IBM – Coursera

Data Science Specialization

Capstone project

Finding location for residential construction investment in Boston

Alexander Iliyashov

Introduction of Business problem:

The aim of the project is to evaluate areas of Boston city and find attractive locations for residential construction. My customer has a strong believe that safety and comfort of the area (availability of infrastructure and entertainment objects) could increase interest to real estate from potential high-income buyers. The main idea of the project will be to cluster Boston city based on the amount of venues and number of criminal cases. This clustering should be used only as a complementary analysis to other factors evaluation that could potentially impacts property prices and attractiveness.

Boston is a capital of USA state Massachusetts. Population of ca 700'000 and city square of 232km^2 , thereof 125km^2 of land area.

Boston is a technological, political, financial, scientific and educational center. It attracts high-paid talents from around the world and it is good candidate for piloting of our methodology.

Data description:

The following data sources were used in this project:

1. ***Crime incident reports*** are provided by Boston Police Department (BPD) to document the initial details surrounding an incident to which BPD officers respond.
   Link:
   https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system
   Data structure:

Field Name, Data Type, Required	Description
[incident_num] [varchar](20) NOT NULL,	Internal BPD report number
[offense_code][varchar](25) NULL,	Numerical code of offense description
[Offense_Code_Group_Description][varchar](80) NULL,	Internal categorization of [offense_description]
[Offense_Description][varchar](80) NULL,	Primary descriptor of incident
[district] [varchar](10) NULL,	What district the crime was reported in
[reporting_area] [varchar](10) NULL,	RA number associated with the where the crime was reported from.
[shooting][char] (1) NULL,	Indicated a shooting took place.
[occurred_on] [datetime2](7) NULL,	Earliest date and time the incident could have taken place
[UCR_Part] [varchar](25) NULL,	Universal Crime Reporting Part number (1,2, 3)
[street] [varchar](50) NULL,	Street name the incident took place

   This data will be used to generate statistic about Boston crime rate in the different neighborhoods.

2. **US Zip codes public open data**

   Link:
   https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/export/?q=boston

   Data structure:

   Json file including information about Boston zip codes and its geometry (coordinates).

   This data will be used for dividing Boston in different neighborhoods.

3. **Foursquare venue data:**

   This data will be used for analysis and classification of neighborhoods based on venues availability.

4. **Nominatim geolocator** was used to find zip codes of crime locations

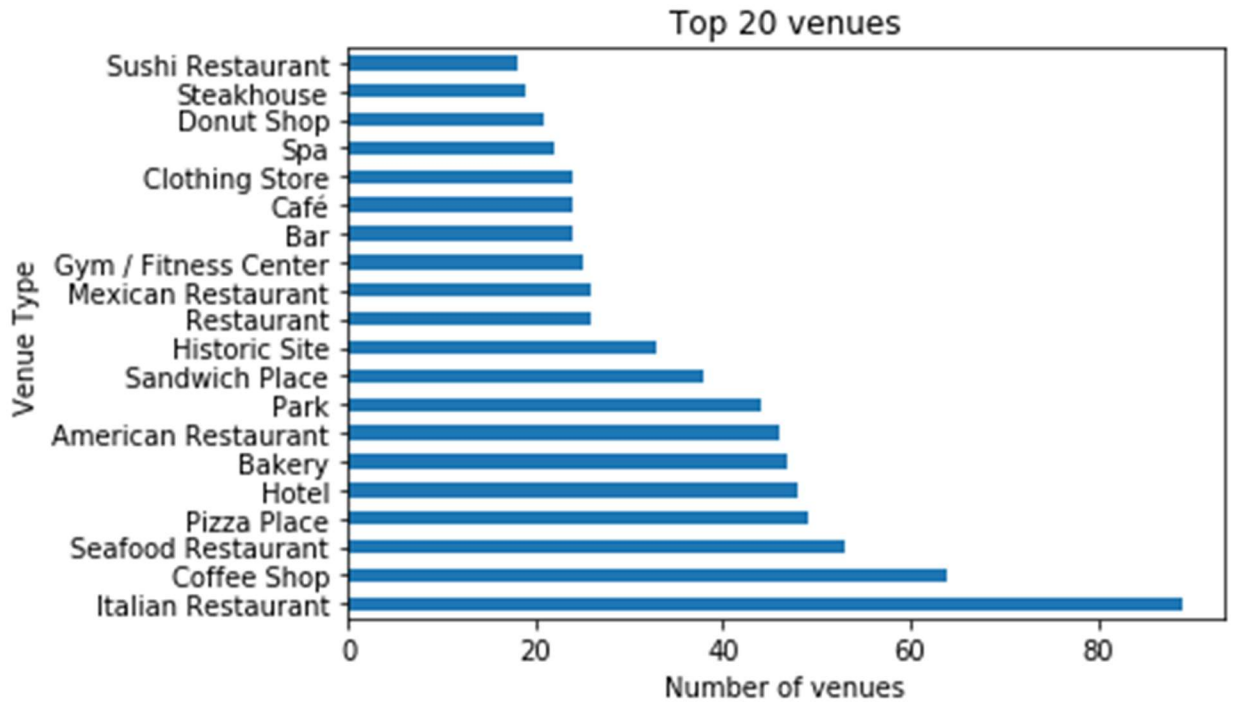
Methodology:

The main idea of analysis is to use k-means clustering to identify clusters with the most developed infrastructure. Then to analyze criminal statistic in each cluster and identify neighborhoods with developed infrastructure and low crime incidents rate. Venue and crime data will be normalized across all Boston neighbourhoods to make it comparable. Before implementation of k-means model, exploratory data analysis will be presented.

3.1 Exploratory Data Analysis

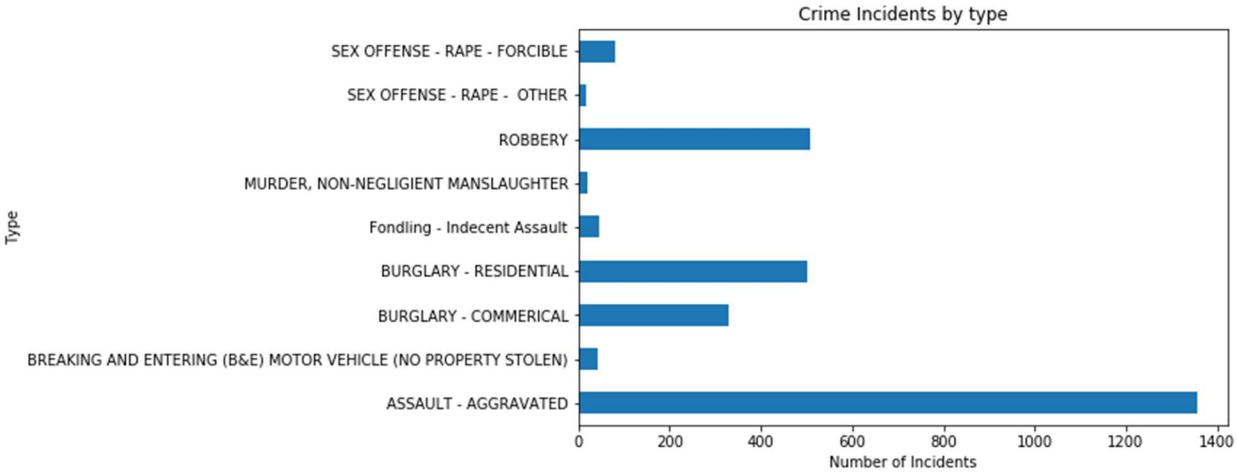
Venues

Total number of venues provided by Foursquare database is ~1’600. Majority of the venues were related to food category.



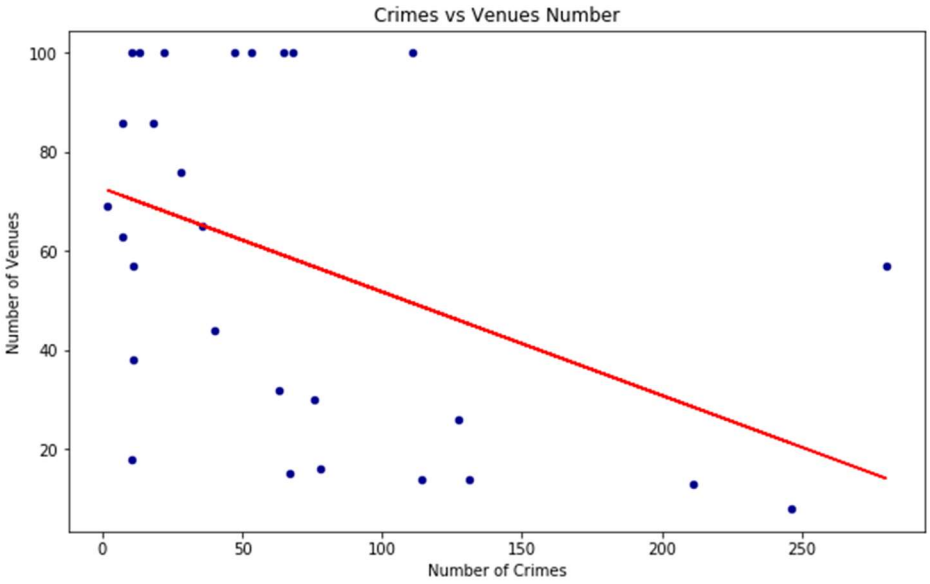
Crime types

Below you can find exploratory analysis of Boston crimes. Total number of the criminal incidents for first seven month of 2020 is ~2’900. Majority of the incidents belong to Assault, Burglary and Robbery categories

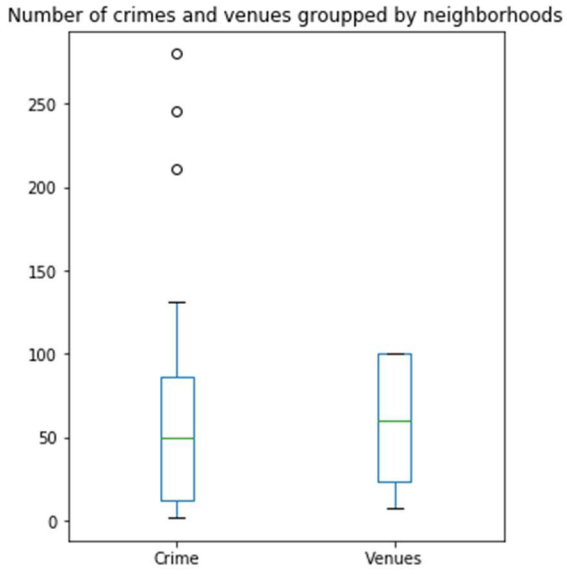


Venues vs Crimes

Based on the scatter plot analysis we can see negative correlation between number of Crimes and Number of Venues grouped by neighborhoods.



Number of venues limited by 100, hence we can see wider distribution of crime cases.



Based on exploratory data analysis we can assume that there are neighborhoods with low criminal rate and developed infrastructure (top left corner of scatter plot). K-means analysis will be implemented to identify these neighborhoods.

3.2 Data preparation

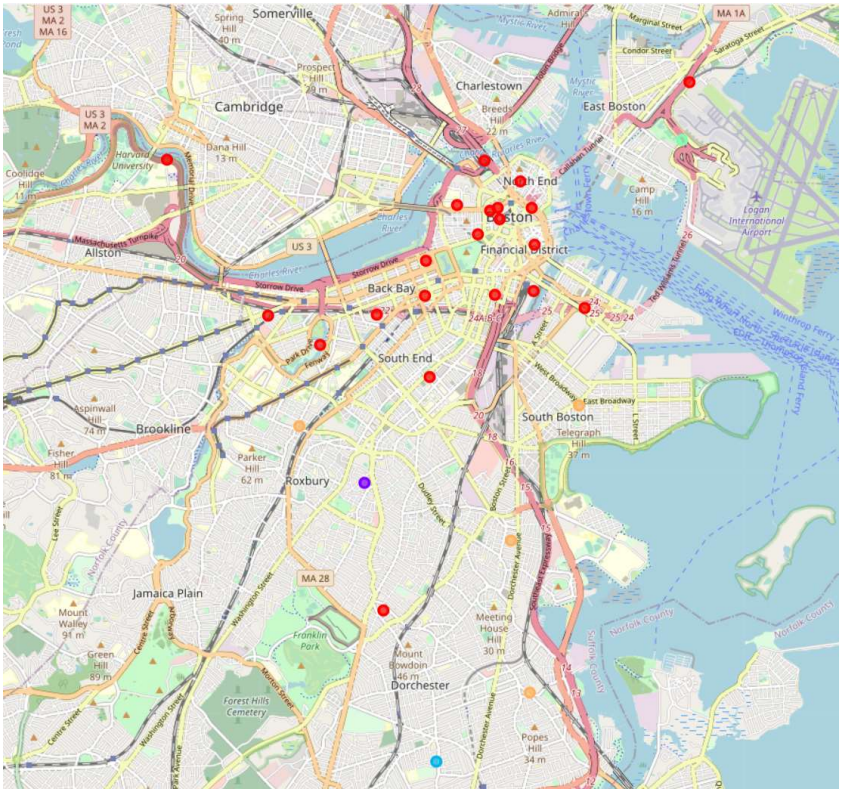
Zip code open data source was used to split Boston to neighbourhoods. During data analysis, one outlier was detected (zip code 02212). Further analysis showed that database had wrong coordinates of the zip code. The coordinates were checked in Google map and manually corrected in the Data Frame.

To analyze criminal statistic in Boston, the most recent criminal data from year 2020 was used and I selected most relevant incidents type groups: killing, rape, robbery, assault, burglary. Zip code data was missing so geolocator Nominatim was used to identify zip codes based on coordinates of the each incident.

Venues information by neighborhoods was retrieved from Foursquare to analyze neighborhood's infrastructure. Request was limited to 100 venues per location in the distance of 500m.

3.3 Clustering

5 clusters K-means neighborhoods classification was implemented based on the venues information.

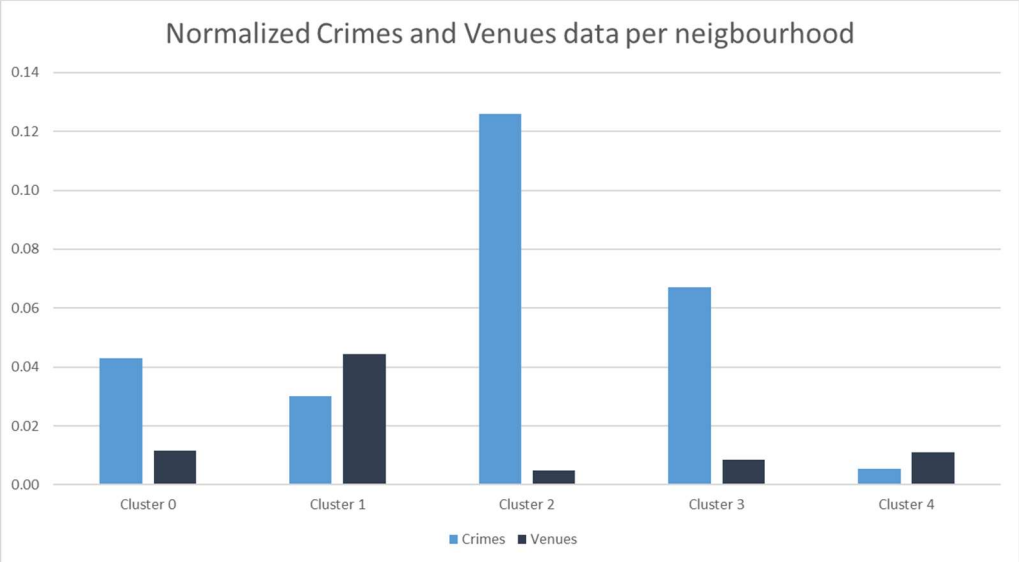


Results analysis:

4.1 Identification of potential target clusters

5 clusters generated by the model have relatively different quantity of neigbourhoods. Two groups of clusters could be identified:

- **Group 1** (Cluster 0 and Cluster 1) have several neigbourhoods, 4 and 21 respectively. My recommendation to make detailed analysis of these clusters.
- **Group 2** (Cluster 2 to Cluster 4) have only one neigbourhood per cluster. All of the clusters have low number of venues. Besides, Cluster 2 and Cluster 3 have high crime rate and both of them will be excluded from our analysis. Cluster 4 could be potentially targeted but it is located very close to airport and I would not recommend investing in this area due to potential noise from the flights.



4.2 Cluster 0 analysis (orange color)

Situated mainly in the South of Boston in the distance of 1-2km from city center. Mainly low-rise residential area with bars, small restaurants and convenience stores.

	Neighborhood	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
19	02122	Pet Store	Donut Shop	Liquor Store	Dog Run	Sandwich Place	Arts & Crafts Store	Chinese Restaurant	Bank	Seafood Restaurant
24	02125	Pub	Bar	Pizza Place	Caribbean Restaurant	Fried Chicken Joint	Deli / Bodega	Yoga Studio	Pharmacy	Convenience Store
27	02120	Pizza Place	Sushi Restaurant	Liquor Store	New American Restaurant	Gym	Furniture / Home Store	Burger Joint	Art Gallery	Convenience Store
52	02127	Bowling Alley	Convenience Store	Sports Bar	Pizza Place	Cosmetics Shop	Juice Bar	Dive Bar	Liquor Store	New American Restaurant

Relatively low number of venues and average crime rate

Boston average crime rate per neigbourhood 0.04%

Number of Crimes and Venues in Cluster 0

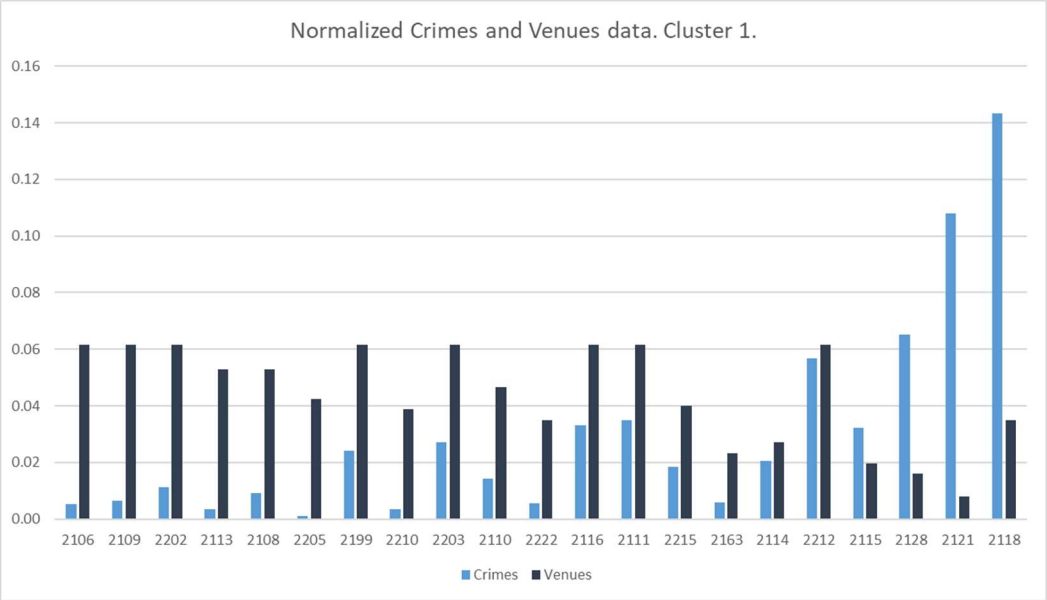
Neighborhood	Crimes	Venues
2127	76	30
2120	67	15
2125	78	16
2122	114	14
Total Cluster 0	**335**	**75**

4.2 Cluster 1 analysis (red color)

Situated mainly in the North of Boston in the Boston city center, close to the city center locations and close to Cambridge area. Mainly high-rise housing area with developed infrastructure: restaurants, shopping malls, banks, offices, hotels, historical places, gyms and Spa centers, etc.

	Neighborhood	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	02111	Asian Restaurant	Bakery	Theater	Sushi Restaurant	Bubble Tea Shop	Café	Coffee Shop	Hotel Bar	Performing Arts Venue
4	02202	Hotel	Seafood Restaurant	Coffee Shop	Mexican Restaurant	Bar	Bakery	Restaurant	American Restaurant	Sandwich Place
6	02109	Seafood Restaurant	Bakery	Historic Site	Park	Café	Pizza Place	Coffee Shop	Hotel	Sandwich Place
9	02110	Seafood Restaurant	Historic Site	Park	Coffee Shop	Salad Place	Harbor / Marina	Burger Joint	Steakhouse	Asian Restaurant
10	02121	Shopping Mall	Pharmacy	Nightclub	Fish & Chips Shop	Men's Store	Farmers Market	Supermarket	Caribbean Restaurant	Donut Shop
11	02113	Pizza Place	Park	Coffee Shop	Seafood Restaurant	Hotel	Café	Sandwich Place	Outdoor Sculpture	Mexican Restaurant
12	02108	Restaurant	Pizza Place	Italian Restaurant	New American Restaurant	American Restaurant	Historic Site	Falafel Restaurant	Plaza	Steakhouse
13	02212	Coffee Shop	Historic Site	Sandwich Place	Italian Restaurant	Bakery	New American Restaurant	Restaurant	American Restaurant	Hotel
15	02203	Historic Site	Seafood Restaurant	Bakery	Coffee Shop	Hotel	Sandwich Place	Park	Salad Place	Pub
17	02128	Chinese Restaurant	Brazilian Restaurant	Metro Station	Liquor Store	Latin American Restaurant	Plaza	Mexican Restaurant	Donut Shop	Diner
20	02215	Café	Lounge	Pizza Place	American Restaurant	Donut Shop	Greek Restaurant	Bakery	Sports Bar	Furniture / Home Store
22	02205	Italian Restaurant	Sandwich Place	Bakery	French Restaurant	American Restaurant	Dive Bar	Cocktail Bar	Food Truck	Museum
23	02222	Hotel	Pizza Place	Sports Bar	Bar	Italian Restaurant	Donut Shop	Harbor / Marina	Gastropub	Brewery
28	02118	Donut Shop	Thai Restaurant	Café	Pizza Place	Coffee Shop	Sporting Goods Shop	Deli / Bodega	Mexican Restaurant	Mediterranean Restaurant
29	02116	American Restaurant	Seafood Restaurant	Italian Restaurant	Gym	Women's Store	Clothing Store	Sandwich Place	Gym / Fitness Center	Cosmetics Shop
30	02115	Art Museum	Coffee Shop	Garden	Sandwich Place	Korean Restaurant	Baseball Field	Lounge	Middle Eastern Restaurant	Burger Joint
35	02210	Italian Restaurant	American Restaurant	Hotel	Coffee Shop	Seafood Restaurant	BBQ Joint	Sandwich Place	Bar	Steakhouse
39	02106	Spa	Jewelry Store	Hotel	Cosmetics Shop	Coffee Shop	Boutique	American Restaurant	Gym / Fitness Center	Breakfast Spot
40	02199	Coffee Shop	Italian Restaurant	Ice Cream Shop	Hotel	Seafood Restaurant	American Restaurant	Spa	Women's Store	Grocery Store
41	02163	Park	Gym / Fitness Center	Japanese Restaurant	Pizza Place	Café	Seafood Restaurant	Breakfast Spot	Men's Store	Burger Joint
53	02114	American Restaurant	Food Truck	History Museum	Pizza Place	Playground	Italian Restaurant	Gourmet Shop	Restaurant	Bistro

Relatively high number of venues and varied crime rate



Number of Crimes and Venues in Cluster 1

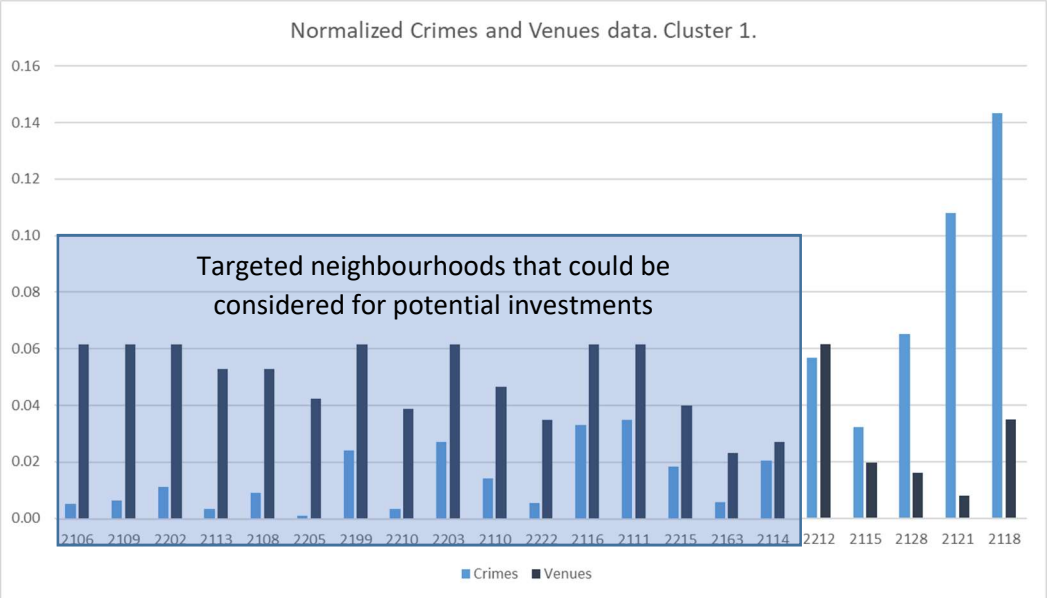
Neighborhood	Crimes	Venues
2106	11	100
2109	13	100
2202	22	100
2113	7	86
2108	18	86
2205	2	69
2199	47	100
2210	7	63
2203	53	100
2110	28	76
2222	11	57

Neighborhood	Crimes	Venues
2116	65	100
2111	68	100
2215	36	65
2163	11	38
2114	40	44
2212	111	100
2115	63	32
2128	127	26
2121	211	13
2118	280	57
Total Cluster 1	**1231**	**1512**

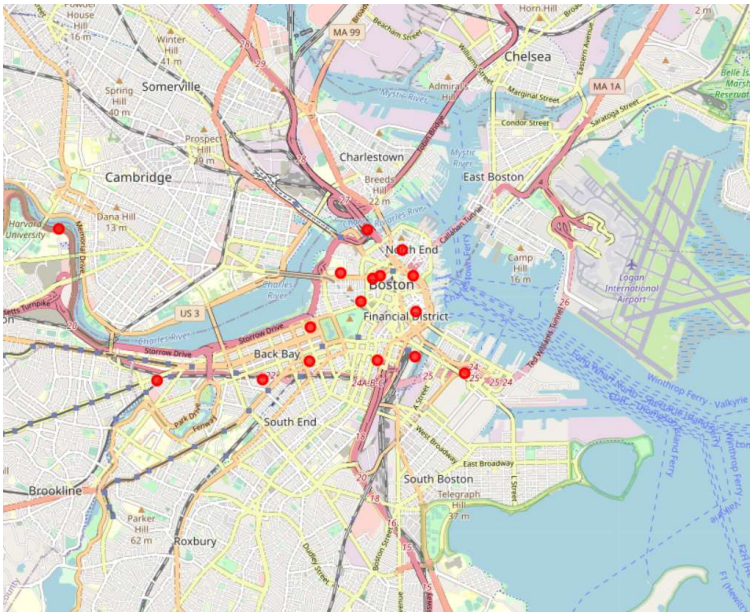
Discussion and recommendations:

My recommendation to invest in the neigbourhoods selected on the next chart.

Selection was based on the relative ratio of venues vs crime rate.



All selected neigbourhoods located in the Cluster 1 and close to Boston city center.



Despite of the good results and clear recommendations, analysis could be further develped and enchanced in the following areas:

1. Enhance analyzed data by
    a. Retrieve more than 100 venues per location
    b. Increase targeted area and include in to analysis Boston surroundings
    c. Test methodology with another big city
2. Change Nominatim for paid geolocator to speed up identification of coordinates
3. Incorporate crime data in k-means analysis and set up one k-means model
4. Analyze further features that could impact attractiveness for investments (hosing price, public transport availability, etc.)

Conclusion:

In my study I analyzed neigbourhoods of Boston city. K-means combined with exploratory analysis was applied to identify potentially attractive neighbourhoods for investments in residential contruction. Potential attractive locations were successfully identified.

This model should be used as one of the component of the more complex analysis that should include other relevant to investment attractiveness parameters.

Besdies, this model could be applied for other location to solve similar problems.