An Analysis and Recommendation of Better Places to Live in Boston

A Report for the final project of IBM Data Science Specialization course hosted on Coursera platform

Submitted by: Shweta Shrestha

Date: June 03, 2019

1. Introduction

1.1 Background

What makes a great place to live and how do we decide that? Is health care most important, or affordable housing? Which matters more, commute times, climate or crime rate?[1] Different people have different preferences and different perspective for better societies. Here, I have tried to narrow down my friend's preferences and tried to analyze better places to reside in Boston.

1.2 Problem

A friend of mine is looking to move and settle down in Boston. Boston, being one of the oldest cities in the United States, it is most populous city of the Commonwealth of Massachusetts. Its rich history attracts many tourists and its many colleges and universities make it an international center of higher education [2]. A friend of mine is looking for a proper neighborhood to reside in Boston. But he has some priorities for the location. His preferences are as follows:

- His younger boy needs special-needs education. He is looking for a neighborhood which has more number of special schools, so that he can choose the nearest or the best among these schools for his kid.
- Neighborhood with low criminality rate, which is safe for them and their kids.
- City which has lots of restaurants/cafes or hotels in town, as he used to work as a chef in a Chinese restaurant and he wants to continue the same in Boston.

But his first priority is the area with more number of special schools.

1.3 Target audience

The target audience for this project are:

- Potential movers, who are willing to settle down in Boston.
- Parents focusing on the schools for their children around Boston.
- Real estate builders and planners who can decide what kind of neighborhoods are more attractive to open their apartments.
- Potential researcher on the diverse information about venues in the neighborhoods of Boston.
- All the students, curious to learn python's panda's dataframe, python's visualization tools and data visualization using python libraries.

2. Data Acquisition and Cleaning

2.1 Data Description

Boston has been chosen as the city of target because, it is one of the oldest city in the United States and the dataset required for our analysis are readily available in the web. The datasets used for the analysis of the problem are as follows:

1. Crimes in Boston - https://www.kaggle.com/ankkur13/boston-crime-data
2. Boston public Schools data - https://www.kaggle.com/crawford/boston-public-schools
3. Boston Police Districts geo location data - https://data.boston.gov/dataset/police-districts
4. Boston Neighborhoods geo location data - https://data.boston.gov/dataset/boston-neighborhoods

5. Boston Police Districts and Neighborhoods served - https://en.wikipedia.org/wiki/Boston_Police_Department

2.2 Data Collection

The datasets collected are mostly from Airbnb's Kaggle portal and geo location data are from official sources of Boston.

The crime dataset includes the crimes reported and recorded at the Boston Police Districts in the year 2018 and the Boston public schools dataset includes the name, address, school type of the public schools in Boston for the year 2018-2019. Those datasets from Kaggle portal are in CSV format which can be readily read into Panda's dataframe. But for my ease, I just downloaded the CSV datasets into my local folder and read into Panda's dataframe, as shown in Figure:1 and Figure:2.

The geo location data available are in the GEOJSON format as shown in Figure:3, is used to render Boston city's map with boundaries according to the Police Districts and according to Neighborhoods in Boston.

List of Boston Police Districts (BPDs), the Neighborhoods served under each police district, the latitude and longitude of each BPD and each Neighborhoods were not easily found in a single dataset. So, these data are entered manually in an excel sheet. The latitude, longitude for each BPD and each Neighborhood are calculated using the online geographical lat-long converter tool (https://www.latlong.net/convert-address-to-lat-long.html), which converts addresses into geographical latitude-longitude. This excel data is later read into a Panda's dataframe.

In order to explore the venues in every Neighborhood and to segment and cluster the Neighborhoods around Boston, I have used the Foursquare API.

```
[5]: schools_df.tail()
```

[5]:

	X	Y	OBJECTID_1	OBJECTID	BLDG_ID	BLDG_NAME	ADDRESS	CITY	ZIPCODE	CSP_SCH_ID	...	SCH_NAME	SCH_LABEL	SCH_TYPE	SHARED	COMPLEX	Label	TLT	
126	-71.092030	42.317660	127	1736	52	Higginson Bldg	160 Harrishof Street	Roxbury	2119	4241	...	Higginson Elementary (K1-2)	Higginson (K1-2)	ES			61	1	M
127	-71.037940	42.371568	128	2136	0	Alighieri Bldg	37 Gove St.	East Boston	2128	4321	...	Alighieri Montessori	Alighieri	ES			2	1	M
128	-71.068150	42.348770	129	2938	0	Church Street Bldg	20 Church Street	Boston	2116	1215	...	Boston Adult Tech Acad	BATA	Special			8	4	Fr
129	-71.145961	42.350441	130	2946	139	Taft Bldg	20 Warren Street	Brighton	2135	1470	...	Boston Green Academy	Boston Green Academy	2012-06-07 00:00:00			11	4	Jon
130	-71.080504	42.326153	131	3346	150	Dearborn Bldg	35 Greenville Street	Roxbury	2119	1260	...	Dearborn Academy	Dearborn Academy	2012-06-07 00:00:00			31	4	An

Figure 1: Dataframe showing Boston public schools data

```
[14]: crime_df.head()
```

[14]:

IDENT_NUMBER	OFFENSE_CODE	OFFENSE_CODE_GROUP	OFFENSE_DESCRIPTION	DISTRICT	REPORTING_AREA	SHOOTING	YEAR	MONTH	DAY_OF_WEEK	HOUR	UCR_PART	STREET	Lat
I182080058	2403	Disorderly Conduct	DISTURBING THE PEACE	E18	495	NaN	2018	10	Wednesday	20	Part Two	ARLINGTON ST	42.262608
I182080053	3201	Property Lost	PROPERTY - LOST	D14	795	NaN	2018	8	Thursday	20	Part Three	ALLSTON ST	42.352111
I182080052	2647	Other	THREATS TO DO BODILY HARM	B2	329	NaN	2018	10	Wednesday	19	Part Two	DEVON ST	42.308126
I182080051	413	Aggravated Assault	ASSAULT - AGGRAVATED - BATTERY	A1	92	NaN	2018	10	Wednesday	20	Part One	CAMBRIDGE ST	42.359454
I182080050	3122	Aircraft	AIRCRAFT INCIDENTS	A7	36	NaN	2018	10	Wednesday	20	Part Three	PRESCOTT ST	42.375258

Figure 2: Dataframe showing Boston's crime data


```
{
   "type": "FeatureCollection",
   "features": [
      {
         "type": "Feature",
         "properties": {
            "OBJECTID": 27,
            "Name": "Roslindale",
            "Acres": 1605.5682,
            "Neighborhood_ID": "15",
            "SqMiles": 2.51,
            "ShapeSTArea": 6.9938272E7,
            "ShapeSTLength": 53563.914
         },
         "geometry": {
            "type": "MultiPolygon",
            "coordinates": [
               [
                  [
                     [
                        -71.12593,
                        42.272015
                     ],
                     [
                        -71.12611,
                        42.27162
                     ],
                     [
                        -71.12603,
                        42.27159
                     ],
                     [
                        -71.12572,
                        42.271523
                     ],
                     [
                        -71.12559,
                        42.27146
```


Figure 3: JSON data showing the geo location of BPDs.

2.3 Data Preparation

Data have been collected from different sources. In the Crime data CSV, the crime reported in the Boston Police Districts were recorded, while in the Boston Public Schools CSV, the schools were listed according to the Neighborhoods in Boston. To link those two dataset, an excel sheet with list of Boston Police Districts (BPDs), the Neighborhoods served under each police district, the latitude and longitude of each BPD and each Neighborhoods have been entered manually in an excel sheet.

The Crime data CSV consisted of a column with garbage value (values as ######). These columns have been removed from the dataset. Also in the Public Schools dataset, some of the schools' school type are missing and instead some date values are inserted. Those entries have also been removed from the dataset for analysis.

3. Methodology

Our problem is to find districts which contains more number of special schools, low crime rate and the district rich in bars and restaurants. Following steps have been carried out for the analysis:

- Downloaded the data from various sources into our panda's dataframe.
- The next step was to do an Exploratory Data Analysis.
    - To begin the exploratory data analysis, we have used df.describe() method to summarize the basic statistical distribution of all numeric variables in the dataframe.
    - We have removed the entries with missing values.
    - To summarize the categorical variables, we have used the value_counts() method and groupby() method to group data into categories.
- Data Visualization – For the data visualization purpose we have use matplotlib's plt() function. We have plotted the distributions in bar charts, histograms. To visualize the distribution of categorical variables we have used Heat maps. We have also visualized the crime rates and venues cluster in the choropleth map from folium library.

Heat map - A heat map is a graphical representation of data where the individual values contained in a matrix are represented as colors. A heat map has been used to represent the number of different types of schools in different neighborhood cities of Boston.

Choropleth maps - A choropleth map is a thematic map in which areas are shaded or patterned in proportion to the measurement of the statistical variable being displayed on the map, such as population density or per capita income. The higher the measurement the darker the color. To generate a choropleth map we need the GeoJSON file of that region's boundaries to represent that region in the map. Choropleth maps have been built to show the crime rates in the Boston city.

To cluster the neighborhoods according to the type of most common venues around, we used the K-means clustering. K-means clustering is an unsupervised algorithm. It is a type of partitioning clustering that divides the data into k non-overlapping subsets or clusters. Objects within a cluster are very similar and objects across different clusters are very different or dissimilar. So, using this algorithm, we distributed the neighborhoods into different clusters to find the similarities and dis-similarities among those neighborhoods in terms of venues around them. We have divided the neighborhoods into 4 different clusters.

4. Results

From the above data visualization and modeling methods used, we have found that the number of different types of school in the whole Boston city is as follows:

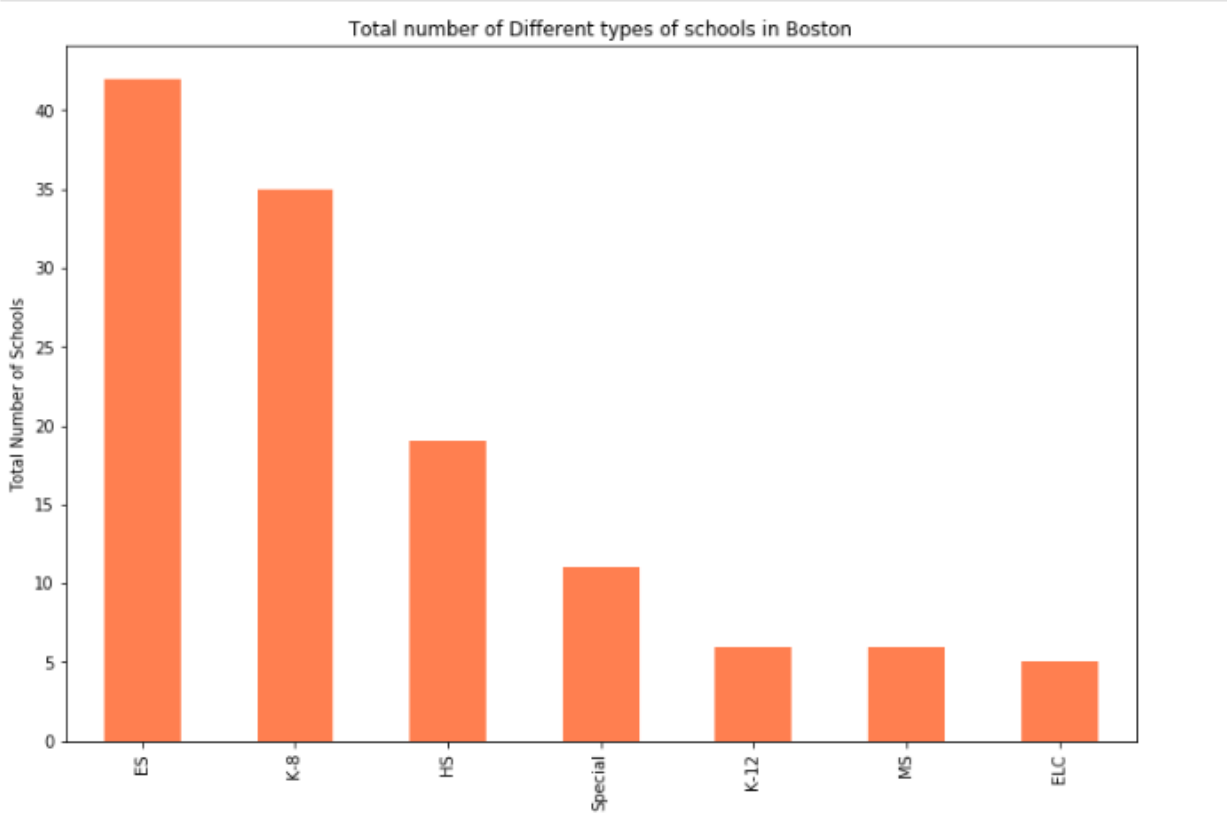

Figure 4: Bar chart representing total number of different types of schools in Boston.

The visualization shows that there are in total 135 schools in Boston city. The bar chart shows that there are more numbers of Elementary Schools (i.e. around 45) and then K-8 and Higher Secondary schools in Boston. The chart also shows, there are only handful numbers of special schools in Boston (i.e. around 10).

Let us now see the distribution of different types of schools along the cities/neighborhoods of Boston. The Heatmap below shows the distribution of different types of schools along the Neighborhoods of Boston.

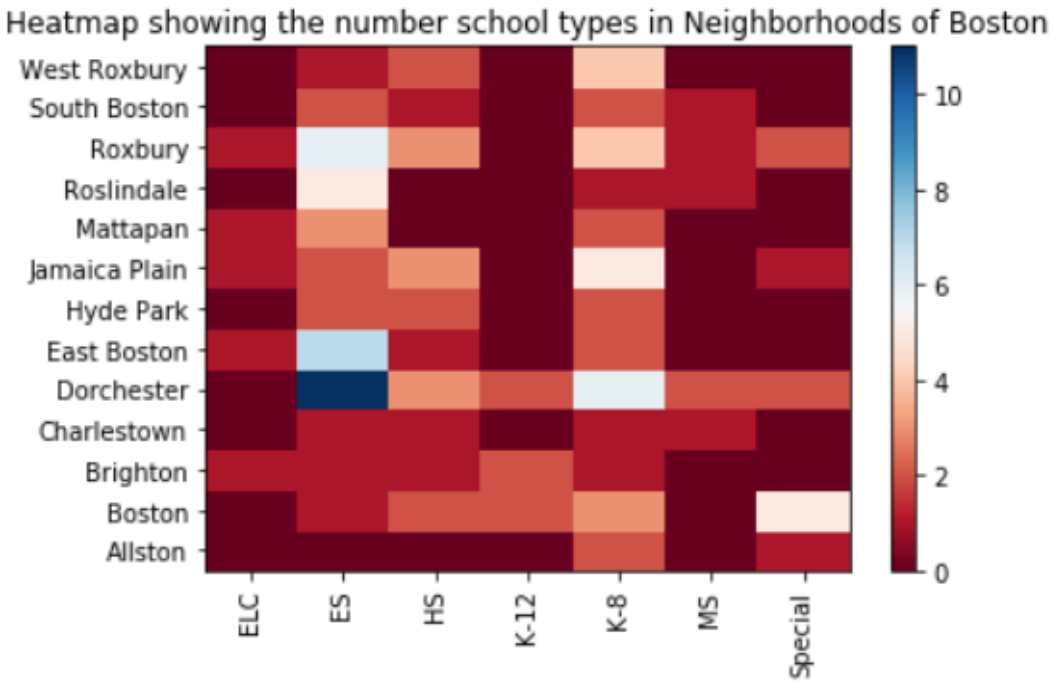

Figure 5: Heat map showing the number of school types in Neighborhoods of Boston.

Looking at the Heat map, East Boston and Dorchester have maximum numbers of Elementary Schools(ES). If we look at the Special schools, which was our priority, Center Boston has maximum number of special schools around 5, Dorchester and Roxbury have 2 numbers of special schools, while other cities either don't have any Special school or have only one of it. So, according to the number of Special Schools, our desired neighborhood will be Center Boston. If not then, should go with Dorchester or Roxbury.

Now let's see Boston from the Crime rate perspective.

The crime data has been recorded according to the Boston Police Districts (BPDs). Each BPD serves 1 or more neighborhoods. The list of BPD and serving neighborhoods can be found here (https://en.wikipedia.org/wiki/Boston_Police_Department). The number of crimes reported in the different Boston Police Districts are as follows:

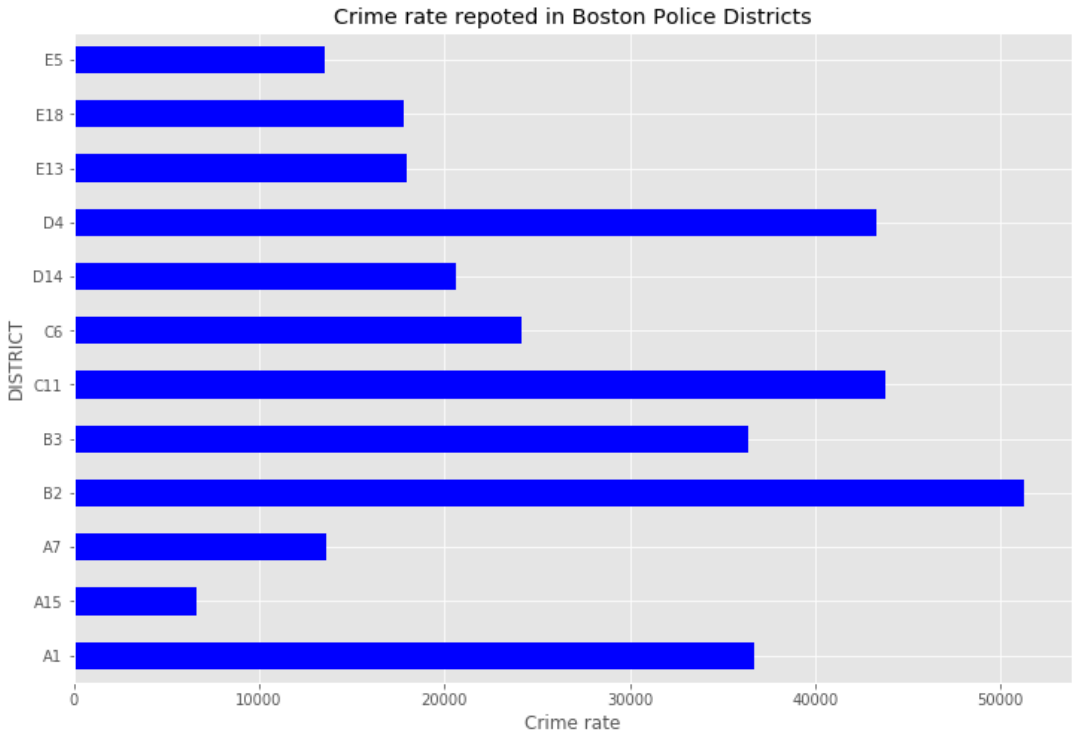

Figure 6: Bar chart showing the number of crimes recorded over each BPDs.

The bar diagram clearly shows Boston Police District B2 as the highest crime reported police district. And the other high crime reported districts are C11 and D4.

Let us visualize the crime reported in Boston Police Districts in a Choropleth map. In the map, the areas are shaded or patterned in proportion to the measurement of crimes reported in the Police Districts. The blue pop up markers represents each of the Police Districts in Boston and the green marker represents the location of special schools. If we look at the map, the highest numbers of special schools lies in the Police District D4. But if we view the crime rate, D4 falls as a second largest crime reported zone.

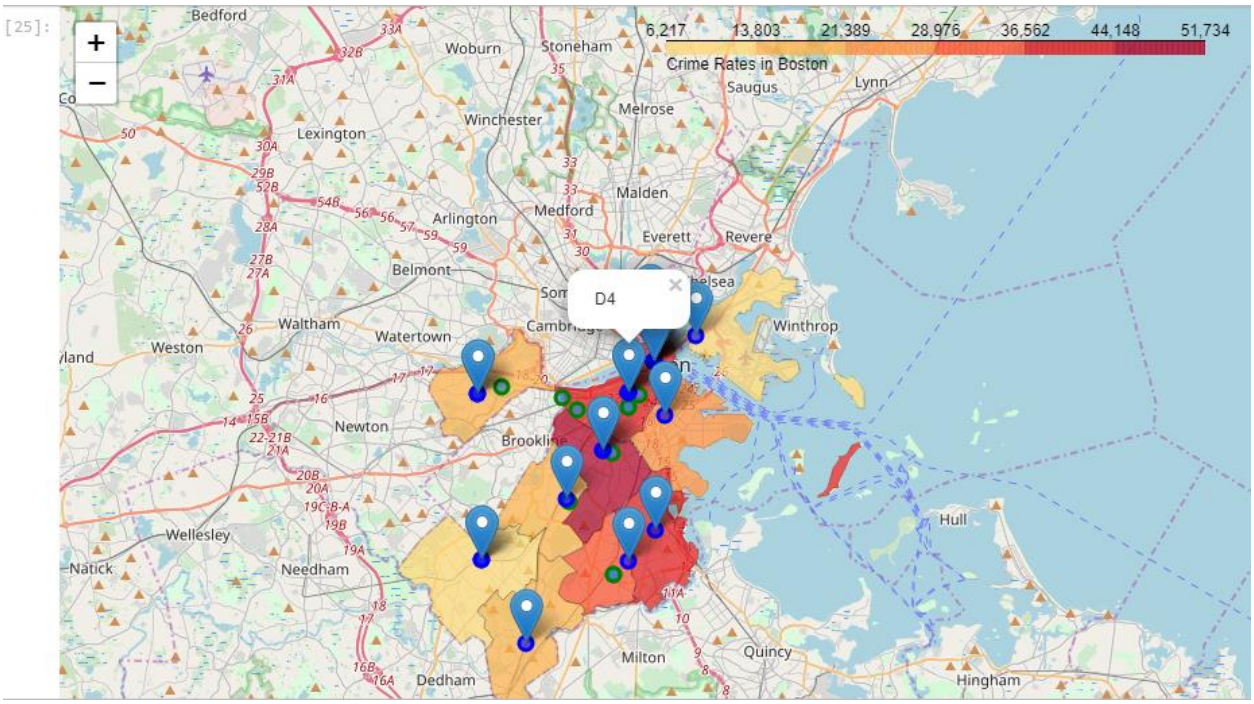

Figure 7: Choropleth map showing the crime density in Boston.

Now let us look at the third criteria of our problem, i.e. the Neighborhoods with large number of restaurants and bars. For this, we have prepared a table with the list of Neighborhoods and their latitude and longitude. This data we have fed to the Foursquare API to find the 100 most venues within the radius of 500km around each neighborhood.

The API requires GPS coordinates of the neighborhoods and returns the JSON objects of venue, its location, venue category. The dataframe below displays the neighborhood and 10 most common venues around it.

```
neighborhoods_venues_sorted.head()
```

[102]:

	Neighborhood	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	Allston	Chinese Restaurant	Korean Restaurant	Bakery	Pizza Place	Mexican Restaurant	Sushi Restaurant	Bar	Spa	Thai Restaurant	Italian Restaurant
1	Back Bay	Chinese Restaurant	Bakery	Asian Restaurant	Theater	Italian Restaurant	Seafood Restaurant	Hotel	Performing Arts Venue	Sushi Restaurant	Spa
2	Bay Village	American Restaurant	Arts & Crafts Store	Assisted Living	Brewery	Bar	Zoo Exhibit	Eastern European Restaurant	Food Truck	Food & Drink Shop	Fish Market
3	Beacon Hill	Italian Restaurant	Pizza Place	Coffee Shop	Bakery	Donut Shop	Hotel	Park	Bar	Sandwich Place	Sports Bar
4	Brighton	Pub	Bakery	Café	Pizza Place	Coffee Shop	Deli / Bodega	Dessert Shop	Smoke Shop	Chinese Restaurant	Donut Shop

Figure 8: Dataframe showing the Neighborhood and 10 most common venues in Boston.

To cluster the neighborhoods according to most common venues, we simply applied the k-means algorithm to the venue dataset. Assuming there are 4 different clusters, the map with the clustered neighborhood looks like as follows:

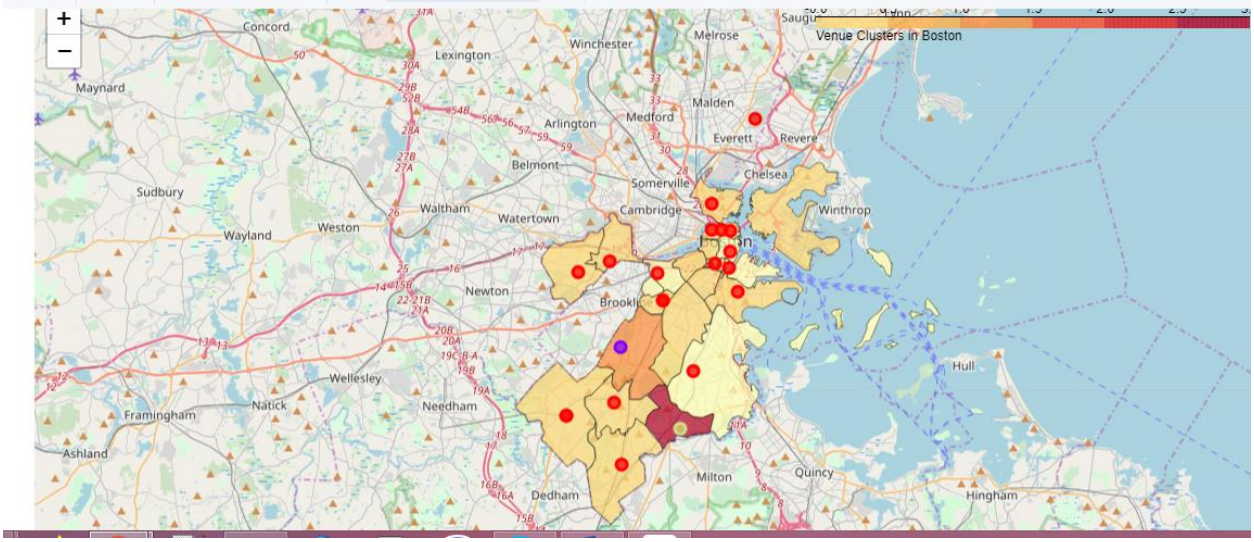

Figure 9: Choropleth map showing the Neighborhood clustering in Boston.

Different colored markers represents different cluster labels. Apparently a lot of the neighborhoods are in the red cluster. When we actually look at the red cluster, it becomes clear that the most common venues in the neighborhoods in that cluster are as:

```
In [49]: list_venue(neighborhoods_venues_sorted,0)
```

Out[49]:

	Venue Type	Count
0	Park	7
1	Liquor Store	7
2	Hotel	5
3	Bakery	5
4	Pizza Place	5
5	Coffee Shop	5
6	Café	5

Figure 10: List of top most venues in Red Clustered Neighborhood.

This clearly shows, most of the neighborhoods in Boston are full of restaurants, café, coffee shops, bars. Only few neighborhoods are a bit different like Mattapan, Jamaica Plain with common venues like Business Services, Markets.

```
In [52]: list_venue(neighborhoods_venues_sorted,3)
```

Out[52]:

	Venue Type	Count
0	Business Service	1
1	Dumpling Restaurant	1
2	Farmers Market	1
3	Fast Food Restaurant	1
4	Fish Market	1
5	Food & Drink Shop	1
6	Mattapan	1

Figure 11: List of top most venues in Mattapan Neighborhood

Most other neighborhoods are centered around eating venues like restaurants, café, coffee shops. So, in terms of availability of restaurants, our friend can choose any neighborhoods other than Mattapan and Jamaica plain.

5. Discussion

If we look upon the criteria of city with lots of restaurants, all the neighborhoods in Boston (except Mattapan and Jamaica plain) are full of restaurants, bars, café and coffee places. Based on this criteria, my friend could choose any neighborhood to settle down in Boston. Depending upon his next two criteria, neighborhood with more number of special schools and low crime rate, Central Boston neighborhood has lots of special school as compared to other neighborhoods. If we look at the crime rate reported, Central Boston neighborhood falls under the D4 Boston Police District (BPD), which is the second highest crime reported BPD. It is hard to meet all of the criteria, as there always comes a trade-off between commercialization and crime rate. More commercialization, more population, more crimes. A better place to live in, could be chosen compromising with any one of the priorities.

6. Conclusion

We have provided here an interactive analysis to meet the desired criteria to find a better place to live in Boston. According to the requirements and the analysis, we have identified Central Boston neighborhood as a good place to move into but has to understand the trade-off of high crime rate. Thus this type of neighborhood exploration and analysis can be a useful tool to recommend better places to live in, as we could anytime include other attributes or adjust the priorities of attributes too for the analysis.

7. Future Enhancements

This analysis, for recommending a better place to live in Boston has been carried out considering only the three above criteria: crime rate, schools and neighborhoods with many restaurants. Other attributes could

be considered too like housing prices, commuting time etc. Also, this analysis has been mostly carried out with categorical variables, analysis with the numerical variables might show even better results.

8. References

1. https://livability.com/best-places/top-100-best-places-to-live/2018
2. https://en.wikipedia.org/wiki/Boston