Capstone Project – Battle of the Neighborhoods

Analysis of the Neighborhoods in Boston, MA

Abby Antoo Antony

Contents

1. Introduction

I recently graduated from my master's program which was in san Francisco and I am looking to start a new chapter in my life in Boston, Massachusetts. An individual moves an average of 11.7 times in their lifetime. In fact, in America, between 2012 and 2013, a total of 35.9 million people aged one year or older moved. There are various reasons for moving like Job Relocations, Changing Neighborhoods, Marriage or Retirement. My primary reason for moving is job hunt. Boston, MA has done a fine job of attracting young professionals in abundance over the years with their reputed educational institutions, diverse communities, and access to world class healthcare

Safety and access to necessities are the biggest concerns when thinking about moving to a new location. The goal of this project is split into two parts. First, to find out the safest neighborhood to live in Boston, MA. Second, to explore the neighborhood to determine the various venues per street and cluster them using k-means clustering. By the end of the analysis, I or any individual should be able to narrow down which street to move into in the neighborhood identified as the safest.

2. Data Acquisition

The data is extracted from Analyze Boston, which is an open data hub of the city of Boston. The data set that will be used is Crime Incidents Report. These are reports provided by the Boston Police Department to document the initial details surrounding an incident to which BPD officers respond. The dataset contains records from the new crime incident report system, which includes a reduced set of fields focused on capturing the type of incident as well as when and when it occurred. The record begins in June 2015. Since the analysis is focused on safety, the latest year i.e 2019 is considered.

***Metadata**:*

- **Incident_num**: Internal BPD report number
- **Offense_code**: Numerical code of offense description
- **Offense_Code_Group_Description**: Internal categorization of [offense_description]
- **Offense_Description**: Primary descriptor of incident
- **District**: What district the crime was reported
- **Reporting_area**: RA number associated
- **Shooting**: indicated a shooting took place
- **Occurred_on**: Earliest date and time the incident could have taken place
- **UCR_Part**: Universal Crime Reporting Part number (1,2,3)
- **Street**: Street name the incident took place

3. Data Analysis and Modelling

3.1 Data Cleaning

This dataset consists of data from 2015 to 2020. Since the analysis requires the most recent year, the 2019 data is selected. This reduces the number of rows and columns to 98082 and 17 respectively. The cleaning process is split into 4 parts. They are as follows:

3.1.1 Adding the District_Names column

The values in the 'DISTRICT' column represents the different neighborhoods in Boston but this is represented by district codes e.g C6, E18 etc. This makes it difficult for an individual who is not from Boston to identify the names of the neighborhood. Therefore, for the sake of simplicity, I added a column, 'dist_names'(District_Names), with the name of the neighborhoods corresponding to their district code. (Appendix Fig 1)

3.1.2 Dropping Columns

There are 17 columns in the dataset. For this analysis, the initial goal is to identify the safest neighborhood in Boston, MA and for that 6 out of 17 columns will only be needed i.e 'District_Names', 'Street', 'Offense_Category', 'Lat', 'Long'. The remaining 11 columns are dropped. (Appendix Fig 2)

3.1.3 Missing Values

One of the main challenges in this analysis are the missing values. All the Columns have missing values but District_Names have missing values that are represented by blanks whereas the other columns have their missing value represented as 'NaN'. The key columns of this analysis are the District_Names and Offense_Category, Therefore, I converted the missing values that are blank in the District_names column to NaN and dropped the NaN's in both District_Names and Offense_Category. (Appendix Fig 3)

3.1.4 Pivot Table – Adding 'Total' Column

Next, I grouped the District_Names and Offense_Category to get the number of different offenses or crimes per district or neighborhood and renamed the column as 'Count of offenses per dist'. Once that was done, I used the pivot_table function to covert the data frame into a pivot table where the rows and columns are 'District_Names' and 'Offense_Category' respectively with the 'Count of offense per dist' as values. Finally, the 'Total' column is added to the table which represents the sum of crimes per neighborhood. (Appendix Fig 4)

3.2 Data Exploration

3.2.1 Districts with the highest and lowest reported crime

Using the seaborn and matplotlib libraries, a bar graph was created to visualize the top four neighborhoods with the highest and lowest number of crimes. Based on the bar chart, Roxbury has the highest number of reported crimes with an astounding 11764 reported cases in year 2019 and Charlestown has the lowest with 1397 reported cases. Therefore, Charlestown is the safest neighborhood based on the total number of crimes per neighborhood. (Appendix Fig 5)

3.2.2 Top 10 offense category in Charlestown

Looking deeper into Charlestown, it's clear that top 10 offenses are motor vehicle accident response, medical assistance, larceny, investigate Person, towed, simple assault, vandalism, Drug Violation and larceny from motor vehicle where motor vehicle accident response and larceny from motor vehicle are the most and least among all of them. (Appendix Fig 6)

3.2.3 Dropping duplicates and using Folium to map Charlestown

The next goal is exploring the different venues around Charlestown by street and segregate these venues into different clusters using k-means. Before that, the data needs to be cleaned a bit more. The reason for not dropping all the NaN during the data cleaning process is because dropping all the NaN would lead to having less amount of observations which would have an impact on the first part of the analysis i.e finding the safest neighborhood. Now since Charlestown is determined to be the safest neighborhood, the next step is dropping the remaining NaN as well as duplicates from the 'Street' column to map Charlestown. (Appendix Fig 7)

3.3 K-means clustering

The final dataset contains the different streets in Charlestown along with their corresponding latitude and longitude. Now, that the final data set is ready, the Foursquare API is connected to find the various venues within a 500-meter radius. A JSON file is returned containing all the venues in each street. To get a clear picture of the data obtained from Foursquare, the data is converted into a dataframe and is assigned to a variable called Charlestown_data.

The next step is One hot encode the Charlestown_data to convert the categorical variables into a binary to perform k-means. The dataset is then grouped by street and the average or mean is calculated which finally leads to listing the top 10 common venues in each street.

K-means clustering is a machine learning algorithm that segregates data and clusters them based on similar traits given that the cluster sizes are predefined. In this case, the predefined cluster size is 5 and the goal is to find similar streets in Charlestown based on the various venues. This gives an individual options in the form of clusters to decide which street to move into based on the venues in those respective streets. (Appendix Fig 8)

4. Results

The five clusters created by running k-means clustering consists of 138 streets in the neighborhood of Charlestown. Let's look into each of the clusters.

4.1 Cluster 1

The first cluster consists of 18 out of the 138 streets. Individuals who are current residents or potential residents will have access to a bunch of pizza places, deli shops and coffee but mainly pizza places as it appears 14 times in the 1st and 2nd most common venues. There are other venues available like music venues, tennis courts and gyms but they are not as common. (Appendix Fig 9)

4.2 Cluster 2

The second cluster is the largest and consists of 60 streets out of the total 138. Individuals moving into these streets or who are currently residents have access to a lot of pubs as it appears 37 times in the 1st, 2nd, and 3rd most common venue columns. The streets also have other venues such pizza places, convenience stores, café's, and pharmacies. The other venues that are not very common are pet stores, gyms and grocery stores to name a few. (Appendix Fig 10)

4.3 Cluster 3

The third cluster is the second largest one and consists of 34 streets. These streets are for individuals who enjoy going to history museums, national parks, going on boats & ferries and having donuts occasionally or every day. The other venues that are not very common are pizza places, gyms and grocery stores to name a few. (Appendix Fig 11)

4.4 Cluster 4

The fourth cluster consists of 25 out of 138 streets and these streets are for individuals who are into fitness as well as for individuals who rely on public transportation for their daily commute or for new residents who are planning to use public transport as their main source of transportation. Other venues are donut shops, cafe's, parks, and liquor stores to name a few. (Appendix Fig 12)

4.5 Cluster 5

The fifth cluster consists of just 1 street out of the 138 making this an outlier. The most 3 most common venues are taxi stands, restaurants and yoga studios. (Appendix Fig 13)

5. Discussion

The goal of this analysis/project was to help me to look into the different neighborhoods of Boston, MA to decide which one to move to but this is an analysis/project that can assist any individual who are planning to move into the city of Boston. The analysis showed that Charlestown was the safest neighborhood having 1397 reported crimes compared to an astounding 11764 reported crimes in Roxbury in the year 2019.

K-means clustering grouped all streets in the neighborhood of Charlestown into 5 clusters with similar venues. Based on the analysis, I would personally look into the streets in cluster 4 because I am a person who enjoys fitness as well as uses public transport as a daily source of transportation. For individuals who like to eat out a lot and have fun with access to convenient stores and pharmacies, clusters 1 and 2 are for you. Individuals who enjoy hikes, history, boats, ferries and bit of surfing, cluster 3 is for you.

6. Conclusion

Moving into a new city or country has so many factors to it. This analysis is an indicator of how technology can be utilized to gain additional insights on the city or country one is planning on moving to. Couple lines of code narrowed down the safest neighborhood as well the venues around that neighborhood making the whole process of moving easier for an individual. They could use their time to look into the other factors that were not taken into consideration like the cost of living and housing, which deserves their own analysis. But as a start, anyone who is planning to move to the city of Boston, Charlestown is the place to be.

7. Appendix

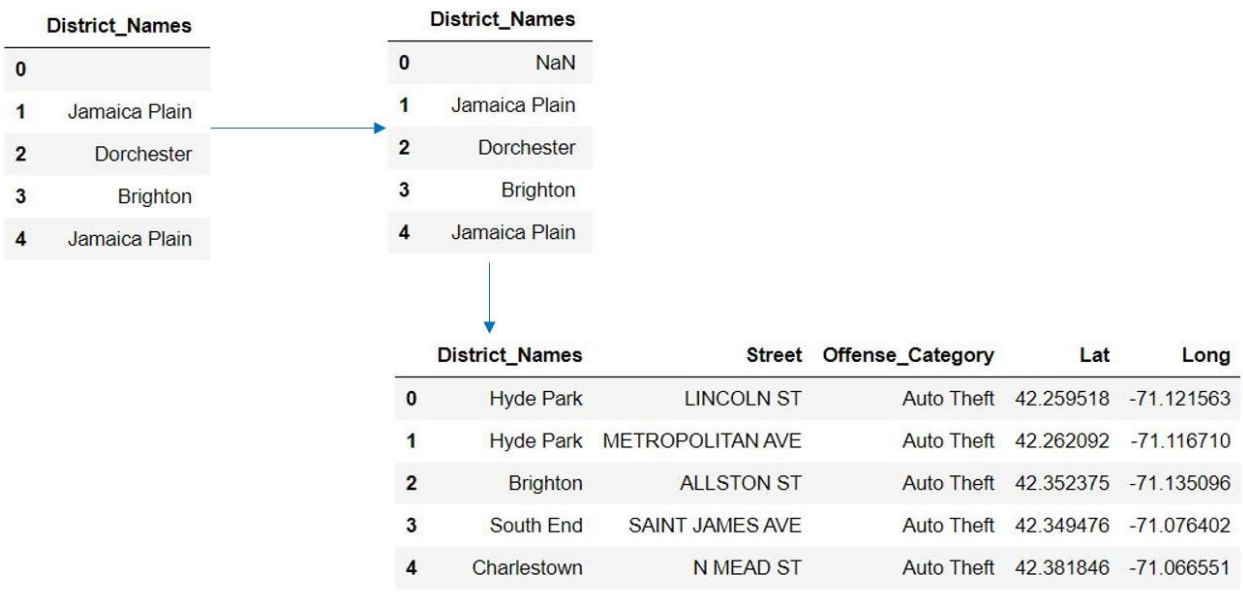
Fig 1:

	HOUR	UCR_PART	STREET	Lat	Long	Location	dist_names
	0	NaN	RIVERVIEW DR	NaN	NaN	(0.00000000, 0.00000000)	
	3	NaN	DAY ST	42.325122	-71.107779	(42.32512200, -71.10777900)	Jamaica Plain
	0	NaN	GIBSON ST	42.297555	-71.059709	(42.29755500, -71.05970900)	Dorchester
	7	NaN	BROOKS ST	42.355120	-71.162678	(42.35512000, -71.16267800)	Brighton
	18	NaN	WASHINGTON ST	42.309718	-71.104294	(42.30971800, -71.10429400)	Jamaica Plain

Fig 2:

	OFFENSE_CODE_GROUP	STREET	Lat	Long	dist_names
0	NaN	RIVERVIEW DR	NaN	NaN	
1	NaN	DAY ST	42.325122	-71.107779	Jamaica Plain
2	NaN	GIBSON ST	42.297555	-71.059709	Dorchester
3	NaN	BROOKS ST	42.355120	-71.162678	Brighton
4	NaN	WASHINGTON ST	42.309718	-71.104294	Jamaica Plain

Fig 3:



	District_Names
0	
1	Jamaica Plain
2	Dorchester
3	Brighton
4	Jamaica Plain

	District_Names
0	NaN
1	Jamaica Plain
2	Dorchester
3	Brighton
4	Jamaica Plain

	District_Names	Street	Offense_Category	Lat	Long
0	Hyde Park	LINCOLN ST	Auto Theft	42.259518	-71.121563
1	Hyde Park	METROPOLITAN AVE	Auto Theft	42.262092	-71.116710
2	Brighton	ALLSTON ST	Auto Theft	42.352375	-71.135096
3	South End	SAINT JAMES AVE	Auto Theft	42.349476	-71.076402
4	Charlestown	N MEAD ST	Auto Theft	42.381846	-71.066551

Fig 4:

Search Warrants	Service	Simple Assault	Towed	Vandalism	Verbal Disputes	Violations	Warrant Arrests	Total
10.0	7.0	200.0	323.0	236.0	114.0	92.0	30.0	4558.0
7.0	0.0	66.0	76.0	64.0	36.0	27.0	9.0	1397.0
27.0	13.0	472.0	254.0	444.0	638.0	172.0	127.0	9464.0
12.0	4.0	629.0	251.0	271.0	42.0	103.0	319.0	8517.0
9.0	8.0	146.0	149.0	161.0	94.0	56.0	32.0	2804.0
23.0	10.0	153.0	111.0	170.0	288.0	103.0	23.0	4304.0
13.0	4.0	175.0	170.0	174.0	167.0	66.0	46.0	4060.0
55.0	5.0	426.0	157.0	415.0	666.0	215.0	69.0	8267.0
56.0	15.0	557.0	250.0	515.0	740.0	370.0	259.0	11764.0
12.0	5.0	293.0	292.0	255.0	141.0	53.0	134.0	5364.0
8.0	6.0	545.0	375.0	433.0	185.0	112.0	237.0	9639.0
10.0	5.0	149.0	72.0	127.0	167.0	49.0	13.0	3106.0

Fig 5:

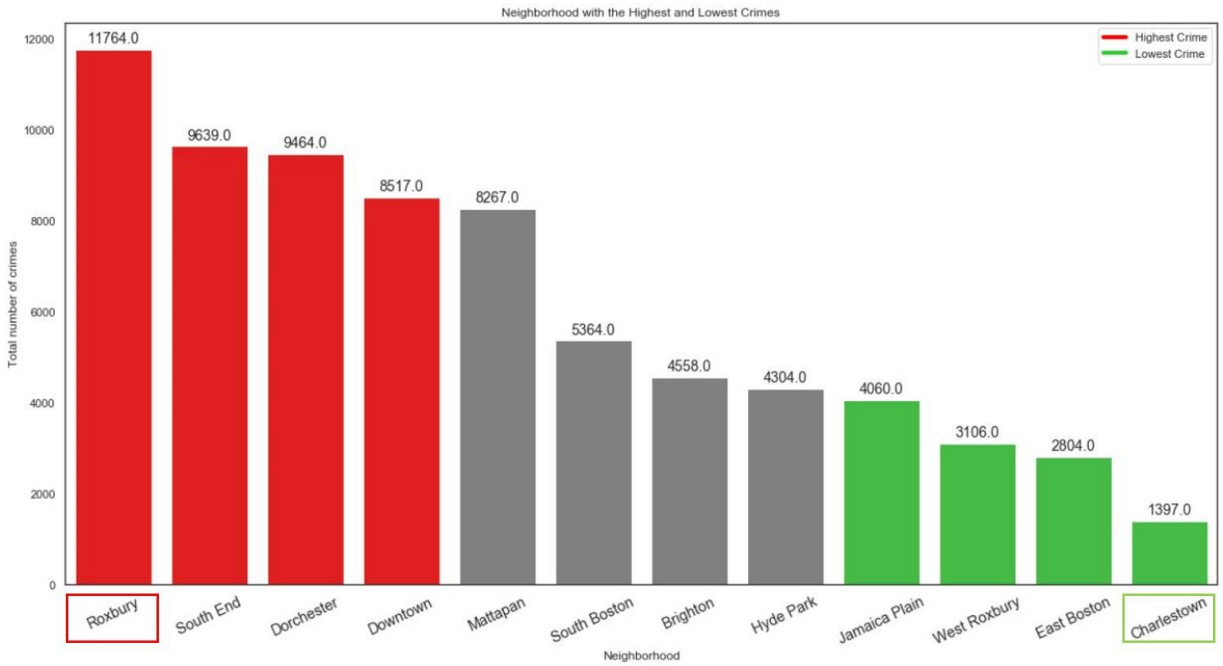

Neighborhood with the Highest and Lowest Crimes
Highest Crime
Lowest Crime
Total number of crimes
Neighborhood
11764.0
9639.0
9464.0
8517.0
8267.0
5364.0
4558.0
4304.0
4060.0
3106.0
2804.0
1397.0
Roxbury
South End
Dorchester
Downtown
Mattapan
South Boston
Brighton
Hyde Park
Jamaica Plain
West Roxbury
East Boston
Charlestown

Fig 6:

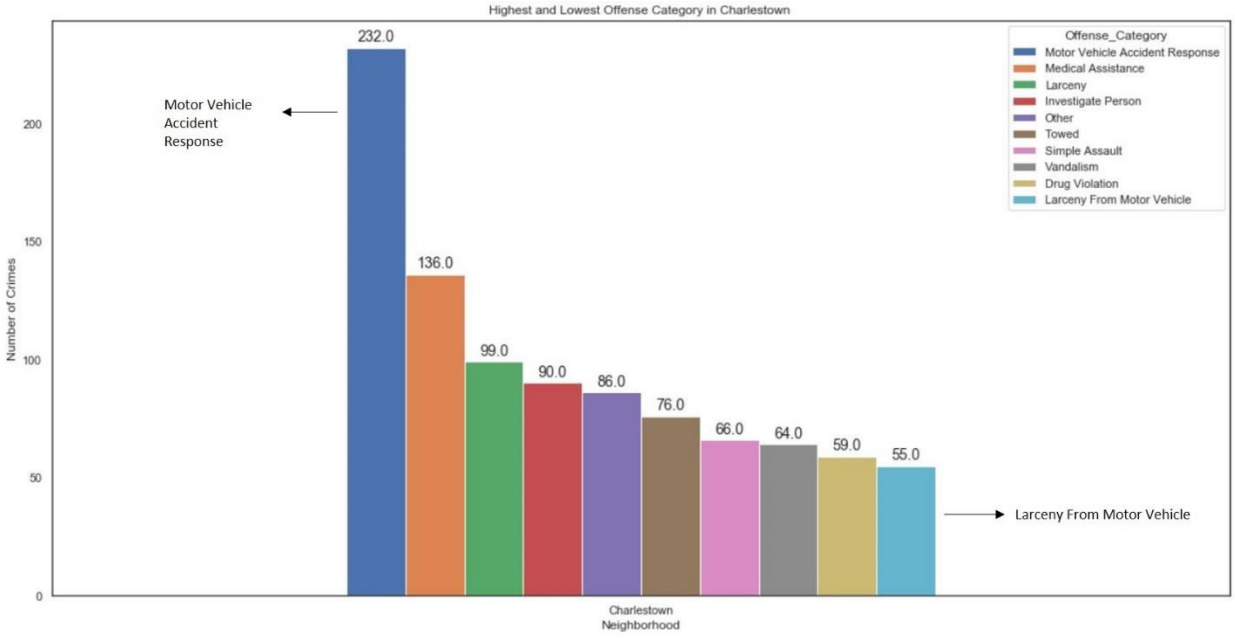

Highest and Lowest Offense Category in Charlestown
Offense_Category
Motor Vehicle Accident Response
Medical Assistance
Larceny
Investigate Person
Other
Towed
Simple Assault
Vandalism
Drug Violation
Larceny From Motor Vehicle
Motor Vehicle Accident Response
Larceny From Motor Vehicle
Number of Crimes
Charlestown Neighborhood
232.0
136.0
99.0
90.0
86.0
76.0
66.0
64.0
59.0
55.0

Fig 7:

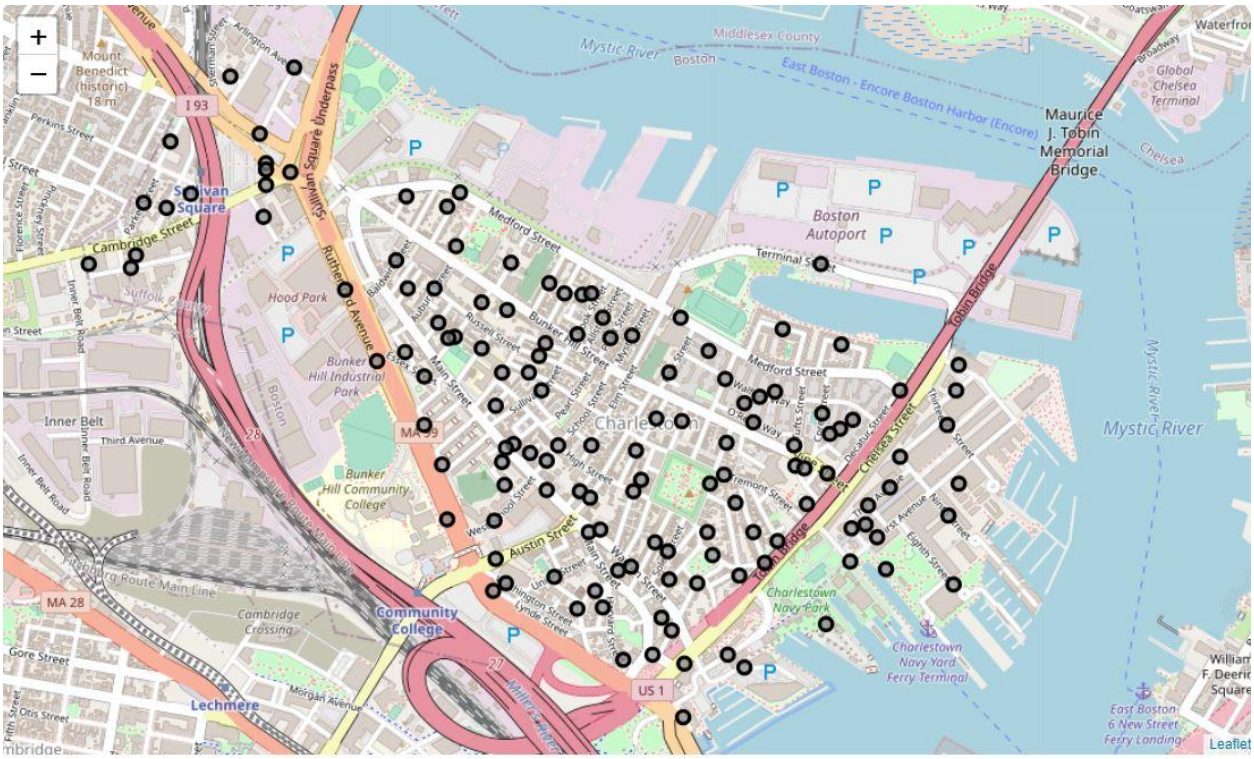

Mystic River
Boston
Middlesex County
East Boston - Encore Boston Harbor (Encore)
Maurice J. Tobin Memorial Bridge
Global Chelsea Terminal
Boston Autoport
Terminal Street
Medford Street
Tobin Bridge
Chelsea Street
Mystic River
Charlestown
Austin Street
Community College
Bunker Hill Community College
Bunker Hill Industrial Park
Hood Park
Sullivan Square
Cambridge Street
Rutherford Avenue
Sullivan Square Underpass
Inner Belt
Third Avenue
Fitchburg Route Main Line
Cambridge Crossing
Lechmere
Morgan Avenue
Gore Street
Otis Street
Charlestown Navy Yard
Charlestown Navy Yard Ferry Terminal
East Boston 6 New Street Ferry Landing
William F. Deering Square
Mount Benedict (historic) 18 m
I 93
MA 28
MA 99
US 1
Leaflet

Fig 8

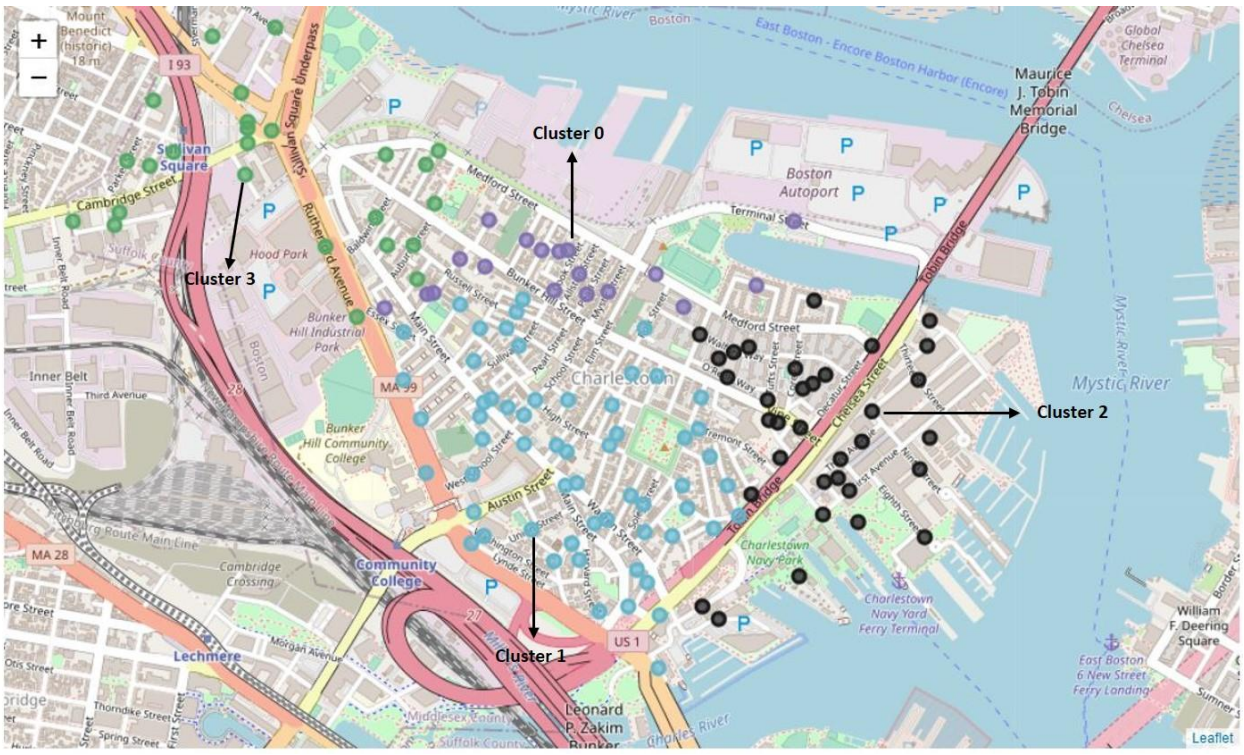

Cluster 0
Cluster 1
Cluster 2
Cluster 3
Mystic River
Boston
East Boston - Encore Boston Harbor (Encore)
Maurice J. Tobin Memorial Bridge
Global Chelsea Terminal
Boston Autoport
Terminal Street
Medford Street
Tobin Bridge
Chelsea Street
Charlestown
Austin Street
Community College
Bunker Hill Community College
Bunker Hill Industrial Park
Hood Park
Sullivan Square
Cambridge Street
Rutherford Avenue
Sullivan Square Underpass
Inner Belt
Third Avenue
Fitchburg Route Main Line
Cambridge Crossing
Lechmere
Morgan Avenue
Charlestown Navy Yard
Charlestown Navy Yard Ferry Terminal
East Boston 6 New Street Ferry Landing
William F. Deering Square
Leonard P. Zakim Bunker
Mount Benedict (historic) 18 m
I 93
MA 28
MA 99
US 1
Leaflet

Fig 9:

	Street	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	N MEAD ST	Pizza Place	Bus Station	Pool	Convenience Store	Theater	Gym / Fitness Center	Gym	Music Venue	Café	Yoga Studio
15	SACKVILLE ST	Convenience Store	Pizza Place	Deli / Bodega	Tennis Court	Pool	Café	Gastropub	Gym	Music Venue	Athletics & Sports
26	MYSTIC ST	Convenience Store	Pizza Place	Deli / Bodega	National Park	Café	Music Venue	Gastropub	Monument / Landmark	Pool	Gym
30	CARNEY CT	Convenience Store	Pizza Place	Boat or Ferry	National Park	Music Venue	Candy Store	Café	Gastropub	Monument / Landmark	Donut Shop
40	MEDFORD ST	Convenience Store	Deli / Bodega	Martial Arts Dojo	Donut Shop	Pizza Place	Candy Store	Café	National Park	Music Venue	Gym
64	OLD IRONSIDES WAY	Convenience Store	Deli / Bodega	Boat or Ferry	Donut Shop	Pizza Place	Park	National Park	Café	Music Venue	Gastropub

Fig 10:

	Street	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
2	LOWNEY WAY	History Museum	Gastropub	Pub	Donut Shop	Park	National Park	Grocery Store	Convenience Store	Coffee Shop	Seafood Restaurant
5	EDEN ST	Pizza Place	Yoga Studio	Gym	Pharmacy	Pet Store	Coffee Shop	Donut Shop	Café	Shopping Mall	Gastropub
6	THOMPSON SQ	Pizza Place	Café	Pub	Coffee Shop	Gastropub	Pet Store	Plaza	Park	History Museum	National Park
11	SALEM ST	Pub	Convenience Store	Coffee Shop	Yoga Studio	Gastropub	Donut Shop	Pharmacy	Deli / Bodega	Pet Store	National Park
14	CHARLES RIVER AVE	Pub	Pizza Place	Donut Shop	Park	Brewery	Moroccan Restaurant	Playground	Plaza	Restaurant	Lounge
18	AUSTIN ST	Coffee Shop	Pizza Place	Pub	Yoga Studio	Shopping Mall	American Restaurant	Bank	Convenience Store	Donut Shop	Gastropub

Fig 11:

	Street	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
4	MONUMENT ST	Boat or Ferry	Surf Spot	Pizza Place	Donut Shop	National Park	Music Venue	Candy Store	Café	Monument / Landmark	Gastropub
7	BUNKER HILL ST	Boat or Ferry	Donut Shop	History Museum	National Park	Monument / Landmark	Playground	Convenience Store	Park	Candy Store	Café
8	DECATUR ST	Park	Surf Spot	Grocery Store	Donut Shop	Tourist Information Center	Martial Arts Dojo	Convenience Store	Gym Pool	Gastropub	Boat or Ferry
10	VINE ST	National Park	Café	History Museum	Boat or Ferry	Donut Shop	Park	Gastropub	Surf Spot	Grocery Store	Monument / Landmark
12	WALFORD WAY	Donut Shop	Café	Boat or Ferry	Gym	Convenience Store	National Park	Candy Store	Gastropub	Monument / Landmark	Grocery Store
16	EIGHTH ST	Café	History Museum	Harbor / Marina	Grocery Store	Park	Seafood Restaurant	Surf Spot	Boat or Ferry	Beer Garden	Convenience Store

Fig 12:

	Street	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	AUBURN ST	Gym / Fitness Center	Bus Station	Pool	Theater	Pizza Place	Gym	Café	Park	Yoga Studio	Dive Bar
9	RUTHERFORD AVE	Bus Station	Gym / Fitness Center	Recording Studio	Food	Bus Stop	Pool	Park	Gym	Café	Bank
13	OAK ST	Gastropub	Bus Station	Pool	Theater	Pizza Place	Gym / Fitness Center	Gym	Café	Park	Dive Bar
25	CAMBRIDGE ST	Bus Station	Gym	Gym / Fitness Center	Donut Shop	Liquor Store	Metro Station	Coffee Shop	Café	Bus Stop	Park
35	BUNKER HILL INDUSTRIAL PA	Bus Station	Pool	Theater	Pizza Place	Metro Station	Food	Gym	Park	Gastropub	Yoga Studio

Fig 13:

	Street	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
146	ALFORD ST	Taxi	Restaurant	Yoga Studio	Grocery Store	Diner	Discount Store	Dive Bar	Dog Run	Donut Shop	Electronics Store