Moving from Manhattan to Cambridge

Jennifer Loutrel

May 5, 2019

Introduction/Business Problem

Colleges and Universities employ a large number of young professionals in both administrative and research positions. As careers grow and change, individuals often change schools to align their research interests and continue to grow professionally. Even though a new school and/or program may be a good fit professionally, it's often hard to determine where to live that is close enough to the school but will also align with personal interests and community expectations.

A young professional working at Columbia University has just accepted a new position at Harvard University. She loves her current community in the Manhattan neighborhood of Morningside Heights, but is now looking to move from New York City, NY to Cambridge, MA. Before she moves, she wants to learn more about where she will be living and find a location to live that is (1) similar to her current Manhattan neighborhood of Morningside Heights and (2) has as many Coffee Shops as possible.

To explore her new home and choose a location, she will:

1. Understand the locations of the Cambridge neighborhoods by viewing them on a map.
2. Determine which Cambridge neighborhoods are the most similar to her Manhattan neighborhood of Morningside Heights
3. Identify the frequency of Coffee Shops within a 500 meter radius in each Cambridge neighborhood.

Data

1. List of New York City neighborhoods with neighborhood name, latitude, and longitude. (This file is available at https://geo.nyu.edu/catalog/nyu_2451_34572)
2. List of Cambridge neighborhoods with neighborhood name, latitude, and longitude. (This file is available at https://geo.nyu.edu/catalog/harvard-cambridge14cddneighborhoods)
3. Latitude and Longitude of Cambridge, MA (obtained via geopy library)
4. Latitude and Longitude of Manhattan, NY (obtained via geopy library)
5. List of venues with categories (Foursquare API data for New York City, NY and Cambridge, MA)

Methodology

A. Neighborhood Data Exploration

Before performing any analysis, the data will be cleaned, organized and examined to understand what data is available. First, imperative data will be pulled from the shapefile datasets (Borough, Neighborhood, Latitude and Longitude) for both New York and Cambridge. The resulting data will be merged into a single dataframe to allow for clustering analysis. To visualize the layout of the neighborhoods, a map will be generated for both Manhattan and Cambridge.

1. The Borough, Neighborhood, Latitude and Longitude will be extracted from the json files of the list of New York City neighborhoods (#1) and placed into a pandas dataframe. A new dataframe that only contains the information for the neighborhood of Manhattan will be created.
2. The Borough, Neighborhood, Latitude and Longitude will be extracted from the json files of the list of Cambridge neighborhoods (#2) and placed into a pandas dataframe. For ease of analysis, it is assumed the the first point listed in the polygon shape is the latitude and longitude point that represents the neighborhood.
3. The Manhattan and Cambridge dataframes will be merged.
4. The latitude and longitude of Manhattan, NY (#4) from geopy will be used to generate a Folium map. The latitude and longitudes of Manhattan neighborhoods will be pulled from the list of Manhattan neighborhoods (#1) and points will be added to the existing Folium map.
5. The latitude and longitude of Cambridge, MA (#3) from geopy will be used to generate a Folium map. The latitude and longitudes of Cambridge neighborhoods will be pulled from the list of Cambridge neighborhoods (#2) and points will be added to the existing Folium map.

B. *Foursquare Data Exploration*

Foursquare credentials allow for the analysis of both Manhattan and Cambridge neighborhoods. A full list of venues will be pulled to provide as much information to the clustering algorithm as possible.

6. A list of venues information will be obtained for all neighborhoods in the merged dataframe by using a GET request through the Foursquare API and passing the latitudes, longitudes, and a radius of 500 meters.
7. The category type for each venue will be extracted and a dataframe containing neighborhood, neighborhood latitude, neighborhood longitude, venue, and venue category will be created.
8. The dataframe will be grouped by neighborhood and the frequency of occurrence of each category will be determined.

C. *K-means Clustering*

K-means clustering is used to create groups of objects whose members are similar in some way. Clustering will allow us to find Cambridge neighborhoods that have similar venues to Morningside Heights.

9. The grouped frequency dataframe will be used to cluster the neighborhoods into 5 clusters using k-mean clustering.
10. The clusters will be searched for the one that contains the Manhattan neighborhood of 'Morningside Heights'.
11. The top 5 most common venues for each Cambridge neighborhood in the cluster containing Morningside Heights will be printed.

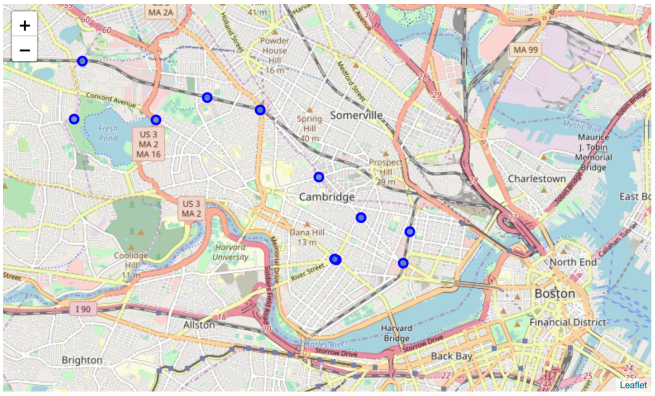
D. *Selecting a Neighborhood*

Using all of the data obtained and cleaned so far, selecting a neighborhood requires joining the frequency of coffee shops with the Cambridge neighborhoods that are most similar to Morningside Heights from clustering.

12. The frequency of coffee shops for all neighborhoods will be printed.
13. The frequency dataframe and common venues dataframe will be joined.
14. A recommendation will be given based on (1) similarity as determined by clustering, (2) the highest frequency of coffee shops, (3) other assumptions as needed.

Results

The downloaded data files were successfully loaded and latitude and longitude information for both Manhattan and Cambridge neighborhoods was extracted and merged into a single dataframe for analysis. Cambridge neighborhoods were visualized in Folium:



Foursquare venue data for all Manhattan and Cambridge neighborhoods was obtained and placed into a frequency table.

	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	Argent Restau
0	Agassiz	0.000000	0.00	0.00	0.000000	0.029412	0.00	0.00	0.000000	0.0000
1	Area 2/MIT	0.000000	0.00	0.00	0.000000	0.025641	0.00	0.00	0.000000	0.0000
2	Area Four	0.000000	0.00	0.00	0.000000	0.000000	0.00	0.00	0.000000	0.0000
3	Battery Park City	0.000000	0.00	0.00	0.000000	0.010000	0.00	0.00	0.000000	0.0000
4	Cambridge Highlands	0.000000	0.00	0.00	0.000000	0.000000	0.00	0.00	0.000000	0.0000
5	Cambridgeport	0.000000	0.00	0.00	0.000000	0.040816	0.00	0.00	0.000000	0.0000
6	Carnegie Hill	0.000000	0.00	0.00	0.000000	0.010000	0.00	0.00	0.000000	0.0100
7	Central Harlem	0.000000	0.00	0.00	0.069767	0.046512	0.00	0.00	0.000000	0.0000
8	Chelsea	0.000000	0.00	0.00	0.000000	0.040000	0.01	0.00	0.000000	0.0000
9	Chinatown	0.000000	0.00	0.00	0.000000	0.040000	0.00	0.00	0.000000	0.0000
10	Civic Center	0.000000	0.00	0.00	0.000000	0.030000	0.01	0.00	0.000000	0.0000

Image represents a subset of the data. See notebook for full table.

The venue data was run through a K-means clustering algorithm which successfully generated 5 clusters of similar neighborhoods. The cluster containing Morningside Heights, cluster number 0, was selected.

22	Manhattan	Morningside Heights	40.808000	-73.963896	0	Coffee Shop	American Restaurant	Bookstore	Park	Fo

The Cambridge neighborhoods in this cluster were extracted and the top 5 venues were obtained.

	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Cambridge	Area 2/MIT	42.364903	-71.089729	0	Café	Bar	Coffee Shop	Italian Restaurant	Hotel
1	Cambridge	Mid-Cambridge	42.378088	-71.107275	0	Grocery Store	Coffee Shop	Climbing Gym	New American Restaurant	Bus Stop
2	Cambridge	Wellington-Harrington	42.371835	-71.098528	0	Coffee Shop	Park	Bar	New American Restaurant	Food & Drink Shop
3	Cambridge	Area Four	42.371835	-71.098528	0	Coffee Shop	Park	Bar	New American Restaurant	Food & Drink Shop
4	Cambridge	East Cambridge	42.369655	-71.088493	0	Café	American Restaurant	Fish Market	Coffee Shop	Italian Restaurant
5	Cambridge	Cambridge Highlands	42.395856	-71.156412	0	Gourmet Shop	Deli / Bodega	Trail	Donut Shop	Art Gallery
6	Cambridge	West Cambridge	42.386807	-71.141027	0	Mattress Store	Bakery	Pet Store	Grocery Store	Gym

This neighborhood cluster was merged with coffee shop frequency to obtain the final result:

	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Coffee Shop
1	Cambridge	Mid-Cambridge	42.378088	-71.107275	0	Grocery Store	Coffee Shop	Climbing Gym	New American Restaurant	Bus Stop	0.080000
2	Cambridge	Wellington-Harrington	42.371835	-71.098528	0	Coffee Shop	Park	Bar	New American Restaurant	Food & Drink Shop	0.076923
3	Cambridge	Area Four	42.371835	-71.098528	0	Coffee Shop	Park	Bar	New American Restaurant	Food & Drink Shop	0.076923
4	Cambridge	East Cambridge	42.369655	-71.088493	0	Café	American Restaurant	Fish Market	Coffee Shop	Italian Restaurant	0.055556
0	Cambridge	Area 2/MIT	42.364903	-71.089729	0	Café	Bar	Coffee Shop	Italian Restaurant	Hotel	0.051282
6	Cambridge	West Cambridge	42.386807	-71.141027	0	Mattress Store	Bakery	Pet Store	Grocery Store	Gym	0.029412
5	Cambridge	Cambridge Highlands	42.395856	-71.156412	0	Gourmet Shop	Deli / Bodega	Trail	Donut Shop	Art Gallery	0.000000

Discussion

The young professional now has an understanding of where Cambridge neighborhoods are as well as the types of venues in each neighborhood. Running a clustering analysis allowed her to discover 7 neighborhoods that were similar to her current neighborhood of Morningside Heights. These 7 neighborhoods had coffee shop frequencies ranging from 0 to 0.08. The Cambridge neighborhood that was most similar to Morningside Heights with the greatest frequency of coffee shops (0.08) is Mid-Cambridge. Therefore, she should look for housing options in Mid-Cambridge to ensure she will love her new home.

Conclusion

This analysis looked for similar neighborhoods in two cities. A strong recommendation for a likely neighborhood, Mid-Cambridge, was generated. Additional analysis could include examination of other venues, similarity of surrounding neighborhoods, distance from work and friends, and accessibility of public transportation. Determining a new place to live is often difficult, but data science can ease this process.