CHICAGO AND DETROIT NEIGHBORHOOD SIMILARITY STUDY

Jimmie Tolliver

June 12, 2019

1. INTRODUCTION

Two of the largest areas in the Great Lakes region of the United States are Detroit and Chicago. Many of the area's larger businesses have ties to both cities and large portions of the population in Chicago originate from Michigan, specifically the metro Detroit region.

Having lived in Chicago for over 10 years, I'm relocating to the Detroit area to be closer to family and friends. I'd like to research neighborhoods in the Detroit area that are comparable to those of Chicago to give insights as to where I'd like to relocate.

This information may also be used to help tourism in the Detroit area or to attract younger talent and recent college graduates to the area. Additionally, it may be useful for anyone else potentially interested in relocating from Chicago to Detroit or vice versa.

2. DATA ACQUISITION AND CLEANING

Since Chicago is substantially larger by population than Detroit and most of the areas of economic growth in recent years have been in Oakland County, MI, this analysis will include neighborhoods from both Wayne (including Detroit) and Oakland counties. Additionally, since many areas in Wayne and Oakland counties are suburban, Cook County, IL has also been included for Chicago. Furthermore, Ann Arbor is relatively close to Detroit and is worth including due to its proximity and economic diversity.

2.1 Neighborhoods

For this study, we needed to find neighborhoods in Ann Arbor, Chicago, and Detroit. Both the cities of Chicago and Detroit maintain this data as GeoJSON files at these respective links (Chicago[1], Detroit[2]) . However, Ann Arbor and Cook, Oakland, and Wayne Counties did not, so they were web scraped from these respective sources (Ann Arbor[3], Cook[4], Oakland[5], Wayne[6]). Once the neighborhood data was collected, a data frame that includes the neighborhoods and geographical coordinates was created.

2.2 Geographical Coordinates

Most of the geographical coordinates in this study were obtained via a geolocator query using the geopy library in Python. However, the geolocator did not return many results for the Detroit neighborhoods, so coordinates from the aforementioned GeoJSON file were used for the Detroit neighborhoods.

---

1 https://data.cityofchicago.org/api/geospatial/cauq-8yn6?method=export&format=GeoJSON
2 https://opendata.arcgis.com/datasets/a25b7114d233496eaece59a23e31f4b2_0.geojson
3 https://annarborobserver.com/cg/t1300.html
4 https://geographic.org/streetview/usa/il/cook/index.html
5 https://geographic.org/streetview/usa/mi/oakland/index.html
6 https://geographic.org/streetview/usa/mi/wayne/index.html

2.3 Foursquare Data

Foursquare[7] data was used to obtain nearby venue information for each neighborhood in the data frame mentioned above. This data was then clustered using Kmeans clustering to group neighborhoods together based on similar nearby venues.

2.4 US Census Demographic Data

Looking at a particular Chicago neighborhood, specifically Lakeview, it was determined that more data was required since the venue based cluster was very large returning 85 neighborhoods in the Detroit area alone. As such, demographic data was determined to be a good way to further cluster the initial venue based cluster. US Census data was used and downloaded from Kaggle[8].

2.5 Walkscore Data

The "Lakeview" cluster was reduced to 31 neighborhoods using the demographic data, which was an improvement, but it was still a rather large cluster of neighborhoods. A walkscore value as provided by the walkscore.com[9] API was then used to further cluster the data. Using the walkscore to cluster the data proved to be quite successful on the already twice clustered data returning only four neighborhoods in the Detroit area that are similar to Chicago's Lakeview neighborhood. Additionally, the data was tested multiple times using different K values and random seeds which regularly returned the same four neighborhoods.

2.6 Cleaning Data

2.6.1 Primary Data Frame

As noted above, the neighborhood data was loaded into a data frame that included geographical coordinates. As such, the geographical coordinates were used to tie everything together since it was common data throughout the sources used in this study. The only data set that didn't include the geographical coordinates was the demographic data from Kaggle. However, Census.gov has an API[10] that returns a "Census Tract" number based on geographic coordinates which in turn can be used with the Kaggle data to get demographic data such as age, income, gender, etc. The primary data frame for this study includes the neighborhood and geographic coordinates as shown in Table 1.

	Neighborhood	Latitude	Longitude
0	Airport, Detroit, MI	42.388475	-83.025065
1	Bagley, Detroit, MI	42.422256	-83.171482
2	Boynton, Detroit, MI	42.264908	-83.164444
3	Brightmoor, Detroit, MI	42.384513	-83.248953
4	Brooks, Detroit, MI	42.344826	-83.204472

Table 1. First five rows of primary data frame.

7 https://foursquare.com/
8 https://www.kaggle.com/muonneutrino/us-census-demographic-data
9 https://www.walkscore.com/
10 https://geocoding.geo.census.gov/geocoder/

2.6.2 Foursquare Data Frame

Two data frame's for the Foursquare data were created using the primary data frame of Table 1. Both data frames were created by cleaning the JSON results from the Foursquare API query. More specifically, venues within 500 meters of the geographical coordinates were categorized and then totaled up for each neighborhood. Next they were one hot encoded and a mean was calculated for each venue in each neighborhood as shown in Table 2. This data was then used for Kmeans clustering to cluster neighborhoods by nearby venues.

	Neighborhood	ATM	Accessories Store	Adult Boutique	African Restaurant	Airport	Airport Lounge	Airport Service	American Restaurant	Antique Shop	...	Video Store	Vietnamese Restaurant	Vineyard	Whisky Bar	Wine Bar
0	ALBANY PARK, Chicago, IL	0.0	0.0625	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.00	0.0	0.0	0.0	0.0
1	ARCHER HEIGHTS, Chicago, IL	0.0	0.0000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.04	0.0	0.0	0.0	0.0
2	ARMOUR SQUARE, Chicago, IL	0.0	0.0000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.00	0.0	0.0	0.0	0.0
3	ASHBURN, Chicago, IL	0.0	0.0000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.00	0.0	0.0	0.0	0.0
4	AUBURN GRESHAM, Chicago, IL	0.0	0.0000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.00	0.0	0.0	0.0	0.0

5 rows × 333 columns

Table 2. First five rows of data frame showing mean of each venue.

The second table was made in a more human readable format to show the top 10 venues within 500 meters of each neighborhood as shown in Table 3.

	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, Chicago, IL	Sandwich Place	Grocery Store	Pizza Place	Bakery	Gas Station	Cocktail Bar	Café	Korean Restaurant	Karaoke Bar	Fried Chicken Joint
1	ARCHER HEIGHTS, Chicago, IL	Mobile Phone Shop	Mexican Restaurant	Gas Station	Grocery Store	Park	Sandwich Place	Rental Service	Bank	Chinese Restaurant	Optical Shop
2	ARMOUR SQUARE, Chicago, IL	Chinese Restaurant	Cosmetics Shop	Asian Restaurant	Sports Bar	Hot Dog Joint	Breakfast Spot	Italian Restaurant	Sandwich Place	Gas Station	Fabric Shop
3	ASHBURN, Chicago, IL	Electronics Store	Cosmetics Shop	Light Rail Station	Construction & Landscaping	Italian Restaurant	Financial or Legal Service	Fabric Shop	Factory	Falafel Restaurant	Farm
4	AUBURN GRESHAM, Chicago, IL	Fast Food Restaurant	Caribbean Restaurant	Greek Restaurant	Cosmetics Shop	Discount Store	Lounge	Pharmacy	Eye Doctor	Fabric Shop	Factory

Table 3. First five rows of data frame showing top 10 venues by neighborhood.

2.6.3 Demographics Data Frame

As noted above, two data frames were created for the demographic information. The first was to correlate geographic coordinates with the "Census Tract" number as shown in Table 4. And the second with the demographic information corresponding with the "Census Tract" number as shown in Table 5. The demographic data included various demographic data categories, but it was decided to use only those deemed to give a neighborhood its general "vibe" such as age, gender, race, and income.

	Neighborhood	Latitude	Longitude	Census Tract
0	Airport, Detroit, MI	42.388475	-83.025065	26163511000
1	Bagley, Detroit, MI	42.422256	-83.171482	26163539400
2	Boynton, Detroit, MI	42.264908	-83.164444	26163524800
3	Brightmoor, Detroit, MI	42.384513	-83.248953	26163543900
4	Brooks, Detroit, MI	42.344826	-83.204472	26163545500

Table 4. First five rows of data frame with census tract identifier.

	Neighborhood	Latitude	Longitude	Census Tract	Voting Age	Men	Women	Hispanic	White	Black	Native	Asian	Pacific	Income	Poverty
0	Airport, Detroit, MI	42.388475	-83.025065	26163511000	0.719390	0.509721	0.490279	0.020	0.018	0.916	0.0	0.029	0.0	17930.0	0.536
1	Bagley, Detroit, MI	42.422256	-83.171482	26163539400	0.769583	0.467026	0.532974	0.009	0.020	0.953	0.0	0.000	0.0	32314.0	0.289
2	Boynton, Detroit, MI	42.264908	-83.164444	26163524800	0.679008	0.393110	0.606890	0.060	0.033	0.896	0.0	0.000	0.0	23430.0	0.513
3	Brightmoor, Detroit, MI	42.384513	-83.248953	26163543900	0.731308	0.434579	0.565421	0.009	0.148	0.819	0.0	0.000	0.0	20500.0	0.527
4	Brooks, Detroit, MI	42.344826	-83.204472	26163545500	0.712465	0.444558	0.555442	0.008	0.177	0.705	0.0	0.000	0.0	20648.0	0.502

Table 5. First five rows of demographic data frame.

2.6.4 Walkscore

The walkscore data was acquired from the walkscore API and loaded into a data frame including neighborhood name, geographic coordinates, and walkscore as shown in Table 6.

	Neighborhood	Latitude	Longitude	Walkscore
0	Grosse Ile, MI	42.138175	-83.154123	22
1	Northville, MI	42.431081	-83.483226	77
2	Plymouth, MI	42.371200	-83.467502	91
3	Berkley, MI	42.503091	-83.183539	64
4	Birmingham, MI	42.546701	-83.211319	94

Table 6. First five rows of walkscore data frame.

3. METHODOLOGY

Although this study compares and clusters all the neighborhoods in Ann Arbor, Chicago, Detroit, and Cook, Oakland and Wayne Counties, it was ultimately decided to look for a neighborhood in the Detroit area that was similar to Chicago's Lakeview neighborhood.

This study makes use of clustering neighborhoods by the aforementioned data sets using a machine learning technique known as Kmeans clustering. Additionally, an iterative approach was used to cluster and then re-cluster the data based on new data sets. More specifically, the data was clustered by a first data set (i.e., venues), those results where then clustered by a second data set (i.e., demographics), and then those results were clustered one more time by a third data set (i.e., walkscore).

After clustering neighborhoods by venue only, which was anticipated to substantially narrow the corresponding neighborhoods, there were still numerous matches (184) with the Lakeview neighborhood as shown below in Figure 1.

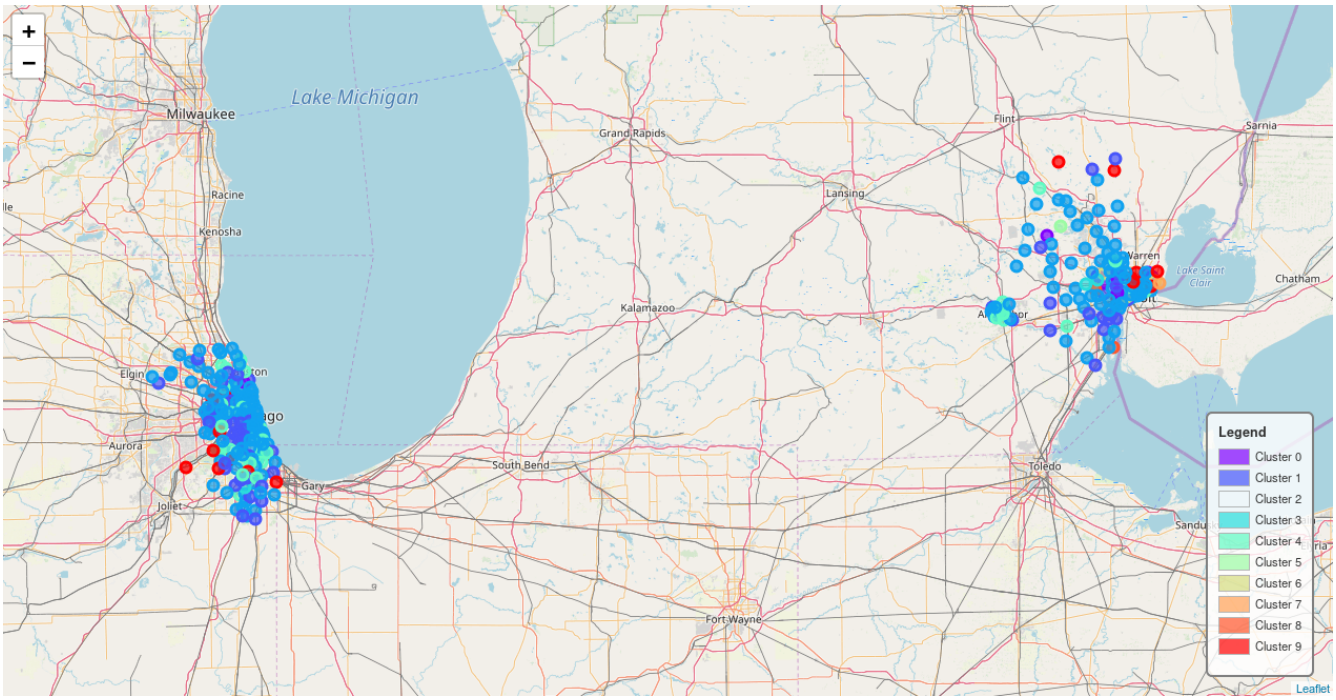

Figure 1. Neighborhood clusters based on venues within 500 meters.

With these results, it was determined that adding the demographic data from Table 5 would help narrow down the neighborhoods. Furthermore, since we're most interested in Detroit area neighborhoods that are similar to the Lakeview neighborhood in this study, the data frame was limited to Lakeview and the Detroit area neighborhoods from the initial cluster (i.e., by venue). Clustering this data set with the demographic data proved quite useful and narrowed the neighborhoods down to 31 as shown in Figure 2. Still more neighborhoods than desired for purposes of this study but if the walkscore factor (see below) is less import to the stakeholder, these would be viable alternatives to Lakeview in the Detroit area.

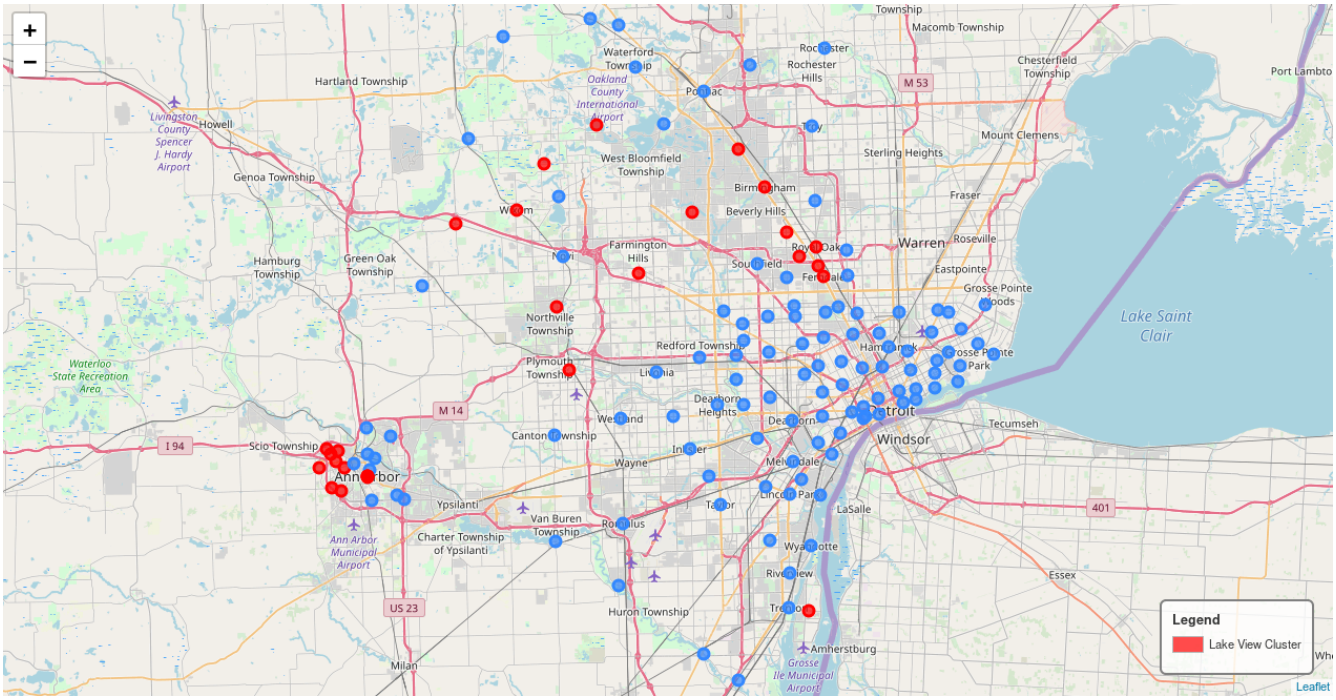

Figure 2. Neighborhoods of original Lakeview clustered based on demographics.

Because we're interested in finding neighborhoods in the Detroit area that are most similar to Lakeview, it was determined that walkscore would be a good measure since many people in Lakeview frequent nearby venues by foot. When factoring in the walkscore with the neighborhoods from the second data set (i.e., demographic cluster), the cluster dropped a quite a bit resulting in four Detroit area neighborhoods that are similar to Lakeview as shown in Figure 3. Based on personal knowledge of the areas and residents of those neighborhoods, this is a viable grouping. As noted above, the clustering algorithm was tested multiple times with different K values and random seeds and produced substantially the same results.

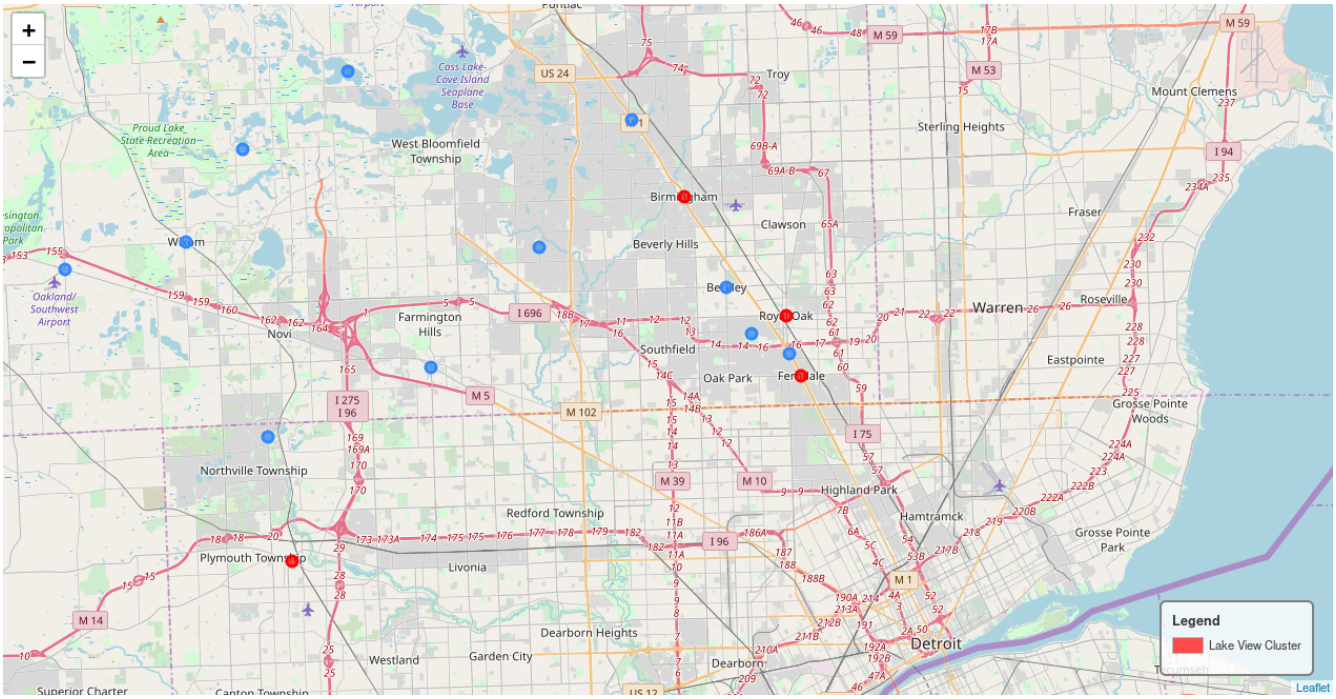

Figure 3. Neighborhoods clustered with Lakeview demographic data clustered again based on walkscore.

4. RESULTS

The results of clustering neighborhoods using three iterations based on venue, demographics, and walkscore, respectively, turned out better than expected. The final cluster comprised four neighborhoods in the Detroit area that are solid candidates for an equivalent of Chicago's Lakeview neighborhood. With this information a stakeholder can be more informed picking a new place to reside when relocating from Lakeview. Table 7 shows the resulting neighborhoods and top 10 venues in each neighborhood to further help a stakeholder make a decision.

	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
73	Plymouth, MI	Italian Restaurant	Coffee Shop	Bakery	Bar	Sandwich Place	Thai Restaurant	Bank	Grocery Store	Greek Restaurant	Farmers Market
87	Birmingham, MI	Spa	Coffee Shop	American Restaurant	Steakhouse	New American Restaurant	Boutique	Middle Eastern Restaurant	Yoga Studio	Italian Restaurant	Bakery
94	Ferndale, MI	Cocktail Bar	Bar	Gym	Gift Shop	Sandwich Place	Sushi Restaurant	Cosmetics Shop	Thai Restaurant	Massage Studio	Food Truck
114	Royal Oak, MI	Brewery	Coffee Shop	Vegetarian / Vegan Restaurant	Sushi Restaurant	Yoga Studio	Italian Restaurant	Lounge	American Restaurant	Café	Seafood Restaurant
188	LAKE VIEW, Chicago, IL	Bar	General Entertainment	Sports Bar	Sandwich Place	Mexican Restaurant	Baseball Stadium	Pizza Place	BBQ Joint	Outdoor Sculpture	Dive Bar

Table 7. Top 10 venues for each neighborhood.

5. DISCUSSION

This study originally called for using solely Foursquare venue and walkscore data. However, the results didn't seem appropriate so other data was sought to try to match the general "vibe' of the neighborhoods. Demographic data was added and seems to have been exactly what was needed.

Other data that could have been used includes population density, public transit data, housing units, housing types, crime statistics, and/or other suitable data.

In addition, if some metrics are more desirable than others to a stakeholder they could be weighted to have a greater or reduced impact on the clustering. For example, if the stakeholder desired to be close to water places, venues such as a Harbor, Lakefront, and/or other data could be boosted by a multiplier to give them a greater weight in the clustering. Likewise, if the stakeholder was less interested in certain aspects of the neighborhood, those aspects could be reduced by a multiplier to give them a reduced weight in the clustering. Other metrics are contemplated.

6. CONCLUSION

This study gathered venue, demographic, and walkscore data and iteratively clustered them to reduce and refine the cluster with each iteration. The resulting four Detroit area neighborhoods are good candidates as being similar to that of the Lakeview neighborhood in Chicago. As such, the model seems to be successful for purposes of this study. For further information please refer to the Jupyter Notebook[11] for this study.

11 https://github.com/JimmieTolliver/Coursera_Capstone/blob/master/AA_Chi_Det_foursquare_walkscore_demographics.ipynb