**The best place to live: Boston neighborhoods crime rate and venue data**

**1. Introduction/Business Problem**

Being inhabited with ~700,000 in 2018, Boston, Massachusetts, is the biggest city of New England and one is the most populous in the USA. Moreover, it is an outstanding educational, science, technical, cultural, health care research centre known all over the world. It makes Boston to be extremely attractive to move there for living from the other states and even abroad. So that why, Boston is pretty high-densely populated and multicultural, forming communities, spread in one or more city neighborhoods.

Different people have various criterias, except cost of living, when determine where they are going to relocate in any kind of place. It can be a nearby school rates for families with children, transport accessibility and highways proximity ratings for commuters, etc. Also a safety is a main priority for many of them, and it will be one of the aims of my research for potential future Bostonians, like my family, now discussing the possibility to relocate to Boston from northern Massachusetts.

Boston is officially divided on 23 neighborhoods, which are pretty different in areas and population. In my research I will take into account another district division using 17 neighborhoods(including Harbor Islands), due to some data limitations, which will be described in the next section.

Accessibility and proximity to the venues, like restaurants, parks, theaters, gyms are the important part of quality of life conception almost for everyone and can't be ignored during making of where-to relocation decision.

A future resident should take into consideration the factors described above and a lot more. I'm going to focus on the crime rates of Boston neighborhoods and analysis of it venues from point of view of potential Boston citizen.

I hope, that my investigation can help people, who are thinking about moving there to insight into the safety and infrastructure.

**2. Data**

The main source of the data used in my work is a portal https://data.boston.gov/. It provides a variety of data, including census, crime, climate and geospatial.

a) <https://data.boston.gov/dataset/planning-districts> geojson file was used to build neighborhoods borders on a folium map.

I've tried to use another geojson file from this site with more districts (23 as mentioned in the section above) but it created a mess in a city centre, where very small districts are located. Due to the similarity of type of buildings and being the most crowded and touristic places I decided to use the bigger map to increase a visibility

b) [https://data.boston.gov/dataset/boston-police-stations](https://data.boston.gov/dataset/boston-police-stations%20) csv file was used to obtain the numbers of police districts, which are not the same with geographical, and map it with geographical districts from the previous dataset

c) [https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system]()%20https:/data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system) csv file was used to receive the data about crimes in 2018-2019 time period

All 3 datasets were used to make a folium map with choropleth, reflecting the total of crime incidents per police districts, including neighborhoods respectively

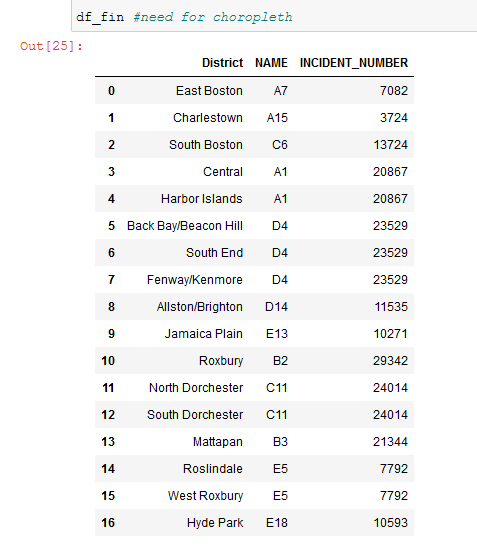
d) <https://data.boston.gov/dataset/public-libraries/resource/ce600f93-20ee-446d-89fd-0f0a92d6f24e> csv file was used to receive coordinates to use it as centers of studied neighborhoods

e) Foursquare API was used to obtain information about venues in Boston neighborhoods, like top 10 venues per neighborhood

The code was created in IBM Watson studio and submitted via Github. The choropleth map was generated using folium package.

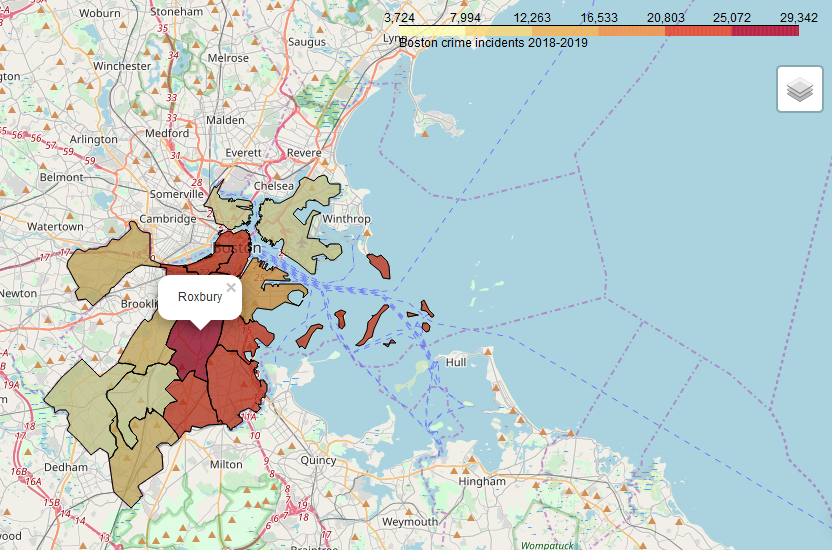
**3. Methodology**

I've started with processing the datasets, containing geodata of Boston neighborhoods ( in code it often called as districts in order to decrease entropy in spelling of this word ) , Boston police districts, which are bigger than 1 geographic neighborhood and crime incidents per police districts. After merging all of them I got the following dataframe:



As I specified earlier, the incident numbers data was bound to police district name, not geographical. So I populated the number of incidents per police district to respective geographical districts, because it could be pretty difficult to derive exact number of incident per geographical district and it's not the main purpose of my research.

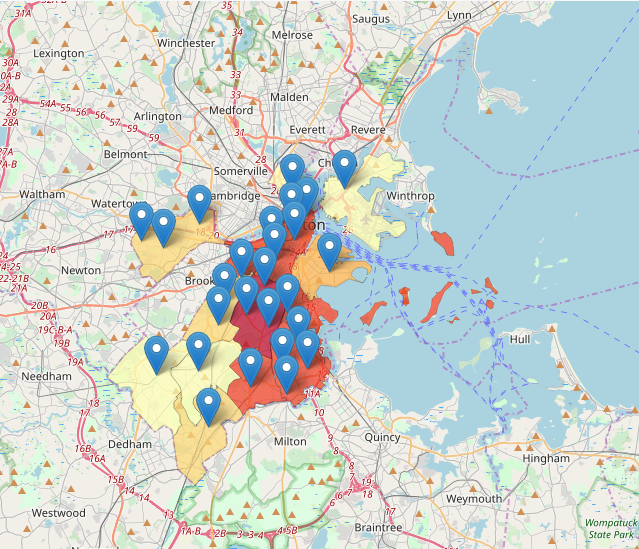
Using the resulting dataframe and already mentioned Boston districts geojson file, I've created the choropleth map:



I managed to highlight the neighborhoods borders and add pop-ups with neighborhood names for increasing map read- and visibility.

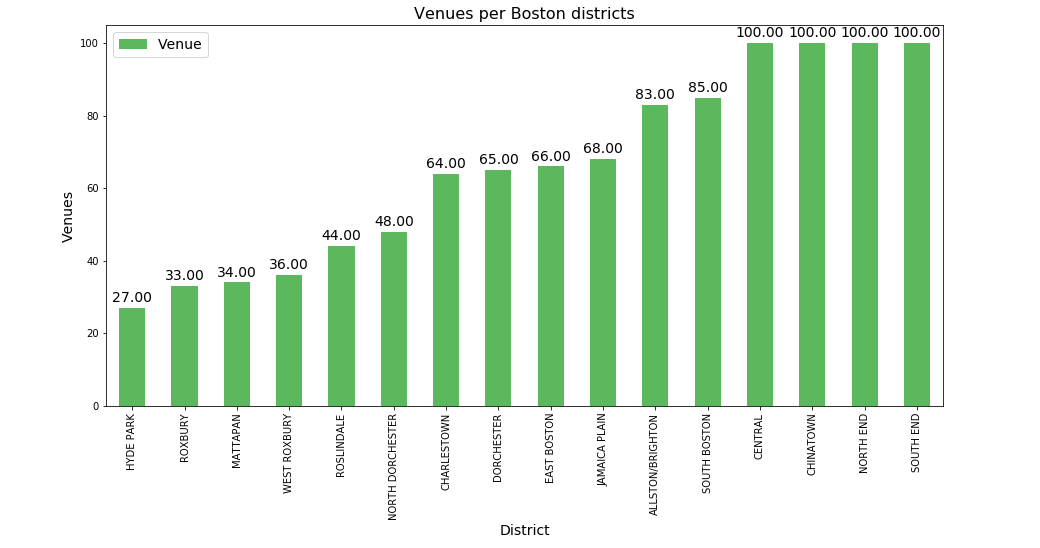
As we can see the noted Roxbury is the most criminal one. Also the number of crime incidents is high in central neighborhoods which can be explained by more populous areas.

Before the next part - to determine the "infrastructure" features of considered neighborhoods - I needed to get the coordinates in every districts, which can be used for retrieval venue data. I decided to obtain such coordinates from the dataset with community centers of Boston neighborhoods. I visualized it on choropleth map as markers:



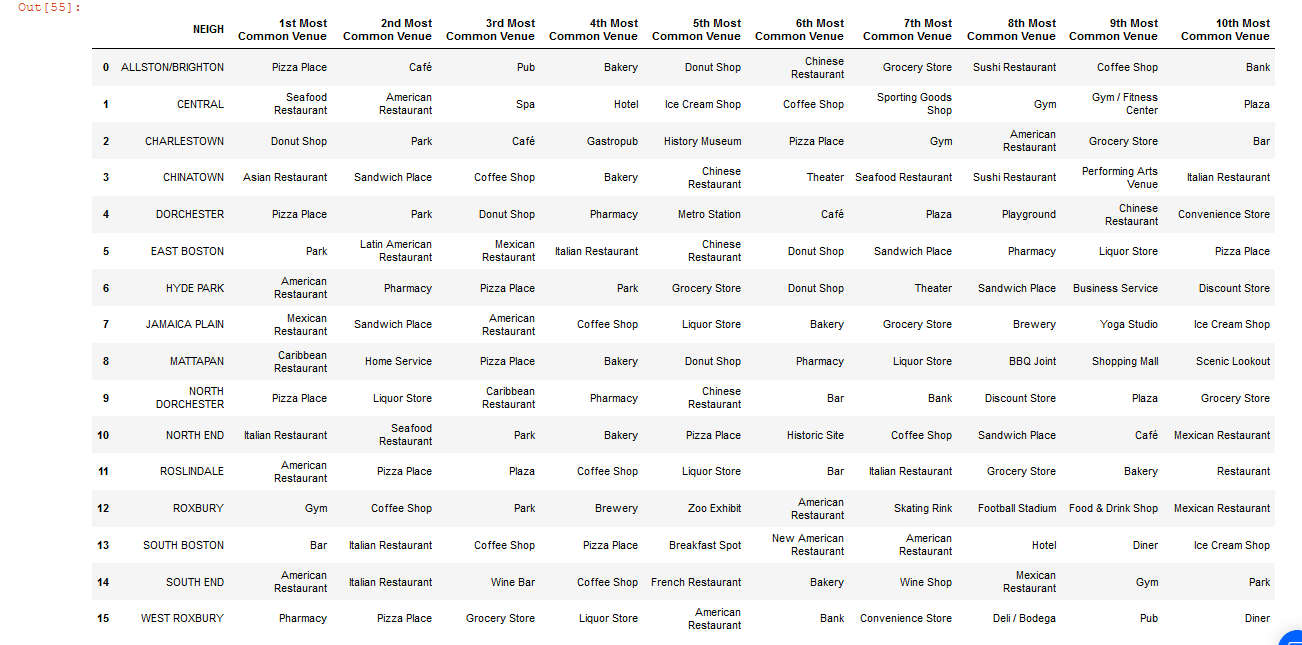
and then removed extra, leaving only one per neighborhood, locating about in its centers.

I fully understand, that it's not optimal and dynamic method, and it could be a good opportunity for my future study, like measuring of polygon centers.

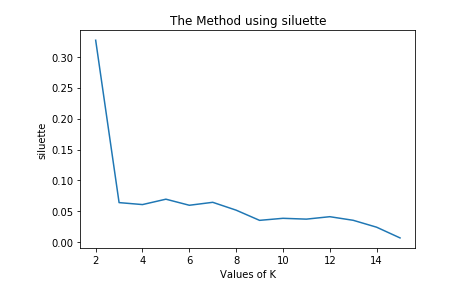
Foursquare API with passed limit venue = 100 and radius = 1000 meters provides the following results for neighborhoods: 

As expected, the neighborhoods, reached the limit of 100 are the central areas with a lot of touristic attractions. Whereas the more remote areas, like Hyde Park, Mattapan, Roslindale has lot less. Together with Roxbury, but do you remember about its crime rate? I think, it can be a root cause...

In summary, the returned venues belongs to 199 categories. Let identify top 10 the most popular per district:



The neighborhoods have a lot common venues, so I can combine it, using k-mean unsupervised learning algorithm, into clusters, which should help me to form the opinion about of the range of suitable neighborhoods, not only looking in every particular one.

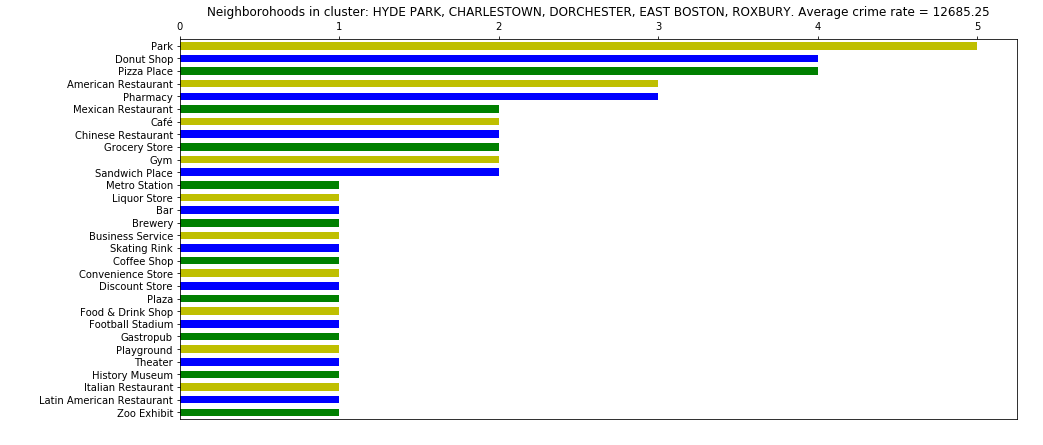
Firstly, I used the most common tool for identifying the number of cluster - elbow method. I tried it, measuring both entropy and distortion parameters, but the results was unclear, and I couldn't even see the expected "elbow". It can be explained by the limitations of data, maybe I need more venue data and more neighborhoods should be taken into account (like mentioned 23, not 16), but I decided to use another method - " silhouette" and it showed things more clear:

I took 4 as a k value, because the graph line is slightly lower than against 3, and the lowest 9 is too much for just 16 input points.

The algorithm provides 4 clusters, which I visualized and analyzed. Let look closer on its:

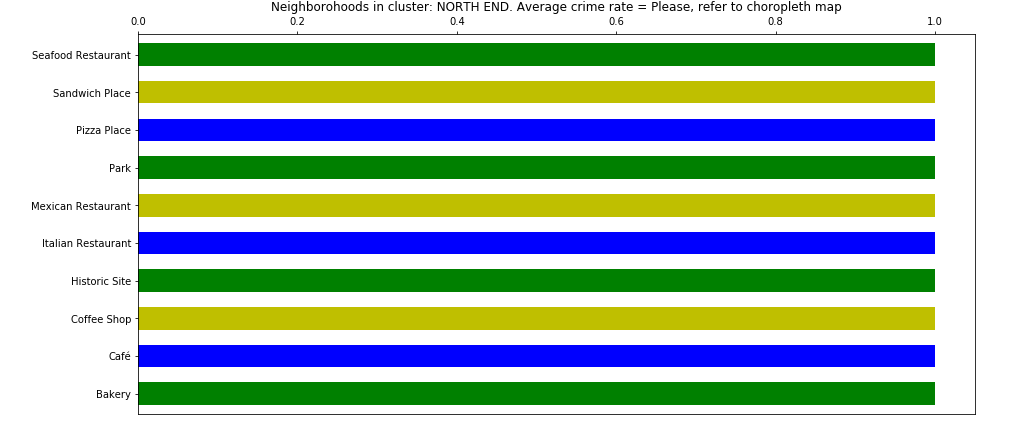
**4. Results:**

Cluster 0:



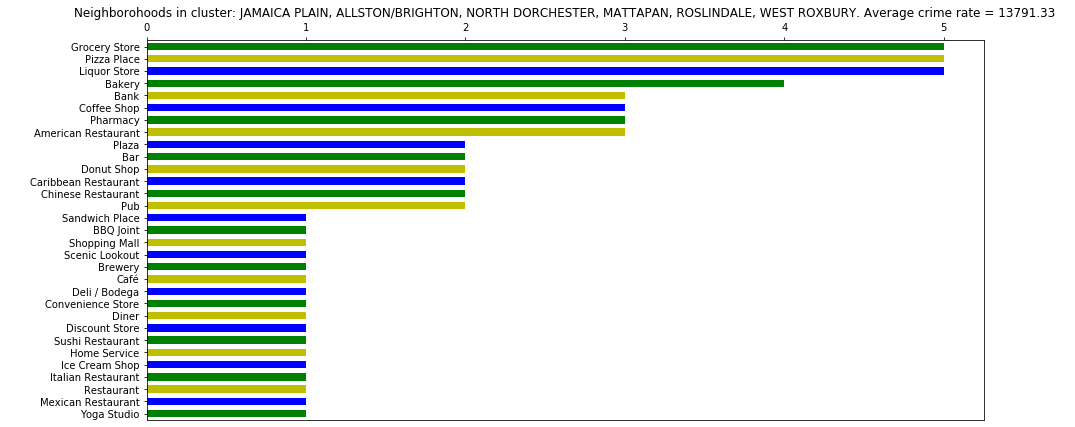
This cluster has the highest number of parks, and a lot of place to eat. Also, a metro station is among popular venues and places to spend free time, like zoo, museum, theatre, football stadium, skating rink. Store and shops, even plaza are there. There are restaurants of different cuisines. It has the lowest average crime rate, in spite of the most criminal Roxbury, which can be explained by low crime rate of East Boston and Charlestown. Looks very convenient to live. I marked it as "Nature/Active outdoor/Shopping/VarietyFood" cluster.

Cluster 1:



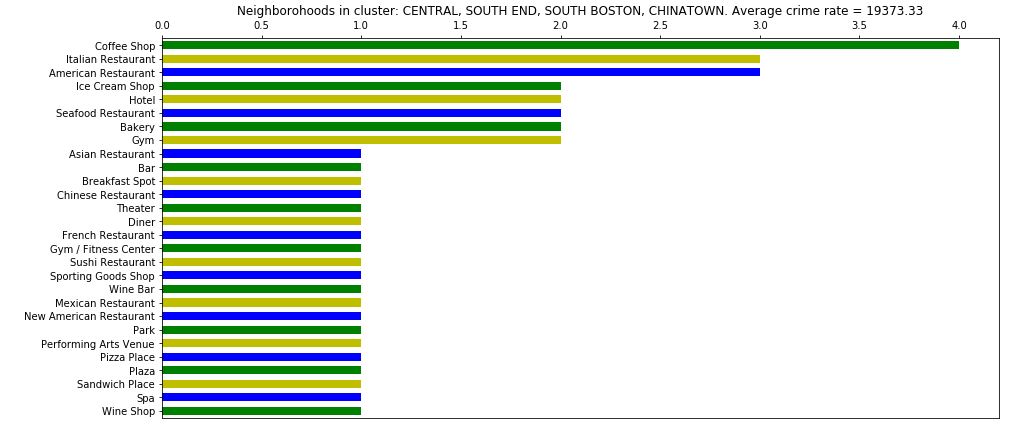
North End was always a lone member of the separate cluster. It can be explained by its busy location near the main Boston transport flow system. It's a historical site, as the same venue displays, with a lot of tourist and point of interests for them. The crime rate is a typical for dense populate city center - between 20 - 25 thousand incident in 2 past years. The most of the venues are restaurants. Taking into account the previous fact, the park and reach history of place, let mark it as "Sights/Dinner"

Cluster 2:



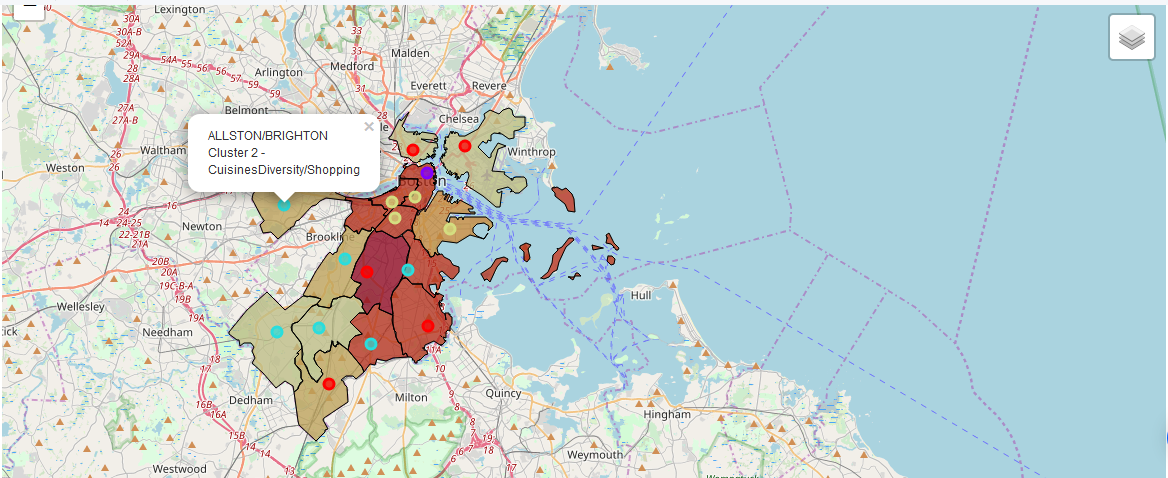
The dominant venues of the cluster are food places, with worldwide national cuisines. Some plazas, pharmacies and mall give an opportunity for shopping and commercial activities together with banks. The average crime rate is a bit higher than in cluster 0. I marked the cluster as "CuisinesDiversity/Shopping"

Cluster 3:



With variety of restaurants, park, theater, art center, gyms and hotels the clustered area looks extremely vibrant and attractive, but locals may suffer due to lack of shopping places, like grocery stores, pharmacies, etc. It's central Boston area with dense population, a lot of tourists and a higher average crime rate. But entertainments are in close proximity! It was marked as "Culture/Food/Center".

All clusters and their labels were visualized on the previously created choropleth map:



And looking in summary, the clustering results look enough convincing: a crime rate, initially separated and not taken part into clustering, distributed evenly between the cluster, relatively their locations central or remote. Each cluster contains life-necessary venues, which can be taken into account under further analysis of particular neighborhood during making a decision of potential moving in Boston.

**5. Discussion and conclusion:**

As stated at the beginning, Boston is a place for living for more than a half million of people and one is the most touristic city in the USA. Choosing a place to live there isn't simple and varies from person to person, taking into consideration a bunch of different circumstances.

During my research, I formed my own view of the city, relying on provided crime rate of the neighborhoods and venue data, clustered across them.

Choropleth map shows descriptive visualization of safety feature of Boston. Also it gives an introduction of the city structure, like probable density of population distributed between districts. For me I decided to avoid Roxbury (because of its crime rate) and central neighborhoods(dense population and touristic flows).

The latter was primarily derived from cluster analysis, which displayed the distribution of different kind of the venues and combines venue-similar districts in order to narrow down possible alternatives and enlighten the choice.

In the aftermath of the research I knew lot more about Boston neighborhoods and started to review such of them, like Hyde Park or Charlestown, as probable place to move.

I hope that my work, having published, will help people to explore Boston and assist to make a successful relocation decision.