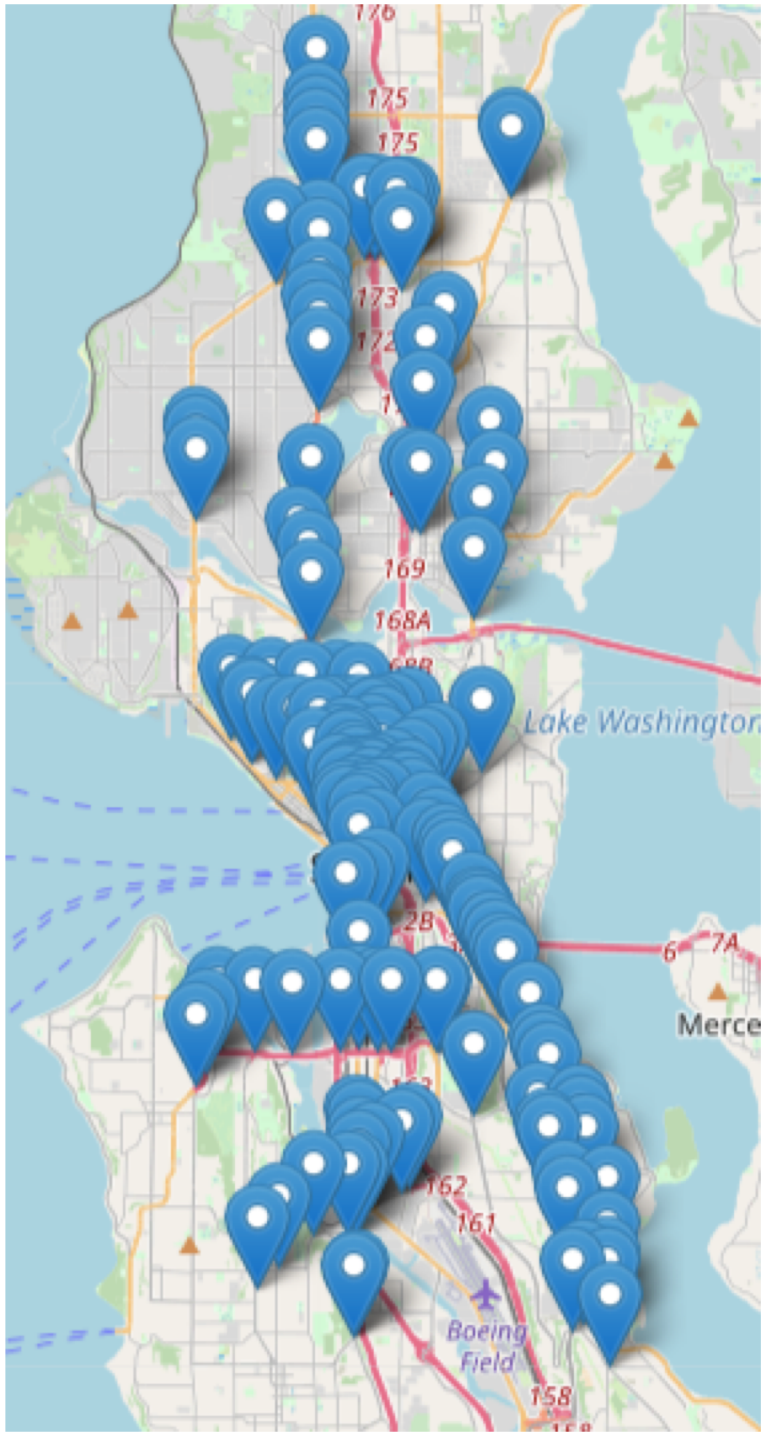


Accidents black-spot

The folium map located in Seattle shows the traffic black-spot using the density-based clustering method.
Each spotted locations represent a dense group of locations.

- Dense group: more than 100 collisions happened within the 200 m radius.
- Total 278 groups were found.
- Black-spots can form another network if the constraints are relaxed.

Summary: Locations is a important factor of collisions
Special care should be made on those spots



Dangerous level prediction

Using similar concept of black-spot, dangerous level prediction algorithm was developed.
Input: location coordinate
Output: number of collisions, dangerous level and most common type of collisions

```
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
```

Two different locations are input
The first one gave dangerous warning with 50(Struck fixed obejct). The second one show no collisions happened before nearby

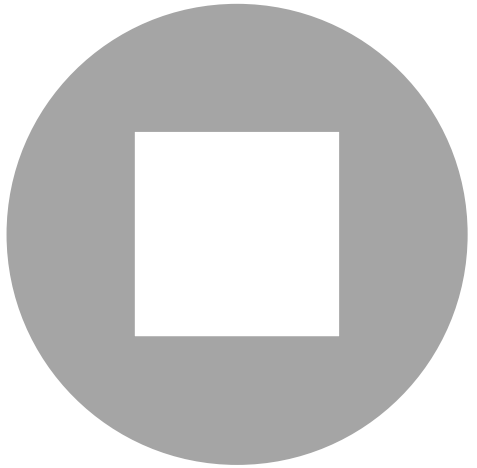
Conclusion

1. The seriousness of collisions are independent of other factors

2. The number of collisions are dominated by few types.
   - Policy focusing on these types can be used to reduce collisions effectively

3. Traffic Black-spots are located
   - Updated road design should be considered in the future

4. Dangerous level prediction was developed
   - More feature can be implemented depend on the need
   - Using Google Map API, app can be developed for application



SPECIAL THANKS TO IBM
AND COURSERA

END