Multiple City K-means Cluster Analysis

Coursera Applied Data Science Capstone Project

RICHARD C. ANDERSON

31 MAY 2020

Introduction

This study was conceived to answer the question:

> *Would a K-means clustering analysis pick up on the differences that I know subjectively about Manhattan, the Boston Metro area, and Central Houston?*

- Expected Manhattan and Boston Metro area neighborhoods to be similar and share cluster types.
- Expected Central Houston to have different cluster types
  - Limited similarity to Manhattan and Boston Metro neighborhoods

Potential applications if differentiation is successful include:

- Locating new target market areas for goods and services
- Tailoring good and services to match neighborhood types
- Personal relocation recommendations

Data Acquisition and Cleaning

Neighborhood Geo-location Data

- Manhattan data from json file
  - https://cocl.us/new_york_dataset
- Boston Metro data manually compiled
  - Wikipedia Neighborhoods in Boston
- Central Houston data manually compile
  - Wikipedia Houston List of Neighborhoods

Neighborhood Venue Data

- Retrieved by Foursquare explore queries

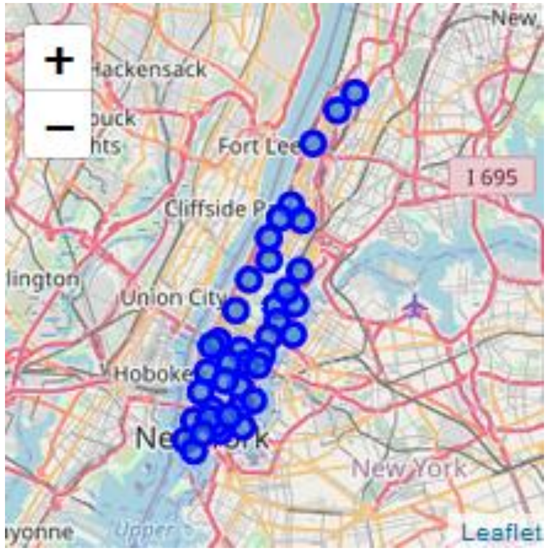
Metro Area Neighborhood Datasets

Metro Area	Neighbor-hoods	Venue Categories	Venues
Manhattan	40	330	3093
Boston Metro	50	301	3936
Central Houston	48	273	3709
Combined Metros	138	431	10378

Data Cleaning

- 1 Houston neighborhood dropped due to consistent cluster outlier status
- ‘Neighborhood’ venue category dropped from Foursquare results

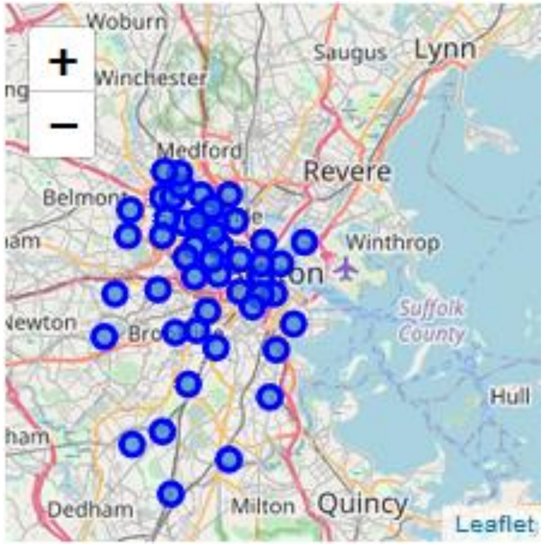
Metro Area Neighborhood Definition



Manhattan

- Borough of Manhattan

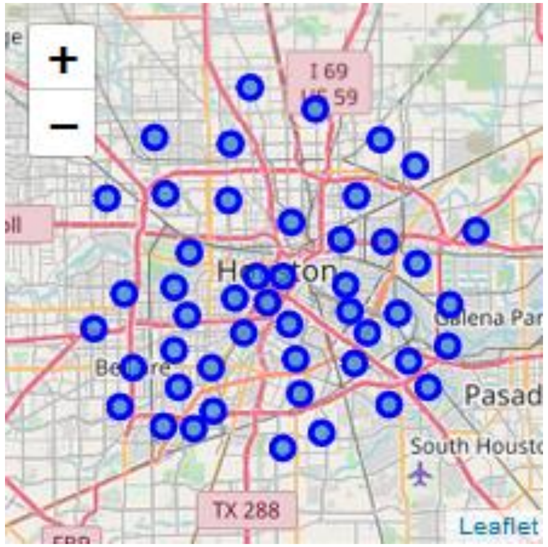
Boston Metro



- Boston
- Brookline
- Cambridge
- Somerville

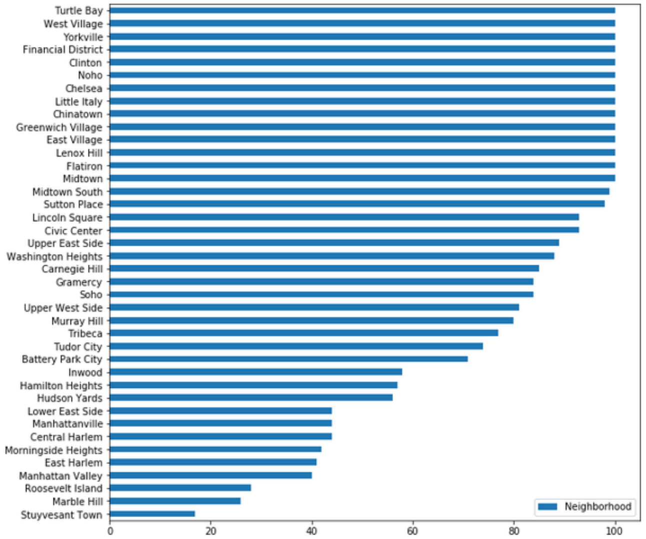
Central Houston

- Inside 610 Loop
- First ring outside 610 Loop
- Bellaire
- West University Place



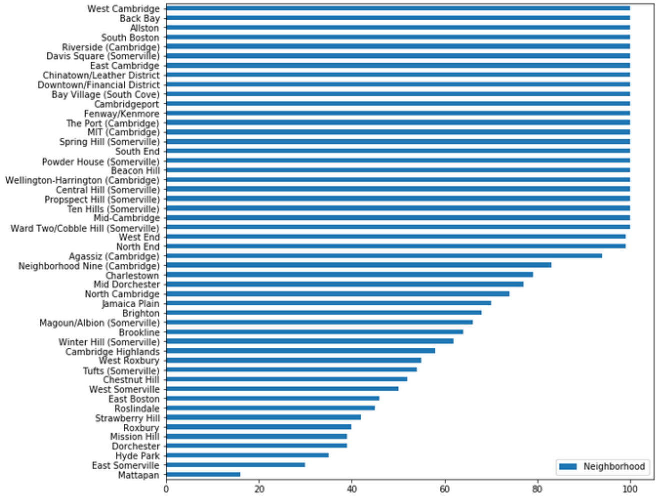
Metro Area Neighborhood Venue Counts

Foursquare search query radius required adjustment relative to city density to obtain similar venue count profiles.



Manhattan

- Search Radius: 500 meters
- 330 Venue Categories
- 3093 Venues

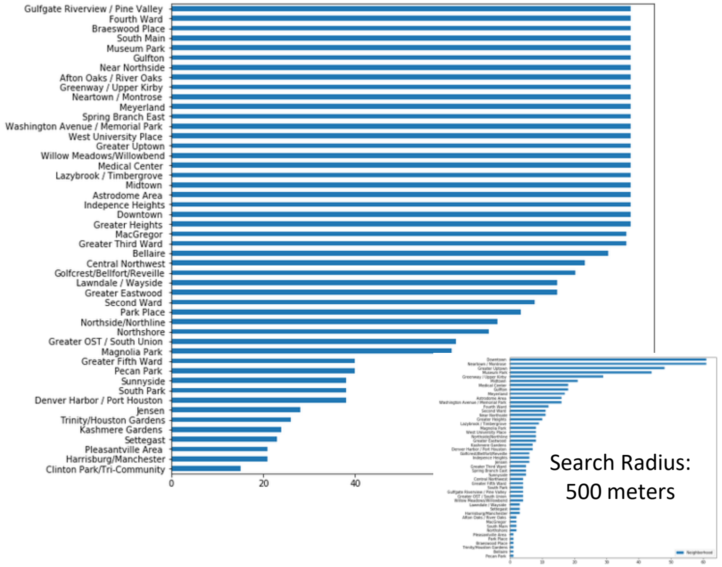


Boston Metro

- Search Radius: 1000 meters
- 301 Venue Categories
- 3936 Venues

Central Houston

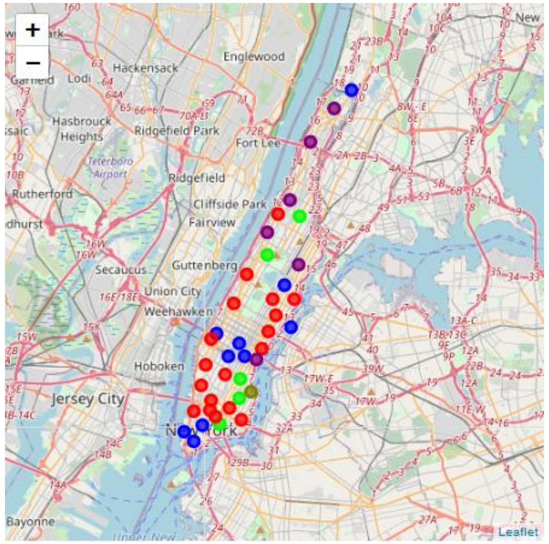
- Search Radius: 2500 meters
- 273 Venue Categories
- 3709 Venues



Search Radius: 500 meters

Metro Area
Neighborhood Venue K-means Clusters

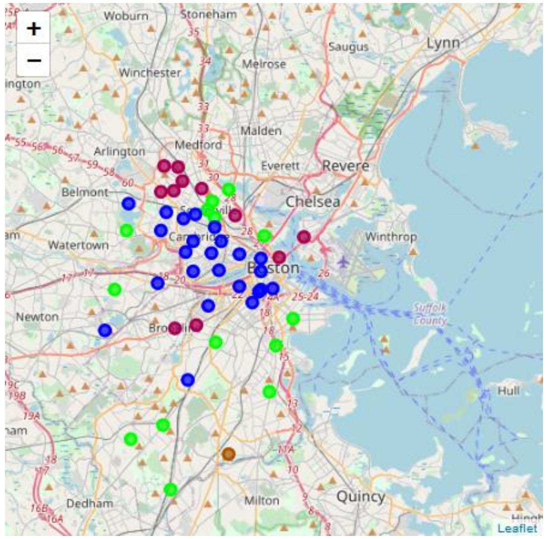
K-means clustering on neighborhood venues was performed for the individual metro areas to aid in data cleansing and to get a feel for how the neighborhoods were clustered within a city.



Manhattan

- 5 Clusters
- 1 Outlier Cluster

  Stuyvesant Town



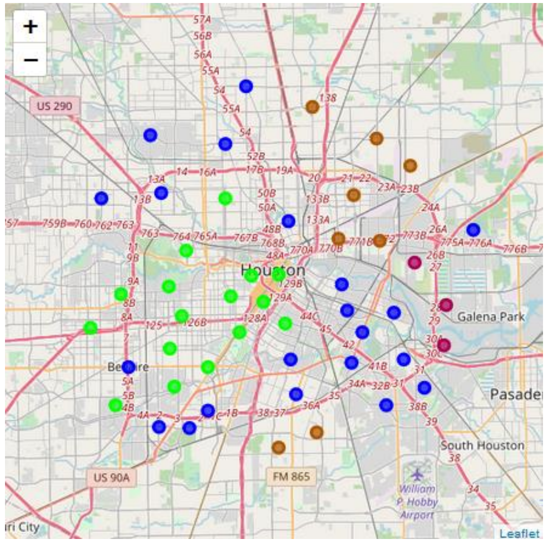
Boston Metro

- 4 Clusters

Central Houston

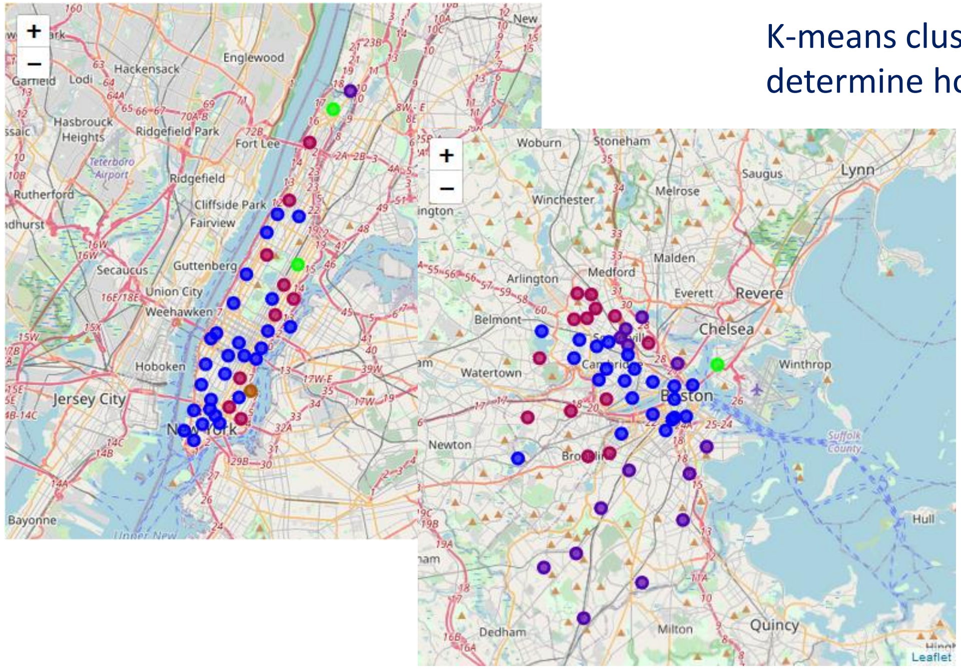
- 4 Clusters
- 1 Outlier Cluster

  - eliminated in data cleansing



Combined Metro Areas
Neighborhood Venue K-means Clusters

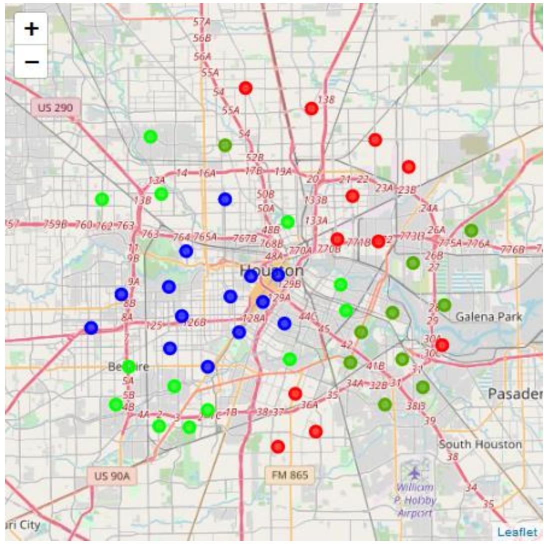
K-means clustering on the combined metro areas venues was performed to determine how the neighborhoods clustered across all three metro areas.



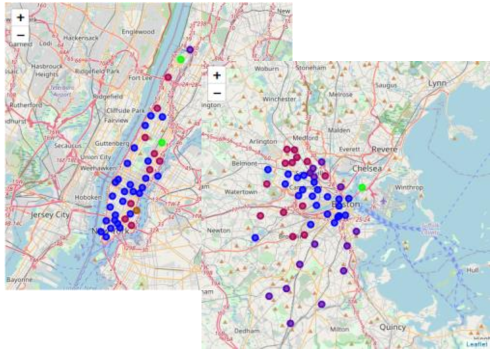
Most Manhattan and Boston Metro neighborhoods grouped into two clusters.

Most Central Houston neighborhoods grouped into three clusters (red, green, and dark green) that shared little membership with Manhattan or Boston Metro.

Central Houston had only one neighborhood cluster (blue) that shared significant membership with Manhattan and Boston Metro.



Results & Recommendations

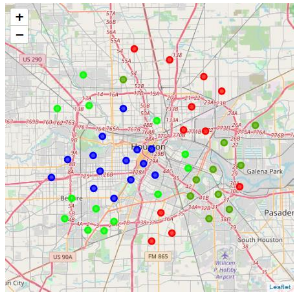


Results

- The combined metro areas K-means clustering results align well with the original assumptions made for the study project. As expected, the neighborhoods of the older cities of Manhattan and Boston neighborhoods differed significantly from the much newer city of Houston.
- As Manhattan and Boston were mostly represented by two neighborhood clusters while Houston required four clusters, there is an implication that the neighborhoods within Houston are less homogenous than those of Manhattan and Boston.

Recommendations

- Investigate using neighborhood specific venue search radius instead of a fixed radius for a metro area.
- Additional data cleansing to aggregate possible redundant venue categories.
- Further investigate impact on K-means clustering results when including outlier neighborhoods in venue datasets.
- Expand venue dataset to include additional older and newer metro areas to confirm the results and conclusions of this study.



Conclusion

This study evaluated whether K-mean clustering could be used to differentiate neighborhoods by their venues across multiple metro 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 that was used to characterize the neighborhoods of each metro area.

The K-means clustering evaluation 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.

Recommended next step is to validate the study results by adding venue data for additional metro areas.