Comparison of the Cities of Nashville, Chapel Hill and Charlotte

Hunter Finger *hufinger*

Introduction

I have lived in two places in my lifetime: Charlotte, North Carolina and Chapel Hill, North Carolina. Recently, I have travelled to Nashville, Tennessee and fell in love with the vibrancy of the city. In order to determine if this is a place that I would like to move to in the future, I want to compare these three cities to each other. My hope is to find that they are similar in the pieces that I value, and if they are not, I have learned something. One of the most important parts of living in a particular location is the venues surrounding your home. If you do not enjoy the activities around the home, you will not enjoy living there. Leveraging Foursquares' API, I hope to find the similarities and differences between these three US cities.

Data

Description and Acquisition

Our dataset will include the venues within a 750 meter radius from the zipcodes from Mecklenburg County (Charlotte, NC), Orange County (Chapel Hill, NC) , and Davidson County (Nashville, TN). The zip codes for these counties will be scraped from http://www.ciclt.net/sn/clt/capitolimpact/gw_ziplist.aspx?ClientCode=capitolimpact after navigating to the correct counties needed for the project. The venues for the zip codes will include all restaurants, shopping centers, and health service capabilities. The venues will be acquired through the Foursquare API used in previous assignments.

Cleaning

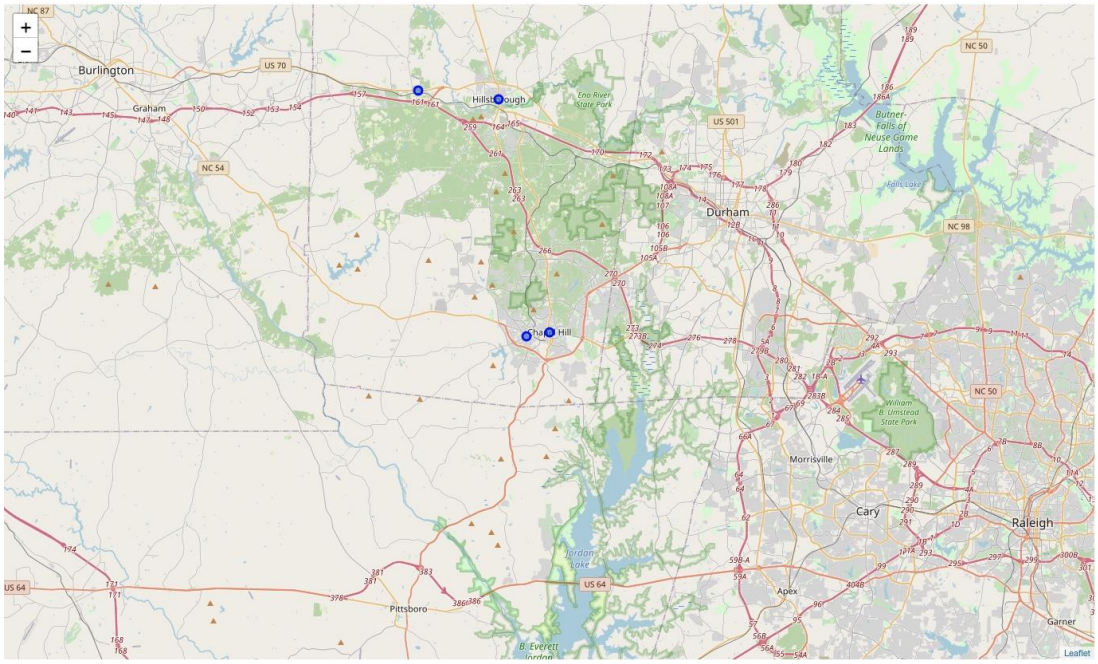
Once the zip code data was pulled from the website, it was converted into a dataframe which included the city, state, zip code and county. Using the first three columns of the dataframe, an address was creating in standard formatting, e.g. "Cornelius, NC 28031." Passing this list into the geocode function resulted in the dataframe including the longitude and latitude of each of the zip codes. After inspecting the dataframe, it was apparent that some cities had two or more zip codes without any difference in location. These duplicate longitude and latitude values were removed. Utilizing the Foursquare API, venues within a 750 meter radius from each longitude, latitude pair were found, and turned into a frequency of each venue type. The eight most frequent venue types were kept in the dataset for clustering.

Exploratory Data Analysis

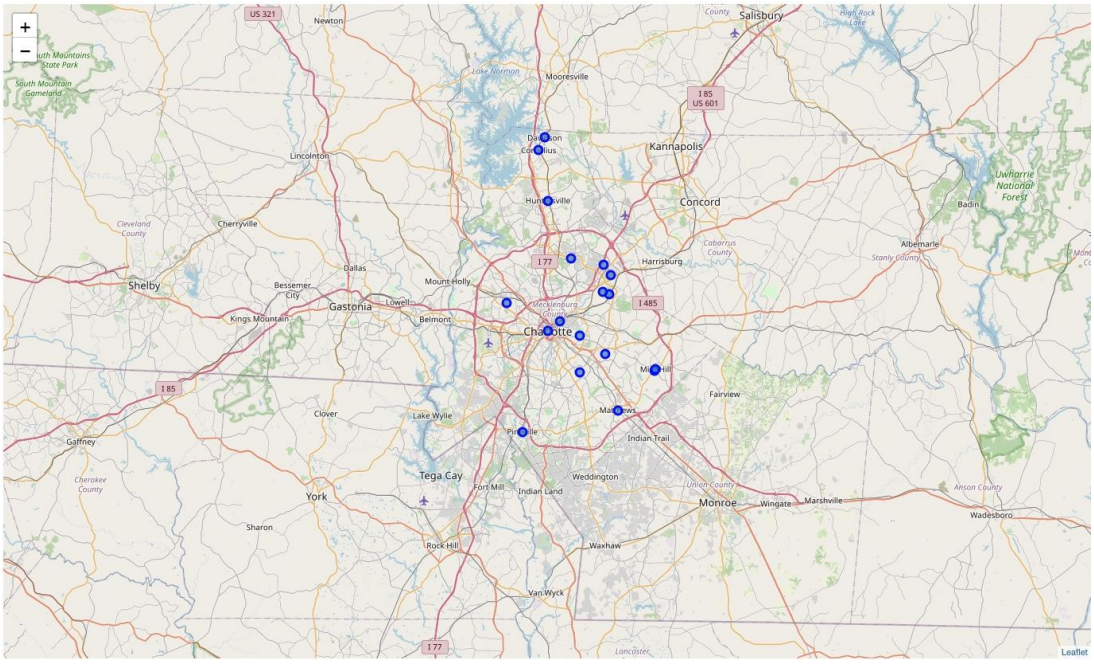
The exploratory data analysis was fairly basic in this project. The total number of zip codes tested was 36 over the three county pool. The maps for where the zip codes are located can be seen below. The total number of unique venue types was 211, with Charlotte having the highest total number of venues at 216. Cedar Grove was the smallest with only one venue

returned from the radius. As expected, the major city centers (Charlotte and Nashville) had the most venues while that number dropped further into the suburbs.

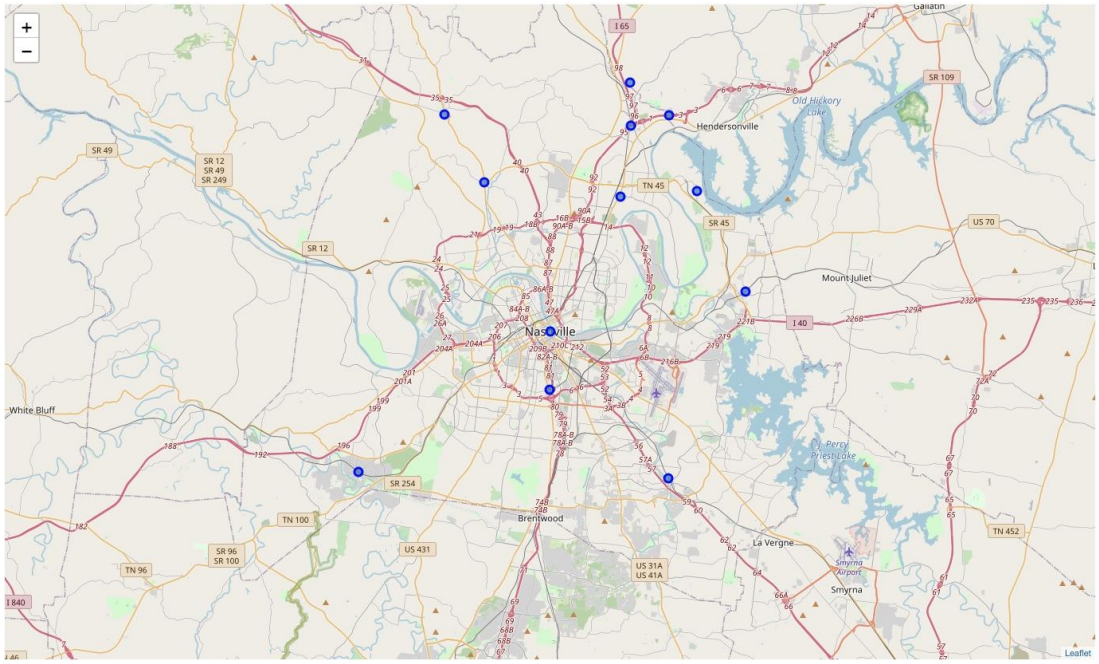
Orange County Map (Chapel Hill)



Mecklenburg County Map (Charlotte)



Davidson County Map (Nashville)



Number of Venues in each City

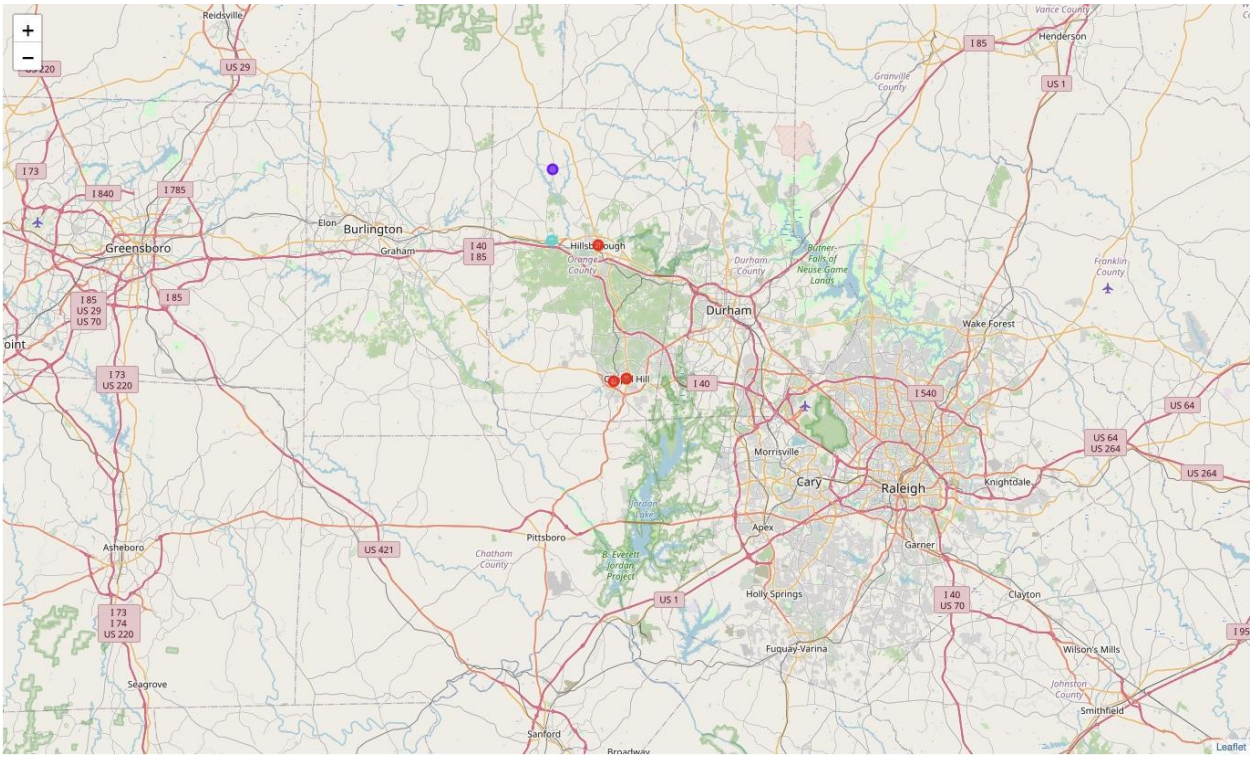
Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Antioch	10	10	10	10	10	10
Bellevue	7	7	7	7	7	7
Carrboro	49	49	49	49	49	49
Cedar Grove	1	1	1	1	1	1
Chapel Hill	70	70	70	70	70	70
Charlotte	219	219	219	219	219	219
Cornelius	19	19	19	19	19	19
Davidson	28	28	28	28	28	28
Efland	3	3	3	3	3	3
Goodlettsville	58	58	58	58	58	58
Hermitage	5	5	5	5	5	5
Hillsborough	22	22	22	22	22	22
Huntersville	13	13	13	13	13	13
Joelton	4	4	4	4	4	4
Madison	28	28	28	28	28	28
Matthews	41	41	41	41	41	41
Melrose	44	44	44	44	44	44
Mint Hill	42	42	42	42	42	42
Nashville	106	106	106	106	106	106
Newell	3	3	3	3	3	3
Old Hickory	15	15	15	15	15	15
Paw Creek	4	4	4	4	4	4
Pineville	23	23	23	23	23	23
Whites Creek	7	7	7	7	7	7

Modeling and Results

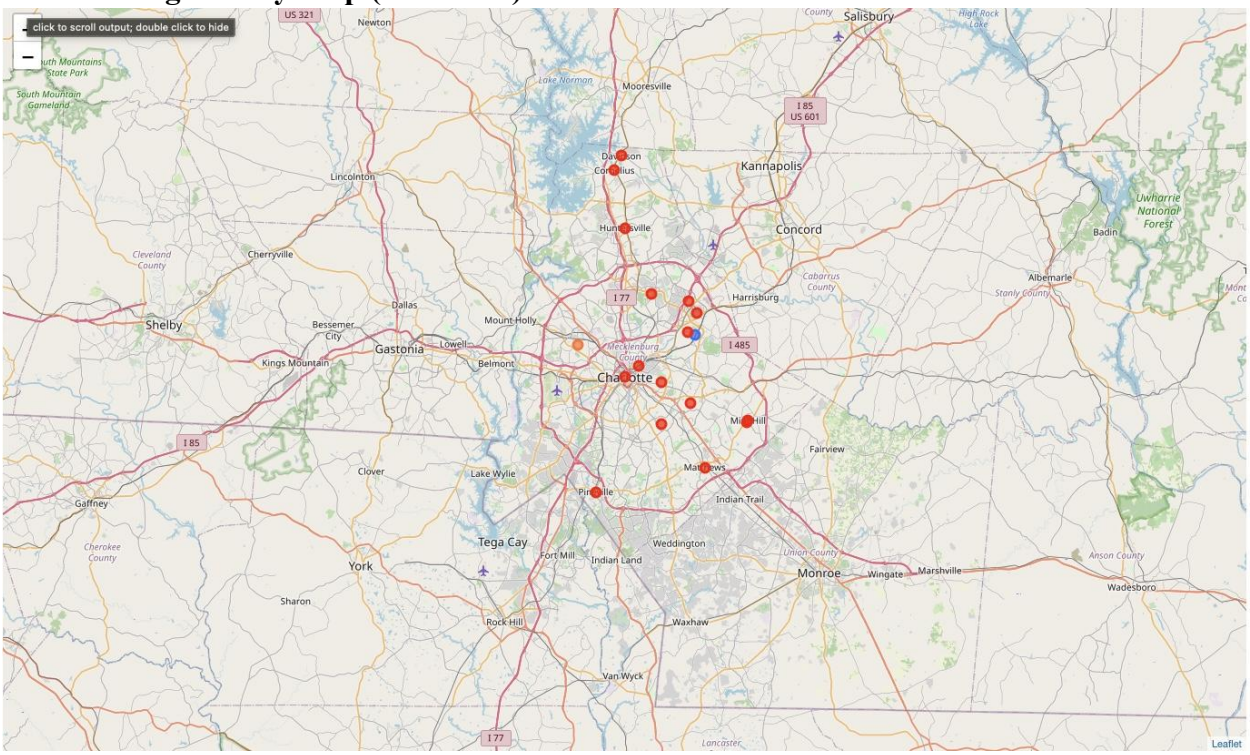
After the exploratory analysis was completed, the K-Means clustering algorithm was used to determine the zip codes that were most similar. A clustering algorithm was chosen because this is an unsupervised machine learning problem, there is not a dataset that defines how similar the two neighborhoods are. Clustering works well for these types of problems because it groups the different data points by their similarities to each other. For this algorithm, K was set to be 7 in order to handle the large number of zip codes given. The results shown below have different colored points on the map to indicate which cluster the zip code belongs. Nashville, Charlotte, and Chapel Hill are all in the same (red) cluster. Looking deeper into the clusters, the main cluster is filled with areas who have lots of food locations. Almost all of the most common venues are food/tourism related. The other 6 groups show venues such as automotive shops, doctor's offices and atms. These groups are also in more rural areas located further away from

the large city center. These indications show that the cities are in fact similar. I would likely enjoy living in the city of Nashville because of how much I enjoyed living in Charlotte and Chapel Hill.

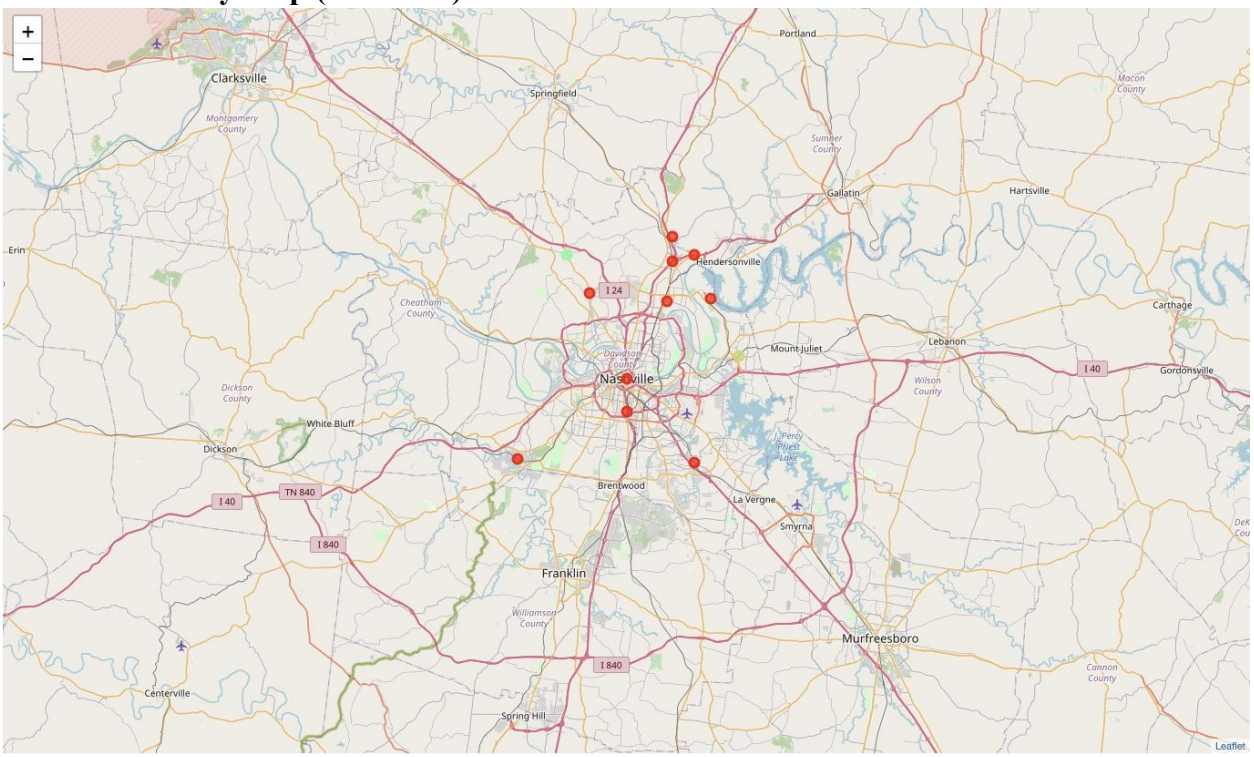
Orange County Map (Chapel Hill)



Mecklenburg County Map (Charlotte)



Davidson County Map (Nashville)



Conclusion and Future Direction

This analysis found that Chapel Hill, Charlotte and Nashville are similar cities based on the venues available in the area. This is a result that I am not entirely confident in because of the large size of the main cluster. For future analysis, I would like to get the coordinates of the major neighborhoods in the large cities of Nashville and Charlotte because each neighborhood has a different feeling to it. This would also allow the ability to remove some of the smaller suburbs in the county in order to find the true differences in the cities. We also proved how different neighborhoods can be in close proximity in the lab about clustering Manhattan neighborhoods. Other data would be useful as well, such as crime rate, housing costs, walkability, and weather patterns. These are all very important to the similarity of two or more places. To address the concerns about the sizing of the clusters, I could see where a questionnaire and a recommendation matrix could be applied to this problem. Ask the user about the things they like and do not like about their current location and generate recommendations or similarity scores based off of these responses.