Analysing Residential Neighborhoods in Boston

Applied Data Science: Capstone Project Report

Sudhir Nair, May 2021

Introduction

Moving to any new city can be quite daunting. Finding a place to live is among the many challenges that newcomers to a city face. It starts with identifying the right area to search for a house. There can be multiple factors that drive this decision but generally people would prefer a neighborhood that allows them access to most amenities while being affordable in terms on the rent. This selection can get quite complicated when one has very little understanding about the new city. My project is attempting to layer in data science to help make this decision making easier. Boston has been considered as the target city in this project but the principles can be replicated for any city of choice.

Boston - The City of Neighborhoods

Boston is a diverse city and the capital of the state of Massachusetts in US. The city covers 48.4 square miles (125 km2) with an estimated population of 692,600 in 2019. Boston is sometimes called a "city of neighborhoods"

because of the profusion of diverse subsections; the city government's Office of Neighborhood Services has officially designated 23 neighborhoods. In this project we will be using the neighborhood definition as per Analyze Boston database.

In this project we will start with ranking of neighborhoods in Boston based on the rental prices and then layer in data of venues in the surrounding to build a blended scoring matrix that allows us to see which neighborhoods offer advantage of having numerous venues in the vicinity while still being affordable.

This analysis will be of interest to anyone looking to move to a new city as it'll help shortlist potential areas that can be considered for house-hunting. Apart from potential residents, this analysis can also be of interest to real estate developers as it could highlight areas where people are interested in living and also areas which are underserved by amenities.

Data

The key to a robust analysis is obtaining reliable data for all the parameters. For this project we will be leveraging data for a diverse set of sources and then integrating them to allow us to see the full picture. We will be referencing the following data sources:

- Geospatial data and neighborhood classification from the Analyze Boston website. Analyze Boston is the City of Boston's open data hub managed by Citywide Analytics Team. Specifically we will be using the Boston Neighborhoods geojson file which will provide the name, polygon coordinates and area for each Boston neighborhood.
- Rental prices in each neighborhood in Boston from Real Estate Boston. This website uses data from the rental website Zumper to present the median rental price for a 1BR apartment in each neighborhood of Boston. The data is presented for multiple time periods but we will be using the latest data (Winter 2020) in the analysis.
- Foursquare Places API to pull the list of venues in each neighborhood. This API returns a set of venues based on the latitude and longitude queried along with other variables such as venue category, radius of search etc.

Methodology

We start with extracting basic information about each neighborhood in Boston. The geojson from Analyze Boston contains information about

the name and area of each neighborhood along with the polygon coordinates which will be used later for plotting choropleth maps. Our initial dataframe thus has two columns: the name of the neighborhood and the area in sq. miles. We will then use geocoder to extract the latitude and longitude data for each of these neighborhoods and add it to the dataframe.

We will then extract the rental prices for each neighborhood to feed into our analysis of how expensive each of them are. We will use web-scraping with Beautiful Soup to extract data from Real Estate Boston website. The target data is contained in the div with class *content-text* so we will start with extracting that into a separate object. From this we will use *next_siblings* to find the neighborhood name and rental price data. The neighborhood names use tag *h2* while prices use tag *p* and with this distinction we can separate out both into individual lists. From the price list we will extract only the latest price data which is from Winter 2020. We will then create a dataframe with the neighborhood names and median rental prices. After some cleaning we can then merge this dataframe to the earlier one to incorporate latitude and longitude data.

To get a list of venues in each neighborhood we leverage Foursquare's Places API. Because there are numerous sub-level categories in Venue data returned by Foursquare, we will extract venue data for each top-level category separately. The top level categories can be found on Foursquare developer documentation page. To do this first we define a function that will make GET requests to the Foursquare API based on the parameters defined. One aspect that we capture in this function definition is to dynamically select the radius of the query. Since the neighborhoods in Boston vary significantly in size (Leather District is only 0.02 sq. miles while Dorchester is 7.29 sq. miles) we will define a range of radius to ensure we are not duplicating results. The radius used will be as follows:

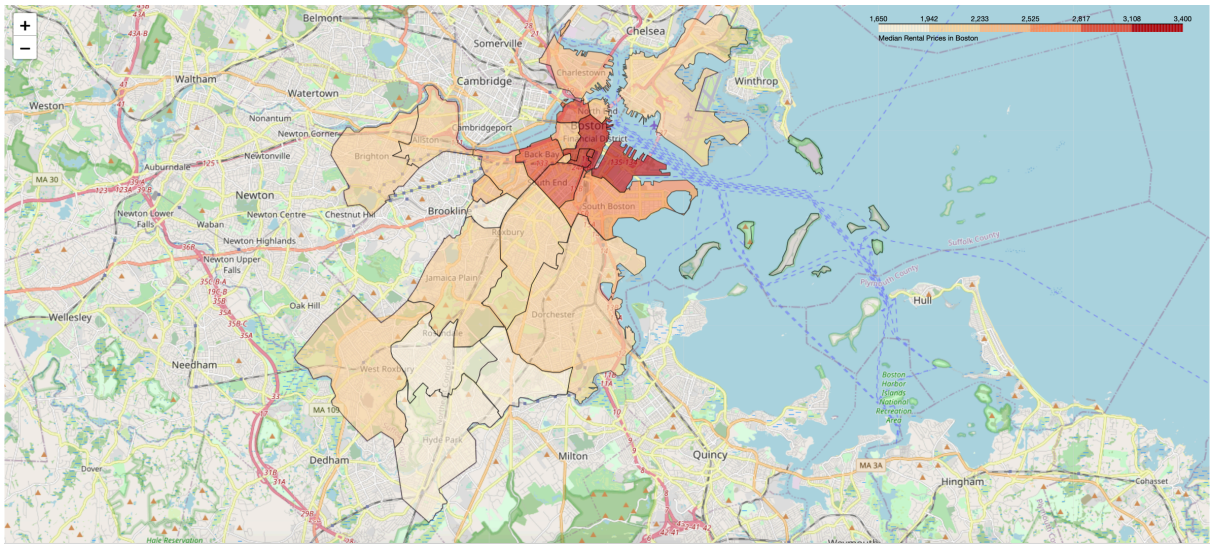
- For small neighborhoods (<0.2 sq. miles): 100 mt.
- For medium neighborhoods (0.2-1.0 sq. miles): 500 mt.
- For large neighborhoods (>1 sq. miles): 1000 mt.

We will then call this function for each top-level category separately and consolidate the results into one dataframe. This data will then be consolidated at neighborhood level and then normalized to obtain a Venue Score.

Finally we will build a dataframe that pulls in relevant information from the rental and venue dataframes. The median rentals will be converted to a Rental Score by inverting and normalizing them. The combined values form the Total Score for each neighborhood.

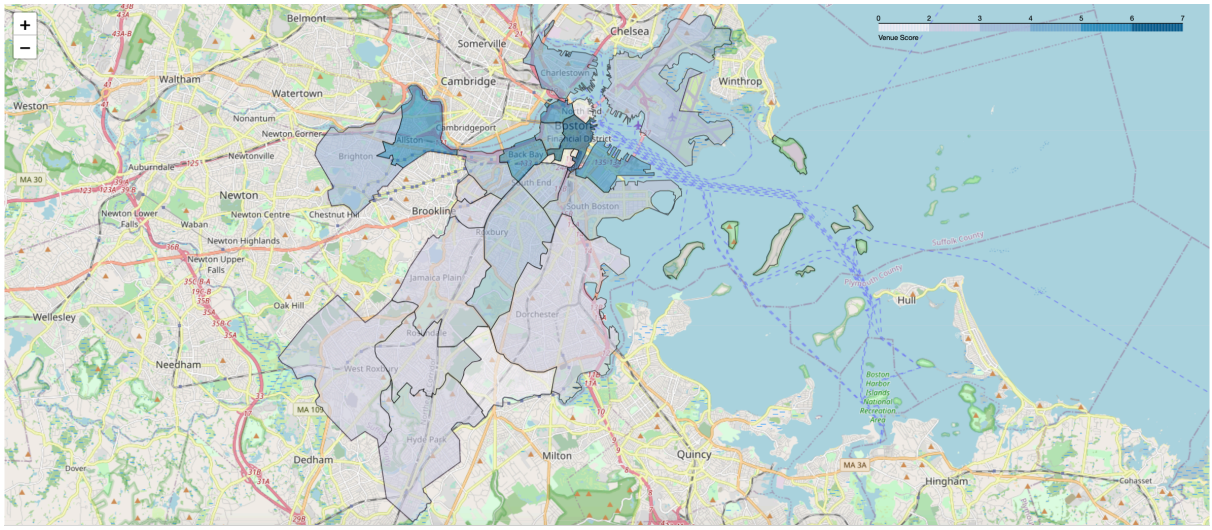
Results

The initial analysis with Rental prices shows the expected results of high rental prices in central neighborhoods and declining rentals as we move away from the center. The median rentals in outlying areas such as Roslindale, Hyde Park & Mattapan are about half of the rentals in central areas like Downton, Chinatown & Leather District.



Mapped by Median Rental Prices

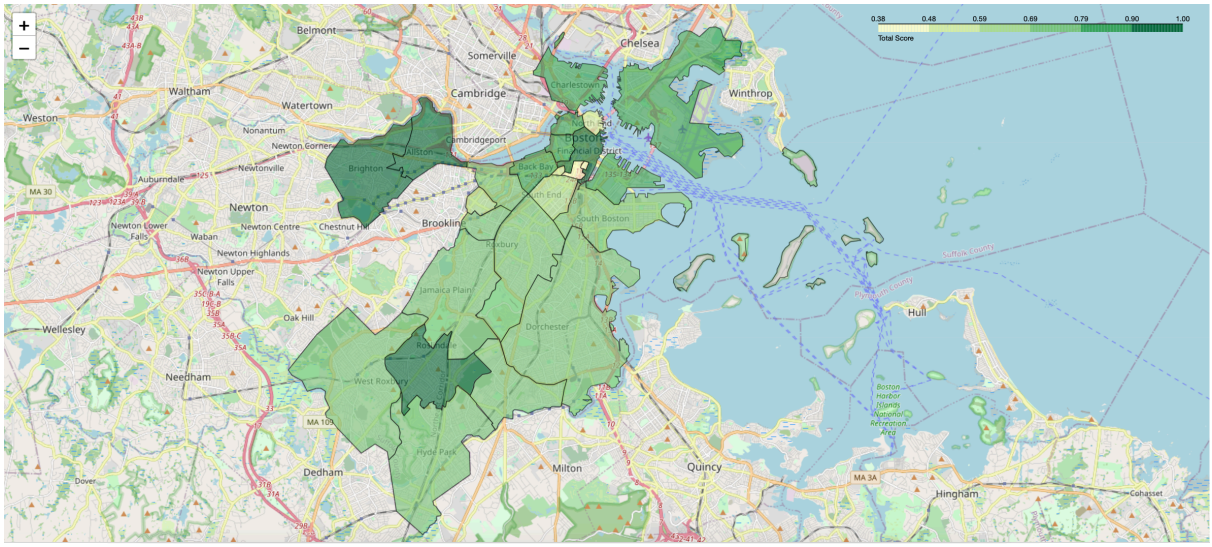
When we layer in the venue data from Foursquare we see that neighborhoods like Downtown, Allston, Back Bay, West End & South Boston Waterfront show up as the ones with the highest amount of nearby venues.



Mapped by Venue Score which captures the amount of venues nearby

Finally on combining the two scores and then mapping the neighborhoods on the Total score gives us a view of which neighborhoods score highly on

availability of diverse venues and also on affordability. In this final map we see that Downtown scores the highest driven by the vast number of venues in the area. However Allston and Brighton show up as affordable alternatives to be considered.



Mapped by the Total Score which combines the Rental Score & Venue Score

Discussion

The results of this analysis bring out some interesting aspects about Boston neighborhoods. The neighborhoods of Allston & Brighton are particularly interesting. Both are highlighted as affordable neighborhoods with a large amount of venues in the vicinity. The presence of Cambridge with its colleges nearby could be a driving factor for the presence of high number of venues in Allston & Brighton. Cambridge was excluded from the analysis as it's not part of the Boston neighborhoods.

Another area which was excluded from analysis as it's not part of the Boston neighborhoods was the the area of Newton which lies between Brighton & West Roxbury. It is however close enough to the city center that it should be considered for inclusion into the analysis.

Finally, in our analysis we have chosen to use different values of radius for each neighborhood when pulling the Foursquare venue data. This was done to avoid any duplication in the venues returned. While this has helped us to look at each neighborhood independently it does lower the attractiveness of small neighborhoods like Chinatown & Leather District that have very few venues returned. One option would be run the analysis again with a common value of radius for all neighborhoods to see how that changes the results.

Conclusion

While this analysis provides a good starting point for evaluating neighborhoods it can be expanded significantly to make it more robust. One potential option would be allow users to define their weights for each of the venue categories and then derive scores based on these user weights. I had tried to implement this within the project but I was unable to find a workable solution and had to defer it for a later version. There is also scope to add additional details about the neighborhood such as commute distances, level of crime, greenery cover etc to make the analysis more holistic.

My ability to execute everything I wanted in this project was certainly limited by my knowledge and experience with the tools. However I do think this approach to analyzing cities has potential to be highly useful. Moving to a new town can be scary but finding the right place to live can go a long way to making it a much more enjoyable experience.