IBM DATA SCIENCE CAPSTONE PROJECT REPORT
Full Report
An Analysis of Boston, MA Neighborhood Restaurant Data

Intro/Business Problem

The city of Boston, Massachusetts is a small city on the northeast coast of the United States. The city is well known for its many cultural and historical attractions that draw in approximately 20 million visitors per year from both the United States and abroad. Additionally the city is home to thousands of college students for much of the year.

The city itself is not very large area-wise and the tourist attractions tend to be contained in one small area of the city with more residential neighborhoods around that. The tourist areas do have some residential areas but these tend to be very upscale and expensive, such as: the North End and Back Bay. The residential areas outside of that zone tend to be more affordable and are where middle-class and college students tend to reside, such as: Allston and Roxbury.

I wish to do an analysis of these neighborhoods to see if there is a difference between the restaurant types that are most prevalent in the tourist areas versus the residential areas. This would assist someone who wishes to open a restaurant in these areas in determining which type of restaurant may be successful.

Data Description

As required by the assignment I will be using restaurant data gathering via a query to the Foursquare API. Foursquare is an American tech company founded in New York City in 2008. According to their Wikipedia page, “The company rose to prominence with the launch of its namesake local search-and-discovery mobile app, now known as Foursquare City Guide, which popularized the concept of real-time location-sharing and checking-in.”

Also used will be zip code (postal code) data for the city scraped from the web page: https://en.wikipedia.org/wiki/Boston#Demographic_breakdown_by_ZIP_Code. From this data I will only use zip code and neighborhood data. I also took a screen capture from here: https://data.boston.gov/dataset/city-of-boston-boundary1/resource/7e6574df-a713-4617-b4f2-f4588fd211d7, but this data is not used in the analysis.

Methodology

In this section I will describe how the data analysis was performed.
The first step was to acquire some neighborhood data for Boston that contained Zip Code (Postal Code) data. To do so I used the Pandas read_url() function on the website: https://en.wikipedia.org/wiki/Boston#Demographic_breakdown_by_ZIP_Code.

The result was a Dataframe:

	Rank	ZIP code (ZCTA)	Per capitaincome	Medianhouseholdincome	Medianfamilyincome	Population	Number ofhouseholds
0	1.0	02110 (Financial District)	$152,007	$123,795	$196,518	1486	981
1	2.0	02199 (Prudential Center)	$151,060	$107,159	$146,786	1290	823
2	3.0	02210 (Fort Point)	$93,078	$111,061	$223,411	1905	1088
3	4.0	02109 (North End)	$88,921	$128,022	$162,045	4277	2190
4	5.0	02116 (Back Bay/Bay Village)	$81,458	$87,630	$134,875	21318	10938

In this Dataframe the zip code and neighborhoods are combined in one field so I had to break those into their own fields using a combination of Python's str.split() and str.replace() functions. Now they have their own fields:

	Rank	ZIP code (ZCTA)	Per capitaincome	Medianhouseholdincome	Medianfamilyincome	Population	Number ofhouseholds	Zip	Neighborhood
0	1.0	02110 (Financial District)	$152,007	$123,795	$196,518	1486	981	02110	Financial District
1	2.0	02199 (Prudential Center)	$151,060	$107,159	$146,786	1290	823	02199	Prudential Center
2	3.0	02210 (Fort Point)	$93,078	$111,061	$223,411	1905	1088	02210	Fort Point
3	4.0	02109 (North End)	$88,921	$128,022	$162,045	4277	2190	02109	North End
4	5.0	02116 (Back Bay/Bay Village)	$81,458	$87,630	$134,875	21318	10938	02116	Back Bay/Bay Village
5	6.0	02108 (Beacon Hill/Financial District)	$78,569	$95,753	$153,618	4155	2337	02108	Beacon Hill/Financial District
6	7.0	02114 (Beacon Hill/West End)	$65,865	$79,734	$169,107	11933	6752	02114	Beacon Hill/West End

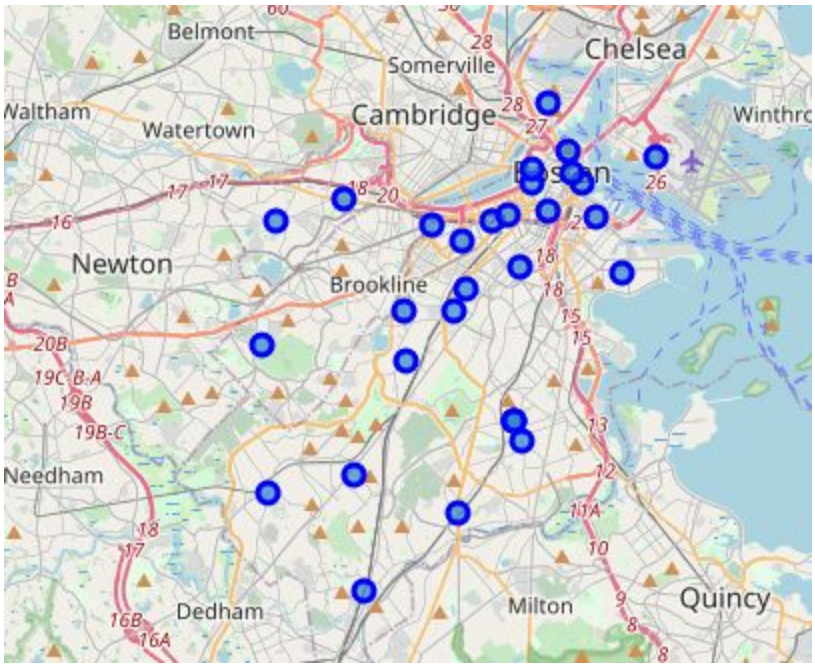
Since I only needed the 2 fields I just created I generated a new Dataframe with just those columns:

	Zip	Neighborhood
0	02110	Financial District
1	02199	Prudential Center
2	02210	Fort Point
3	02109	North End
4	02116	Back Bay/Bay Village

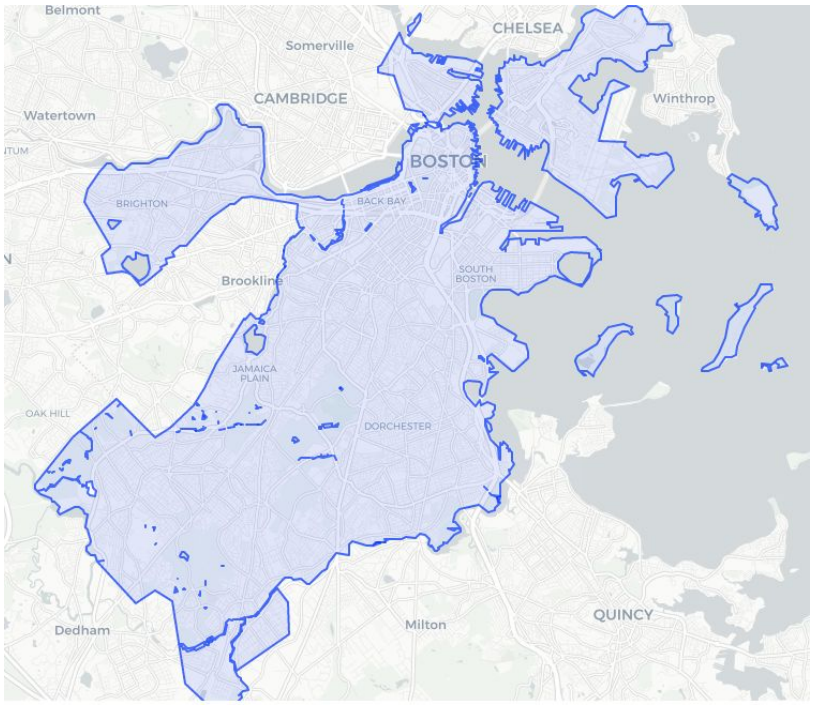
I used the geopy Nominatim() function to add latitude and longitude coordinates based on the neighborhoods zip code and added them to the Dataframe:

	Zip	Neighborhood	Latitude	Longitude
0	02110	Financial District	42.3576	-71.0514
1	02199	Prudential Center	42.3479	-71.0825
2	02210	Fort Point	42.3489	-71.0465
3	02109	North End	42.3600	-71.0545
4	02116	Back Bay/Bay Village	42.3492	-71.0768
5	02108	Beacon Hill/Financial District	42.3576	-71.0684
6	02114	Beacon Hill/West End	42.3611	-71.0682
7	02111	Chinatown/Financial District/Leather District	42.3503	-71.0629
8	02129	Charlestown	42.3778	-71.0627

Next up was to use Folium to map out the neighborhoods:



For reference, here is an outline of the Boston city limits.As you can see the city has an irregular shape which did have an effect on the analysis, which will be discussed later:



The next step was to query the Foursquare API and get data for the restaurant types within my neighborhoods:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Financial District	42.3576	-71.0514	Kane's Donuts	42.356209	-71.052895	Donut Shop
1	Financial District	42.3576	-71.0514	State Street Provisions	42.359507	-71.051386	American Restaurant
2	Financial District	42.3576	-71.0514	James Hook & Company	42.354960	-71.050911	Seafood Restaurant
3	Financial District	42.3576	-71.0514	Casa Razdora	42.358231	-71.054741	Italian Restaurant
4	Financial District	42.3576	-71.0514	Legal Sea Foods	42.359411	-71.050847	Seafood Restaurant

I performed a one-hot encoding (0/1) on the retrieved data then used groupby() and mean() functions to obtain the frequency of each restaurant category as shown in these 2 screenshots:

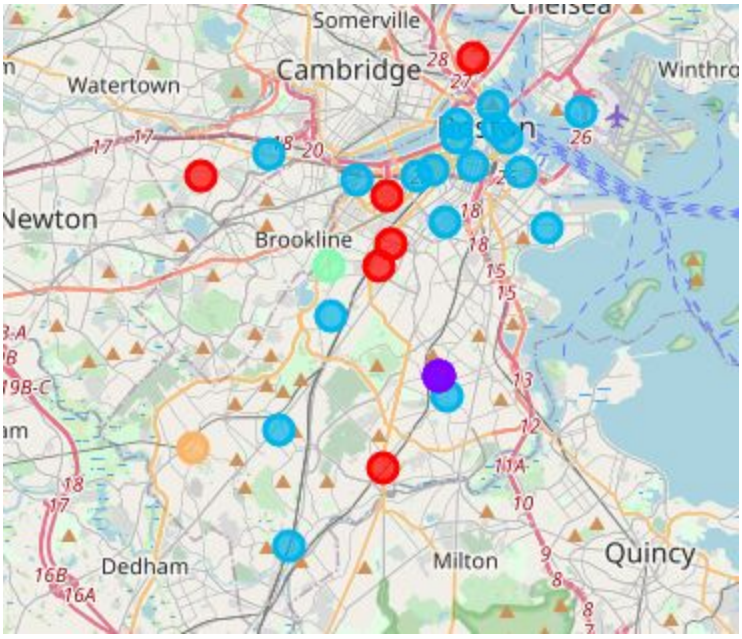
	Neighborhood	African Restaurant	American Restaurant	Arepa Restaurant	Asian Restaurant	Australian Restaurant	BBQ Joint	Bagel Shop	Bakery	Belgian Restaurant	...	Sushi Restaurant	Szechuan Restaurant	
0	Financial District	0	0	0	0	0	0	0	0	0	...	0	0	
1	Financial District	0	1	0	0	0	0	0	0	0	...	0	0	
2	Financial District	0	0	0	0	0	0	0	0	0	...	0	0	
3	Financial District	0	0	0	0	0	0	0	0	0	...	0	0	
4	Financial District	0	0	0	0	0	0	0	0	0	...	0	0	

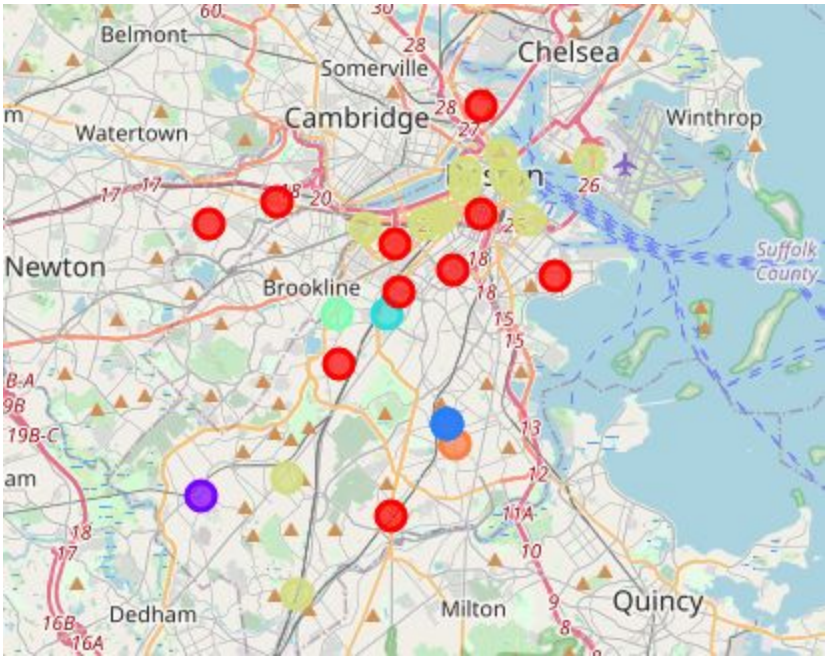
	Neighborhood	African Restaurant	American Restaurant	Arepa Restaurant	Asian Restaurant	Australian Restaurant	BBQ Joint	Bagel Shop	Bakery	Belgian Restaurant	...	Sushi Restaurant	Szechuan Restaurant	Taco Place	
0	Allston	0.000000	0.012500	0.000000	0.050000	0.00	0.000000	0.000000	0.050000	0.000000	...	0.050000	0.000000	0.012500	
1	Allston-Harvard Business School	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	
2	Back Bay, Longwood, Museum of Fine Arts/Sympho...	0.000000	0.115385	0.000000	0.000000	0.00	0.000000	0.000000	0.115385	0.000000	...	0.076923	0.000000	0.000000	
3	Back Bay/Bay Village	0.000000	0.114286	0.000000	0.042857	0.00	0.000000	0.014286	0.042857	0.000000	...	0.014286	0.000000	0.000000	
4	Beacon Hill/Financial District	0.000000	0.161290	0.000000	0.000000	0.00	0.000000	0.032258	0.032258	0.000000	...	0.064516	0.000000	0.000000	
5	Beacon Hill/West End	0.000000	0.093750	0.000000	0.000000	0.00	0.031250	0.031250	0.031250	0.000000	...	0.031250	0.000000	0.000000	
6	Brighton	0.000000	0.050000	0.000000	0.000000	0.00	0.000000	0.000000	0.150000	0.000000	...	0.050000	0.000000	0.000000	
7	Charlestown	0.000000	0.050000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	
8	Chinatown/Financial District/Leather District	0.000000	0.022989	0.000000	0.103448	0.00	0.000000	0.000000	0.091954	0.000000	...	0.045977	0.011494	0.000000	

I used this information to create a Dataframe that ranks the top 5 restaurant categories per neighborhood:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Allston	Korean Restaurant	Asian Restaurant	Pizza Place	Chinese Restaurant	Bakery
1	Allston-Harvard Business School	Ethiopian Restaurant	Spanish Restaurant	Gastropub	Sandwich Place	Latin American Restaurant
2	Back Bay, Longwood, Museum of Fine Arts/Sympho...	Pizza Place	American Restaurant	Bakery	Restaurant	Food Truck
3	Back Bay/Bay Village	American Restaurant	Italian Restaurant	Seafood Restaurant	Mexican Restaurant	New American Restaurant
4	Beacon Hill/Financial District	American Restaurant	Pizza Place	French Restaurant	Italian Restaurant	Restaurant

Next I did a cluster analysis using the k-means unsupervised machine learning cluster algorithm from the scikit-learn package. I initially did 5 clusters but the results (first image below) didn't seem specific enough for me given the analysis I wanted to do. I felt that the areas of blue and red could be differentiated more. So I raised k to equal 7 and that gave more differentiated clusters in that area (second image below):



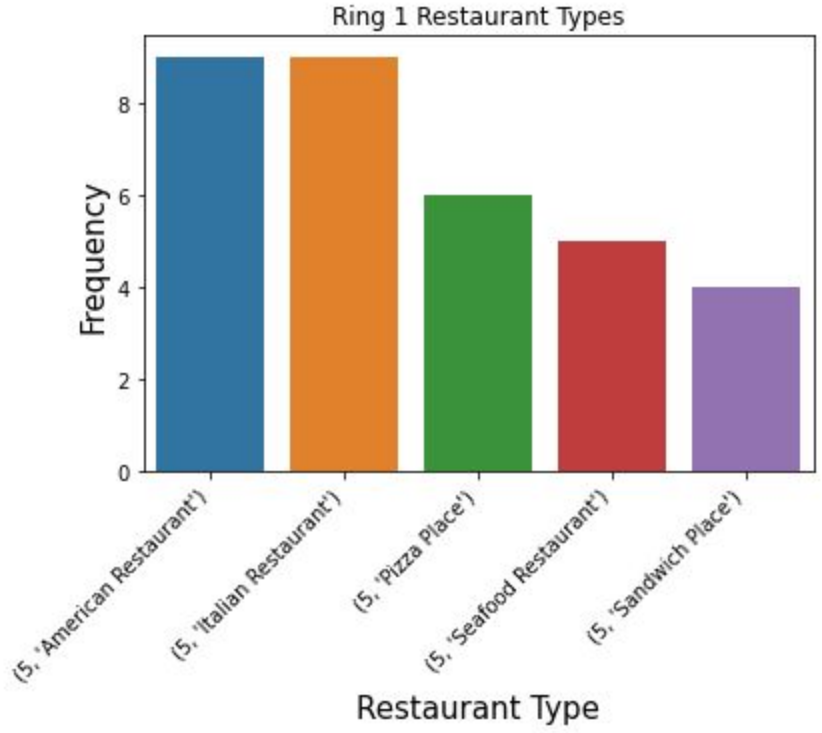


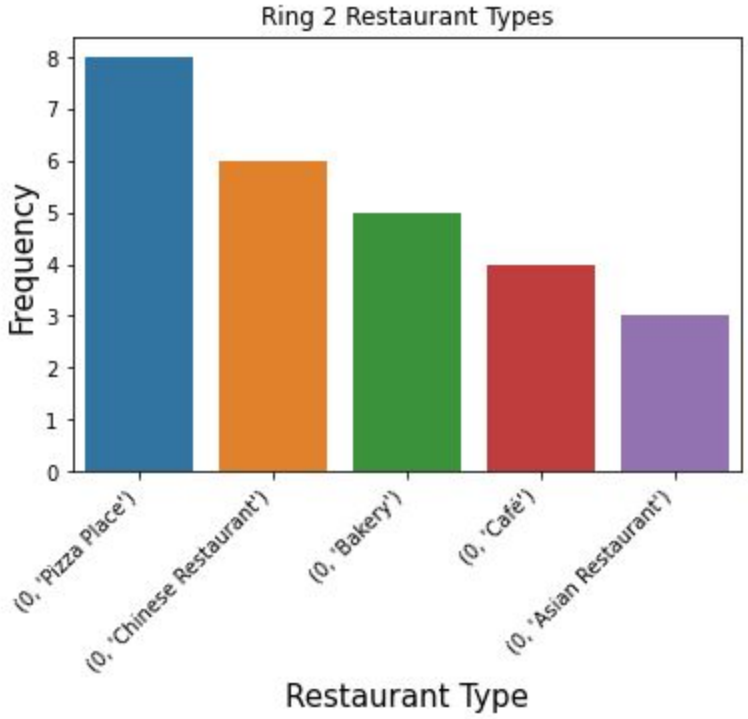
I grouped and counted the restaurant types into single Dataframes for the 2 clusters (5 and 0) I was interested in analyzing. Since the clusters appear to form concentric rings I named cluster 5 as Ring 1 and cluster 0 as Ring 2:

		Count	Venue
Cluster Labels	Venue		
5	American Restaurant	9	(5, American Restaurant)
	Italian Restaurant	9	(5, Italian Restaurant)
	Pizza Place	6	(5, Pizza Place)
	Seafood Restaurant	5	(5, Seafood Restaurant)
	Sandwich Place	4	(5, Sandwich Place)

		Count	Venue
Cluster Labels	Venue		
0	Pizza Place	8	(0, Pizza Place)
	Chinese Restaurant	6	(0, Chinese Restaurant)
	Bakery	5	(0, Bakery)
	Café	4	(0, Café)
	Asian Restaurant	3	(0, Asian Restaurant)

Finally, I created bar plots for Ring 1 and Ring 2 based upon the restaurant category groups and count:



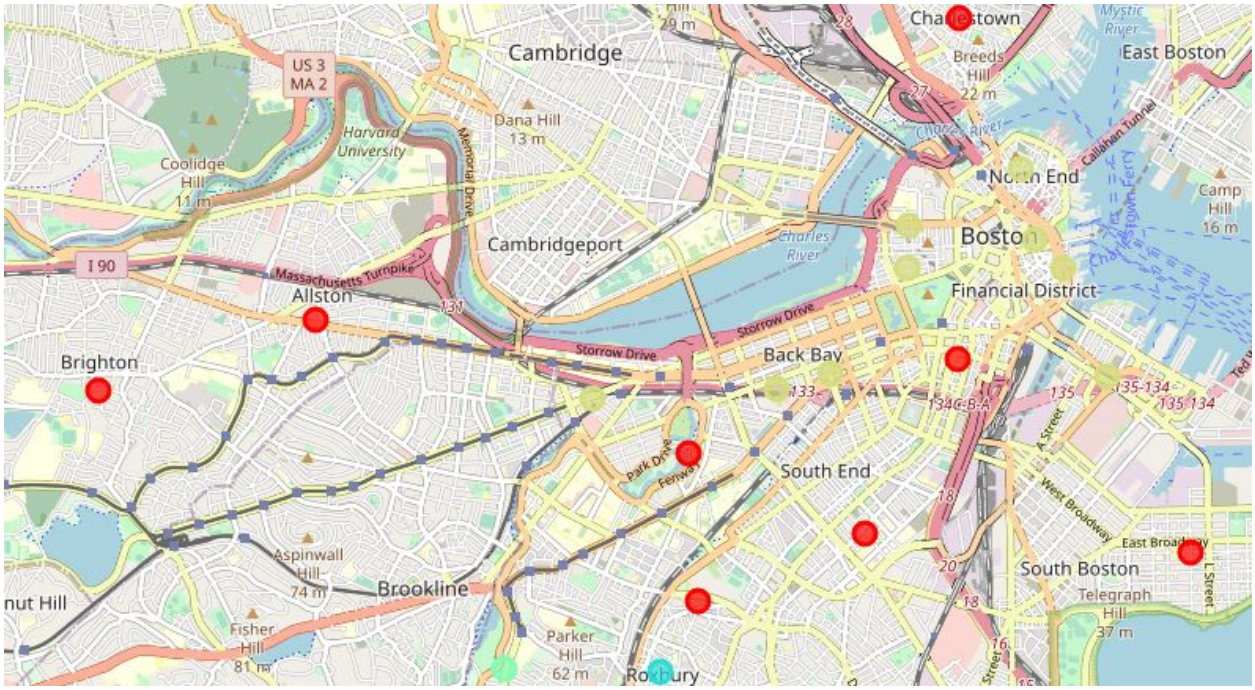


Results

The results show a very clear delineation of restaurant types between the 2 areas of the analysis. This would imply that there are a different set of customers that are eating in these 2 areas. The other clusters did not upon cursory inspection have the same patterns.

Discussion

Using my knowledge of the city as a reference these results appear to make sense. In the zoomed-in map below, the gold colored dots (Ring 1) are the tourist areas whereas the red dots (Ring 2) are the residential neighborhoods. As shown in the bar charts above the restaurant types in the tourist (and upscale residential) area tend to be more upscale, expensive, sit-down type establishments such as American, Italian and seafood. The residential areas appear to be more lower priced categories and seemingly would have take-out options such as pizza and Chinese. Also notable are more bakeries and cafes where college students would study and commuters would stop on the way to work to grab a coffee. A person or company wishing to open a restaurant in these areas could use this information to determine the type they should open in order to be most successful. For example, an upscale seafood restaurant may not be successful near apartments occupied by college students.



Conclusion

This basic analysis seems very accurate in its ability to cluster these 2 areas correctly. As shown in a previous map, Boston has irregular city limits and I believe may have hindered the analysis somewhat. My other clusters tended to be more isolated. I believe that including the cities of Cambridge and Brookline could help the model cluster areas more efficiently as they would provide a more regular-shaped area for analysis. At the same time, Brookline is a very expensive area wedged between 2 of Boston's residential areas and might skew the model even worse. Additionally, population, demographic and income data could be used to expand the model input and possibly provide a more detailed conclusion.