RISKY COMMUTES FOR CYCLISTS IN BOSTON

PAUL BASTIDE

Coursera Capstone | 28 February 2019

Introduction/Business Problem

Background

The City of Boston is the largest city in Massachusetts with 800,000 inhabitants who travel through the city every day of the year - work, life, pleasure. [1] The inhabitants walk, take mass transit, drive and bicycle to their destinations. These inhabitants are often joined by metro-area commuters increasing the number of trips taken by the people in Boston to: 40,000 bike trips a day, 131,000 drive alone and 1.3 million take transit. [2][3][4]

From 2016 to 2017, the number of bike trips a day rose from 30,000 to 40,000 trips; similar increases are found in New York City. [5] The increase in the number of cyclists and other metro-area commuters has resulted in *injuries and deaths* for cyclists.[7][8]

The City of Boston provides raw data related to traffic accidents, neighborhood information and fatalities. There is an opportunity to identify incident hotspots, opportunities to improve street safety, allocate resources, and improve commuter safety.

Problem

The raw data from the City of Boston provides insights on the locations of neighborhoods, fatalities and incidents. This project identifies incident hotspots and neighborhoods and suggests less incident prone neighborhoods of travel for users.

The project aims to help Boston cyclists be better prepared to travel in the neighborhoods they pass through. Given the appropriate data, the problem is:

- What level of risk is there in commuting through the Boston neighborhoods and streets to avoid?

Audience

The target audience is cyclists who would benefit from knowing what are the best neighborhoods to commute through. The report improves the understanding a cyclists traveling in Boston has for safety, and potentially the better neighborhoods to travel through.

The report does not distinguish between the individuals who are bicycling for fun, business or on a commute.

Data

To answer the question proposed in the introduction - "What level of risk is there in commuting through the Boston neighborhoods and streets to avoid?". The analysis uses data sets from the City of Boston open data. This section describes the data sets, the formats, the data quality and the reasons they are included with the analysis.

The data sets used in the analysis are:

- Vision Zero Crash Records
  - Crash https://data.boston.gov/dataset/vision-zero-crash-records
  - Fatality https://data.boston.gov/dataset/vision-zero-fatality-records
- Boston Neighborhoods
  - https://data.boston.gov/dataset/boston-neighborhoods

Vision Zero Crash Records

The Crash Record dataset records vehicular/bike incidents which result in injury or fatality. The data is comprised of date, time, location and the type of incident. The data is from the start of 2015 to the end of 2018.

The data is in the CSV format, and contains the following fields:

- dispatch_ts text of the dispatch time
- mode_type text Type of incident (ped = pedestrian; mv = motor vehicle; bike = bicyclist)
- location_type text Location "street", "intersection", "other"
- street text Street Name
- xstreet1 text Cross-street 1
- xstreet2 text Cross-street 2
- x_cord text X coordinate
- y_cord text Y coordinate
- lat text Latitude
- long text Longitude

Example of the data

```
_id     date_time       mode_type       location_type   street  xstreet1
        xstreet2        x_cord  y_cord  long    lat
1       2015-01-21 15:07:00     ped     Street  CENTRE ST       RITCHIE ST
        LAMARTINE ST    764320.75       2942908.64      -71.1000694606
        42.3227879775
2       2015-04-09 16:00:00     ped     Intersection    None    MELNEA CASS BLVD
        MASSACHUSETTS AVE       771859.87       2946498.89      -71.0721241969
        42.33254716
3       2015-04-25 17:48:00     mv      Street  MASSACHUSETTS AVE       MAGAZINE
ST      PROCTOR ST      773210.41       2944795.5       -71.067586742
        42.3276600882
```

There are three incident types (Ped = Pedestrian, mv = motor vehicle, bike = cyclist)

```
mv      12489
ped      3131
bike     1741
```

The location type in the neighborhoods are:

```
Intersection    8698
Street          7238
Other           1425
```

There are 17361 dispatches in total from 2015 to end of 2018. These dispatches involve 7238 streets.

The analysis uses lat, long to position to crash and location_type these features are extracted to identify each incident type and location. The dispatch_ts is used to split the data into partitioned sets. The partitioned sets showing a trend over the months of data collection, and the changing environment safety.

Vision Zero Fatality Records

The Crash Fatality dataset records incidents which result in fatality. The data is comprised of date, time, location and the type of incident. The data is from the start of 2015 to the end of 2018.

The data is in the CSV format, and contains the following fields:

- dispatch_ts text of the dispatch time
- mode_type text Type of incident (ped = pedestrian; mv = motor vehicle; bike = bicyclist)
- location_type text Location "street", "intersection", "other"
- street text Street Name
- xstreet1 text Cross-street 1
- xstreet2 text Cross-street 2
- x_cord text X coordinate
- y_cord text Y coordinate
- lat text Latitude
- long text Longitude

Example of the data

```
_id     date_time       mode_type       location_type   street  xstreet1
        xstreet2        x_cord  y_cord  long    lat
2015-01-21 15:07:00,ped,Street,CENTRE ST,RITCHIE ST,LAMARTINE
ST,764320.75,2942908.64,-71.1000694606,42.3227879775
2015-04-09 16:00:00,ped,Intersection,,MELNEA CASS BLVD,MASSACHUSETTS
AVE,771859.87,2946498.89,-71.0721241969,42.33254716
```

The analysis uses lat, long to position to crash and location_type these features are extracted to identify each incident type and location. The dispatch_ts is used to split the data into partitioned sets. The partitioned sets showing a trend over the months of data collection, and the changing environmental safety.

There are three incident types (Ped = Pedestrian, mv = motor vehicle, bike = cyclist)

```
ped     38
mv      21
bike     6
```

The location type in the neighborhoods are:

```
Intersection    36
Street          29
```

Boston Neighborhoods

The Boston Neighborhoods data set is a GeoJSON definition file which includes neighborhood boundaries and name labels. This data is current as of April 28, 2017

Example of the Data showing Chinatown border in Boston


```
{"type":"Feature","properties":{"OBJECTID":33,"Name":"Chinatown","Acres":76.3
2440999,"Neighborhood_ID":"26","SqMiles":0.12,"ShapeSTArea":3324678.018460806
5,"ShapeSTLength":9736.590412617801},"geometry":{"type":"Polygon","coordinate
s":[[[-71.0579055147603,42.35237863170756],[-
71.05810830329557,42.35237200984217],[-71.05840144237023,42.352397321360491]]}
```


From the GeoJSON, the neighborhood name *Name* and the shape of the data are used to enable a visualization of the neighborhoods with the most risk. The coordinates in the *geometery* of the data establishes a boundary and aggregate within the boundaries the total incidents in each of the neighborhood. The aggregation visualizes the trends within the neighborhoods.

Total Neighborhoods

```
from pandas.io.json import json_normalize
df_neighborhood = pd.read_json(file_boston_geo)
df_neighborhood_norm = json_normalize(df_neighborhood['features'])
df_neighborhood_norm['properties.Name'].unique()
```

There are 26 unique neighborhoods in Boston.

```
array(['Roslindale', 'Jamaica Plain', 'Mission Hill', 'Longwood',
       'Bay Village', 'Leather District', 'Chinatown', 'North End',
       'Roxbury', 'South End', 'Back Bay', 'East Boston',
'Charlestown',
       'West End', 'Beacon Hill', 'Downtown', 'Fenway', 'Brighton',
       'West Roxbury', 'Hyde Park', 'Mattapan', 'Dorchester',
       'South Boston Waterfront', 'South Boston', 'Allston',
       'Harbor Islands'], dtype=object)
```

From this date, the venues array is extracted and counted to indicate the number of FourSquare venues in the location. The venues are indicative of a more public nature using the absence of the isFuzzed value from the results in order to calculate the degree of commercial venues. The *lat*/*lng* are used to position the venue within a neighborhood, and the id is used to keep each query unique.

The data is to be joined on latitude and longitude in order to correlate data and answer the question under study.

Methodology

The crash data is augmented with inferential neighborhood information and limits the dataframe to crash data only. The latitude is used to position a point within a specific polygon, if it is contained within the polygon, it belongs to the neighborhood. If the point is on the border and thus in two possible neighborhoods, the code assigns the last neighborhood processed.

```
df_crash_bike = df_crash.loc[df_crash['mode_type'] == 'bike']
df_crash_bike['neighborhood'] = numpy.nan
# Creates a new column and then check each incident's column
for name, geo in zip(df_geo['Name'], df_geo['geometry']):
    col_name = name
    for index, row in df_crash_bike.iterrows():
        lat_inc = row['lat']
        long_inc = row['long']

        incident = Point(long_inc, lat_inc)
        if(geo.contains(incident)):
            df_crash_bike.loc[index,'neighborhood'] = name
```

Each row is formatted like this:

	dispatch_ts	mode_type	location_type	street	xstreet1	xstreet2	x_cord	y_cord	lat	long	bad	neighborhood
1	2015-01-01 18:23:57	bike	Intersection	NaN	OLNEY ST	INWOOD ST	772710.48	2936614.62	42.305413	-71.069163	1	Dorchester
2	2015-01-02 22:27:44	bike	Street	KENSINGTON ST	DEAD END	ELMORE ST	766791.85	2942087.73	42.320627	-71.091100	0	Roxbury

The traffic incidents are shown on the map with all locations of incidents, the locations on the Islands and the Airport show that there is no issue in the parts where there is no logical street traffic.

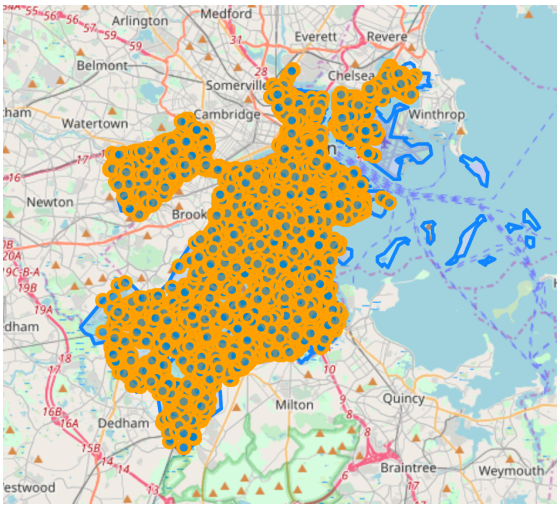


Figure 1 Traffic Incidents in Boston

The incidents are filtered on the mode to 'bike' and plotted. Interesting, there are no incidents on the islands of Boston (no Bike paths) and none at the airport. Filtering the traffic incidents, one sees the primary cyclist routes evolve on the map.

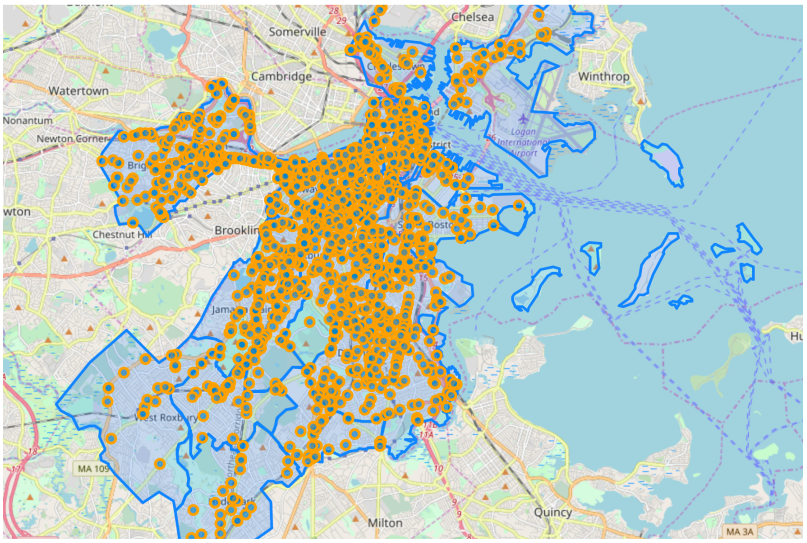


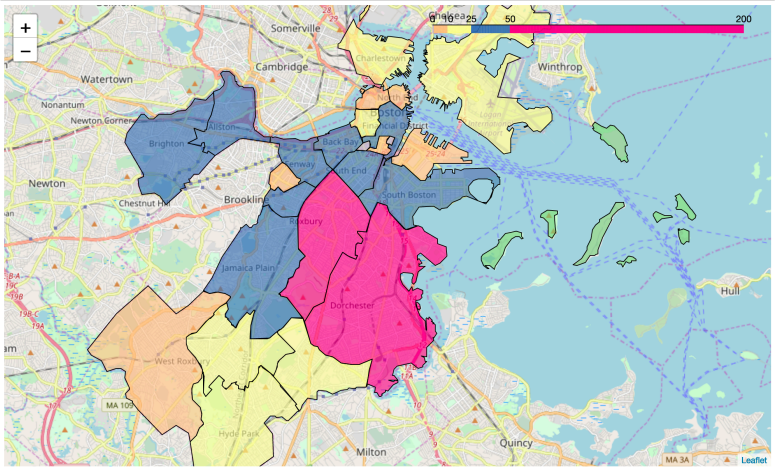
Figure 2 Cyclist Traffic

Incidents in the areas which are specific to cyclists were calculated using value counts converted to a data frame.

```
# Calculate the number of crashes involving bikes and in which
neighborhood these occured.
df_value_counts = df_crash_bike['neighborhood'].value_counts(sort=True)
df_value_counts = df_value_counts.reset_index()
df_value_counts.columns = ['neighborhood', 'count']
df_value_counts.head(30)
```

```
Dorchester                 280
Roxbury                    204
Jamaica Plain              139
Allston                    136
Back Bay                   131
Fenway                     115
Brighton                   105
Downtown                    87
South End                   80
South Boston                75
Mission Hill                57
Charlestown                 43
Beacon Hill                 37
Roslindale                  35
East Boston                 34
Mattapan                    30
Hyde Park                   26
West End                    24
South Boston Waterfront     22
West Roxbury                21
North End                   17
Chinatown                   16
Longwood                    15
Bay Village                  3
Leather District             3
Name: neighborhood, dtype: int64[i]
```

A choropleth map is used to highlight the areas in the City of Boston which have the highest risks. Dorchester and Roxbury show the concentration of incidents in the area.



There is a clear benefit to find the clusters, and the clusters were generated around the longitude and latitude in a supporting data frame.

```
df_cluster = df_crash_bikex.drop('dispatch_ts',1)
df_cluster = df_cluster.drop('mode_type',1)
df_cluster = df_cluster.drop('location_type',1)
df_cluster = df_cluster.drop('street',1)
df_cluster = df_cluster.drop('xstreet1',1)
df_cluster = df_cluster.drop('xstreet2',1)
df_cluster = df_cluster.drop('x_cord',1)
```

```
df_cluster = df_cluster.drop('y_cord',1)
df_cluster = df_cluster.drop('bad',1)
df_cluster.head()
```

The preprocessing leaves two columns:

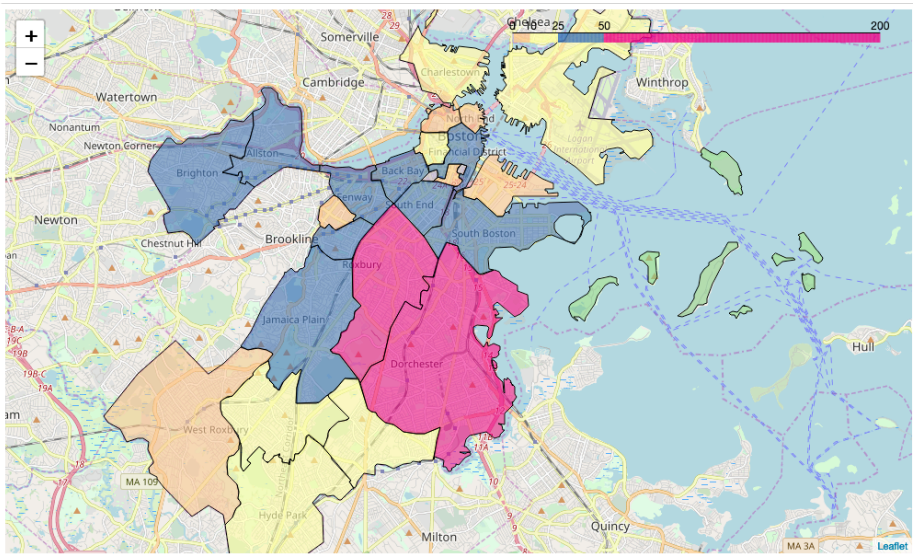
	lat	long
3	42.305413	-71.069163
12	42.320627	-71.091100

Then runs the clustering algorithm on top of the data, and assign the labels for all 1700 incidents:

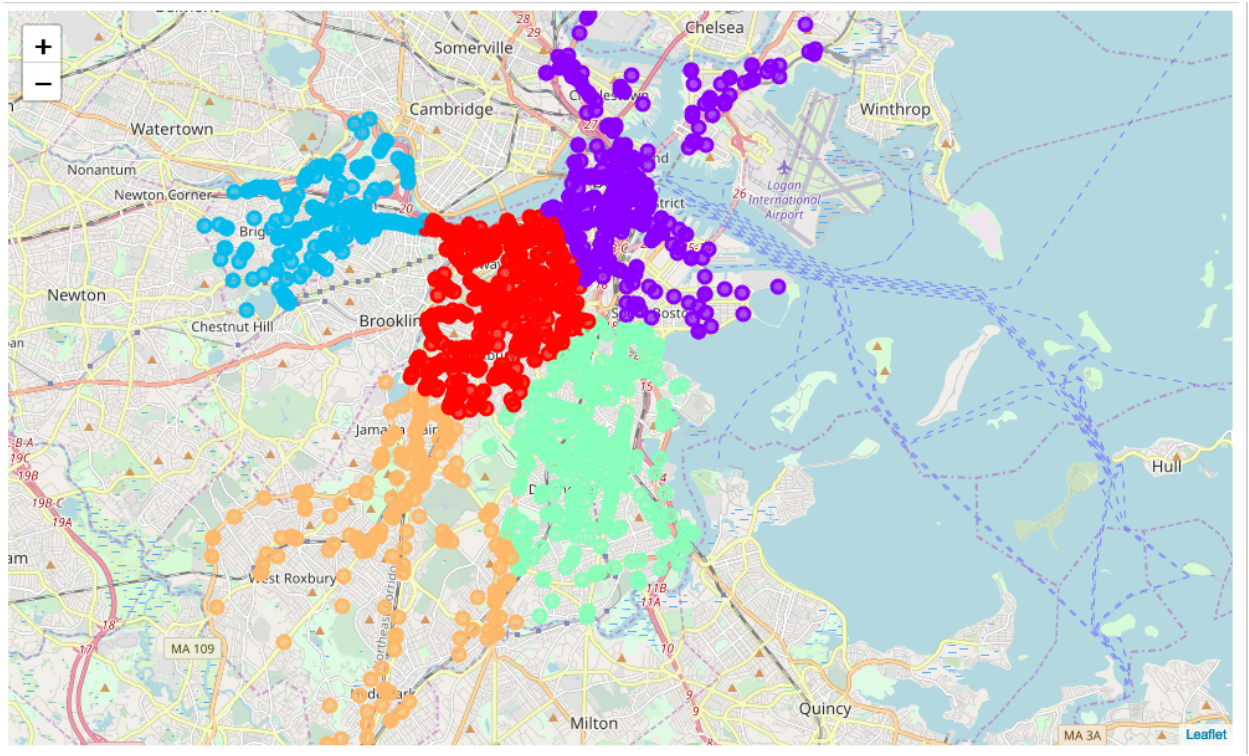
```
# cluster numbers
kclusters = 50

# k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(df_cluster)

# Size of the Shape (results 0 to the)
kmeans.labels_[0:1713]
```

Results

The Dorchester and Roxbury neighborhoods are very risk for travelers, more than 50 incidents since 2015.

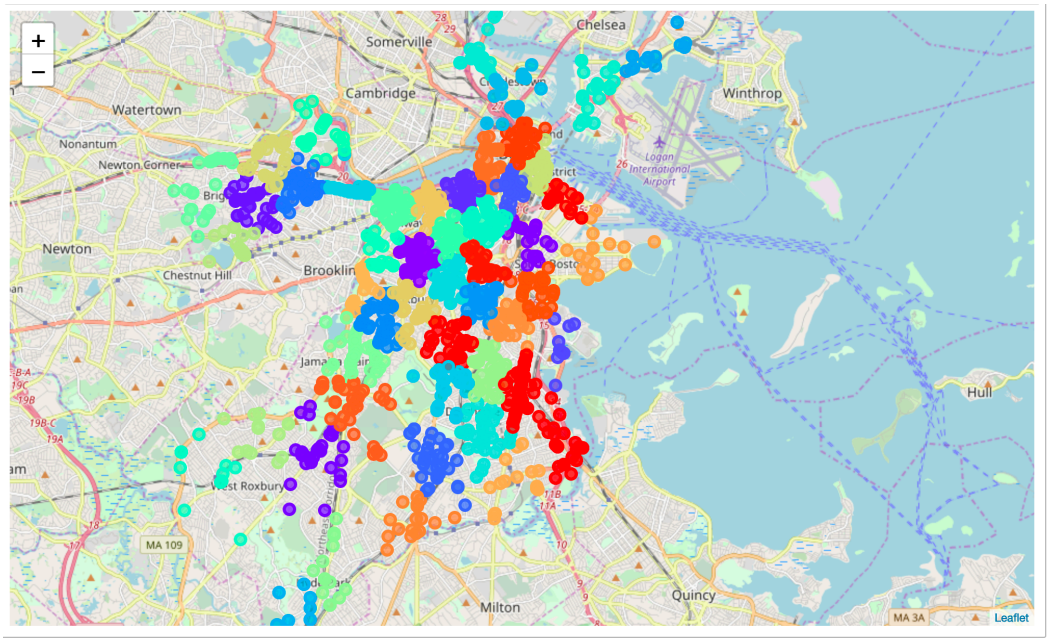


Using a cluster size of 5 one sees the five generated groupings and clusters.



5 Clusters

Using a cluster size of 25, one sees the 25 generated groupings and clusters. The intense groupings around the center of Boston show a conglomeration of incidents in that area – Tremont and Dorchester Avenue.



25 clusters

Discussion

With the 25 clusters, one is able to see the risks centered around the Dorchester and Roxbury neighborhoods. Further, there are streets which span many neighborhoods. The neighborhoods have a level of risk, and the streets have a next level of risk as shown in the clusters.

Conclusion

The City of Boston has an active population who travel through the neighborhoods for pleasure and the commute. The cyclists take risks moving in traffic to their destinations. The results of this analysis "What level of risk is there in commuting through the Boston neighborhoods and streets to avoid" discovers incident hotspots in Dorchester and Roxbury neighborhoods and a number of streets to avoid Dorchester Avenue and Tremont Street. Armed with this information cyclists are better prepared to travel to their destinations, and the City of Boston is able to make investment decisions on infrastructure and policing to better support cyclists.

While the model addresses risk for cyclists using geocoordinates, there is an opportunity for further work, and the risks for cyclists may further be refined and more precisely calculated. To determine the human traffic, the model may use social popularity data (FourSquare Check-in data, Facebook geolocation tags) to calculate the frequency and popularity; vehicular traffic would be hard to estimate using this technique. To determine street level congestion, as a feature of the risk model, the model may use Waze Data Feeds to account for 'Moving Vehicles' – the data would need to be aggregated over time and used to predict when and where risk exists, and how-to re-pattern the cyclist commute habits. There are many opportunities to extend the model and the precision of the model.

References

- [1] https://en.wikipedia.org/wiki/Boston
- [2] https://www.boston.gov/departments/boston-bikes/bike-data#citywide-bike-counts
- [3] https://www.boston.gov/sites/default/files/document-file-03-2017/go_boston_2030_-_3_boston_today_spreads.pdf
- [4] https://en.wikipedia.org/wiki/Massachusetts_Bay_Transportation_Authority
- [5] https://www.boston.gov/departments/boston-bikes/bike-data/2016-boston-bicycle-counts
- [6] https://www1.nyc.gov/html/dot/html/bicyclists/bikestats.shtml
- [7] https://www.bostonglobe.com/metro/2016/04/30/boston-duck-tours-vehicle-hit-pedestrian-police-say/rQli0qL5NwwIrD6cENHedN/story.html
- [8] https://www.bostonglobe.com/metro/2018/11/09/pedestrian-hit-truck-seriously-hurt-near-museum-science/RO8nIZod95n7WOSEqybClK/story.html
- [9] https://news.harvard.edu/gazette/story/2016/12/better-days-for-boston-cyclists/

- [10] https://data.boston.gov/dataset/vision-zero-crash-records/resource/e4bfe397-6bfc-49c5-9367-c879fac7401d
- [11] https://developers.google.com/waze/data-feed/overview

[i] Note, the Harbor Islands has zero incidents (it has no vehicular traffic).