Comparison of neighborhoods of Manhattan vs. Boston

Introduction

For this project, I would like to utilize a similar approach taken to evaluate the neighborhoods of Manhattan/Toronto towards comparing two highly prominent cities in the USA, Boston, MA and Manhattan, N.Y. While both cities are known throughout the world as bustling cities, Boston is largely looked upon as a moderately populated city, with an economy driven by biotech innovations and harboring some of the greatest educational and medical institutes in the world. In comparison, Manhattan is generally seen as extremely populated city, the financial center of the USA and harboring some highly recognized educational and medical institutes in the USA. However, overall, while Manhattan portrays a picture of bustling city with great ethnic diversity, in the eyes of the world, Boston portrays a somewhat more subdued, more conservative and more monochromatic demeanor.

Therefore, I am interested in exploring the neighborhoods of these two boroughs and identify those having similar attributes in terms of their dwellers' personal and professional lifestyles. For this analysis, Foursquare location and venue data will be utilized. The data will be evaluated on the basis of making a call on which neighborhoods are similar to one another or whether neighborhoods of these two cities are somewhat unique.

Those that benefit from this analysis

I believe that such analysis will have an impact on evaluations such as which promising businesses could be brought into particular neighborhoods, which neighborhoods in the other city may match the needs of those seeking to relocate and which cultural attractions could enhance the quality of life of a particular neighborhood. **Therefore, such comparisons could benefit professionals involved in city planning, business development as well the general public in making informed decisions when facing key decisions such as determining relocation venues, real estate purchases and others.

Data To Be Utilized For This Study

For this study, I will use the following datasets:

Lists of neighborhoods of Manhattan and Boston- https://en.wikipedia.org/wiki/Neighborhoods_in_Boston and https://geo.nyu.edu/catalog/nyu_2451_34572 - These data enable the extraction of neighborhoods of the two cities (boroughs) to submit towards the Foursquare database to extract venues that are present within these neighborhoods.

Foursquare location data- Location data will be extracted by contacting the Foursquare databases and providing the identities of the neighborhoods within the city limits. This data will enable the defining of the neighborhoods to extract venue info from the Foursquare database. Also, will help in mapping the neighborhoods on maps to visualize the neighborhoods and clustering patterns.

Foursquare venue data - Venue data will be extracted by contacting the Foursquare databases and providing the locations of the neighborhoods to extract the venues present within the neighborhoods. Also, one can extract information regarding the popularity of the venues. This data will enable identifying the personal and professional lifestyle of the neighborhood dwellers including visitors by evaluating the types of businesses, cultural attractions etc., that are present within the neighborhoods as well as most frequented by the inhabitants along with visitors.

During the study, I wanted to expand and associate the median income and the median rental prices for each of these neighborhoods. Therefore, I extracted the data from the following pages.

Manhattan neighborhood median income and median rental price sources

https://ny.curbed.com/2017/8/4/16099252/new-york-neighborhood-affordability

Manhattan neighborhood median income and median rental price sources

https://www.homesnacks.net/richest-neighborhoods-in-boston-128969/

https://www.rentcafe.com/average-rent-market-trends/us/ma/boston/

Sites used to fill the gaps in the above data

https://en.wikipedia.org/wiki/List_of_Massachusetts_locations_by_per_capita_income

https://bostonpads.com/average-rent-prices-boston-by-town/

https://www.rentcafe.com/average-rent-market-trends/us/ma/boston/

Methodology

Assembling of the dataset for the Manhattan neighborhoods along with their GPS coordinates.

For the Manhattan data-set, I used the one that was provided for the class tutorial on New York neighborhoods. I extracted the dataframe with Manhattan neighborhoods along with the GPS coordinates.

Assembling of the dataset for the Boston neighborhoods along with their GPS coordinates.

For the Boston neighborhoods, I extracted the list of Boston neighborhoods by web scraping and then assembled the GPS coordinates to that manually because of the difficulty in finding one source that contained that data within my programming skills. I assembled this data into a .csv file and then exported that into the IBM skills lab and used the read_csv command to embed the data into the Jupyter notebook file.

Data wrangling to clearly label the neighborhoods to enable easy identification with respect to different cities (States – ex: Manhattan vs. Boston neighborhoods)

Added the state (NY or MA) to the dataframes containing the neighborhoods and their GPS coordinates to clearly label each neighborhood with the respective state and then joined the two dataframes into one for clustering purposes.

Use of Foursquare data to find the most frequented venues in each neighborhood

I followed the data evaluations used in the capstone neighborhood exploration exercises to obtain the Foursquare venue data for each of the neighborhoods.

Data wrangling to find the top10 venues for each of the neighborhoods

The top 10 venues for each of the neighborhoods were identified by data wrangling methods

Neighborhood clustering

Then I used the machine learning method, K-means clustering to cluster the neighborhoods **to find out similar neighborhoods in Manhattan and Boston.**

Dataframe manipulation to add the neighborhood income and rental cost information

The separate dataframes containing the top10 destinations for each of the **clusters** were modified with the addition of the income and rent information by joining the dataframes. The income and rental cost information was brought in as .csv files and were added to the existing cluster dataframes using the common column named, 'Neighborhood'.

Visualizing of the clustering of neighborhoods of Manhattan, NY vs. Boston, MA

The main dataframe with the clustering labels and the top 10 venues for each of the neighborhoods was modified with the addition of the income and rent data as two new columns. Then Folium maps were generated with the labels designed to display the following – Neighborhood, Cluster Label, Income and Rent.

Enabling the neighborhood evaluating starting from the income or rental perspective

Also, separate Bar-charts were also assembled to make it easier for people to compare their income and current rents and easily find similar neighborhoods to explore as potentials from the venue (reflecting lifestyle habits) perspective.

Results

The link to the Jupyter file in Github is:

https://github.com/serendib3000/Coursera_Capstone/blob/master/Workbook_ass3.ipynb

However, when I tried to see the file, Folium maps were not displayed. Therefore, please use the above link to view the file in nbviewer. File displays properly with nbviewer. Thanks!

Results of Neighborhood Clustering using K-Mean clustering method

The application of K-means clustering method using the most frequented venues of the neighborhoods resulted in the formation of 6 distinct clusters. The clusters varied in size with two comprising of a single neighborhood each (Clusters 0, 4) and others comprising 38 (Cluster 1), 9 (Cluster 2), 2(Cluster 3) neighborhoods.

Cluster 1 appears to be more"happening/tourist frequented" areas with dining (restaurants/coffee shops) and lodging venues being most popular. Cluster 2 in contrast seems more oriented towards theatre/arts over restaurants (dining). Cluster 3 is bit of a surprise because it is similar to Cluster 1 but may have been separated out due the presence of a cemetery and other oddball venues such as an outdoor sculpture. Cluster 4 seems to be predominantly a recreational area with a trail being the most frequented. Therefore, overall neighborhoods bear many similarities but also have distinct venues that appear to bestow uniqueness to the neighborhoods.

The neighborhoods within these two cities that make up the respective clusters are detailed in Table 1.

Table 1: Distinct neighborhoods clustering together. NY and MA, respectively identify Manhattan and Boston neighborhoods.

Cluster_0	Cluster_1	Cluster_2	Cluster_3	Cluster_4
West Roxbury, MA	Marble Hill, NY	Upper East Side, NY	West Village, NY	Mattapan, MA
	Chinatown, NY	Lincoln Square, NY	North End, MA	
	Washington Heights, NY	Clinton, NY		
	Inwood, NY	Midtown South, NY		
	Hamilton Heights, NY	Stuyvesant Town, NY		
	Manhattanville, NY	Beacon Hill, MA		
	Central Harlem, NY	Downtown, MA		
	East Harlem, NY	Hyde Park, MA		
	Yorkville, NY	West End, MA		
	Lenox Hill, NY			
	Roosevelt Island, NY			
	Upper West Side, NY			
	Midtown, NY			
	Murray Hill, NY			
	Chelsea, NY			
	East Village, NY			
	Lower East Side, NY			
	Tribeca, NY			
	Little Italy, NY			
	Morningside Heights, NY			
	Gramercy, NY			
	Battery Park City, NY			
	Carnegie Hill, NY			
	Turtle Bay, NY			
	Flatiron, NY			
	Hudson Yards, NY			
	Allston, MA			
	Brighton, MA			
	Charlestown, MA			
	Chinatown, MA			
	Dorchester, MA			
	East Boston, MA			
	Jamaica Plain, MA			
	Mission Hill, MA			
	Roslindale, MA			
	Roxbury, MA			
	South Boston, MA			
	South End, MA			

<u>Visualization of the targeted two cities to demonstrate the physical distance despite the similarities</u> (Figure 1)

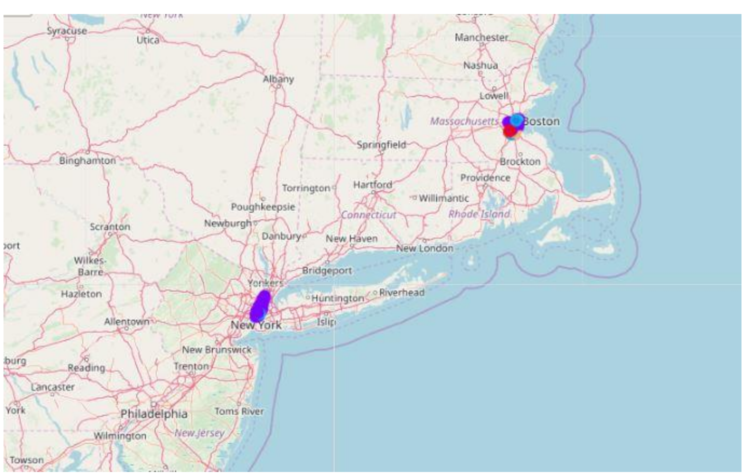

Figure 1: Folium map depicting the neighborhood clusters on a map of the USA demonstrating the distance between the two evaluated cities despite similarities of the neighborhoods of each of these cities.

<u>Closer visualization of the targeted neighborhoods in Manhattan, NY and Boston, MA (Figure 2)</u>

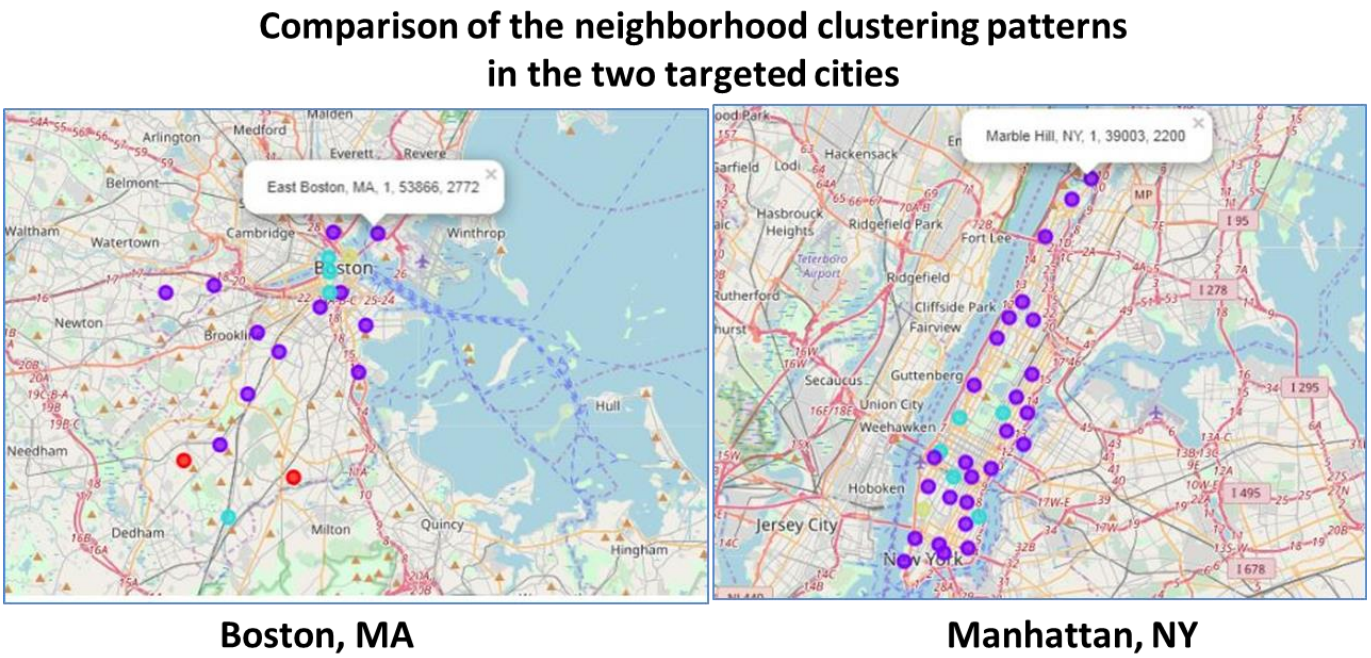

Figure 2: Side-by-side comparison of the clustering pattern of the neighborhoods in the two targeted cities, Boston, MA and Manhattan, NY. Similar colors indicate neighborhoods clustering together. The labels indicate the neighborhood name, cluster number, median income of neighborhood and the median rent of a 2-bedroom apartment. Cluster 0 = purple circles; Cluster2=teal circles; Cluster 3=beige green circles, Clusters 0 and 4=red circles with cluster 0 on right side and Cluster 4 on the left side.

<u>Income and rent info of the clustered neighborhoods for easy reference and targeting of closer evaluation of particular neighborhoods</u>

Income and rental data were also associated with the neighborhood clustering to enable those who may seek to re-locate from one city to the other. This data will provide the following advantages for individuals with multiple interests. Two such are detailed below.

(1) With this information plotted in bar-charts, one can easily target one's income in any of the clusters (if present) and then identify neighborhoods with similar income levels to explore as possible areas for re-location. Similarly, rental data will enable the exploration of the various venues relative to one's current rental costs or the desired neighborhood's rental costs.
(2) Also, this data will be useful for those interested in starting a business to evaluate the neighborhoods with similar characteristics and patterns relative to the income level of occupants or patrons to narrow down the suitable neighborhoods for one's business interests.

Income and rental cost of neighborhoods of Cluster_0

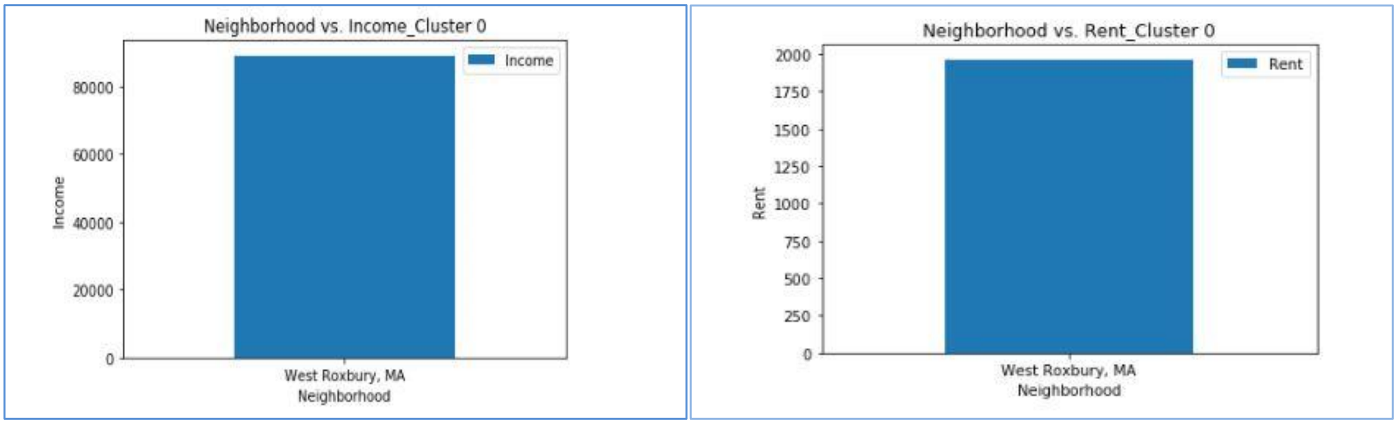

Figure 3: Median house-hold income and rental cost info for Cluster_0 neighborhood.

Income and rental cost of neighborhoods of Cluster_1

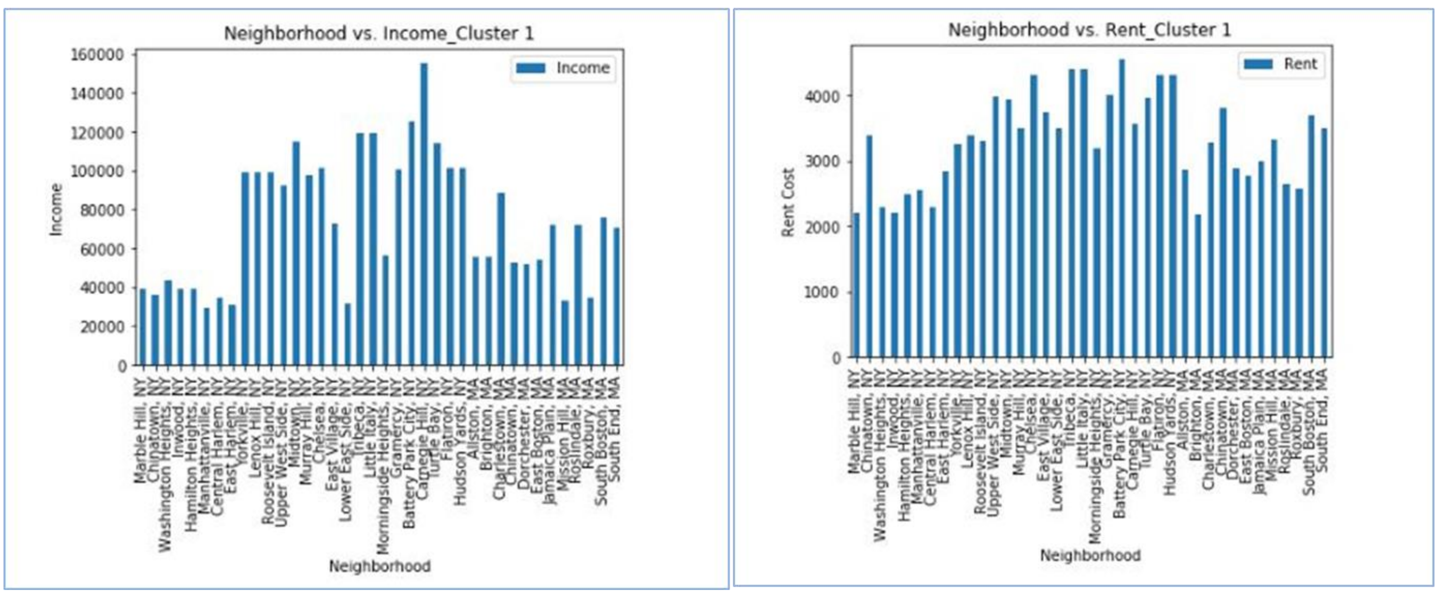

Figure 4: Median house-hold income and rental cost info for Cluster_1 neighborhoods.

Income and rental cost of neighborhoods of Cluster_2

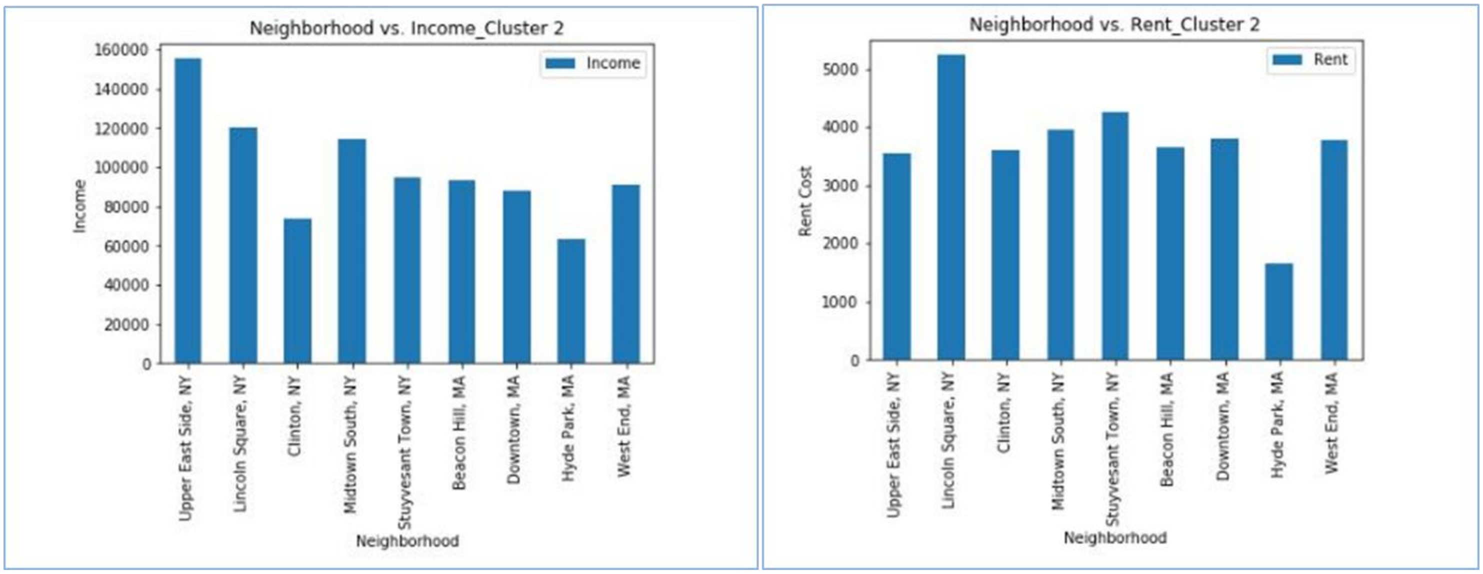

Figure 5: Median house-hold income and rental cost info for Cluster_1 neighborhoods.

Income and rental cost of neighborhoods of Cluster_3

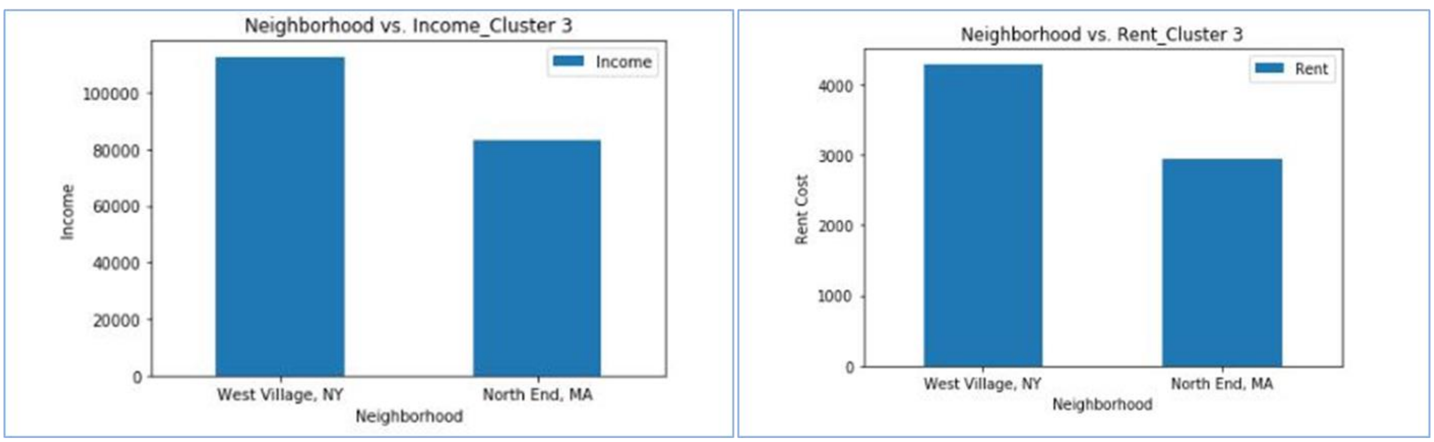

Figure 6: Median house-hold income and rental cost info for Cluster_2 neighborhoods.

Income and rental cost of neighborhoods of Cluster_4

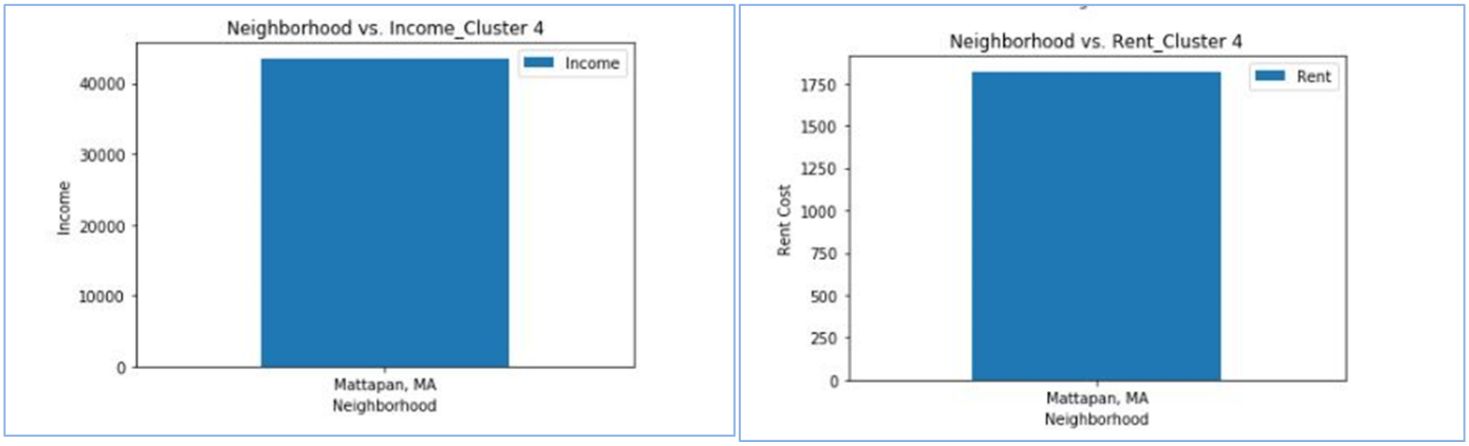

Figure 7: Median house-hold income and rental cost info for Cluster_3 neighborhood

Discussion

The data presented enabled the identification of similar neighborhoods in two geographically distinct major cities of the USA, Boston, MA and Manhattan, NY. The data clearly indicate the following.

(1) Neighborhoods in Boston, MA and Manhattan, NY are quite similar in terms of people's lifestyles with many neighborhoods clustering together with respect to restaurant/hotel visits (Cluster 1). Also, certain neighborhoods cluster together with respect to restaurant/hotel visits but shows distinction in its theatre and arts patronizing patterns (Cluster 2). Clusters 0 and 4 tend to orient toward outdoor activities which also may be reflected by their placement at the outer boundaries of Boston. In contrast, no neighborhoods similar to those in Clusters 0 and 4 were found in Manhattan_NY most likely due to the compactness of space and the lack of such natural out-door activity areas.

(2) Also, when evaluated from the income and rental cost perspective, most neighborhoods do not show any association between income and rental costs. In Manhattan especially, the comparison of income to rent indicate that rental costs are relatively similar in all neighborhoods despite income levels. For example, higher income does not result in higher rental costs because there are multiple neighborhoods with higher rental costs but with lower income levels. This data can be further evaluated using regression evaluations but the time allocated to finish up this report is not sufficient to take such deep evaluations, especially for those new to programming. Therefore, this study could be extended in the future with such evaluations.

Recommendations:

(1) Cluster 0 and 4 neighborhoods may benefit from businesses such as coffee shops, ice cream parlors etc., especially close to the recreational facilities or businesses such as lawyer's offices. However, the income in Cluster 4 is lower compared to other clusters such as Cluster 0. Therefore, further careful analysis is necessary.

(2) Evaluation of why certain neighborhoods with higher income levels have lower rent, as in Upper east side, NY, would be interesting for further study. Such study may benefit to lower rents in other neighborhoods.

(3) Further parse the data in Clusters 1 and 2 to evaluate whether the neighborhoods in these clusters can be further clustered to intimately identify those neighborhoods with even closer matches. May help users to target

the desired neighborhoods more readily. Also, businesses that can be targeted to be initiated in such neighborhoods may be projected in a more accurate fashion.

(4) Extend the study to other cities with one of these cities being the seed to enable most accurate comparisons among all cities. Once such data are available, perhaps one can project similar neighborhoods just by using Foursquare data on venue visits without carrying out a lengthy study. This is a long term objective.

Conclusion

The study indicates that similar neighborhoods (reflecting inhabitants' and visitors' lifestyle choices) exist in cities that are geographically highly distinct. This type of study can be extended to other cities as well using one of these cities as the seed so that comparisons can be made with higher precision. This study need to be further evaluated with criteria such as income and rent costs to truly provide users with precise guidance.

Also, neighborhood clustering data can be utilized to find any missing businesses in similar neighborhoods because they may make up an unmet need in the area. This data can be further analyzed to cover such evaluations.