Coursera Capstone Project

IBM Applied Data Science Capstone

Picking Neighborhood for Prospective Resident of Boston

By: Chang Liu

January 2020

Introduction

This report is for the IBM Capstone Project of Applied Data Science. A 9-courses series created by IBM, hosted on Coursera platform. The problem and the analysis approach are left for the learner to decide, with a requirement of leveraging the Foursquare location data to explore or compare neighborhoods or cities of your choice or to come up with a problem that you can use the Foursquare location data to solve.

The main goal will be exploring the neighborhoods of Boston in order to extract the correlation between the average rent and its number of surrounding venues. In other words, a hypothesis that more surrounding venues will lead to a high average rent price for a neighborhood will be tested.

Besides, neighborhoods in Boston will be clustered based on similarities of the most common top 10 venues in each Boston neighborhoods.

The idea comes from the process of a normal family finding a place to stay before moving to another city. It is common that the owners or agents advertise their properties are closed to some kinds of venues like supermarkets, restaurants or coffee shops, etc.; showing the "convenience" of the location in order to raise their house's value.

So, can the level of "convenience" (the amount of venues in a neighborhood) affect the price of the rent? If so, how strong this relationship between the number of venues and rent price will be?

Target Audiences of this project are listed as below:

1. Prospective residents for Boston who want to pick certain neighborhood for staying with concerns of the conveniences and rent price.
2. Current residents who want to pick a similar neighborhood for moving into a new place.
3. Real estate companies who want to optimize their advertisement and provide suggestions for helping their customers pick a new or a similar neighborhood with the neighborhood where they are living for now.

Business Problem

The objective of this capstone project is to guide potential residents of Boston to pick their most suitable neighborhood before moving in. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In the city of Boston, MA, US, if a potential resident for Boston is looking for renting a apartment, where should we suggest him/her to rent in terms of the level of “convenience ” of certain neighborhood and how many alternative neighborhoods are there based on the neighborhood groups?

Data

To solve the problem, we will need the following data:

- List of neighborhoods in Boston.
- Latitude and longitude coordinates for each neighborhood in Boston.
- The average rent price for each neighborhood in Boston.
- Venue data for each neighborhood in Boston.

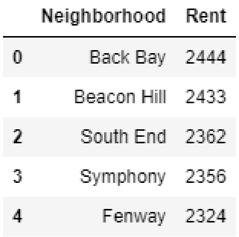
Sources of data and methods to extract them:

1. The data of a list of neighborhoods in Boston can be easily extracted in a type of a CSV file from “https://data.boston.gov/dataset/boston-neighborhoods/resource/c46fae56-956b-44c1-9454-c16cc2ddf270”
2. Then the latitude and longitude of each neighborhoods can be obtained by using Python Geocoder package.
3. The average rent price for each neighborhood in Boston can be web scraped from “https://bostonpads.com/average-rent-prices-boston-by-town/”. In this project, only the table of one-bedroom rent price will be considered.
4. Venue data of each neighborhood in Boston can be obtained by taking the power of Foursquare API.

Methodology

1. **Collecting data from internet**

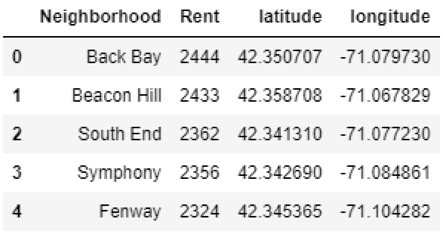
   First, we downloaded data from “https://data.boston.gov/dataset/boston-neighborhoods/resource/c46fae56-956b-44c1-9454-c16cc2ddf270”, and then web scraped data from “https://bostonpads.com/average-rent-prices-boston-by-town/”. After merging data from two online source, a head of a data frame can be shown as below:



	Neighborhood	Rent
0	Back Bay	2444
1	Beacon Hill	2433
2	South End	2362
3	Symphony	2356
4	Fenway	2324

2. Adding geo-location data based on the neighborhood

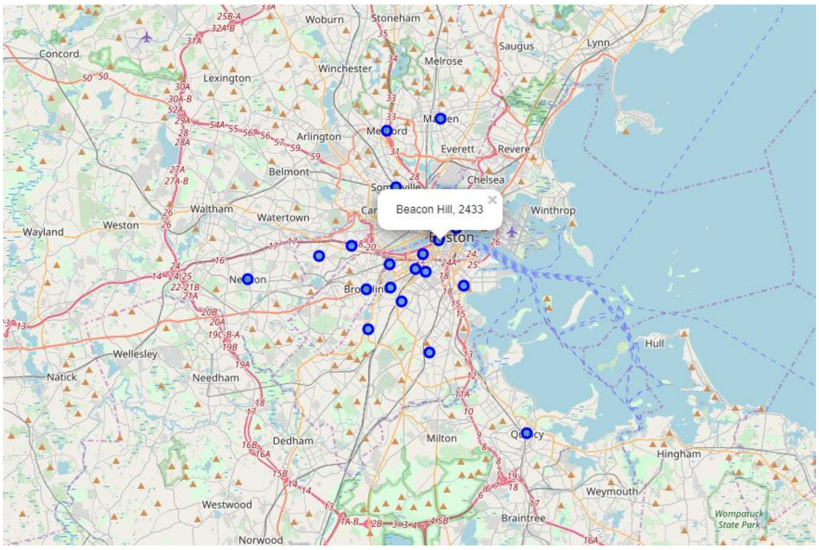
Here we utilized Python Geocoder Package to obtain latitude and longitude for each neighborhood in the list. A head of the data frame with geo-location data is shown as below:



	Neighborhood	Rent	latitude	longitude
0	Back Bay	2444	42.350707	-71.079730
1	Beacon Hill	2433	42.358708	-71.067829
2	South End	2362	42.341310	-71.077230
3	Symphony	2356	42.342690	-71.084861
4	Fenway	2324	42.345365	-71.104282

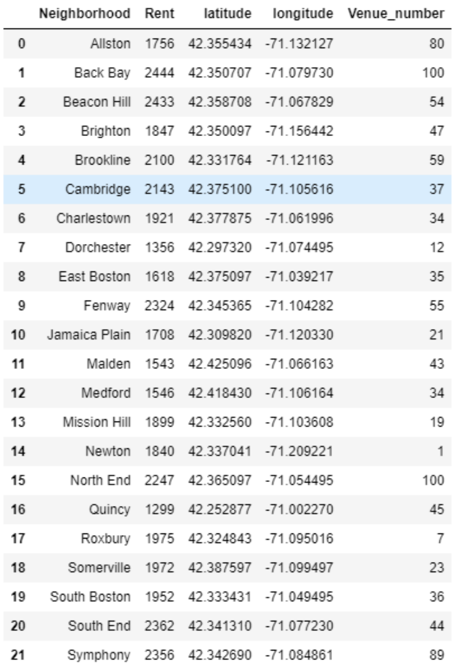
3. Generate a map

Here Folium Package was used to generated map and all popup labels with popup markers of name of Neighborhood and rent price.



4. Explore Venue Data

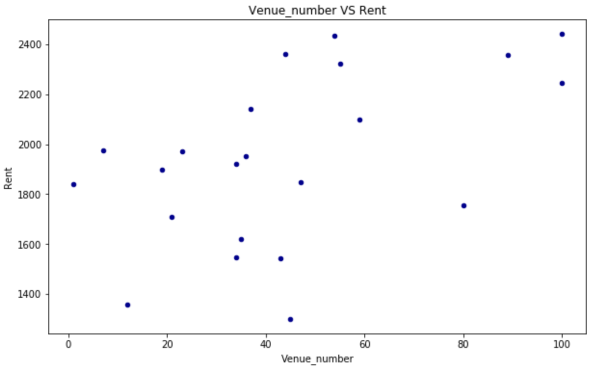
Here Foursquare API was used to obtain venue data for each neighborhood. After that, we used a “get category” function to extract the more information from the venue data belonging to its neighborhood. Then, we also count the number of venues that each neighborhood has and put the result into the data frame.

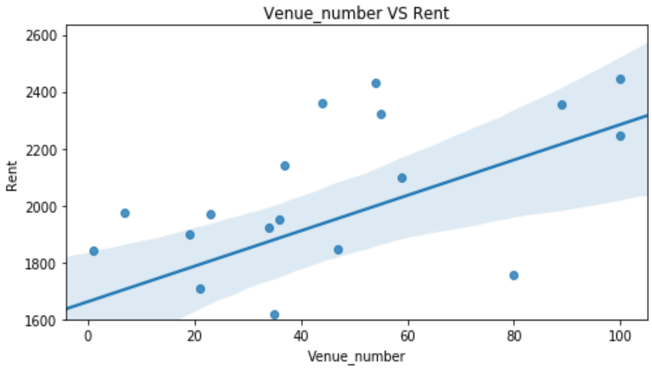


	Neighborhood	Rent	latitude	longitude	Venue_number
0	Allston	1756	42.355434	-71.132127	80
1	Back Bay	2444	42.350707	-71.079730	100
2	Beacon Hill	2433	42.358708	-71.067829	54
3	Brighton	1847	42.350097	-71.156442	47
4	Brookline	2100	42.331764	-71.121163	59
5	Cambridge	2143	42.375100	-71.105616	37
6	Charlestown	1921	42.377875	-71.061996	34
7	Dorchester	1356	42.297320	-71.074495	12
8	East Boston	1618	42.375097	-71.039217	35
9	Fenway	2324	42.345365	-71.104282	55
10	Jamaica Plain	1708	42.309820	-71.120330	21
11	Malden	1543	42.425096	-71.066163	43
12	Medford	1546	42.418430	-71.106164	34
13	Mission Hill	1899	42.332560	-71.103608	19
14	Newton	1840	42.337041	-71.209221	1
15	North End	2247	42.365097	-71.054495	100
16	Quincy	1299	42.252877	-71.002270	45
17	Roxbury	1975	42.324843	-71.095016	7
18	Somerville	1972	42.387597	-71.099497	23
19	South Boston	1952	42.333431	-71.049495	36
20	South End	2362	42.341310	-71.077230	44
21	Symphony	2356	42.342690	-71.084861	89

5. **Explore relationship between the number of venues and the rent price**

Before exploring the relationship, we make a simple hypothesis that more venues will lead to higher price of rent because more venues can indicate a higher level of convenience. Thus, we applied a linear regression method to verify our hypothesis.

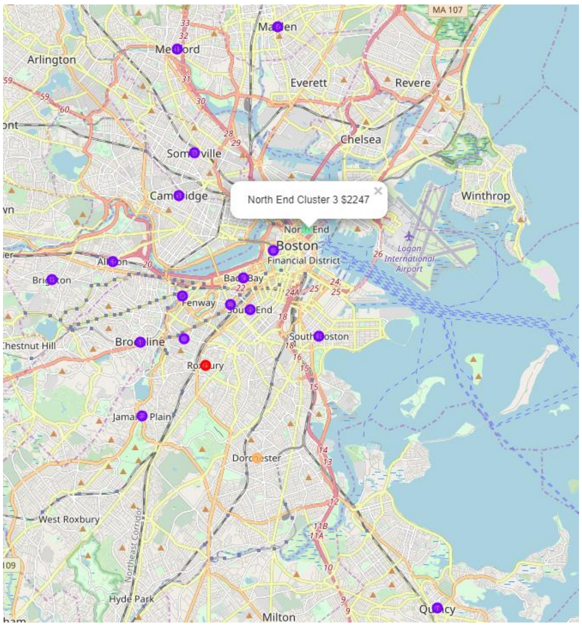




This linear regression does show that the price of rent has a linear relationship with the number of venues, but this relationship is not strong. The weakness of this linear relationship can also be reflected by a low R^2 value, 0.26.

6. **Neighborhood Clustering**

We followed the instructions in the course. First, we applied one-hot encoding to treat our data and clustered our neighborhood into 5 clusters based on the most common top 10 venues of each neighborhood. The result is shown below as a map with popup labels showing special colors indicating cluster label. The popup markers also include the information of neighborhood name, cluster label and rent price.



Result and Discussion

In this section, we can discuss some results we got from our project. First, our linear regression plot which is already shown above indicate that a neighborhood with a large number of venues tends to have a higher rent price for an apartment, but this relationship is weak which can also be reflected by a small R square value of only 0.26. This might be because we do not actually take the categories of venues into account.

From the machine learning for neighborhood clustering, we can easily see that most of neighborhoods belong to cluster 2. This means that it has a large possibility for someone to find a similar neighborhood with where they are living now.

Conclusion

In this report, we explored the relationship between rent price and the number of venues for neighborhoods in Boston. The weak linear relationship suggests that the model might need the information with venue categories. Besides, from the neighborhood clustering result, we can see most of neighborhoods having similar venues which indicates a similar level of convenience. However, the exact pattern for the cluster segmenting is not very clear.