Corporate Neighborhood Recommender

1. Introduction

a. A corporation has offices in Boston, Chicago, Atlanta, and Denver. Employees often transfer between the locations, with stays typically lasting 2 years. To ease the transition, the human resources department would like to be able to recommend areas to look for housing based on where the employee previously lived. By analyzing the area where the employee previously lived, the HR department hopes to be able to recommend an area where the employee will feel "right at home".

2. Data

a. On the assumption the employee is happy with the area where they currently live, HR hopes to find a similar area near the office they will be moving to. "Similar" in this case means it has the same types of venues (shopping, dining, entertainment, etc) nearby. Foursquare will provide venue data both for the employee's current location and the areas around the office they will be moving to.

b. We will also need a source to identify specific neighborhoods around the offices. ~~This may include scraping Wikipedia.~~ Luckily Chicago (via Chicago Data Portal), Boston (via Boston Open Data), Atlanta (via Atlanta Region Commission), and Denver (via Denver Open Data catalog) all supply this data in geojson and / or shapefile formats.

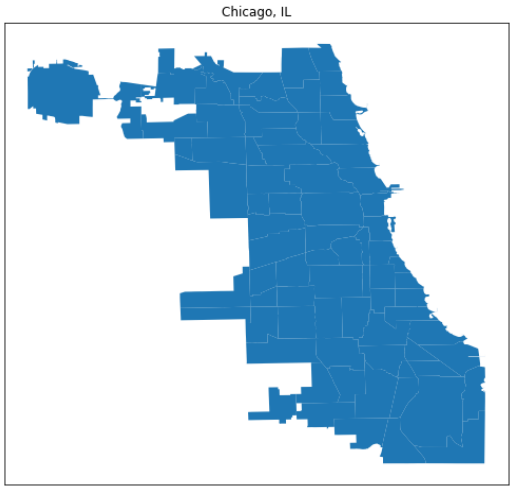
3. Methodology

To determine similarity between neighborhoods, I used clustering analysis, based on the most common types of venues in the area. This entailed the following steps.

a. Get geojson or shapefile for each city. Here's a sample from Chicago.

community	area	shape_area	perimeter	area_num_1	area_numbe	comarea_id	comarea	shape_len	geometry
DOUGLAS	0	46004621.16	0	35	35	0	0	31027.05451	MULTIPOLYGON (((-87.60914087617894 41.84469250265398, -87.60914874757808 41.84466159842403, -87.60916112041259
OAKLAND	0	16913961.04	0	36	36	0	0	19565.50615	MULTIPOLYGON (((-87.59215283879394 41.81692934626684, -87.59230805083372 41.81693210894971, -87.59489183437287
FULLER PARK	0	19916704.87	0	37	37	0	0	25339.08975	MULTIPOLYGON (((-87.62879823733725 41.80189303368919, -87.62879419577438 41.80175909929755, -87.6287945369763
GRAND BOULEVARD	0	48492503.16	0	38	38	0	0	28196.83716	MULTIPOLYGON (((-87.6067081256125 41.81681377057218, -87.60670480954393 41.81657908503414, -87.60670022649295
KENWOOD	0	29071741.93	0	39	39	0	0	23325.16791	MULTIPOLYGON (((-87.59215283879394 41.81692934626684, -87.59214911362939 41.8168007663423, -87.59215726739471
LINCOLN SQUARE	0	71352328.24	0	4	4	0	0	36624.60308	MULTIPOLYGON (((-87.6744075678037 41.97610340441675, -87.67440395329793 41.97601526547697, -87.67439621658549
WASHINGTON PARK	0	42373881.48	0	40	40	0	0	28175.31609	MULTIPOLYGON (((-87.60603749217005 41.7858740649942, -87.60626871776799 41.78587042782912, -87.6064051372113
HYDE PARK	0	45105380.17	0	41	41	0	0	29746.7082	MULTIPOLYGON (((-87.58037662085418 41.80252502205126, -87.58039585121138 41.80244312058704, -87.58041099059199
WOODLAWN	0	57815179.51	0	42	42	0	0	46936.95924	MULTIPOLYGON (((-87.57714456891335 41.78614641013439, -87.5771445356807 41.78614630508401, -87.5770936284558

b. Use the geometry feature of the geojson to map the city.



c. Get venue information from Foursquare for each neighborhood. Parameters from the venue search were a radius of 2.5K (~1.5 mi) from the neighborhood center, and a maximum of 100 venues per neighborhood.

```
LIMIT = 100
venues_corp = getNearbyVenues(cities = df_corp['Metro'],
                              names=df_corp.Neighborhood,
                                latitudes=df_corp.Latitude,
                                longitudes=df_corp.Longitude,
                                radius=2500)
```

```
Found 32171 venues in 425 neighborhoods.
```

Metro	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Atlanta	15431	15431	15431	15431	15431	15431	15431
Boston	2418	2418	2418	2418	2418	2418	2418
Chicago	7001	7001	7001	7001	7001	7001	7001
Denver	7321	7321	7321	7321	7321	7321	7321

d. Find the most common types of venue in each neighborhood. I decided to use the top 20 categories.

Metro	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Atlanta	Sherwood Forest	_Coffee Shop	_Southern / Soul Food Restaurar	_American Restaurant	_Pizza Place	_Park
Atlanta	West Highlands	_Park	_Convenience Store	_Construction & Landscaping	_Discount Store	_CafÃ©
Boston	Chinatown	_Coffee Shop	_Italian Restaurant	_French Restaurant	_Gym	_Sandwich Place
Chicago	AUSTIN	_Fast Food Restaurant	_Fried Chicken Joint	_Park	_Discount Store	_Donut Shop
Chicago	HEGEWISCH	_Bar	_Harbor / Marina	_Park	_Pizza Place	_Discount Store
Chicago	OAKLAND	_Beach	_Sandwich Place	_Fast Food Restaurant	_Art Gallery	_BBQ Joint
Denver	Barnum	_Mexican Restaurant	_Convenience Store	_Vietnamese Restaurant	_Dim Sum Restaurant	_Fast Food Restaurant
Denver	Globeville	_Brewery	_Coffee Shop	_Bar	_Diner	_Mexican Restaurant
Denver	Ruby Hill	_Mexican Restaurant	_Vietnamese Restaurant	_Convenience Store	_Coffee Shop	_Brewery

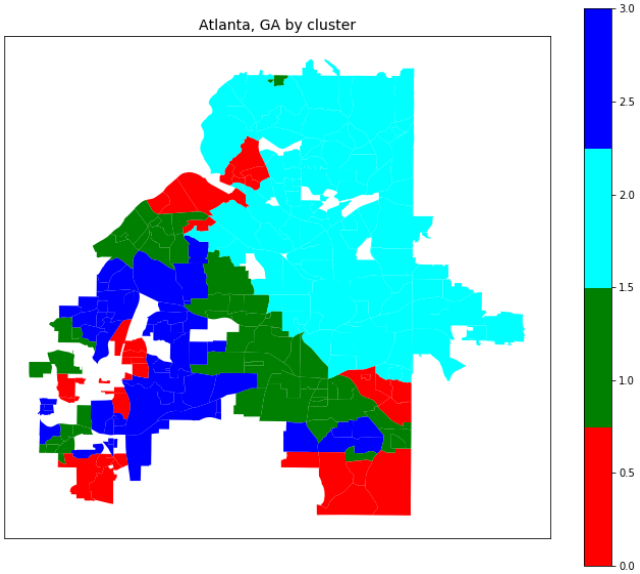
e. I used K-Means clustering to group neighborhoods with similar venues. I tested the number of clusters and the random seed to use, to see how they impact the division of neighborhoods into clusters. The best result seemed to come using 4 clusters and a random state of 4. That led to these splits:

Cluster#	Count	%
0	101	23.76%
1	55	12.94%
2	217	51.06%
3	52	12.24%

Note – K-Means clustering uses a 0-index labeling method. So the 1st cluster is cluster #0, and cluster #3 is the 4th cluster. When discussing the clusters, I will use this same labeling.

f. Running the K-Means clustering with these settings, I find the top overall venue types in each cluster. I scored each category on a 20 point scale (20 for most common category, 19 for 2nd most common,… 1 for 20th most common). The scores are aggregated as a weighted average (score divided by # of neighborhoods in the cluster). I used weighted average instead of straight average (mean) because a mean would give a higher score to a category that was the most common in one neighborhood but not in the top-20 of the others.

g. With each neighborhood assigned to a cluster, we can add that cluster information to the city's geometry and map it again.



4. Results

a. Atlanta was the only city that contained all 4 clusters. Conversely, cluster #3 was only found in Atlanta. I guess there are some things unique to that area.

b. **Cluster breakdown**:

i. Cluster #0: The most common venue type in cluster #1 is Fast Food Restaurants, followed by Sandwich Places. Fast Food Restaurants scored nearly 15 points in cluster 1, meaning they were consistently the 6th (or better) most common venue type. Sandwich places scored 13 points, putting them as 8th or better.

ii. Cluster #1: Parks were the most common in cluster #2. They scored almost 17 points. Also in the top-5 in cluster#2 were trails, scoring 9.5 points. Contrasting this outdoorsy feel though, the 2nd highest score in the cluster were gas stations, scoring 15.6.

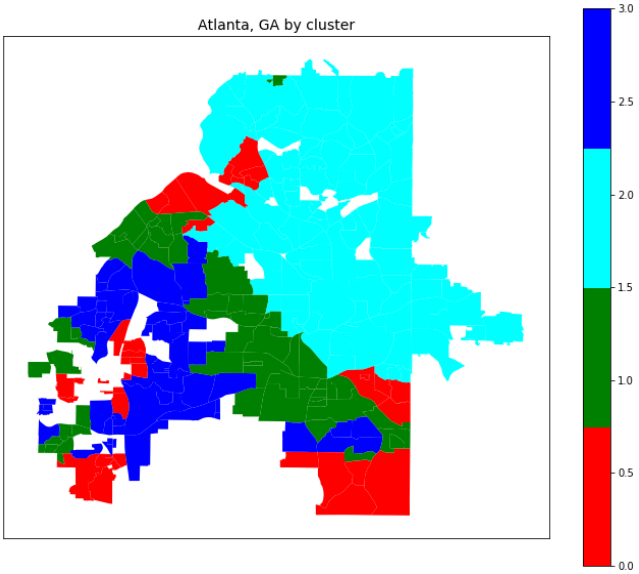
iii. Cluster #2: Cluster 2 is characterized by food and beverage. Coffee Shops had the highest score, but the top-10 is also dotted with restaurants (American, Italian, and Mexican) and Bars & Breweries.

iv. Cluster #3: Found only in Atlanta, cluster 3 is led by Discount Stores, Gas Stations, and Fast Food Restaurants. While these are found in the first 3 clusters, the difference in scoring for those categories versus the others is much higher in cluster 3.

c. **City breakdown**:

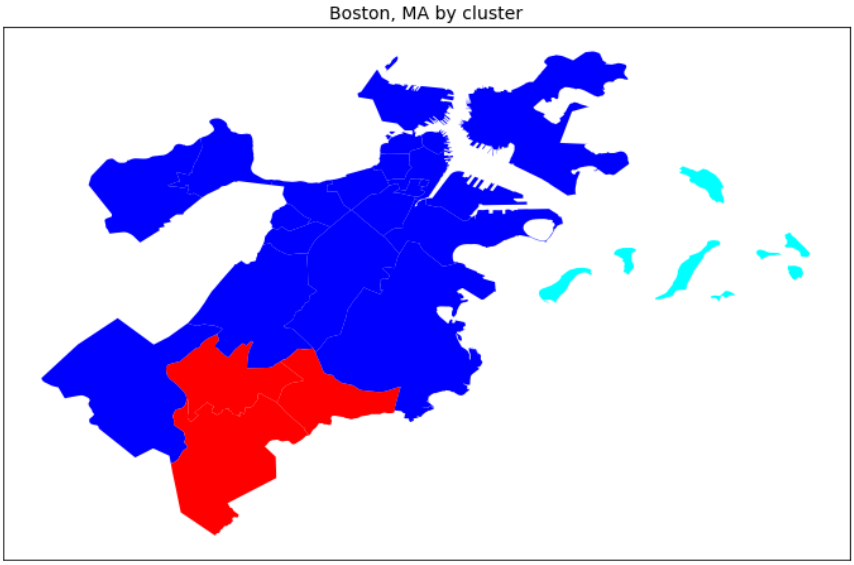
i. <u>Atlanta</u>:

1. Cluster 0 (Fast Food / Sandwiches) neighborhoods are located in the southeast corner, as well as a couple of pockets in the southwest and northwest.
2. Cluster 1 (Parks / Trails) cuts a path through central Atlanta.
3. Cluster 2 (Food & Beverage) is entirely located in the northern part of the city.
4. Cluster 3 (Discount Shopping) lies more in the central / southwest area.



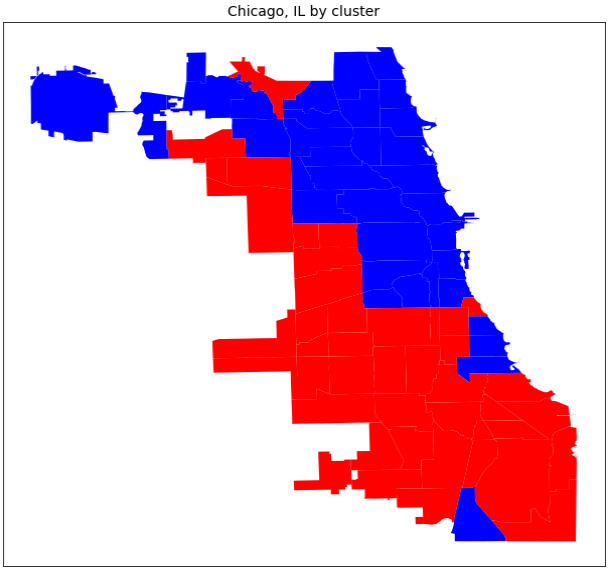
ii. <u>Boston</u>:

1. Cluster 0 (Fast Food / Sandwiches) lies to the southern part of Boston.
2. Cluster 1 (Parks / Trails) are found out on the Harbor Islands.
3. Cluster 2 (Food & Beverage) makes up the vast majority of Boston.
4. Cluster 3 doesn't occur in Boston.



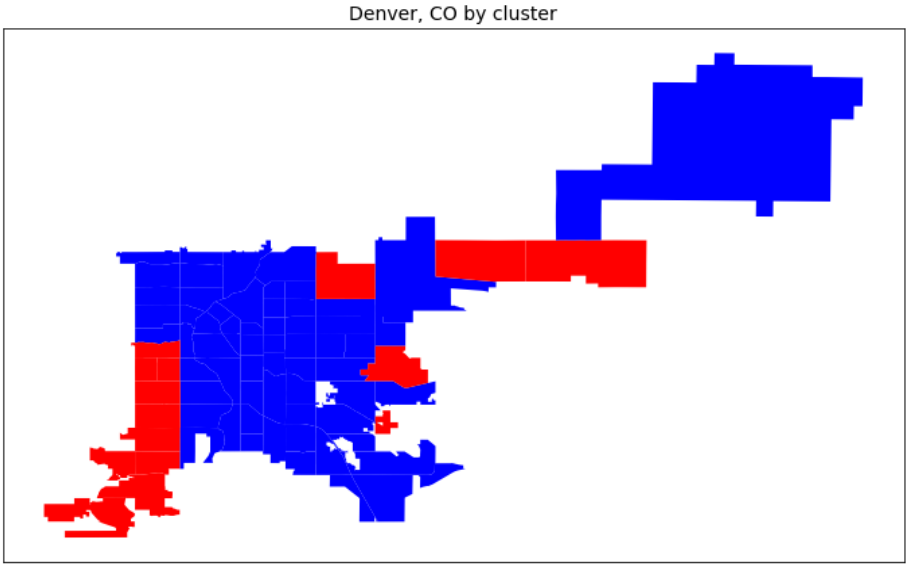
iii. Chicago:

1. Chicago is split between cluster 0 (Fast Food) to the west, and cluster 2 (Food & Beverage) to the east, along Lake Michigan.



iv. Denver:

1. Denver is made up mostly of cluster 2 (Food and Beverage). It covers almost the entire city.
2. Cluster 0 does cover a few areas in the southwest corner, and a couple of other areas in the north.



5. **Discussion**

A few observations:

a. Going in, the expectation was that the clusters would be more scattered across the cities, not all (or most) of them *geographically* clustered as well. This may be a result of the radius used to select venues from Foursquare. If the same venue is showing up in multiple neighborhoods' result set, it would influence which cluster the neighborhood goes into. On the other hand, 1.5 miles from the center of the neighborhood isn't that big of a range.

b. Cross indexing the cities and clusters, the groupings would appear to be
   i. Southeast Atlanta, South Boston, West Chicago, and Southwest Denver
   ii. Central Atlanta and Boston's Harbor Islands.
   iii. North Atlanta, most of Boston (outside of South Boston and the islands), eastern Chicago, and most of Denver.

c. Cluster 1 is only in central Atlanta and Boston's Harbor Islands. The most common venue in the cluster is Parks. The next 2 venue categories are Gas Stations and Discount Stores. Those are the top 2 categories in cluster 3, suggesting they can reasonably be blended together. Unfortunately that still includes only Atlanta and Boston. However that mix of venues, at least at the top of the cluster, aligns well with cluster 0. So those living in central and southwest Atlanta may want to look at South Boston, western Chicago, and southwest Denver.

6. Conclusion

a. It would be incredibly insulting to call these 4 cities interchangeable. Each has its own distinct characteristics, and each neighborhood within the city is unique. But everyone should be able to find a place where they feel comfortable and at home.

b. For all of the (even more) nitty gritty, including a list of every neighborhood in each cluster, view the entire notebook on GitHub.