Segmenting and Clustering Neighborhoods in Boston

John F. McCarthy

Coursea Capstone Project

August 23, 1019

Introduction

Hypothesis

Health status indicators directly correlate with the frequency of specific venues within 500 meters of the centroids in 24 Boston neighborhoods

Value of Analysis

This analysis should be 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

Geospatial Data

Latitude and Longitude of each Boston neighborhood was derived from the shape files at https://data.boston.gov

The centroids of each neighborhood were calculated using Arcgis online and the tools provided

Health Status Data

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

Processing Geospatial Data

- Pandas data frame created from CSV file at https://data.boston.gov
- Python used to filter and relabel some columns
- Geocoder library (geopy.geocoders) was used to extract longitude and latitude for each neighborhood
- Folium library used to create a map of the Boston neighborhoods with superimposed centroids
- Foursquare API was employed to generate a list of venues in each Boston neighborhood
- Resultant JSON file was processed to create a new data frame containing the category name of each venue and one-hot encoding was then used to obtain the frequency of venue occurrence by neighborhood
- KMeans clustering was applied using the scikit-learn library (sklearn.cluster) with cluster size of 5 (k=5)
- Resultant clusters were superimposed on the Boston neighborhood map using the Folium library and the top 10 venue categories in each of the clusters were examined

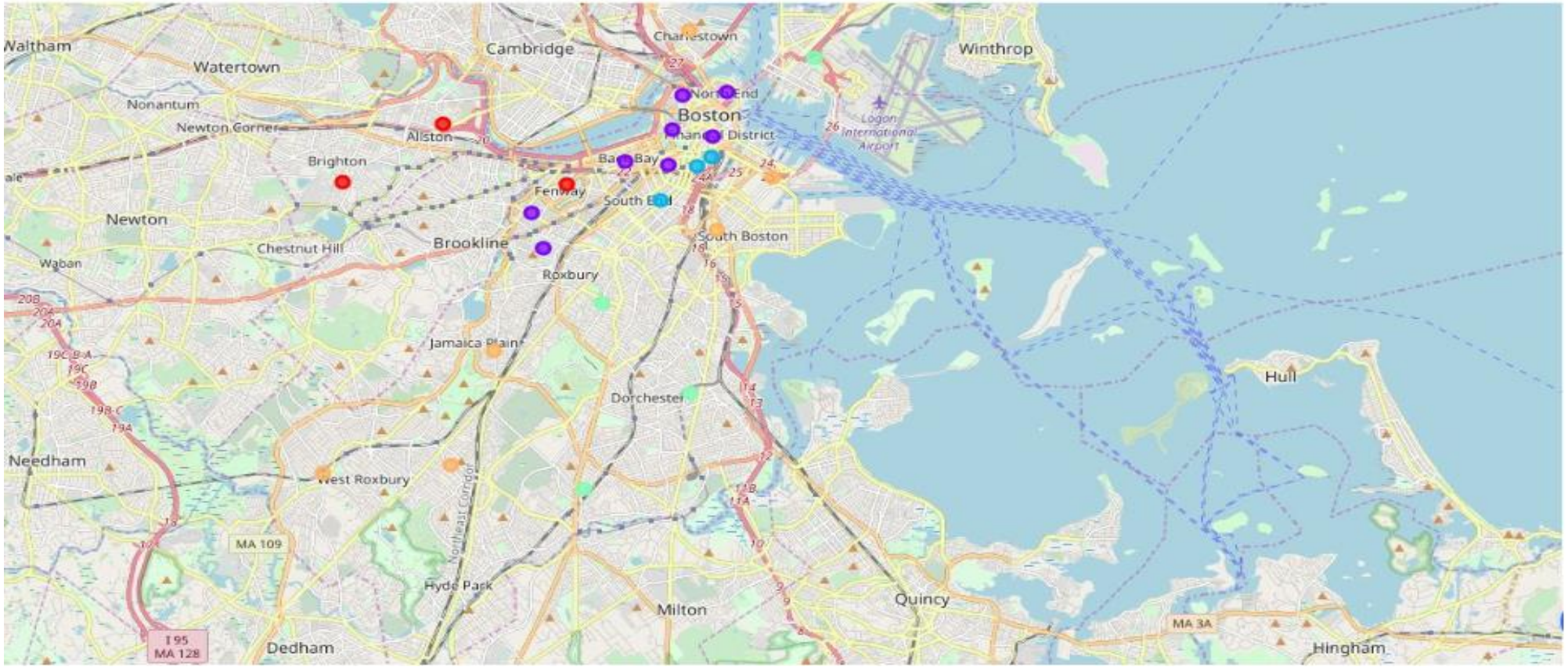
Methodology

Processing Health Status Data

- A CSV file was created and populated using data extracted from neighborhood maps contained in the online document *"Health of Boston 2016-2017"* and loaded into a Pandas data frame
  - 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)
- KMeans clustering using the *s*cikit-learn library with 5 clusters (k=5) was performed
- Clusters were summarized by taking the mean across each neighborhood health metric within a cluster
  - Each cluster summary metric was reported on a -1, 0, 1 scale for below, equal, or above the cluster mean
  - The degree of neighborhood overlap between each vendor cluster and its corresponding health status cluster was determined
- Each health status cluster was examined, aligned with the corresponding venue cluster, and speculations proposed regarding the impact of specific vendor categories on the heath status of neighborhoods in a given cluster

Results

Clustering by Health Status



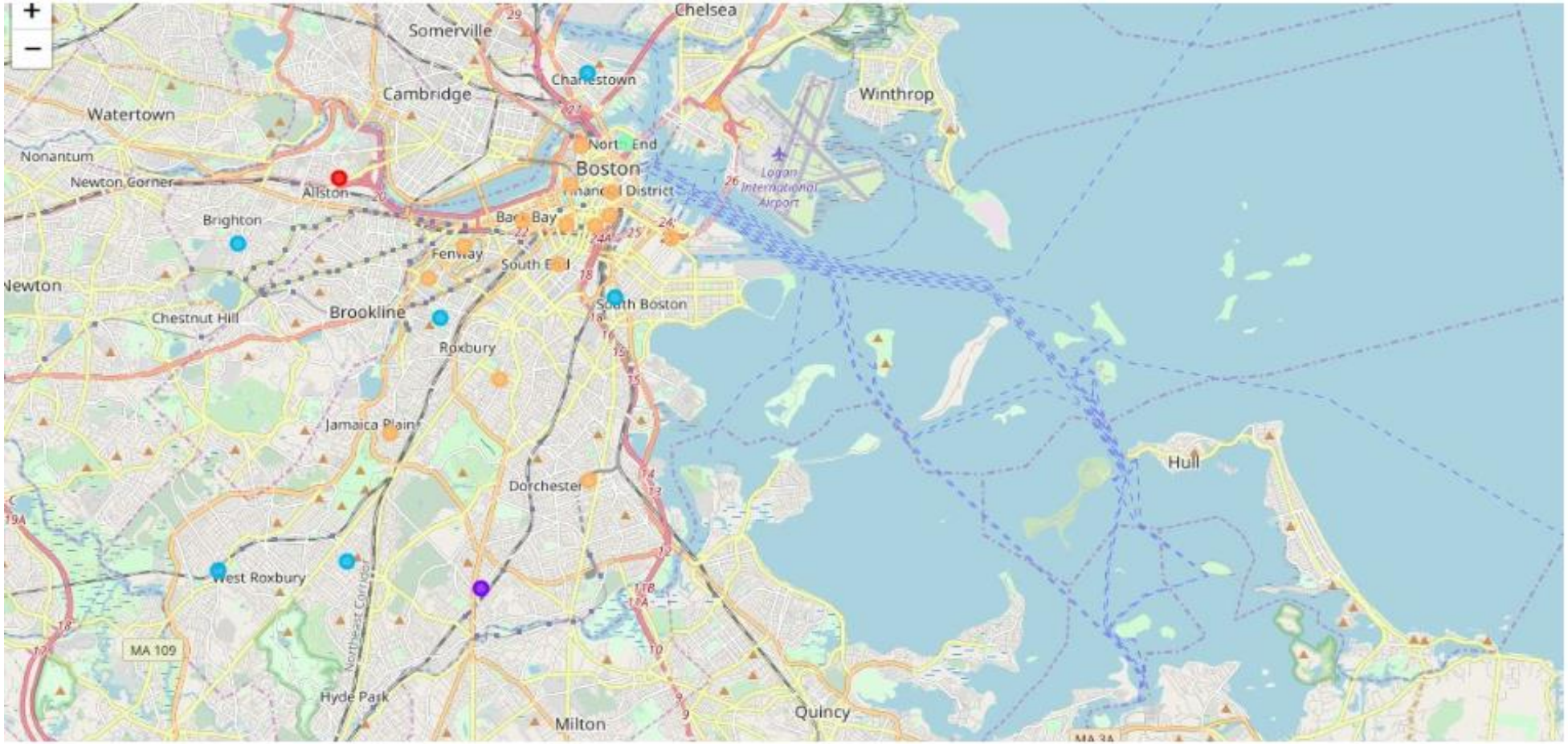
Results

Health Status Clusters

- **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

Results

Clustering by Venues Within 500 M of Neighborhood Centroids



Results

Neighborhood Venue Clusters

- **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.

Results

Venue and Health Status Cluster Overlap

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

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

- Results of this cluster comparison analysis suggest that 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
    - 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

- No significant overlap between the way Boston Neighborhoods cluster by top Venues and Health Status indicators
- 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
- The granularity of neighborhoods in the venue clusters was greater than in the health status clusters
  - Result of having to combine some neighborhoods together due to unavailability of neighborhood specific health metrics possibly introducing some bias in the results of this study.
- Further 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.