Comparison Between Two Metropolitan Cities of US: Applied Data Science Capstone Project

Atlanta, GA vs Chicago, IL

Author: Pavithra Raman

Date: August 2020

1.Introduction

1.1. Background

Chicago is the most populous city in the U.S. state of Illinois, and the third-most-populous city in the United States. With an estimated population of 2,693,976 in 2019, it is also the most populous city in the Midwestern United States. Chicago is an international hub for finance, culture, commerce, industry, education, technology, telecommunications, and transportation. Atlanta is the capitol of the US state of Georgia and is the most populous city in the state with an estimated 498,044 residents. Atlanta is a culturally and economically diverse city with dominant economic sectors including aerospace, transportation, professional and business services, media and medical operations, and information technology.

1.2. Business Problem

The goal of the project is to explore the neighborhoods in both cities and group them by common nearby venues. This will assist anyone visiting or relocating between the cities to consider which areas are most like their current neighborhood and therefore might offer their preferred range of amenities. This information is very useful when moving to an unknown city and will help narrow down the list of areas to search for a new home, thus speeding up the relocation process and avoiding overly long and potentially pricey stays in hotels or other temporary living arrangements.

2. Data

The following data sources were used to complete this project.

Data required for Chicago, IL

1. Cross referenced Chicago Neighborhoods and Zip code postal data from US map guide :https://www.usmapguide.com/illinois/chicago-zip-code-map/
2. Chicago Zip codes & Neighborhood data from local real estate company : https://www.seechicagorealestate.com/chicago-zip-codes-by-neighborhood.php
3. Foursquare API

Data required for Atlanta, GA

1. Atlanta Zip codes & Neighborhood data from local real estate company : https://www.realsourcebrokers.com/atlanta-zip-code-map/
2. Cross referenced Atlanta Neighborhoods and Zip codes postal data from US map guide : https://www.usmapguide.com/georgia/atlanta-zip-code-map/
3. Foursquare API

2.1. Atlanta Neighborhood Data Cleaning and Sourcing

The data set (1) for Atlanta was the most complete and included zip code data, neighborhood names, and the corresponding latitude and longitude coordinates for all Atlanta zip codes, which covers the entire county. The data was in the form of a downloadable excel spreadsheet, which I then cleaned and formatted to include only Atlanta city zip codes, neighborhood names and map coordinates. Finally, I reduced the list of neighborhoods by removing duplicate values so that there would only be one occurrence of each neighborhood and corresponding data. It should be noted that this method randomly dropped duplicates so the remaining full zip codes corresponding to each neighborhood were one of many possible options. Different zip code choices would have had slightly differing latitude and longitude coordinates. This may have affected the resulting venue data sourced from Foursquare and skewed the results. I then uploaded this data set to my Jupyter notebook and used the function to transform it into a pandas data frame.

	Zipcode	Neighborhoods	Latitude	Longitude	City	State
0	30303	Downtown - Central Business District - Fairlee...	33.752856	-84.39013	Atlanta	GA
1	30305	Buckhead - Garden Hills - Haynes Manor - Peach...	33.830054	-84.38472	Atlanta	GA
2	30306	Virginia Highlands - Morningside/Lenox Park - ...	33.786755	-84.35149	Atlanta	GA
3	30307	Candler Park - Druid Hills - Edgewood - Emory ...	33.768205	-84.33786	Atlanta	GA
4	30308	Midtown - Old Fourth Ward	33.771755	-84.38065	Atlanta	GA

2.2 Chicago, IL Neighborhood and School Data Sourcing and Cleaning

The data sets (2)(3) used to source a list of Chicago neighborhoods and corresponding zip codes were simply lists from an Chicago real estate website and a US map guide website respectively. I manually copied and input this data into an excel spreadsheet and added any differences between the data (missing or additional neighborhoods or zip codes) to ensure a more complete breakdown. Unfortunately, data available from local city government sources was not in the required format so I could not use more authoritative sources. Therefore, the breakdown of neighborhoods to zip codes in this data set should be taken as advisory only and may differ between data sets. I found and downloaded a spreadsheet of all US zip codes and corresponding latitude and longitude coordinates (4) from the Open Soft Data website. I manually filtered this excel spreadsheet to list only Chicago zip codes and map coordinates. I uploaded both excel sheets to my Jupyter notebook using the function to transform them into Pandas data frames, dropping any unnecessary columns. Finally,

	COMMUNITY_AREA_NAME	Zipcode	Latitude	Longitude	City	State
0	Rogers Park	60626	42.009731	-87.66938	Chicago	IL
1	West Ridge	60645	42.008956	-87.69634	Chicago	IL
2	Uptown	60640	41.973181	-87.66650	Chicago	IL
3	Lincoln Square	60625	41.971614	-87.70256	Chicago	IL
4	North Center	60618	41.945681	-87.70480	Chicago	IL

The resulting dataset for Atlanta had 28 neighborhoods and the dataset for Chicago had 75 neighborhoods.

```
In [9]: atl_zip.shape
   Out[9]: (28, 6)

In [20]: chicago_zip.shape
   Out[20]: (75, 6)
```

2.3. Final List of Neighborhoods used in this Project

Atlanta Neighborhoods

Downtown - Central Business District - Fairlee Poplar
Buckhead - Garden Hills - Haynes Manor - Peachtree Battle - Peachtree Hills - Tuxedo Park
Virginia Highlands - Morningside/Lenox Park - Poncey-Highland - Druid Hills
Candler Park - Druid Hills - Edgewood - Emory - Inman Park - Lake Claire - Little Five Points
Midtown - Old Fourth Ward
Midtown - Ansley Park - Brookwood Hills - Loring Heights
Adair Park - Capitol View - Oakland City - West End
Cascade Real Estate
Downtown Atlanta - Grant Park
Downtown Atlanta - Castlebury Hill
Vines City - Mozely Park
Grant Park - Peoplestown - Lakewood
Cabbagetown - East Atlanta Village - Ormewood Park - South DeKalb
East Lake Real Estate - Kirkwood Real Estate - Edgewood Real Estate
Home Park - Northwest Atlanta - Collier Hills - Underwood Hills - Midtown West
Brookhaven - North Atlanta - Dunwoody
Morningside/Lenox Park - Piedmont Heights - Lenox -Lavista Park
Lenox
Buckhead - North Atlanta
Emory - Toco Hills - Briarcliff
College Park Real Estate
Vinings Real Estate
 Northlake - Tucker
Chamblee
Buckhead - North Buckhead - Chastain Park - North Atlanta
East Point
Briarcliff Woods - Oak Grove - Northlake
Hapeville

Chicago Neighborhoods

Rogers Park
West Ridge
Uptown
Lincoln Square
North Center
Lake View
Lincoln Park
Near North Side
Edison Park
Norwood Park
Jefferson Park
Forest Glen
North Park
Albany Park
Portage Park
Irving Park
Dunning
Belmont Cragin
Hermosa
Avondale
Logan Square
Humboldt park

West Town
Austin
West Garfield Park
East Garfield Park
Near West Side
North Lawndale
South Lawndale
Lower West Side
Loop
Near South Side
Armour Square
Douglas
Oakland
Fuller Park
Grand Boulevard
Kenwood
Washington Park
Hyde Park
Woodlawn
South Shore
Chatham
Avalon Park
South Chicago
Burnside
Calumet Heights
Roseland

Pullman
South Deering
East Side
West Pullman
Riverdale
Hegewisch
Garfield Ridge
Archer Heights
Brighton Park
McKinley Park
Bridgeport
New City
West Elsdon
Gage Park
Clearing
West Lawn
Chicago Lawn
West Englewood
Englewood
Greater Grand Crossing
Ashburn
Auburn Gresham
Beverly
Washington Height
Mount Greenwood
Morgan Park
O'Hare

3. Methodology

Before sourcing the venue data, I completed initial visual analysis of the neighborhood data for both cities to view the layout of the neighborhoods on a map. This was to ensure the coordinates were initially generally correct and to see the spread of the neighborhoods across each city, as they vary significantly in geographical size.

Using the Nominatum tool in Geopy, I calculated the latitude and longitude coordinates of both cities.

```
]: address = 'Atlanta,GA'

geolocator = Nominatim(user_agent="Atl_explorer")
location = geolocator.geocode(address)
latitude_ATL = location.latitude
longitude_ATL = location.longitude
print('The geograpical coordinates of Atlanta,GA are {}, {}.'.format(latitude_ATL, longitude_ATL))
```

```
The geograpical coordinates of Atlanta,GA are 33.7490987, -84.3901849.
```

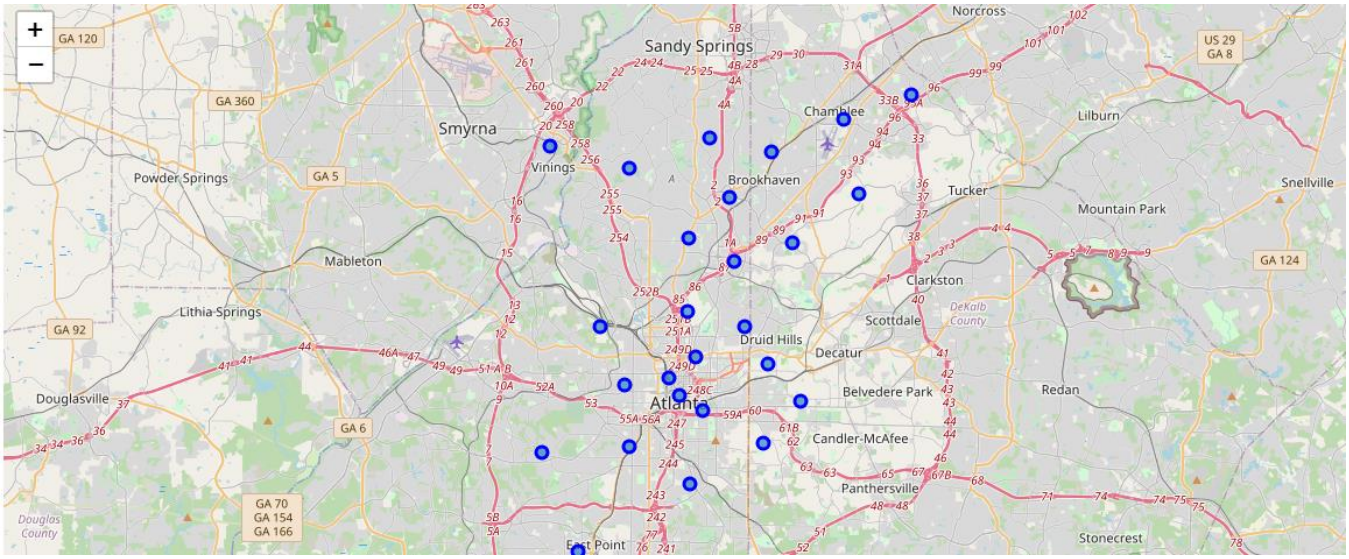
```
]: address = 'Chicago,IL'

geolocator = Nominatim(user_agent="Chi_explorer")
location = geolocator.geocode(address)
latitude_CHI = location.latitude
longitude_CHI = location.longitude
print('The geograpical coordinates of Chicago,IL are {}, {}.'.format(latitude_CHI, longitude_CHI))
```

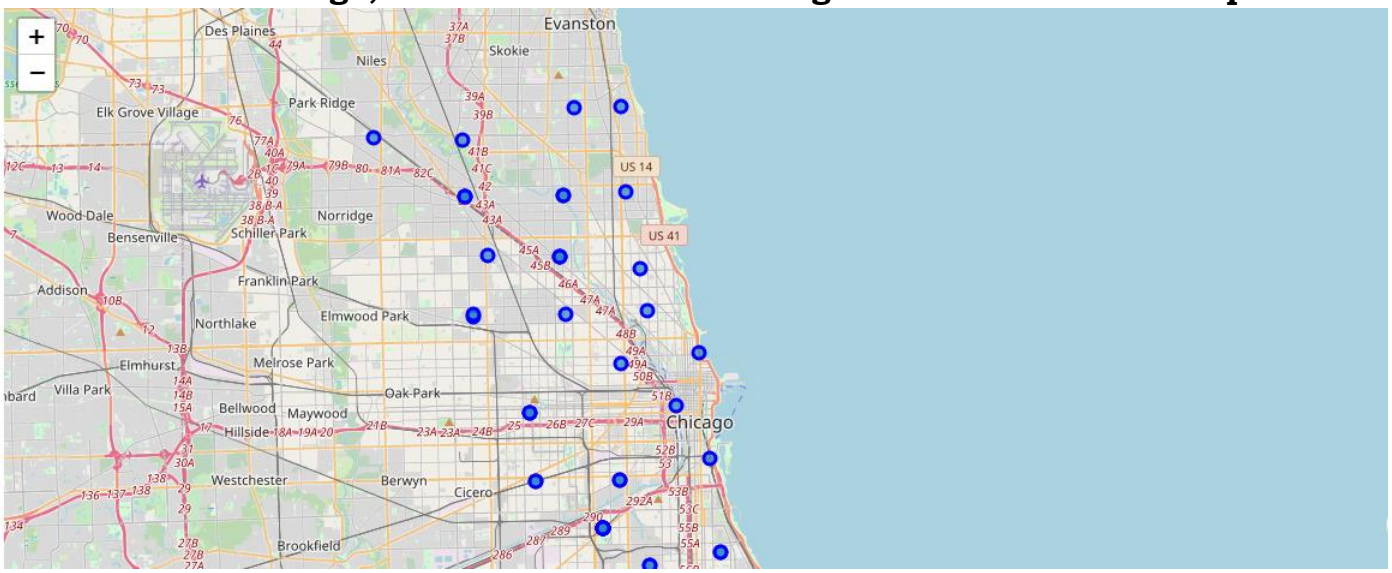
```
The geograpical coordinates of Chicago,IL are 41.8755616, -87.6244212.
```

I then used Folium to create maps of the two cities using the above generated coordinates. Finally, I was able to code markers onto each city map of the corresponding neighborhood coordinates using the data from the previously created data frames.

3.1 Atlanta, GA Neighborhoods Map



3.2. Chicago, IL Neighborhoods Map



3.3 Foursquare API: Venue Data

Using Foursquare, I was able to generate a list of venues by category in each neighborhood based on the corresponding map coordinates in the data sets for both cities. I set the radius to 500 and limited the venue results to 100 per neighborhood or set of coordinates. I then transformed this venue data into Pandas data frames (see below example of Atlanta neighborhood venue data generated using Foursquare API.) The process was repeated for Chicago neighborhoods.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Downtown - Central Business District - Fairlee...	33.752856	-84.39013	Ebrik Coffee Room	33.753897	-84.388782	Coffee Shop
1	Downtown - Central Business District - Fairlee...	33.752856	-84.39013	Walgreens	33.754345	-84.389484	Pharmacy
2	Downtown - Central Business District - Fairlee...	33.752856	-84.39013	The Masquerade	33.751720	-84.389739	Music Venue
3	Downtown - Central Business District - Fairlee...	33.752856	-84.39013	Dua Vietnamese Noodle Soup	33.755610	-84.389530	Vietnamese Restaurant
4	Downtown - Central Business District - Fairlee...	33.752856	-84.39013	Blossom Tree	33.755496	-84.389006	Korean Restaurant

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Rogers Park	42.009731	-87.66938	Morse Fresh Market	42.008087	-87.667041	Grocery Store
1	Rogers Park	42.009731	-87.66938	The Common Cup	42.007797	-87.667901	Coffee Shop
2	Rogers Park	42.009731	-87.66938	Glenwood Sunday Market	42.008525	-87.666251	Farmers Market
3	Rogers Park	42.009731	-87.66938	Smack Dab	42.009291	-87.666201	Bakery
4	Rogers Park	42.009731	-87.66938	Rogers Park Social	42.007360	-87.666265	Bar

3.4. Atlanta Top 10 Venues by Neighborhood

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Northlake - Tucker	Construction & Landscaping	Asian Restaurant	Sandwich Place	Gas Station	Zoo Exhibit	Food	Fruit & Vegetable Store	Frozen Yogurt Shop	Fountain	Food Truck
1	Adair Park - Capitol View - Oakland City - Wes...	Brewery	Pop-Up Shop	Trail	Gastropub	Boutique	Beer Store	Market	Paper / Office Supplies Store	Thrift / Vintage Store	Liquor Store
2	Briarcliff Woods - Oak Grove - Northlake	Lake	Food	Zoo Exhibit	Flower Shop	Furniture / Home Store	Fruit & Vegetable Store	Frozen Yogurt Shop	Fountain	Food Truck	Food Service
3	Brookhaven - North Atlanta - Dunwoody	American Restaurant	Ice Cream Shop	Sandwich Place	Mexican Restaurant	Burger Joint	Salon / Barbershop	Sporting Goods Shop	Southern / Soul Food Restaurant	Soup Place	Shopping Mall
4	Buckhead - Garden Hills - Haynes Manor - Peach...	Italian Restaurant	Chinese Restaurant	Salon / Barbershop	Sushi Restaurant	Pharmacy	Tanning Salon	Shipping Store	Farmers Market	Basketball Court	Cosmetics Shop

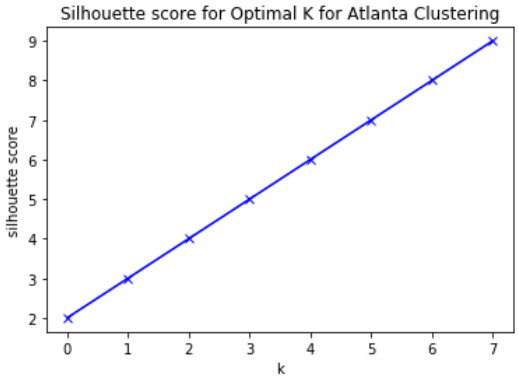
3.5. Chicago Top 10 Venues by Neighborhood

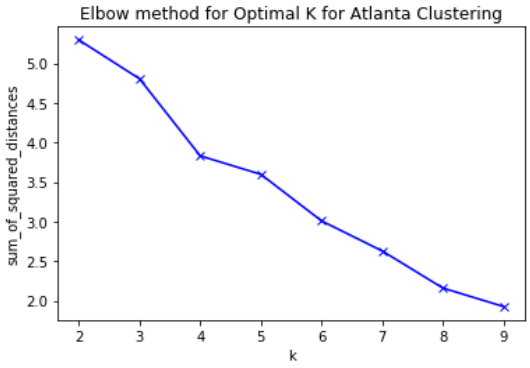
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Albany Park	Bar	Ice Cream Shop	Sushi Restaurant	Coffee Shop	Bakery	Donut Shop	Filipino Restaurant	Chinese Restaurant	Light Rail Station	Furniture / Home Store
1	Archer Heights	Diner	Video Store	Thrift / Vintage Store	Mexican Restaurant	Train Station	Supermarket	Fast Food Restaurant	Ice Cream Shop	Fried Chicken Joint	Bakery
2	Ashburn	Diner	Video Store	Thrift / Vintage Store	Mexican Restaurant	Train Station	Supermarket	Fast Food Restaurant	Ice Cream Shop	Fried Chicken Joint	Bakery
3	Auburn Gresham	Diner	Video Store	Thrift / Vintage Store	Mexican Restaurant	Train Station	Supermarket	Fast Food Restaurant	Ice Cream Shop	Fried Chicken Joint	Bakery
4	Austin	Diner	Video Store	Thrift / Vintage Store	Mexican Restaurant	Train Station	Supermarket	Fast Food Restaurant	Ice Cream Shop	Fried Chicken Joint	Bakery

3.6 Finding the best K for K-Means Clustering

K-Means is one of the most common methods of unsupervised machine learning for clustering. Using one hot encoding and mean frequency on the new data frames, I was able to then apply algorithms from the SciKit Learn library to calculate the best K value for K-means clustering of the neighborhoods in each city. I initially used the Silhouette method, but the results were inconclusive. I therefore tried the Elbow method (sum of squared distances) and achieved slightly better results in both cases. I used Matplotlib to plot the results.

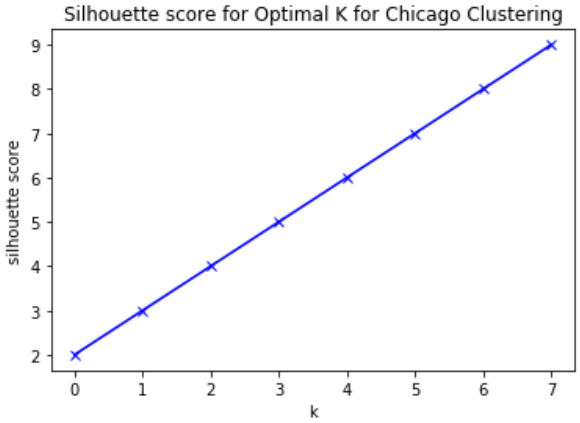
3.7. Finding K for Atlanta Clustering

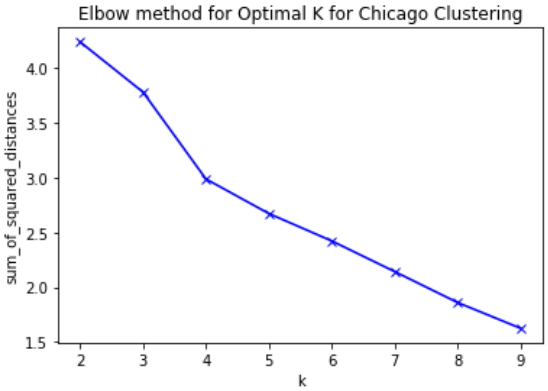




I determined the best K could be 4 for the Atlanta venue data. However, I felt that was a bit low for clustering 28 neighborhoods and wanted there to be at least as many clusters in Atlanta as in Chicago. I implemented clustering using 5 and 6 and ultimately choose 6 as a good option for K in this case.

3.8. Finding K for Chicago Clustering





I determined the best K would be either 5 or 6 for the Chicago venue data. However, after implementing both, it was clear the neighborhood clustering stopped at 5.

3.9. K-Means Clustering Neighborhood

Using the K-means algorithm, I clustered the neighborhoods in both cities and merged this data with the Top 10 Venue data frames. I also cleaned the data to ensure the clusters were integers and not floats, as otherwise they would not show up properly on the maps using Folium.

Chicago Clustered Neighborhoods Pandas Data Frame

	COMMUNITY_AREA_NAME	Zipcode	Latitude	Longitude	City	State	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Rogers Park	60626	42.009731	-87.66938	Chicago	IL	0	Mexican Restaurant	Pizza Place	American Restaurant	Bakery	Donut Shop
1	West Ridge	60645	42.008956	-87.69634	Chicago	IL	0	Bar	American Restaurant	Animal Shelter	Park	Dessert Shop
2	Uptown	60640	41.973181	-87.66650	Chicago	IL	0	Gym / Fitness Center	Middle Eastern Restaurant	Adult Boutique	Bike Rental / Bike Share	Bookstore
3	Lincoln Square	60625	41.971614	-87.70256	Chicago	IL	0	Mexican Restaurant	Bus Station	Park	Bank	Gym
4	North Center	60618	41.945681	-87.70480	Chicago	IL	0	Bus Station	Currency Exchange	Bank	Chinese Restaurant	Park

Atlanta Clustered Neighborhoods Data Frame

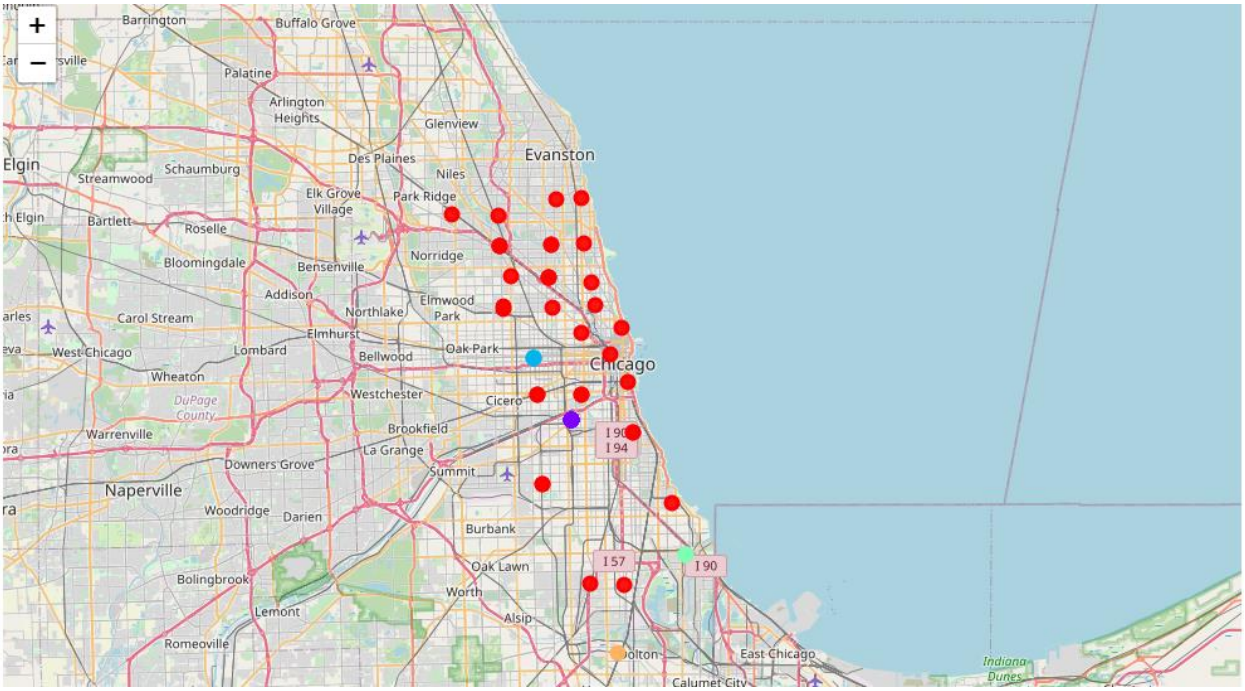
	Zipcode	Neighborhoods	Latitude	Longitude	City	State	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
0	30303	Downtown - Central Business District - Fairlee...	33.752856	-84.39013	Atlanta	GA	0	Sandwich Place	Theater	Mexican Restaurant	Deli / Bodega	Caribbean Restaurant	Coffee Shop
1	30305	Buckhead - Garden Hills - Haynes Manor - Peach...	33.830054	-84.38472	Atlanta	GA	0	Italian Restaurant	Chinese Restaurant	Salon / Barbershop	Sushi Restaurant	Pharmacy	Tanning Salon
2	30306	Virginia Highlands - Morningside/Lenox Park - ...	33.786755	-84.35149	Atlanta	GA	0	Boutique	Massage Studio	Yoga Studio	Café	Pet Store	Burger Joint
3	30307	Candler Park - Druid Hills - Edgewood - Emory ...	33.768205	-84.33786	Atlanta	GA	0	Playground	Athletics & Sports	Church	Golf Course	Tennis Court	Basketball Court
4	30308	Midtown - Old Fourth Ward	33.771755	-84.38065	Atlanta	GA	0	Southern / Soul Food Restaurant	Hotel	Lounge	Pizza Place	Park	Cuban Restaurant

4.Results and Discussion

4a. Mapping the Neighborhoods by Clusters

Using Folium once again and the new data frame including the top 10 venues in each neighborhood and the Cluster labels, I mapped out the neighborhoods in both cities. The neighborhoods are color coded by cluster to show the cluster grouping visually.

Map of Chicago Neighborhoods (Color Coded by Cluster)



4b. Labelling and Initial Analysis by Cluster:

Chicago Cluster 4 (Orange): Food, Home service store. This is the least cluster with one neighborhood.

Cluster 3 (Light Green): Bar, Gym, Cafes and Parks. This cluster has 3 neighborhoods.

Cluster 2 (Light Blue): Small Shops, Food and park.This cluster also has 3 neighborhoods.

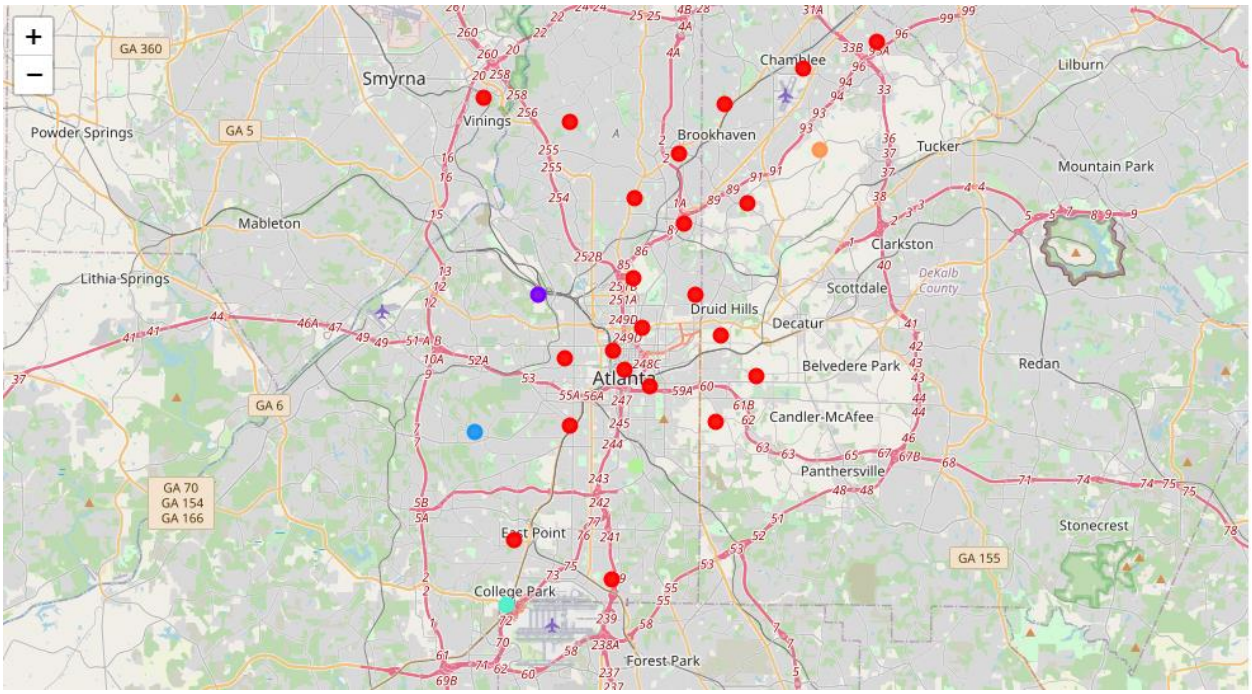
Cluster 1 (Purple): Restaurant, small stores and Train station. This Cluster is the second largest with 31 neighborhoods.

Cluster 0 (Red): Pubs, Shopping Mall, Restaurants, coffee shop Museums and Bars. This cluster is the largest by a significant margin and includes 64 of the 78 Chicago Neighborhoods. This could be due to several factors including the range of venue types returned by Foursquare. As mentioned in the data section of this report, the venue list generated relies on the latitude and longitude coordinates provided for each neighborhood. If these coordinates are not the optimal choice, then the venue data may be inaccurate, and this could have skewed the cluster results.

It may be that these areas did not have enough venues to properly cluster them or there were very distinctive venues. However, looking at the top three venues

listed for clusters 4, 3 and 2, this does not seem likely. It is also possible they are heavily residential or zoned for business.

Map of Atlanta Neighborhoods (Color Coded by Cluster)



4c. Labelling and Initial Analysis by Cluster:

Atlanta Cluster 0 (Red): Restaurants, Businesses, Tourist Attractions, Hotels, Breweries, Music Venues, Bars This is by far the largest cluster of neighborhoods and we can see that neighborhoods across all areas of Atlanta have been included in this group. 21 of the 28 neighborhoods in Atlanta have been assigned to this cluster. As with cluster 4 from the Chicago data, it may be that the neighborhoods in this cluster have too wide a range of venue results to be very useful as a measure of similarity. Clustering based on other data or a subsection of the venue data could be required to better categorize these neighborhoods and break them down into smaller and more distinct clusters. It may also be that the radius needs to be changed when generating the venue lists from Foursquare.

Cluster 1 (Purple): Event Venues, Zoo Exhibits, and Fish Market

Cluster 2 (Light Blue): Gyms, Fast Food and Sports Stadiums

Cluster 3 (Teal): Nature/Parks, Zoo and Fast Food

Cluster 4 (Lime green): Residential Apartments, Gay Bars, and Smoke shops

Cluster 5 (Orange): Discount shops, Playgrounds and Southern/Soul Food Restaurants

The remaining clusters have only one neighborhood each. Again, this may be due to inaccurate or incomplete venue data or it may be the result of better clustering than the above Cluster 0.

4d. Comparing Neighborhood Clusters Between Cities

For both cities we see a similar results pattern in the clustering of neighborhoods. Both have returned one cluster comprising the majority of the neighborhoods, with the remaining clusters generally having one neighborhood each. The most similar clusters between the two cities are these large clusters, Cluster 0 in Atlanta and Cluster 0 in Chicago. However, it is more clear clustering analysis on the basis of other data beyond nearby venues will be required to more accurately group similar neighborhoods in each city. Even if this is accomplished, the results may still show that there are many neighborhood clusters that do not have direct comparison between these two cities. This could be due to a number of factors, such as the geographical size and layout of the neighborhoods and differences in culture and lifestyle between the Atlanta and Chicago. Further analysis and investigation is required.

It may also be necessary to better clean the venue data returned by Foursquare API. As we can see below, some of the top venues listed and used in the clustering analysis include uninformative categories such as 'Bus Stop' or 'Miscellaneous Shop' or 'Discount Store'. This may or may not be a significant venue and could be excluded for more statistically significant venues. This is something to consider if this project were to be replicated.

Top Five Venues in Each Cluster: Chicago vs Atlanta

Cluster Labels	CHI Cluster 0	CHI Cluster 1	CHI Cluster 2	CHI Cluster 3	CHI Cluster 4	ATL Cluster 0	ATL Cluster 1	ATL Cluster 2	ATL Cluster 3	ATL Cluster 4	ATL Cluster 5
1st Most Common Venue	Mexican Restaurant	Diner	Shoe Store	Wine Bar	Pizza Place	Sandwich Place	Event Service	Business Service	Hotel	Discount Store	Lake
2nd Most Common Venue	Pizza Place	Video Store	Fast Food Restaurant	Mexican Restaurant	Home Service	Theater	Zoo Exhibit	Gastropub	Sports Bar	Southern / Soul Food Restaurant	Food
3rd Most Common Venue	American Restaurant	Thrift / Vintage Store	Sandwich Place	Park	Lounge	Mexican Restaurant	Food	Gas Station	Tram Station	Gas Station	Zoo Exhibit
4th Most Common Venue	Bakery	Mexican Restaurant	Cosmetics Shop	Yoga Studio	Dry Cleaner	Deli / Bodega	Furniture / Home Store	Furniture / Home Store	Parking	Fruit & Vegetable Store	Flower Shop
5th Most Common Venue	Donut Shop	Train Station	Park	Eastern European Restaurant	Flower Shop	Caribbean Restaurant	Fruit & Vegetable Store	Fruit & Vegetable Store	Rental Car Location	Frozen Yogurt Shop	Furniture / Home Store

4.Conclusions

This project has given us some insight into the amenities in the selected neighborhoods in both Chicago and Atlanta, which partially fulfills the intended purpose of the exercise. The information garnered provides a useful, albeit cursory and broad, snapshot of each neighborhood. However, based on the results it is clear we need more holistic data to improve the accuracy and usefulness of our neighborhood clustering. If I were to redo this project, I would consider including data on population, cost of living, demographics, schools and transportation. I would also better clean the venue data and ensure that the best map coordinates were being used to represent each neighborhood in order to improve the accuracy of venue results. Finally, I would consider whether factors such as culture or geographical size and spread are impacting the results and how these could be minimized to better standardize the data and subsequent results to ensure more accurate comparison.

Thank you for reading! This project was created for my Coursera capstone course to complete my IBM