Segmenting and Clustering Neighborhoods in Boston

Introduction

I will investigate the relationships between 24 Boston neighborhood venues within 500 meters of their centroids and the health status of these neighborhoods after clustering based on both available venue types as well as health status indicators. Several different health status indicators will be selected based on availability of data on a neighborhood level. Such indicators include demographics, social determinants of heath, community assets, environmental heath, access to care, maternal and child health, health related behaviors, chronic disease, infectious disease, sexual health, injury and exposure to violence, mental health, substance misuse, and death. Significant differences between indicators in any two clusters will be determined and an attempt will be made to correlate the differences with the availability of specific venue types in a specific neighborhood. This information should of interest to both public health officials and city planners, as well as developers, in attempting to remediate health related disparities and improve community health.

Data

I will be using data kindly provided in a CSV file from the Boston Planning & Development Agency (BPDA) which contains the latitude and longitude of the centroids of each Boston Neighborhood derived from the neighborhood shape files as provide in Analyze Boston (https://data.boston.gov). The shape files were run through Arcgis online and the tools they provided were used to calculate the latitude and longitude of the Boston neighborhood centroids. Health status indicators were obtained by manually populating a CSV file from data contained within the online document "Health of Boston 2016-2017" at http://www.bphc.org/healthdata/health-of-boston-report/Documents/_HOB_16-17_FINAL_SINGLE%20PAGES.pdf.

Methodology

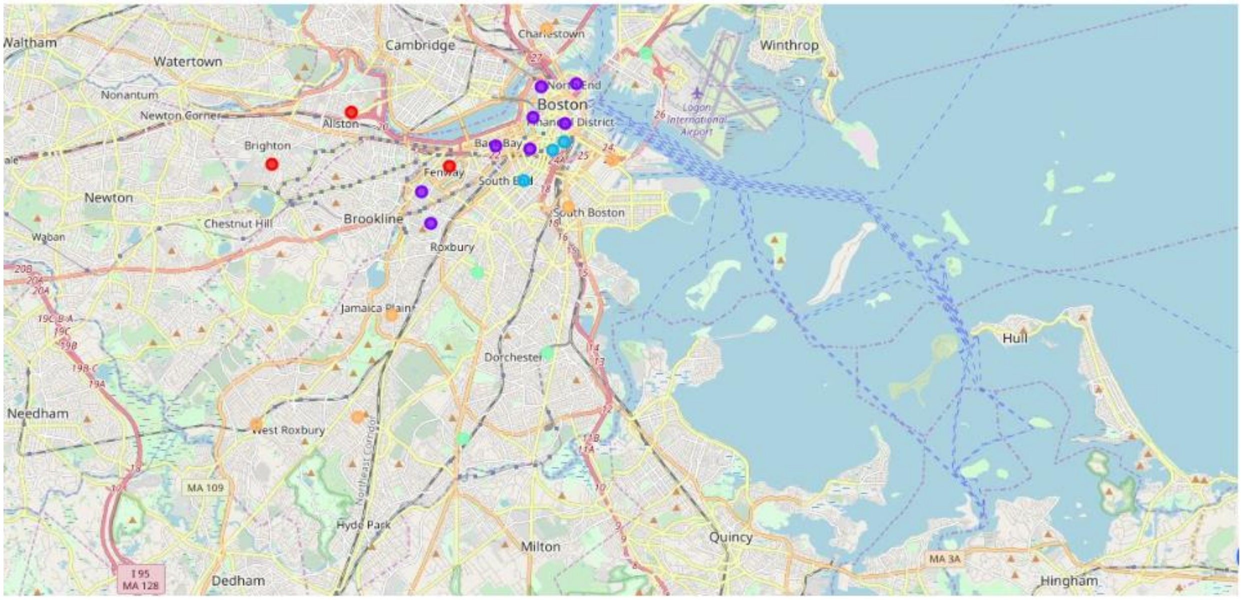
First the CSV file from BPDA was loaded into a Pandas data frame and the Harbor Island neighborhood was removed as this centroid did not correspond to a physical land mass. Next, a new data frame was created by filtering and renaming some columns. A Python geocoder library (geopy.geocoders) was used to extract the longitude and latitude values for each neighborhood. The Folium library was then used to create a map of the Boston neighborhoods with superimposed centroids. The Foursquare API was employed to generate a list of venues in each Boston neighborhood. The resultant JSON file was then processed using a Python library (pandas.io.json) to create a new data frame containing the category name of each neighborhood venue. The number of unique venue categories for each neighborhood was then counted and onehot encoding was applied followed by grouping by neighborhood mean frequency. The top 10 venue categories for each neighborhood was then determined. KMeans clustering was then applied using the scikit-learn library (sklearn.cluster). After trial and error using many different cluster sizes, it was determined that 5 clusters gave a reasonable clustering without too many or too few neighborhoods in any given cluster. These clusters were once again superimposed on the Boston neighborhood map using the Folium library. Finally, the top 10 venue categories for neighborhoods in each of the clusters wwere displayed and examined.

A neighborhood health status CSV file was manually created from each of the neighborhood maps contained in the “Health of Boston 2016-2017” document. Each health status indicator was coded as less than the neighborhood average (-1), equal to the neighborhood average (0), or greater than the neighborhood average (1). Ethnicity was coded as white (1), mixed (0), or black (-1). Population density was coded as within the IQR (0), above the IQR (1), or below the IQR (-1). This information was manually entered into a CSV file which was then imported into a Pandas data frame. This data frame was once again clustered into 5 clusters using KMeans clustering from the scikit-learn library and each cluster summarized by taking the mean across each neighborhood health metric within the cluster. Once again, each cluster summary metric was reported on a -1, 0, 1 scale for below, equal, or above the cluster mean. Each health status cluster was then examined, aligned with the corresponding venue cluster, and speculations were then proposed regarding the impact of specific vendor categories on the heath status of neighborhoods in a given cluster. The degree of neighborhood overlap between each vendor cluster and its corresponding health status cluster was also determined.

Results

https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/6e071441-50b2-4929-a02a-8ab3a62d57ce/view?access_token=34354aa0dafb5b42073f8fedf1d2d515e35b640d6877523aa6e166a986183648

Clustering by Health Status



Cluster 1:

Size=3

Neighborhoods: Alston, Brighton, Fenway

Health Status: White, Age>65, HS Educated, Low Unemployment, Much Open Space, Physically active, Low Uninsured, Nonsmoking, Low sugary drink consumption, Good nutrition, Low obesity, Low chronic and infectious disease mortality, Low substance and opioid mortality, High psychiatric hospitalization, High life expectancy.

Cluster 2:

Size=8

Neighborhoods: Back Bay, Bay Village, Beacon Hill, Downtown, Longwood, Mission Hill, North End, West End

Health Status: White, HS Educated, Low Unemployment, Much Open Space, Low Uninsured, Nonsmoking, Low sugary drink consumption, Good nutrition, Low obesity, Low chronic and infectious disease mortality, High life expectancy.

Cluster 3:

Size=3

Neighborhoods: Chinatown, Leather District, South End.

Health Status: White, Not HS Educated, High hepatitis B & C, Gonorrhea incidence, High ER visits for injury and assault.

Cluster 4:

Size=4

Neighborhoods: Dorchester, East Boston, Mattapan, Roxbury

Health Status: Mixed, Low income, Low hepatitis B & C, High psychiatric hospitalization, Low salmonella incidence, Low substance and opioid mortality.

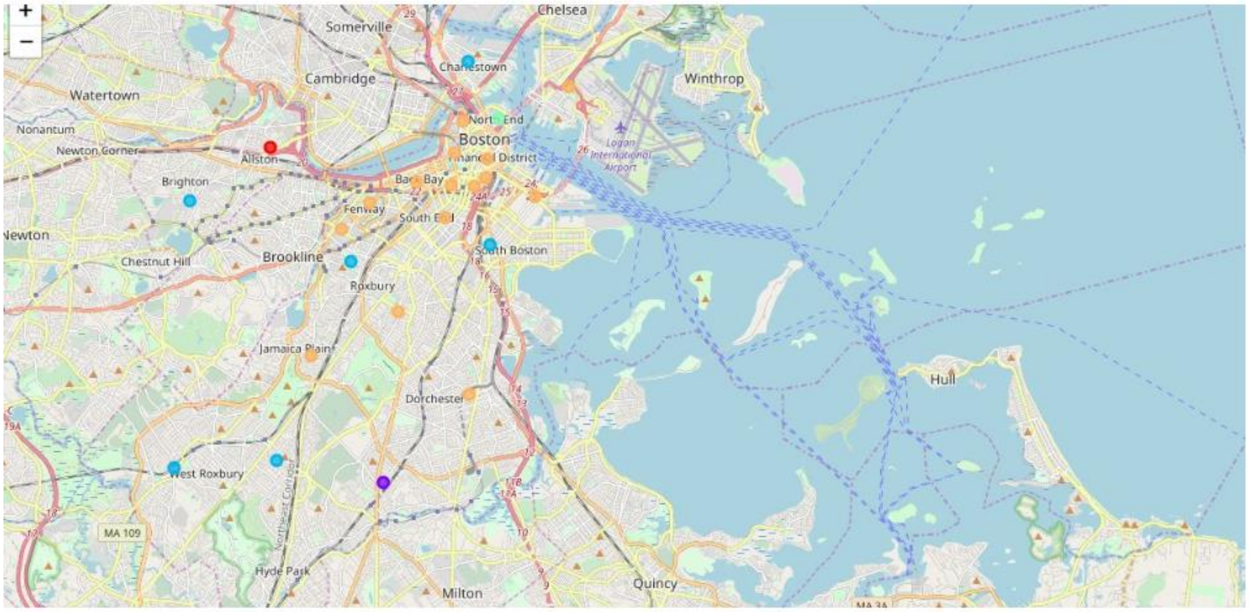
Cluster 5:

Size=6

Neighborhoods: Charlestown, Jamaica Plain, Roslindale, South Boston, South Boston Waterfront, West Roxbury

Health Status: White, High income, HS Educated, Low unemployment, Low uninsured, Low hepatitis B, Low Chlamydia incidence, Low ER visits for assault.

Clustering by Venues Within 500 M of Neighborhood Centroids



Cluster 1:

Size=1

Neighborhoods: Alston

Venues: Rental Car Location, Donut Shop, Convenience Store, Thrift Vintage Store, Smoke Shop, Liquor Store, Plaza, Pizza Place, Dance Studio, Deli.

Cluster 2:

Size=1

Neighborhoods: Mattapan

Venues: Ice Cream Shop, Furniture Home Store, French Restaurant, Food Court, Flower Shop, Fish Market, Fast Food Restaurant, Farmers Market, Falafel Restaurant, Event Space.

Cluster 3:

Size=6

Neighborhoods: Brighton, Charlestown, Mission Hill, Roslindale, South Boston, West Roxbury

Venues (Top 5): Pizza Place, Ice Cream Shop, Climbing Gym, Grocery Store, Convenience Store, Sandwich Place, Train, Coffee Shop, Plaza, Chinese Restaurant, Pharmacy, Thai Restaurant, Farmers Market, Pub, Pet Store, Bar, Park.

Cluster 4:

Size=1

Neighborhoods: North End

Venues: Italian Restaurant, Park, Seafood Restaurant, Pizza Place, Bakery, Wine Shop, Café, Market, Grocery Store, Coffee Shop.

Cluster 5:

Size=15

Neighborhoods: Back Bay, Bay Village, Beacon Hill, Chinatown, Dorchester, Downtown, East Boston, Fenway, Jamaica Plain, Leather District, Longwood, Roxbury, South Boston Waterfront, South End, West End.

Venues (Top 5): American Restaurant, Theater, Pizza Place, Chinese Restaurant, Vietnamese Restaurant, Coffee Shop, Park, Bakery, Coffee Shop, Fried Chicken Joint, Sandwich Place, Spa, Playground, Asian Restaurant, Café, Pizza Place, Fast Food Restaurant, Wine Bar, Italian Restaurant, Hotel, Pub, Gym, Falafel Restaurant, Basketball Court, Mexican Restaurant, Clothing Store, Sushi Restaurant, Airport Terminal, Electronics Store, Donut Shop, Seafood Restaurant, Plaza, Bubble Tea Shop, Cajun Creole Restaurant, Bus Stop, Museum, Pet Store, Bar.

Venue and Health Status Cluster Overlap:

Venue Cluster #	Health Status Cluster #
1	**1 (33%)**
2	**4 (25%)**
3	**5 (67%)**
4	**2 (13%)**
5	**2 (75%)**
	3 (100%)
	4 (75%)
	5 (33%)

Discussion

The results of this cluster comparison analysis suggest that, except in a few cases, there is not significant overlap between neighborhood clustering of venues and neighborhood clustering of health status indicators. While it is difficult to infer, base solely on correlation, a causal relationship between a given neighborhood venue and a specific health status indicator, several plausible relationships between each of the 5 pairs of neighborhood cluster categories seem to make sense. In Venue Cluster 1, the 10 most frequent venues among which are Rental Car Location, Convenience and Thrift Vintage Stores, as well a Deli and Dance Studio, in addition to plenty of Open Space, seem appropriate for a white, older, well educated, healthy and physically active population as described by Health Status Cluster 1. The high rate of psychiatric hospitalization may simply be due to the aged population and high rates of health insurance including Medicare. In Venue Cluster 2, the top 10 venues appear to be food establishments which may support the healthy, but otherwise low income, mixed population of food workers as described in Health Status Cluster 4. In Venue Cluster 3, the top 10 venues include Ice Cream Shops, Climbing Gyms, Grocery Stores, Sandwich and Coffee Shops, as well Pet Stores and ethnic restaurants. These are venues which would be largely consistent with the white, high income, healthy, and well-educated population of Health Status Cluster 5. In Venue Cluster 4, the top 10 venues appear to be high end restaurants and specialty stores which are supported by the white and well-educated population of Health Status Cluster 2. In Venue Cluster 5 the wide variety of ethnic restaurants may provide employment for primarily the white, poorly educated workers in Health Status Cluster 3. The high rates of Hepatitis B & C seem consistent with spread by food contamination from untrained food preparers. The high rate of Gonorrhea may also be linked to lack of education on safe sexual practices. In addition, and to a lesser extent, populations from Health Status Clusters 2, 4, and 5 may be consumers of goods and services produced in Venue Cluster 5.

Conclusion

The lack of significant overlap between the way Boston Neighborhoods cluster by top Venues and Health Status indicators is disappointing but, strongly suggests that the venues in a neighborhood may not be a significant contributor to the heath status of a neighborhood and that other factors may be more influential in impacting neighborhood health. While we can speculate on several potential correlations between health status indicators and venues in a neighborhood, this is difficult to prove and causality cannot be inferred. Since the granularity of neighborhoods in the venue clusters was greater than in the health status clusters, as a result of having to combine some neighborhoods together due to unavailability of specific neighborhood health metrics, this could possibly introduce some bias in the results of this study. Other limitations of the present study include absence of additional significant health indicators and lack of significant differences in the venues available in the immediate neighborhoods surrounding Boston. Undertaking a similar study across more diverse regions and populations may produces more insightful findings.