# Bike Lane Data Report

**Project:** Bike Lane Blocks (DS）

**Team Member:** Mingdao Che

Yucheng Zhu

Zhengyang Tang

### Introduction:

This project is designed for analyzing the bike blockages over the years. According to the dataset given by Boston City and other supporters, we mainly focus on: the reasons that lead to bike blocks, block changes over years, the correlation between bike lane block and crashes and bike rider counts with bike lane blocks.

To solve these problems, we mainly used Boston and Cambridge's 311 data, Boston's bike crash data and Boston's bike count data. We also use Google trend to analyse people's search of bike lane closures in Boston. The following sections show the procedure and method we used to solve these problems.

### 311 Request Service Data:

In Boston and Cambridge's 311 data, each row is a reported case corresponding to public utilities. The hard part is, both cities' 311 data are pretty rough and have no accurate/enough bike lane block labels for us to analyse or train a model.

To extract these useful data from all data related with the bike lanes, we first combined the two cities' data. And we kept searching "bike lane", finding these blockage report mainly belong to "Parking Enforcement", "Request for Snow Plowing", "Request for Street cleaning" etc. We also considered the "repaint bike lanes" as a kind of blockage. There will be pieces of data related to the bike block on next page.

Once we determined which rows are reports for bike blockage, we can start finding relations between bike lane blockage and bike lane crashes, the number of bike riders, seasons, nearby businesses and neighborhoods. Useful columns for the analysis are **open_dt**(the time when bike lane blockages occur), **ontime**(whether bike lane blockages are handled on time), **case_title**(case details), **closure_reason**(case details), **type**(case details), **neighborhood**, **latitude**, **longitude**.

With **open_dt** associated with **dispatch_ts** and **mode_type** in the bike crash dataset, we can determine the relationships between bike lane blockage and bike crashes; with **latitude** and **longitude** associated with Google Map API, we can determine the relationship between bike lane blockage and nearby businesses; with **open_dt,** we can determine the relationship between bike lane blockage and seasons; with **open_dt** associated with the bike count dataset, we can determine the relationships between bike lane blockage and bike riders; finally, with **open_dt** and **neighborhood,** we can determine the relationship between bike lane blockage and neighborhoods.

### Problem 1: What businesses have the highest correlation with complaints about bike lane blockages on 311?

By classifying Boston and Cambridge's 311 data, we found the highest reason of bike lane blockage is **illegal parking** and the method to find the highest reason will be interpreted in problem 3.

To analyzing the highest correlation between illegal parking and businesses, we get the businesses near the illegal parking places  by Google Map API with the coordination of those places. As shown in the Figure.1 below, **health care**(hospitals, pharmacies, etc) and **food**(restaurant, cafe, etc) businesses have the strongest correlation with illegal parking.

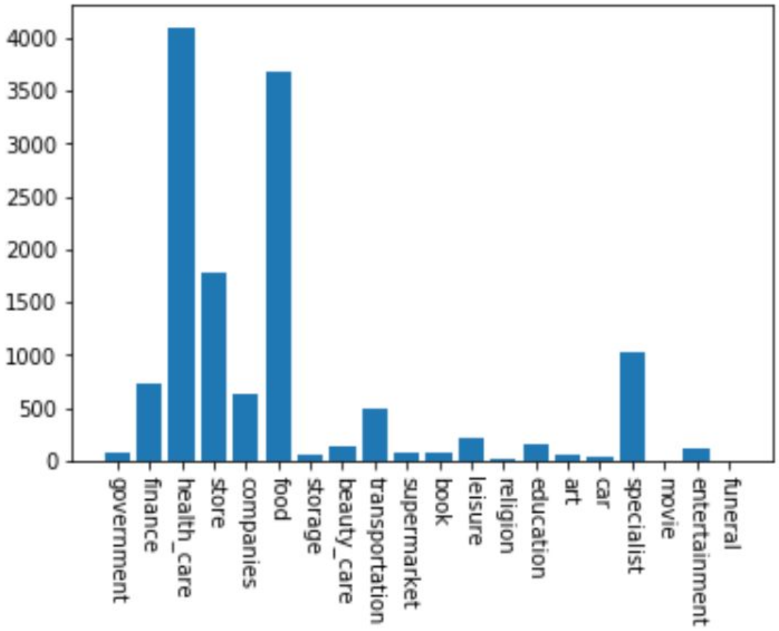


**Figure.1 The correlation between businesses and illegal parking**

**Problem 2: Which neighborhoods in Boston have the highest number of 311 calls/tickets related to bike lane/parking blockages?**

Back Bay has the highest number of 311 calls related to bike lane blockages.

To get the result, we first extracted rows for bike blockage from the 311 calls dataset, and put the corresponding data in the “neighborhood” column in a dictionary, where keys are the neighborhoods’ names and values are how many times this neighborhood appeared in bike blockage data. At last, we got the dictionary like this:

{'Boston': 72, 'Back Bay': 110, 'Jamaica Plain': 23, 'Roslindale': 5, 'Allston / Brighton': 43, 'Dorchester': 14, 'Hyde Park': 5, 'Fenway / Kenmore / Audubon Circle / Longwood': 63, 'South Boston / South Boston Waterfront': 33, 'Charlestown': 6, 'West Roxbury': 1, 'Roxbury': 30, 'South End': 35, 'Greater Mattapan': 5, 'Downtown / Financial District': 42, 'East Boston': 6, 'Mission Hill': 4, 'Beacon Hill': 1, 'South Boston': 1} and a plot for this:

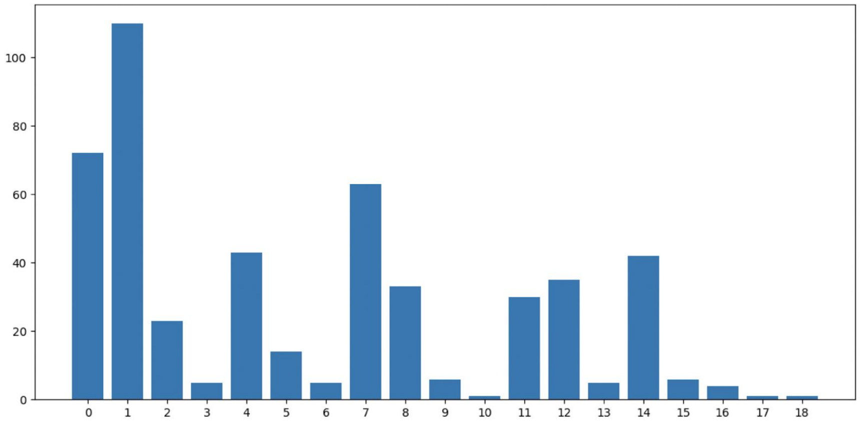


**Figure.2 The calls of neighborhoods**

where X coordinates are the index of the dictionary keys and Y coordinates are how many times this neighborhood appeared in bike blockage data.

Thusly we got the conclusion that Back Bay has the highest number of 311 calls related to bike lane blockages.

**Problem 3: What is the breakdown of reasons cited for bike lane blockages? e.g. construction, double-parked vehicles, etc.**

When we tried to find the specific reason for every block issue, we first used Boston’s 311 data. And the difficult part is the data is rough, basically has all types of complaints, not just bike lane block. Also the bike lane we got after filtered by keywords is so little, about 300.

Then we used Cambridge's 311 data, Cambridge, MA, which is already filtered and all the data belong to bike lane block, and there are 1342 messages in total, so it's easier to classify the reasons. And we combined the two 311 data together and tried to breakdown the block reasons based on the column '**closure_reason**'.

We use 8 classes to define the reasons: Snow, Repair, Garbage, Uber/Lyft, Tree/Branch, illegal parking, Blank and Others. For the first 6 classes, we set some keywords for them separately, for example, we set a list: ['car', 'vehicle', 'park', 'truck', 'SUV', 'bus', 'shuttle', 'driver', 'delivery', 'taxi'] to extract "illegal parking ".

"Blank" here means there are 226 rows' description are blank. "Others" here means the data we cannot tell the class based on the description: 234 rows' description don't have the keywords we set for other classes.

Following is the histogram for the classes.

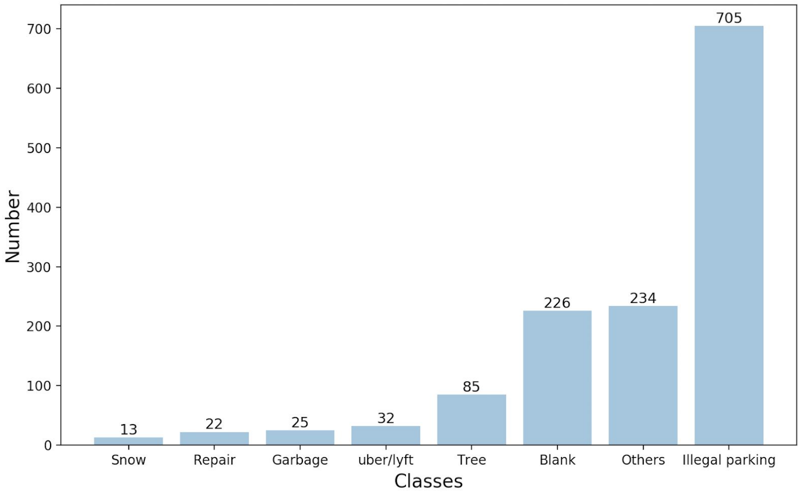


**Figure.3 Bike Lane Blockage Classes**

This method is not accurate enough, but it was the best way we can find now. We also tried some NLP library functions like Text-Grocery and TfidfVectorizer, which will predict the labels based on the frequency of each word, but the result is not good. The reasons are:

1. We cannot give accurate labels for these model (Manual classification is tiring and not encouraged in this course).
2. The total size of bike blockage data (1861) is too small to train an accurate model.

**Problem 4: Have 311 calls related to this issue changed over the past 5 years? Seasonally?**

There have been some changes in 311 calls related to this issue over the past few years. As you can see in Figure. 4 shown below, the number of bike lane blockage remains relatively low through 2011-2015, and suddenly start increasing from 2015 to 2018, and the decreased a little in 2019, but mostly because the data for 2019 is not complete since this year is not finished yet, so bike lane issue has been increasing fast these years, which is shown in Figure. 4,

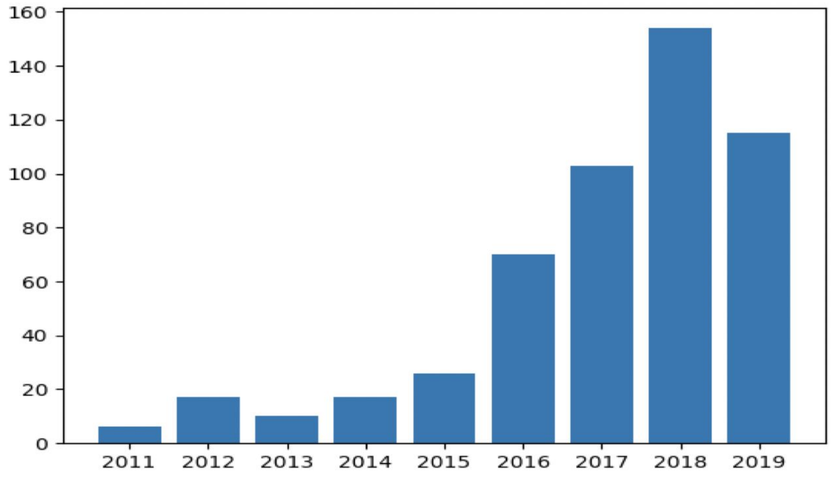


**Figure.4 Bike Lane Issues from 2011 to 2019**

We also counted the ontime-solving rate (number of issues solved on time/number of issues) of bike lane blockage, finding that although the number of issues keeps increasing, the percentage of issues solved on time is not decreasing, in fact, it even increased a little bit.

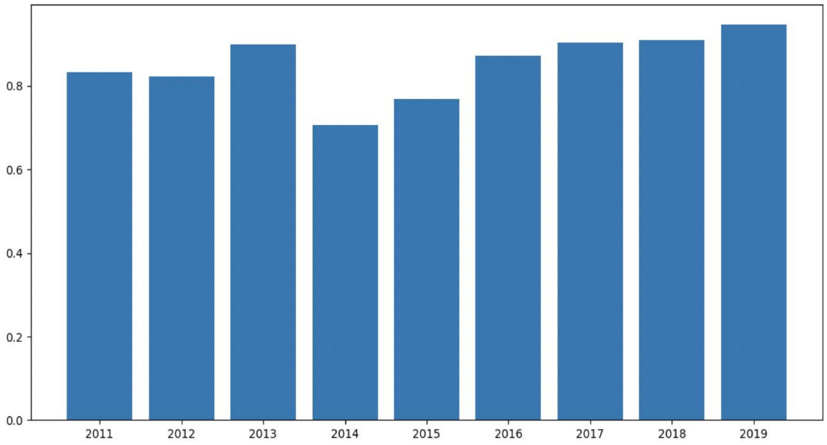


**Figure.5 Time-solving Rate from 2011 to 2019**

We also tried to find the relations of issues and months (or seasons), and as figure 6 shows, it turns out that fewer blockages happen in winter and early spring (November to February), and mostly happens at the mid of summer (June), but the ontime-solving rate doesn’t seem to be different through every month.

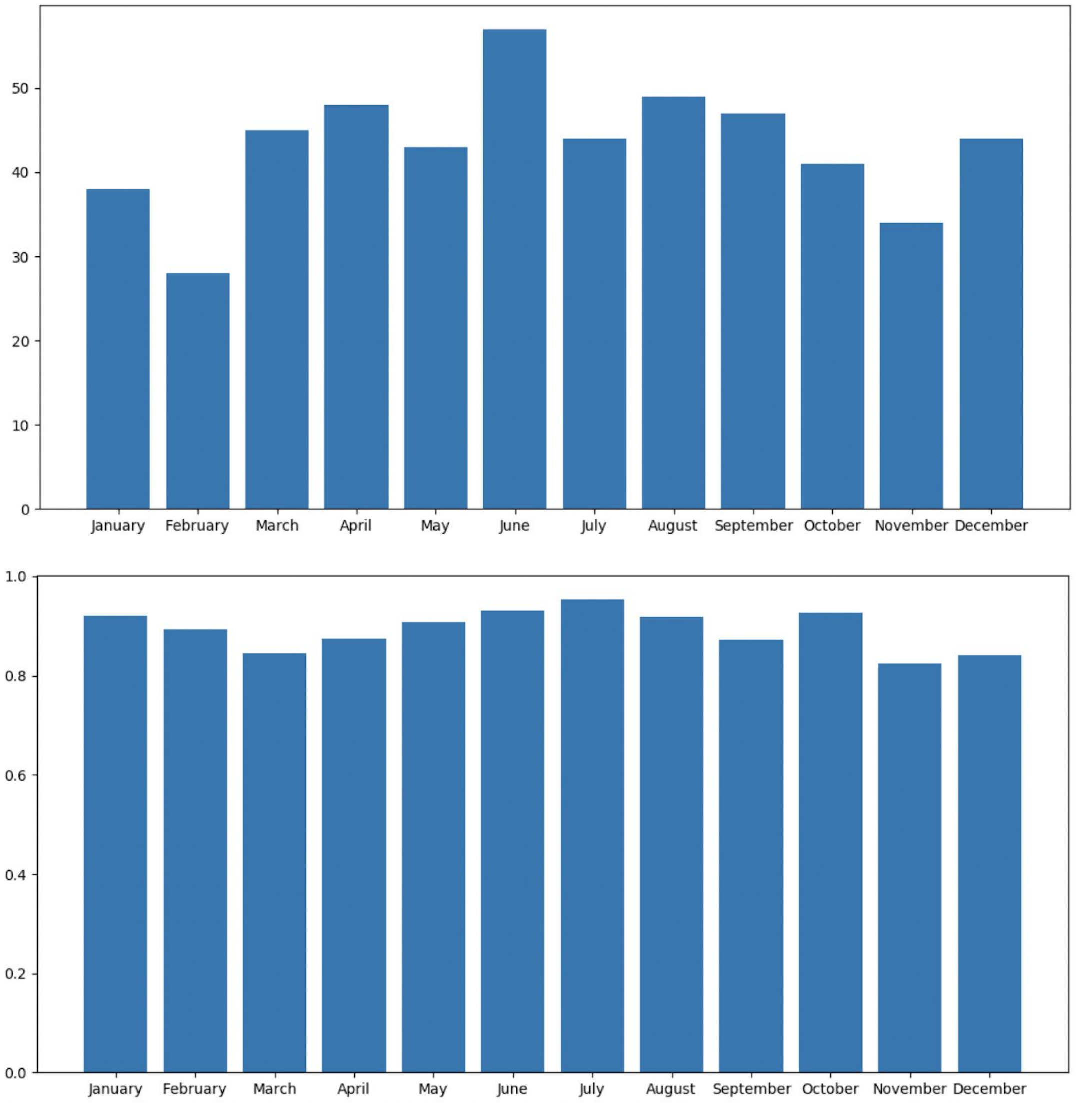


**Figure.6 Ontime-solving Rate Seasonally Changes**

To get these results, we first extracted rows for bike blockage in 311 data, and split the “open_dt” column into year and month, then threw them into two dictionaries. The year dictionary has keys of years and values of the number of issues that happened and ontime-solving rate in that year, and the month dictionary has keys of months and values of the number of issues that happened and ontime-solving rate in that month.

## Bike Crash Data:

The data posted by the Boston government is with some common information about crashes that happened from 2015 to 2019. This dataset contains some columns named **mode_type**, **dispatch_ts, location_type, street, xstreet1, xstreet2, x_cord, y_cord, lat and long.**

For this project, **dispatch_ts, mode_type, lat and long** provide our team with basic information about the crash. The **dispatch_ts** is the dispatch time of the police which can be considered as the time when the crash happened. The **mode_type** is the type of crash and 'bike' means a bike crash. The **location_type** means whether this crash happened in Intersection or somewhere else. The **lat and long** offer geographical information about the crash. The **street, xstreet1, xstreet2, x_cord, and y_cord** are not that useful which may not be included in this project.

When obtaining **dispatch_ts, mode_type, lat and long,** we can establish the correlation between the crashes and bike lane blockages over the years.

The 311 Boston Data is from 2011 till now, but the crash data is from 2015 till now. So our team designed to utilize the data beginning on 01/01/2015 to analyze the correlation between bike blockages and crashes. As the first concern, we visualized the number of crashes by years. The trend is shown in Figure. 7,

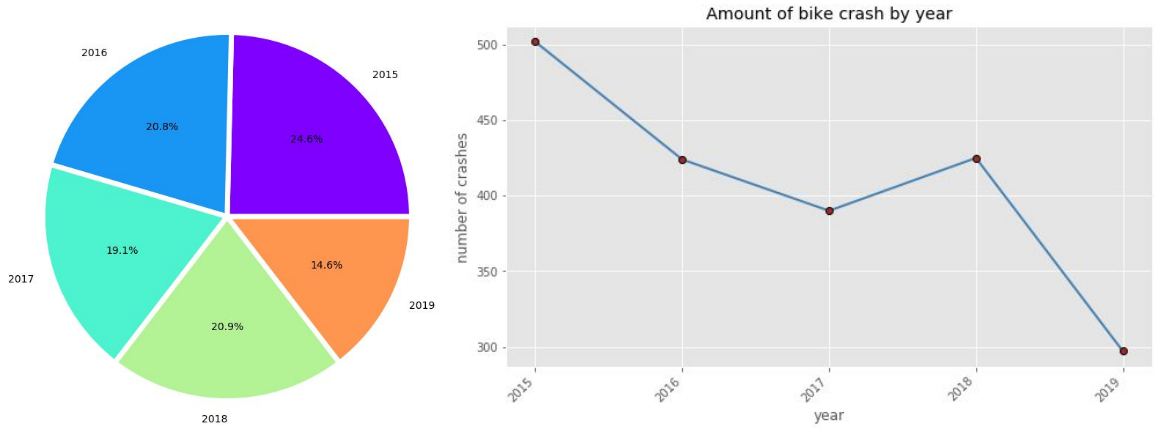


**Figure.7 Percentage and Trend of crashes over years**

According to the Figure., the highest number of crashes is in 2015 and the lowest is in 2019, and we can find there exists a decrease in crashes in Boston through the five years via the line chart, which is absolutely a good situation for the residents in Boston.

To observe it more clearly, We concentrated on the 2015 crash data to plot the data on Google Map in Figure. 8,

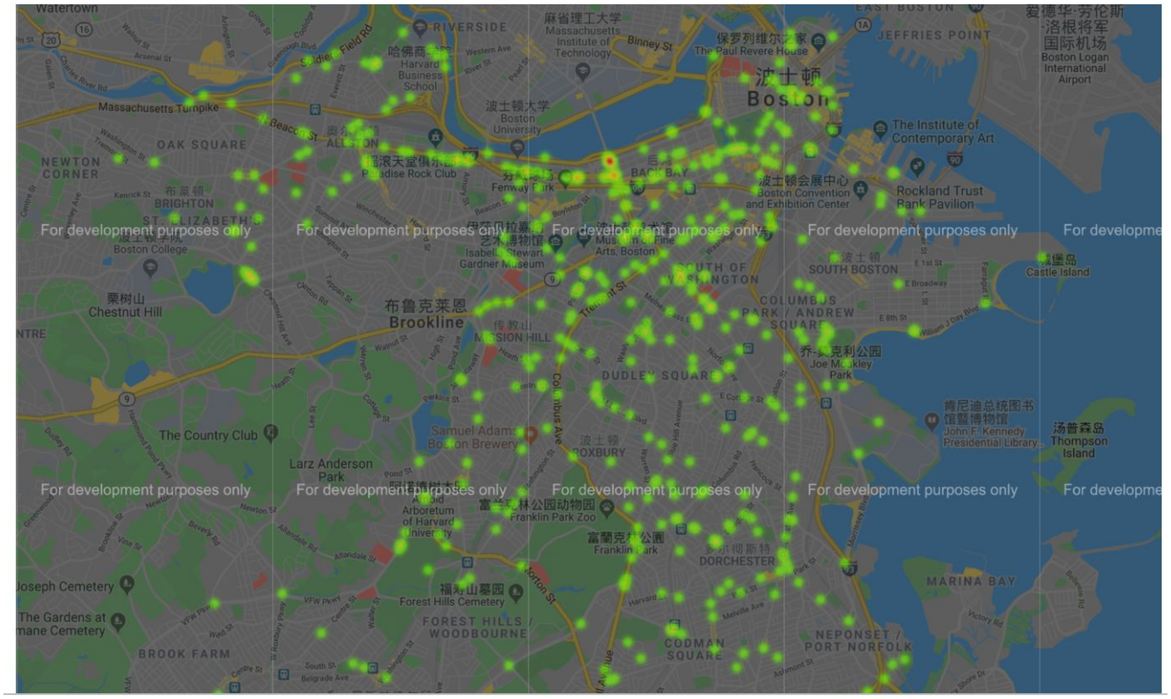


**Figure.8 Heatmap of bike crashes of Boston in 2015**

The main bike lane blockages are surrounding the commonwealth.
And we also analyze the HSIP crash data. Here in Figure.9 is the heatmap of clusters of HSIP crash data, which encourages us to get access to its geo API to get its coordinations. After primary recognization of the data then we moved to establish the correlation between the crashes and blocks.

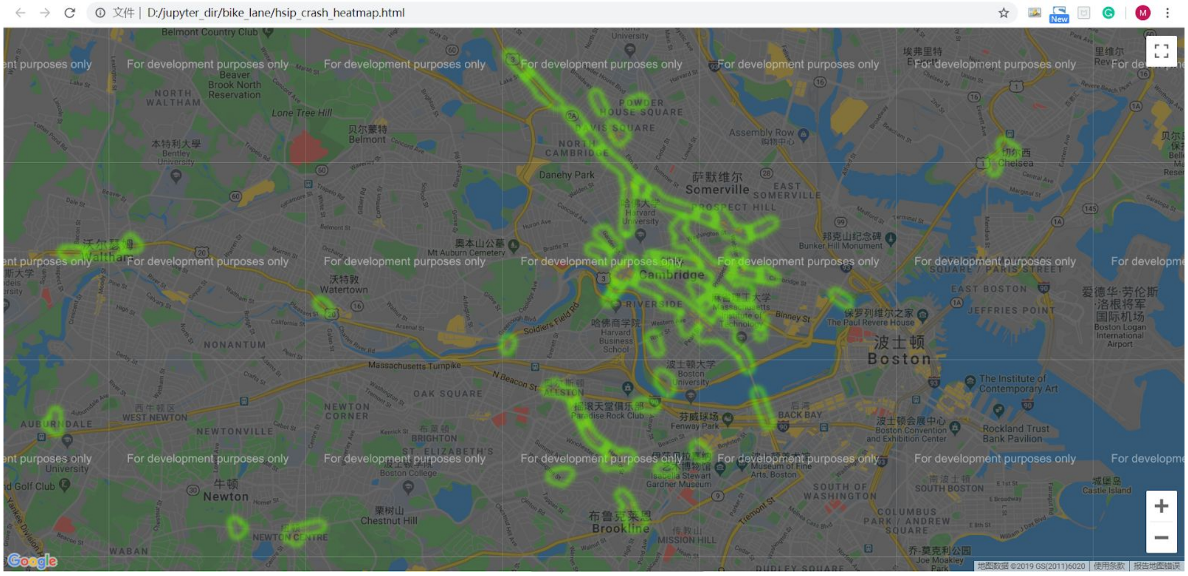


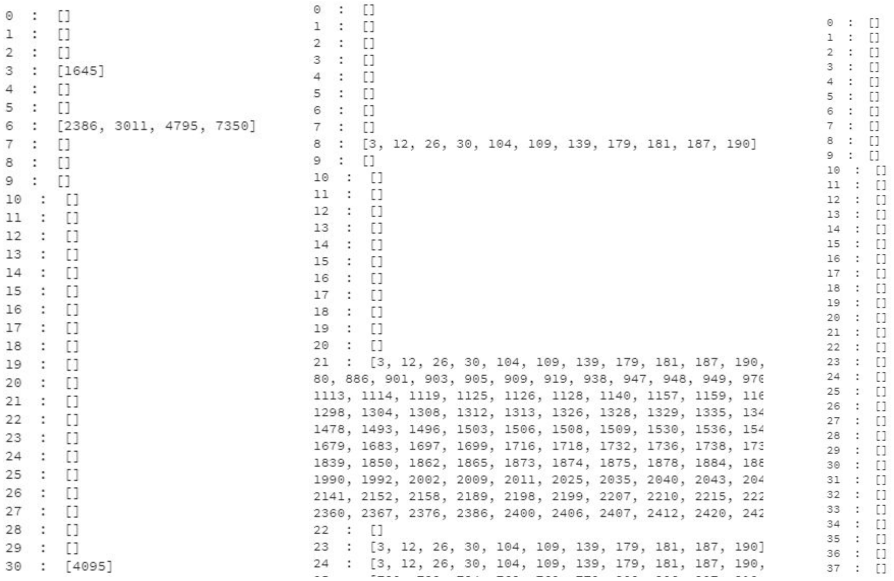
**Figure.9 Heatmap of HSIP crash data**

**Problem 5: What is the correlation of bike lane closures to reported bike crashes?**

The crash dataset contains 20590 records of the crashes that happened in Boston from 2015 till now and 2038 of them are related to bicycles. By computing the date and coordination, we can figure out the relationship between the blocks and crashes. One thing that needs to notice is that the date in the block dataset is with the time zone of  US Eastern, but the date in crashes is based on none time zone. As a result, our team needs to set the time zone to None when processing the data.

However, when we implemented the code for this question, the obstacle which stopped our steps is the size of blocks in Boston is too limited to build the relationship between blocks and crashes. We tried to match the crashes and blocks according to coordinates and date, then calculate the intersection of them.

The method is, we use a dictionary whose keys are the index of the blocks and each key has a list of crashes' indexes, but the list we got is empty when respected to both location and date. The problem is shown in  Figure. 10.



```
0  :  []
1  :  []
2  :  []
3  :  [1645]
4  :  []
5  :  []
6  :  [2386, 3011, 4795, 7350]
7  :  []
8  :  []
9  :  []
10  :  []
11  :  []
12  :  []
13  :  []
14  :  []
15  :  []
16  :  []
17  :  []
18  :  []
19  :  []
20  :  []
21  :  []
22  :  []
23  :  []
24  :  []
25  :  []
26  :  []
27  :  []
28  :  []
29  :  []
30  :  [4095]
```

```
0  :  []
1  :  []
2  :  []
3  :  []
4  :  []
5  :  []
6  :  []
7  :  []
8  :  [3, 12, 26, 30, 104, 109, 139, 179, 181, 187, 190]
9  :  []
10  :  []
11  :  []
12  :  []
13  :  []
14  :  []
15  :  []
16  :  []
17  :  []
18  :  []
19  :  []
20  :  []
21  :  [3, 12, 26, 30, 104, 109, 139, 179, 181, 187, 190,
80, 886, 901, 903, 905, 909, 919, 938, 947, 948, 949, 97
1113, 1114, 1119, 1125, 1126, 1128, 1140, 1157, 1159, 11
1298, 1304, 1308, 1312, 1313, 1326, 1328, 1329, 1335, 13
1478, 1493, 1496, 1503, 1506, 1508, 1509, 1530, 1536, 15
1679, 1683, 1697, 1699, 1716, 1718, 1732, 1736, 1738, 17
1839, 1850, 1862, 1865, 1873, 1874, 1875, 1878, 1884, 18
1990, 1992, 2002, 2009, 2011, 2025, 2035, 2040, 2043, 20
2141, 2152, 2158, 2189, 2198, 2199, 2207, 2210, 2215, 22
2360, 2367, 2376, 2386, 2400, 2406, 2407, 2412, 2420, 24
22  :  []
23  :  [3, 12, 26, 30, 104, 109, 139, 179, 181, 187, 190]
24  :  [3, 12, 26, 30, 104, 109, 139, 179, 181, 187, 190,
```

```
0  :  []
1  :  []
2  :  []
3  :  []
4  :  []
5  :  []
6  :  []
7  :  []
8  :  []
9  :  []
10  :  []
11  :  []
12  :  []
13  :  []
14  :  []
15  :  []
16  :  []
17  :  []
18  :  []
19  :  []
20  :  []
21  :  []
22  :  []
23  :  []
24  :  []
25  :  []
26  :  []
27  :  []
28  :  []
29  :  []
30  :  []
31  :  []
32  :  []
33  :  []
34  :  []
35  :  []
36  :  []
37  :  []
```

**Figure.10  Dict 1 with location            Dict two with date            Dict 3 with both**

According to the Figure above, we can match several blocks and crashes based on coordinations or dates, but we did not obtain a good result when respect to coordinates and dates. Even though we assumed the same location means that location +- 0.0001 of their coordinate, we can not find the obvious correlation between them based on the 311 data.

## Google Trend :

### Problem 6: How have google searches of bike lane closures in Boston changed over the years?

One mainly usage of Google Trend is to show the changes in Google searches over the years. As there do not exist enough “bike lane block” or “bike lane closure” searches to plot. We scaled the searches plot to “Bike Lane” in Figure.11.

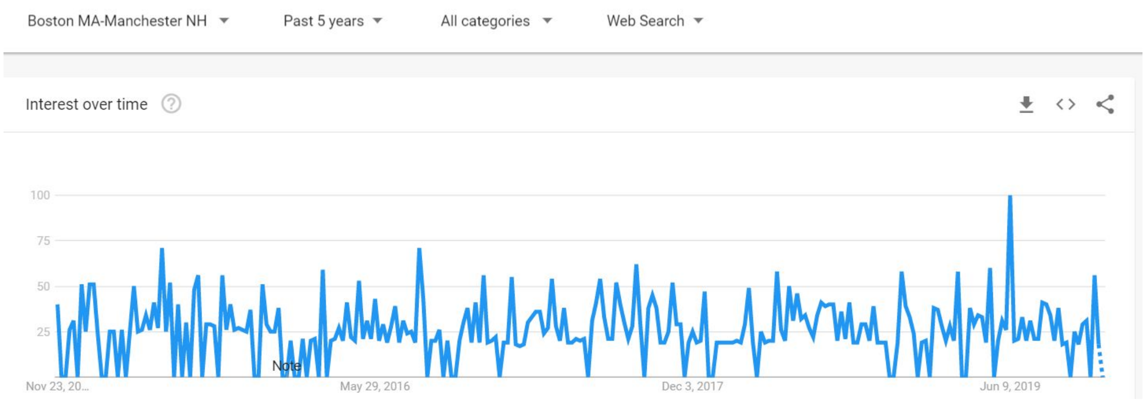


**Figure.11 Bike Lane Searches over the past 5 Years**

Then we download the CSV file for bike lane searches, from which we calculate the overall search number for every year and convert them into a dictionary: {'2014': 89, '2015': 1513, '2016': 1513, '2017': 1656, '2018': 1165, '2019': 1490}. Since we just have December data for the year 2014, we can ignore the data for 2014. Therefore, we can see that in 2018 we got the lowest number of searching, and search numbers didn’t vary much in other years.
We can also see that there are some 0 points in figure 9. 0 points represent weeks whose data are not recorded, and we also have a dictionary for 0 points in each year: {'2014': 2, '2015': 14, '2016': 9, '2017': 9, '2018': 18, '2019': 9}. We ignore the year 2014 for the same reason, then we can see 2018 has the largest number of 0 points. Maybe that’s part of the reason why 2018 got the smallest number of searching.

## Bike Count Data:

This dataset provides very detailed information about bike counts. Our team can easily find out the number of bicycles in a specific street or neighborhood.  Corresponding to the 311 location data, it provides an opportunity for our team to build a direct correlation between the number of bike riders + 311 calls for lane closures.

### Problem 7: Is there a direct correlation between the number of bike riders + 311 calls for lane closures?

Before deeply mining the data in the Bike count dataset, our team assumed the days which were observed in the bike count data could be regarded as a representative of the counts in these locations annually.

The initial work is to obtain the coordination in Google as there only contains the street names in the dataset. Our team chose the coordinate by its instruction in the dataset and use them to match the blocks. When respect to the estimation error, we slightly scaled the matching condition from 0.0001 to 0.001 based on the coordinations.

By matching coordinations, there hardly exists the matching data between the 311 data and the count data. The bike data of matching locations are lacking in significant patterns. The bike count of a specific matching location is shown in Figure. 12.

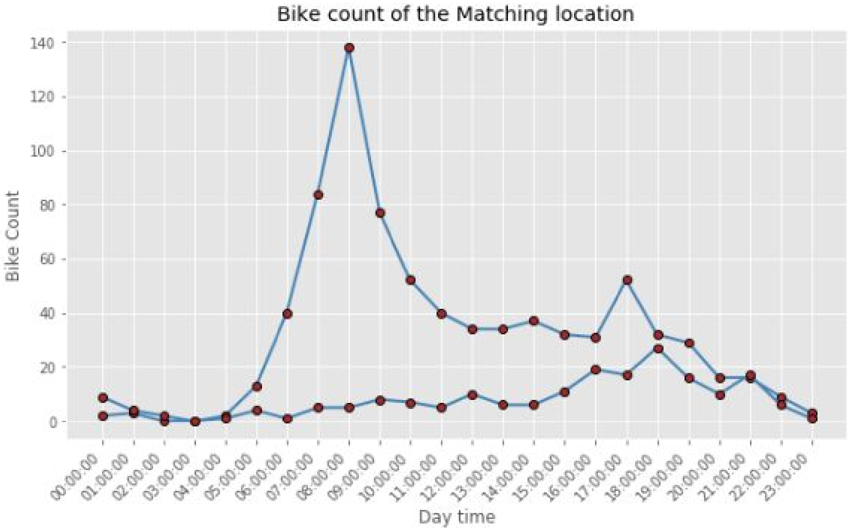


**Figure.12 Bike count Data of a specific matching place**

## Epilogue:

The whole project is very practical and challenging for our team due to the lack of useful block data and we spent a huge effort in the data processing. The result would be more accurate if we have some better labeled data to train a model. But it’s our pleasure to answer all these questions and we feel fulfilled to make it. Thanks for reading.

**Github Link:**

https://github.com/mdche001/CS506_Bike_Lane

**Data Link:**

1. 311 Service Requests
2. Bike Count Data
3. Cambridge, MA 311 Cambridge Data
4. Bike Crash Data
5. HSIP Bicycle Crash Clusters 2007 - 2016