Multiple City K-means Cluster Analysis

Coursera Applied Data Science Capstone Project

Richard C. Anderson

31 May 2020

Introduction

Background

While I worked through the New York and Toronto clustering exercises for the Applied Data Science course , I found myself wondering how a K-means clustering analysis of the neighborhood venues of New York's Manhattan would compare with the neighborhood venues of the Boston metro area, where I now live in the suburbs, and with central Houston, where I have lived in the past.

Problem

Would a K-means clustering analysis pick up on what I know subjectively about the three cities? I would expect Boston and Manhattan neighborhoods to be similar, as both cities are older, denser, pre-date the automobile, and have extensive mass-transit options. Houston, on the other hand, has developed entirely with the automobile as its primary transportation method and, in part due to cheaper land prices, has a much lower population density. Houston's central downtown is all business and mostly deserted after 8pm as very few people live there. Its mass transit system is mostly bussing with an emphasis on workers commuting to and from the downtown area.

The central question is whether venues in the Houston neighborhoods will cluster with those of Boston and Manhattan. I theorize that Houston might contain mostly independent venue clusters while Boston and Manhattan share similar cluster types. This study will perform multiple K-means investigations. First, the project will replicate the Manhattan neighborhoods venue clustering evaluation from the course exercises and develop new venue clustering evaluations for Boston and Houston neighborhoods. Second, the Manhattan, Boston, and Houston neighborhood venue data will be aggregated for a multi-city K-means venue clustering evaluation. The multi-city evaluation will compare the neighborhood venues of the older cites of Manhattan and Boston with the much newer Houston.

Potential Applications

The multi-city evaluation performed for this study project is an approach that has potential marketing and operational benefits for businesses. For instance, a venue evaluation can help locate both saturated and underserved market areas. For a business considering opening into to completely new markets, the multi-city evaluation can provide clues for new site location as well as how a business might need to adapt its product offerings to compensate for differences in predominate neighborhood characteristics. The multi-city evaluation approach could also be useful for more personal use as well. As an example, someone relocating from one city to another could start by determining which neighborhoods in a new city most resemble (or differ!) from their current neighborhood.

Data Sources and Acquisition

Data Requirements

This project will require the following information for the neighborhoods of Manhattan, the Boston metro area, and central Houston:

- List of neighborhoods
- Geo-location data for each neighborhood
- Set of venues for each neighborhood

Data Sources

The neighborhood list and geo-location data for Manhattan will come from the json dataset that was used for the earlier course exercises. Unfortunately, a google search for tabular data of Boston and Houston neighborhoods and their geo-locations did not yield any directly usable results. However, it was possible to use Wikipedia to manually construct CSV files with neighborhood and geo-location information. Foursquare will be used for gathering the venue data for the city neighborhoods.

Data Acquisition

The manual construction of neighborhood names and geo-locations for Boston and Houston required several subjective judgements, as there did not appear to be a single definitive neighborhood list on Wikipedia and other websites for either Boston or Houston. I used my personal familiarity with both cities to determine suitable neighborhood lists.

Another issue for compiling the neighborhood lists for Boston and central Houston is how to define their metro area boundaries for comparison with the borough of Manhattan. Central Houston is typically defined as the neighborhoods inside the 610 Loop freeway. However, there are two independent cities, Bellaire and West University Place, that are fully contained in this area and will be included as part of central Houston. Also, I chose include the first ring of neighborhoods just outside the 610 Loop as part of central Houston for the purpose of this study. I made this decision in part to give each metro area approximately the same number of neighborhoods. Defining the Boston metro area also required some subjective adjustments as there are independent major suburbs, particularly Brookline, Cambridge, and Somerville that are sufficiently close enough to Boston to be considered part of its central metro area.

Foursquare was used for gathering the venue data for all neighborhoods. However, the developmental differences between Houston and the older cities of Boston and Manhattan had an impact on the queries used for gathering the venue data. In the original course exercise for Manhattan the venues were pulled from Foursquare using a 500-meter radius around the center geo-location of each neighborhood. The population density differences between the three cities suggested that a different radius setting would be required for each city to gather sufficient venue data.

Iterative venue count evaluations were made for Boston and Houston to determine reasonable radius settings. At the initial setting of 500 meters, many Boston neighborhoods and most Houston neighborhoods returned less than 20 venues within a neighborhood. Houston, using the initial 500-meter radius setting, had several neighborhoods that returned fewer than 5 venues. The search radius for both Boston and Houston were increased until each city had a least 10 neighborhoods that returned 100 venues (max limit) and the sparser neighborhoods returned 10 venues. As suspected, Houston required a much larger search radius to obtain reasonable venue lists.

The search radius settings used to acquire the venue information were set as follows:

• Manhattan:	500 meters
• Boston:	1000 meters
• Houston:	2500 meters

Data Cleaning

Two adjustments were made to clean the raw neighborhood and venue data.

First, during the aggregation of the neighborhood geo-location data with the venue data, it was discovered that both Houston and Boston return results for a venue category called 'Neighborhood', causing some interesting conflicts with the 'Neighborhood' name column in the geo-location data. Subsequently, the 'Neighborhood' venue results were filtered out of the Foursquare data prior to joining it with the geo-location data.

Second, during the K-means clustering analysis of Houston, one neighborhood always formed a cluster of one, no matter how few or many clusters were chosen for the K-means evaluation. The outlier neighborhood had only 10 venues returned at the search radius of 2500 meters. Further investigation revealed that the neighborhood was extremely industrial, with little residential or commercial presence. Therefore, the neighborhood was removed from the Houston neighborhood list.

Data Analysis

Visual Verification of Neighborhood Geo-location Data

The neighborhood geo-location data was plotted using folium. All three maps are presented using the same magnification. Note the much lower density of neighborhoods in Houston.

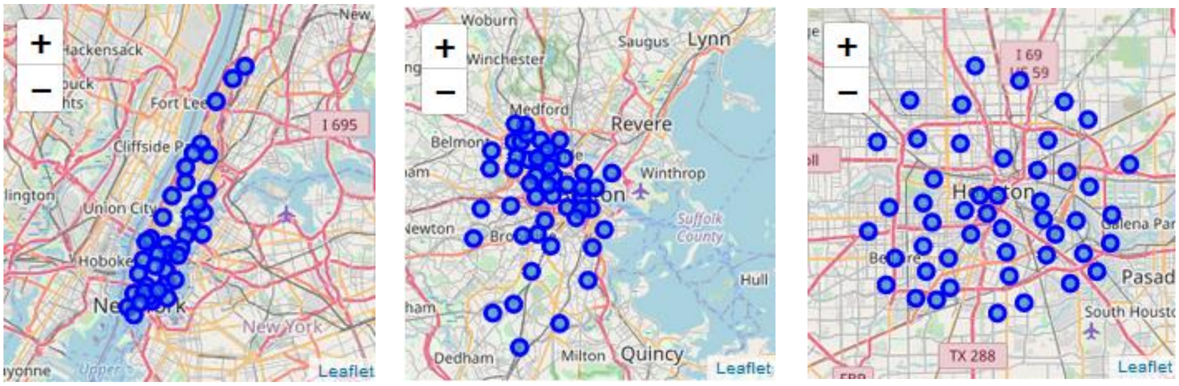

Figure 1 Manhattan, Boston Metro, and Central Houston Neighborhoods

Visual Verification of Neighborhood Venue Data

The difference in neighborhood density required iteration of the venue search radius for Boston and Houston. The goal of the iteration was to achieve similar venue count profiles for the three cities.

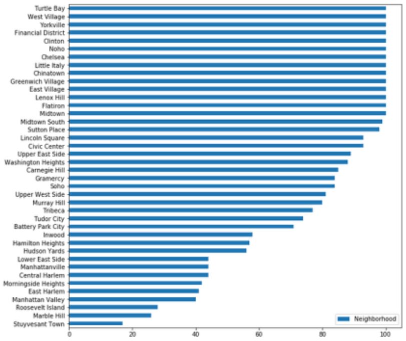

Figure 2 Manhattan Neighborhood Venue Counts, Search Radius: 500 m

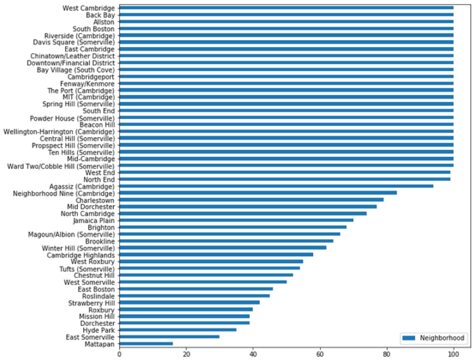

Figure 3 Boston Neighborhood Venue Counts, Search Radius 1000 m

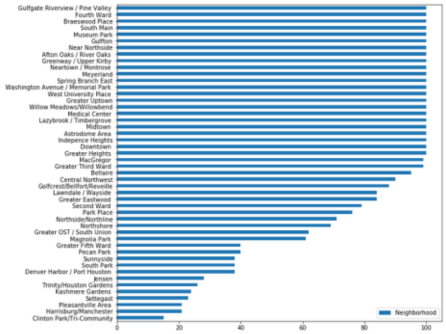

Figure 4 Central Houston Neighborhood Venue Counts, Search Radius 2500 m

For contrast, below is the original venue count profile for Central Houston using the same 500-meter search radius as Manhattan:

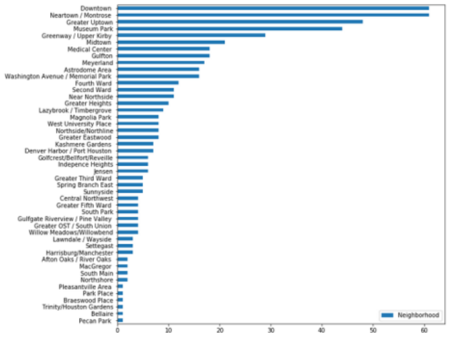

Figure 5 Central Houston Neighborhood Venue Counts, Search Radius 500 m

Venue Category and Count Summary

Summarizing the venue data, we see that each metro area has roughly the same number of venue categories and counts.

	Unique Categories	Venue Count
Manhattan	330	3093
Boston Metro	301	3936
Central Houston	273	3709
Combined Metros	431	10738

Note that the Combined metro areas number of unique categories is much larger than that of any individual metro area. While this discrepancy was not researched in detail, a quick investigation showed some weakness in the Foursquare data. For example, Boston and Houston returned venues labeled as 'Neighborhood', where as Manhattan did not. Also, there appears to be some redundancy in venue categories; 'Gym' and 'Gym/Fitness Center', 'Zoo' and 'Zoo Exhibit', etc. The choice of category for a venue could therefore impact the results of the K-means clustering analysis.

Individual Metro Area K-mean Clustering Analysis

K-means clustering was applied to the venue count data each metro area using a range of cluster values to determine a best cluster size. The goal for the best cluster size was to find highest number of clusters that did not have more than one outlier cluster (with only one or two members).

Manhattan – 5 Neighborhood Clusters

Note that there is an outlier cluster with only 1 member.

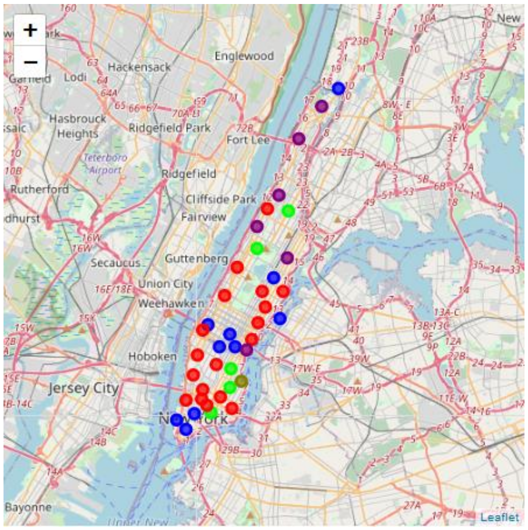

Figure 6 Manhattan Neighborhood Clusters

Boston Metro – 4 Neighborhood Clusters

Cluster sizes of 4, 5, and 6 all appeared to be viable choices, each with one outlier cluster. A visual review indicated that 4 clusters made the most sense from subjective knowledge of the area.

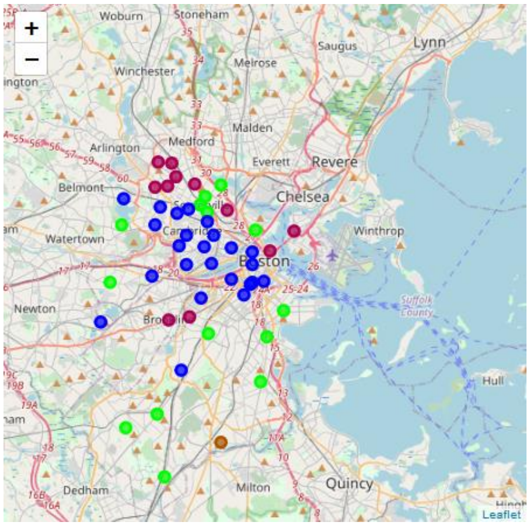

Figure 7 Boston Metro Neighborhood Clusters

Central Houston – 4 Neighborhood Clusters

Cluster sizes of 4, 5, and 6 all appeared to be viable choices, each with one outlier cluster. A visual review indicated that 4 clusters made the most sense from subjective knowledge of the area.

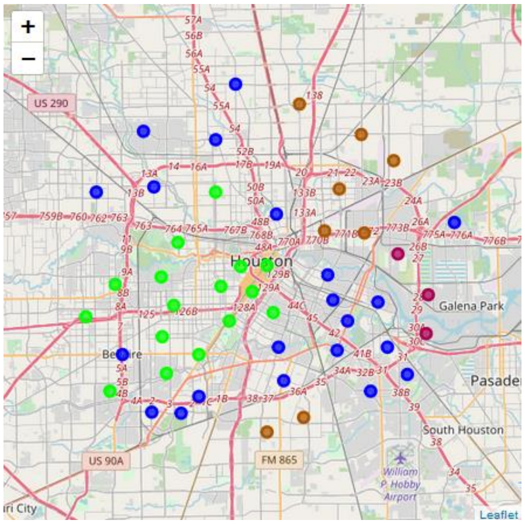

Figure 8 Central Houston Neighborhood Clusters

Combined Metro Areas K-mean Clustering Analysis

Like the individual metro areas, K-means clustering was applied to the venue count data the combined metro areas using a range of cluster values to determine a best cluster size. Cluster sizes of 5, 6, 7, and 8 were viable candidates. A cluster size of 7 was chosen after a visual review.

Visualizing the combined cluster data required plotting each of the metro areas using a common color palette for the clusters.

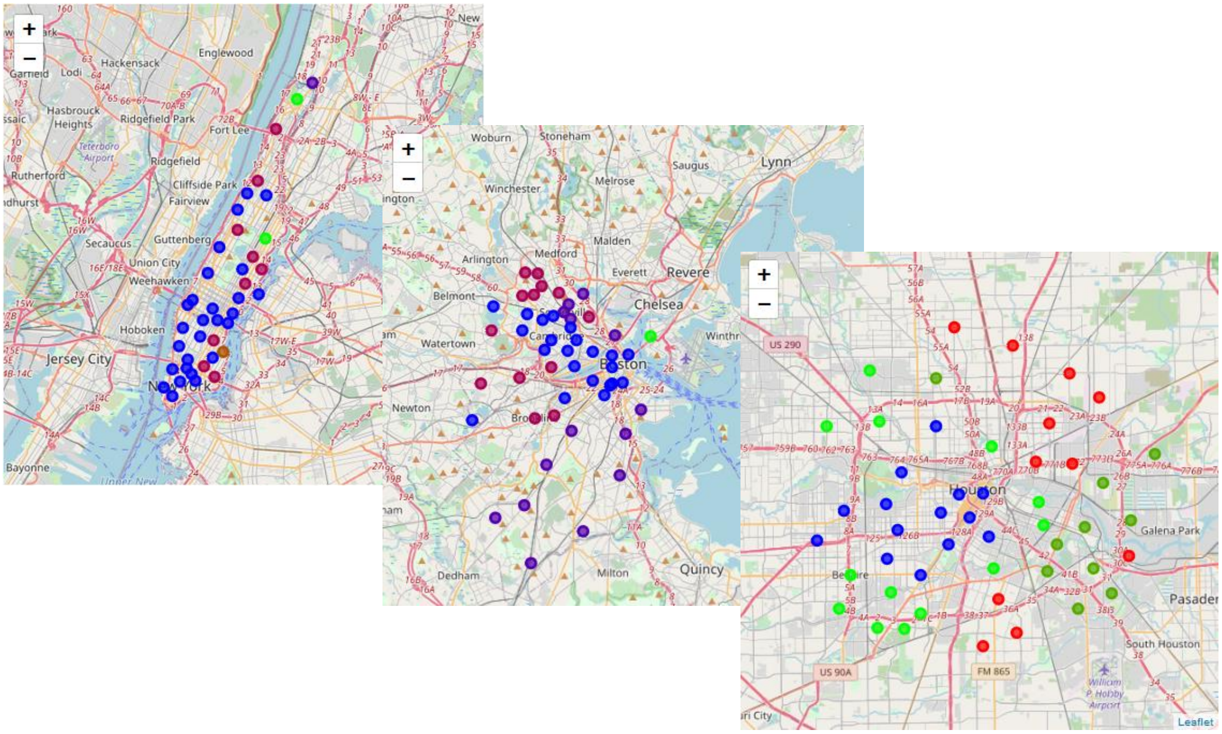

Figure 9 Combined Neighborhood Clusters

Results

The combined metro areas K-means clustering results align well with the original assumptions made in the Introduction. As expected, many Manhattan and Boston neighborhoods share similar characteristics. Only a small set of Houston neighborhoods are comparable with those of Manhattan and/or Boston.

Manhattan and Boston share significant membership in two neighborhood clusters. Houston, out of four neighborhood clusters, shares significant membership in only one cluster with Manhattan and Boston. Also, it is of note is that the Manhattan outlier neighborhood is still an outlier for the combined metro areas. The Boston outlier neighborhood has been assimilated into a larger cluster.

Recommendations

There are several ways that the analysis from this study could be improved.

First, the methodology of using a fixed neighborhood search radius for a given metro area assumes that all neighborhoods have a similar physical range. This potentially allows the smaller neighborhoods to “borrow” venues from other close-by neighborhoods. The searches could be modified so that each neighborhood uses a distinct search radius based on factors such as physical area, population density, and/or proximity to its next nearest neighborhood.

Second, a quick evaluation of the combined metro venue data set showed several potentially redundant venue categories as well as some outright questionable categories. More needs to be known about how venues get categorized by Foursquare for different cities. As noted in the Data Analysis section, can the counts for categories like ‘Zoo’ and ‘Zoo Exhibit’ or ‘Gym’ and ‘Gym /Fitness Center’ be combined or are they truly separate categories that are applied consistently. Some categories require further understanding and potentially removed from the venue dataset. For example, there is a category ‘Border Crossing’ in the combined metro venues dataset, yet not one of the cities is on a international border. What is a ‘Border Crossing’ venue?

Third, should outlier neighborhoods be eliminated or left it the venue dataset? As noted, one neighborhood from Houston was eliminated due to its outlier status. The subsequent K-means clustering results were quite different with the outlier eliminated. If time permitted, it would have been interesting to eliminate Manhattan’s outlier neighborhood, Stuyvesant Town, to see what impact it would have had on the remaining clusters.

Fourth, the datasets should be expanded to include additional older and newer metro areas to see if the results and conclusions hold up beyond the small sample used for this study.

Conclusions

This study evaluated whether K-means clustering could be used to differentiate the neighborhood characteristics across different metros areas. Two similar metro areas, Manhattan and Boston, were chosen to be contrasted with Houston. Geo-location data was used to obtain Foursquare venue data to characterize the neighborhoods of each metro area. The results of K-means clustering clearly showed that Houston neighborhood characteristics were different from those of Manhattan and Boston.

The K-means clustering approach can potentially be applied to a variety of problems where it is desirable to understand similarities or differences between neighborhoods of different cities.