Comparing Neighborhoods of Houston and Chicago

By: Nick Adelberger
March 3, 2020

1. Introduction

1.1 Background

Chicago, on Lake Michigan in Illinois, is the 3rd largest city in the United States and is home to more than 2.7 million people. Known as "The Windy City", Chicago is an international hub for finance, culture, commerce, industry, education, technology, telecommunications, and transportation (reference: https://en.wikipedia.org/wiki/Chicago). Houston, popularly known as "The Bayou City", is set to become the 4th largest city in the United States by the second half of 2020. Houston's economy has a broad industrial base in energy, manufacturing, aeronautics, and transportation. It has become a global city, with strengths in culture, medicine and research (reference: https://en.wikipedia.org/wiki/Houston). As a current resident of Chicago, I have been offered a job in Houston as a Data Anaylist. Before accepting the position, I would like to discover more insight into the city of Houston to find any similarities or dissimilarities to Chicago.

1.2 Problem

Since there are a lot of neighborhoods within both cities, I will first try to detect the **distinct venues** that make up the **North Side of Chicago** since I currently live within that area. I will then break down the City of Houston into similar areas to compare and contrast so that I can **find the area of Houston that is similar to where I live in Chicago.**

I will use my data science powers to generate a few most promising neighborhoods based on the above criteria. Advantages of each area will then be clearly expressed so that best possible final location can be chosen by anyone who is considering moving to Houston.

2. Data Acquisition and Cleaning

2.1 Data Sources

Based on definition of our problem, factors that will influence our decision are:

- Neighborhoods and Areas within Houston
- Neighborhoods in Chicago that are located on the North Side
- The top 10 venues that make up the neighborhood areas for comparison

Following data sources will be needed to extract/generate the required information:

- List of neighborhoods in both Chicago & Houston using **Google Search**
    - Chicago Neighborhoods - https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Chicago
    - Houston Neighborhoods - https://www.houston.org/living-in-houston/neighborhoods-communities
- Coordinates (latitude & longitude) of each neighborhood using **Python Geopy Geocorders**
- Types of venues located within each neighborhood using **Foursquare API**

2.2 Data Cleaning

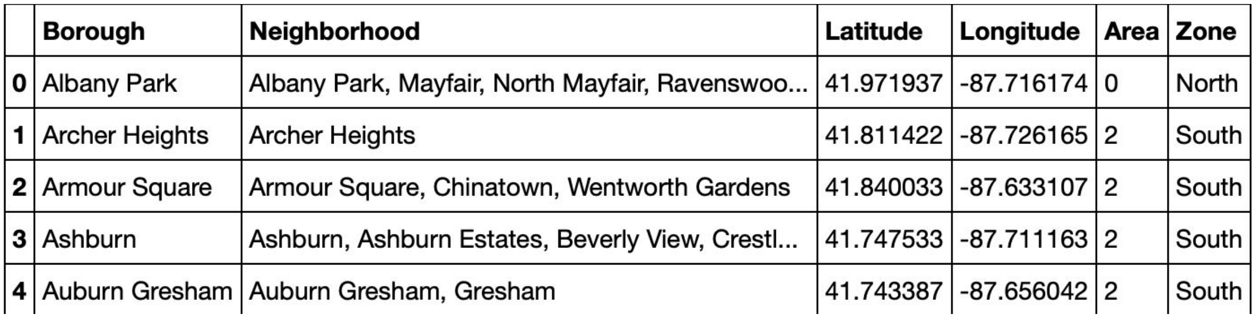
Data scraped from the webpages using BeautifulSoup were both placed into separate tables, one for the neighborhoods in Houston and one for the neighborhoods in Chicago. Since the data for both tables contained only the neighborhood names, I needed to make sure to get the geographical coordinates of each neighborhood to utilize the Foursquare API for analysis.

For the Chicago data frame, I created a list from the column Neighborhoods. From there I created two empty lists, one for each Longitude and Latitude. Creating a for loop to use the Geopy package to loop through the neighborhood list and append the resulting longitude and latitude values, I now had a data frame consisting of Neighborhood, Latitude and Longitude coordinates.

From there, I added an Area column to place the neighborhoods based on their location surrounding Chicago. The area classifiers are as follows:

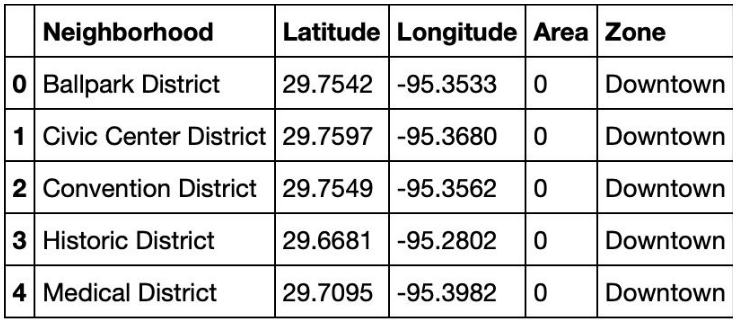
- North = 0
- West = 1
- South = 2
- Downtown = 3

I then added those values into the corresponding neighborhoods and ran a where clause on a new column to add the string of those values. The final data frame for Chicago neighborhoods is shown below.



	Borough	Neighborhood	Latitude	Longitude	Area	Zone
0	Albany Park	Albany Park, Mayfair, North Mayfair, Ravenswoo...	41.971937	-87.716174	0	North
1	Archer Heights	Archer Heights	41.811422	-87.726165	2	South
2	Armour Square	Armour Square, Chinatown, Wentworth Gardens	41.840033	-87.633107	2	South
3	Ashburn	Ashburn, Ashburn Estates, Beverly View, Crestl...	41.747533	-87.711163	2	South
4	Auburn Gresham	Auburn Gresham, Gresham	41.743387	-87.656042	2	South

The same process was done to the data containing neighborhoods in Houston. The final data frame for Houston neighborhoods is shown below.



	Neighborhood	Latitude	Longitude	Area	Zone
0	Ballpark District	29.7542	-95.3533	0	Downtown
1	Civic Center District	29.7597	-95.3680	0	Downtown
2	Convention District	29.7549	-95.3562	0	Downtown
3	Historic District	29.6681	-95.2802	0	Downtown
4	Medical District	29.7095	-95.3982	0	Downtown

2.3 Feature Selection

Since I currently live on the north side of Chicago, I am only interested in comparing the neighborhoods of Houston to those located on the north side of Chicago. To get that data ready for analysis I spliced the data frame of Chicago to consist of only those neighborhoods located in the Zone equal to North.

I then used join to combine the chicago and houston dataframes on Neighborhood, keeping all rows from both tables. I dropped both the area and zone columns and was left with a data frame containing the Neighborhoods, Latitude and Longitude values. To make sure I knew which neighborhood belonged to which city, I ran a lambda function on the column Latitude that entered either ‘Houston’ or ‘Chicago’ into a new column named City depending on whether the value in Latitude was less than 40. I used the value of 40 for my function because the Chicago latitudes were all above 40 while the Houston latitudes were around 29-31.

3. Methodology

3.1 Foursquare API

Foursquare offers data analysts location data based upon geographical coordinates. I used this to generate a list of venues within a radius of 500 meters for each neighborhood in my dataframe. I determined that getting venue information would help compare the neighborhoods to find similarities between Houston and Chicago.

After getting the list of venues, my search returned 1,776 total venues throughout all the neighborhoods. To make the data easier to compare, I ran one hot encoding and got dummy variables for the different types of venues. I then grouped by neighborhood and sorted the venues in descending order to get the top 10 for each neighborhood.

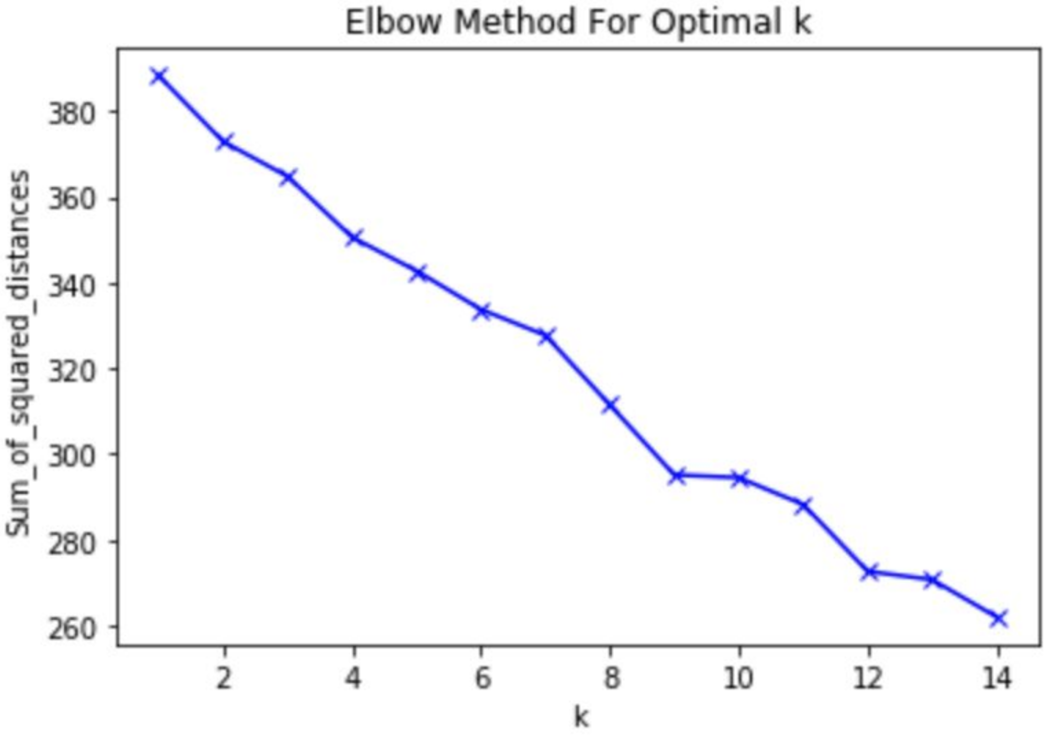
3.2 K-Means Clustering

To help determine which neighborhoods were similar, I used a form of unsupervised machine learning for analysis. K-Means clustering is a method used to partition *n* observations into *k* clusters to where each observation belongs to the cluster with the nearest mean. By getting the dummy variables of the top 10 venues for each neighborhood, K-means would be able to cluster the neighborhoods that are similar to each other.

It is important to find the optimal value of *k* when running clustering as to not over or underfit the data. This would cause certain neighborhoods to either be included or not included within certain clusters which would make analyzing the neighborhoods that were similar difficult.

To find the optimal value for *k*, I ran what is known as the Elbow Method. The Elbow Method looks at the percentage of variance explained as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn't give much better modeling of the data. More precisely, if one plots the percentage of variance explained by the clusters against the number of clusters, the first clusters will add much information (explain a lot of variance), but at some point the marginal gain will drop, giving an angle in the graph.

After running the Elbow Test, it was determined that the optimal number of clusters would be 9.



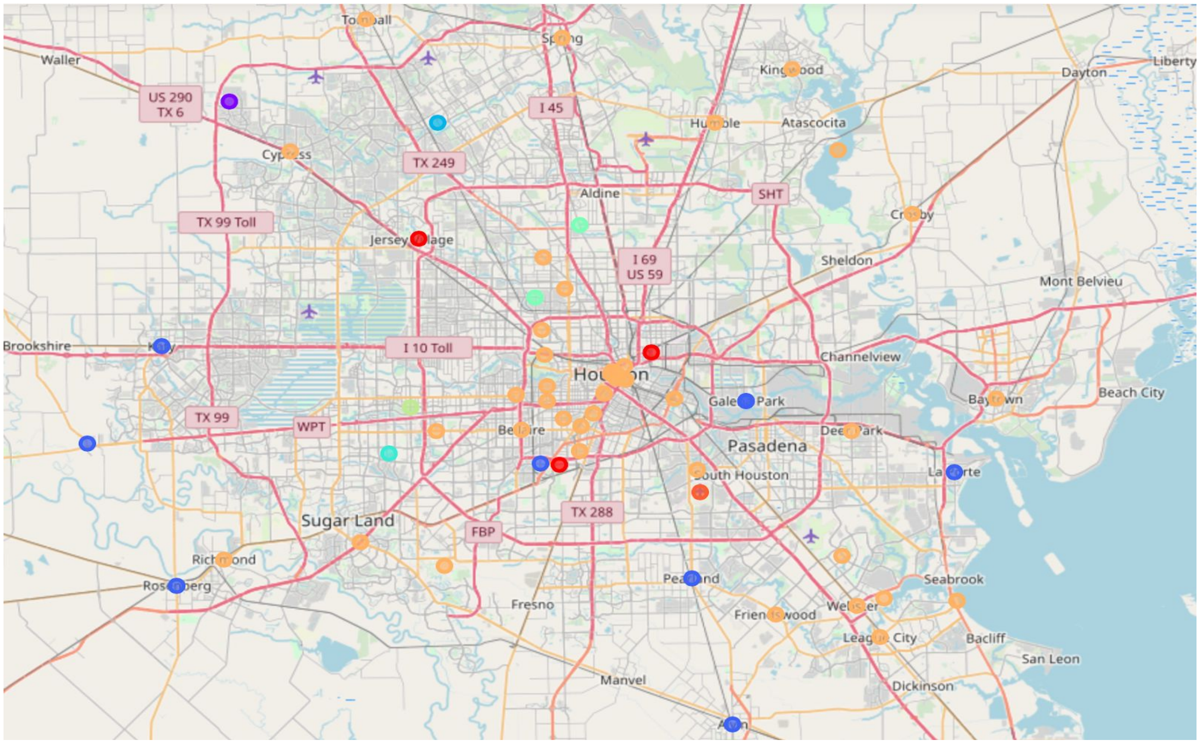
4. Results

After running the K-Means clustering, I had 9 clusters of neighborhoods with similar top 10 venues within each. Cluster 7 was the only cluster that contained both Chicago and Houston neighborhoods. Therefore, it was determined that the Houston neighborhoods within that cluster were the most similar to that of Chicago and would be used for further analysis before deciding whether to move to one of those neighborhoods or not.

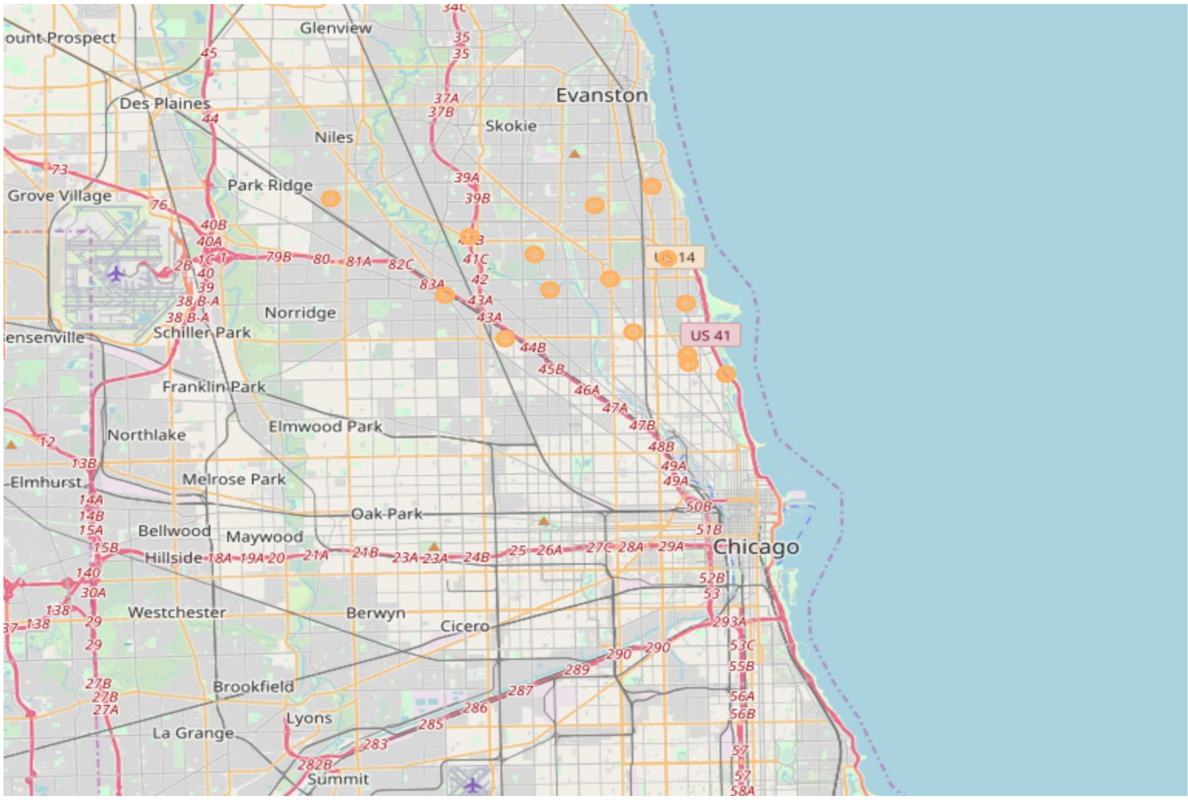
The list consists of 43 neighborhoods in Houston that were similar to those of Chicago. These neighborhoods were:

> Ballpark District, Civic Center District, Convention District, Historic District, Medical District, Shopping District, Skyline District, Theater District, Warehouse District, Bellaire, EaDo, Uptown, Heights, Medical Center, Memorial Park, Midtown, Museum District, River Oaks, Upper Kirby, West University, Northside, Conroe, Crosby, Cypress, North Houston District, Humble, Kingwood, Lake Houston, Spring, Tomball, The Woodlands, Missouri City, Richmond, Sugar Land, Baytown, Clear Lake, Deer Park, Friendswood, Kemah, League City, Nassau Bay, Webster, Chinatown.

After viewing a map of the neighborhoods and their associated clusters, you can see that the majority of the neighborhoods in Houston that belong to cluster 7 (orange dots on maps below) were located Downtown or within the Inner Loop. I determined this to be due to the way Houston is laid out. Chicago has a high population density located close to the center of downtown while Houston is more spread out and its overall population density is less than Chicago. The areas of Downtown Houston, along with Cypress and Sugar Land are very densely populated areas of greater Houston and the venues within those areas were similar to those of Chicago.



Map of Houston Neighborhood Clusters (above)
Map of Chicago Neighborhood Cluster (below)



5. Discussion

As previously mentioned in the result section, the neighborhoods in Houston that were similar to Chicago were located in densely populated areas of Houston. Our analysis shows that there are many neighborhoods in the Houston area (43 neighborhoods in total) that are similar to Chicago. Since we lived on the North Side of Chicago, we focused our attention on finding the venues that made up those neighborhoods. There are hundreds of venues located around each Neighborhood which would have made our data large and messy for analysis. Therefore, by taking the Top 10 venues, we were able to analyze and compare the data across multiple neighborhoods.

Since we were looking to find any neighborhoods in Houston that were similar, we got the Top 10 venues for every neighborhood in Houston. This way, we were able to combine the results and run our cluster model to fit the neighborhoods in Houston to those in Chicago. To figure out the optimal amount of clusters, we ran an Elbow Method test and got a result of 9 clusters.

We got our list of Neighborhoods similar to Chicago where we could further examine and determine which neighborhood/s would be the best fit. This of course, does not imply that those neighborhoods are the best fit for us. While they may have similar venues, neighborhoods are also defined by their culture and the residents living within. You will find all types of venues in every neighborhood, and even though these neighborhoods contained similar venues, I would suggest doing further analysis into each of them before committing to move there.

While we created a recommended list of Neighborhoods, these neighborhoods give us a great starting point where one can start looking into other attributes such as average age of the population, cost of living, crime data and other demographics.

6. Conclusion

As you can see, the purpose of this project was to identify the neighborhoods in Houston that were similar to the neighborhoods on the north side of Chicago to help myself determine if I should accept a job and move to Houston. By determining the Top 10 venues within each neighborhood by using the Foursquare data, I was able to identify the the neighborhoods in Houston that were similar to Chicago by using k-means clustering. The list of neighborhoods generated can be used as starting points for final analysis to determine where I should move.

Final decision on neighborhood location will be made based on specific characteristics of neighborhoods and taking into consideration other additional factors like attractiveness of each location (proximity to park or water), cost of living (rent/own property), social and economic dynamics of every neighborhood etc.

This concludes my report.