IBM Applied Data Science Capstone : Similar Neighbors in Nashville and New York

Introduction

The problem : People are always moving homes from one places to the other, maybe due to finding a new job or some other reasons. When people are moving from one place to the other. Sometimes they may want to move to a similar neighborhood to the one they are currently living in. By similar I mean maybe for both neighbors there are shopping malls nearby, similar kind of resturants. Currently I live in Nashville, TN, USA. Supposed that I find a new job in New York City and want to find a similar neighbor to which I am currently living. How can I do this?

This project will also help a lot of other people with similar concerns, so this study will be very helpful.

Data used in this project

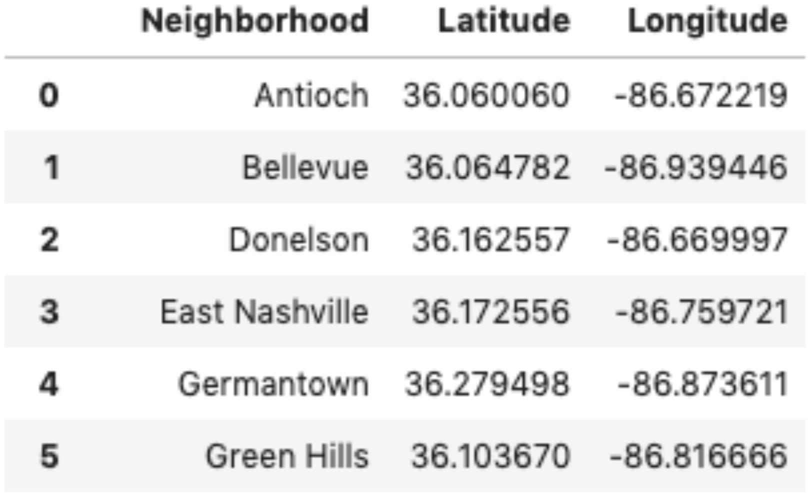
The data we used for this project include the New York neighborhoods information from the previous project and the Nashville Neighborhood information I get from the wiki page :
https://en.wikipedia.org/wiki/Nashville,_Tennessee

I will use the name of the neighborhood as input and use geo library got the latitude and longitude of each neighborhood. Then I will use the Foursquare API to get the venues around each neighbor. Combine this information with the venues in the New York City neighborhoods and use a cluster algorithm to cluster the neighbor in Nashville and New York together. Then we can see which neighbor in Nashville and New York belongs to the same cluster and I would know which neighbor to move to.

One thing to notice is that since we are going to cluster the neighborhood of Nashville and New York together, we need to make sure that the venues in the Nashville neighbors and New York neighbors are the same to make sure that the cluster algorithm are not affected by some venues unique to some areas. So after getting the venues for each Nashville neighbors and New York neighbors, we need to have a study of the neighbors of each city, find the common venues and remove the redudent venues to make sure the cluster works fine.

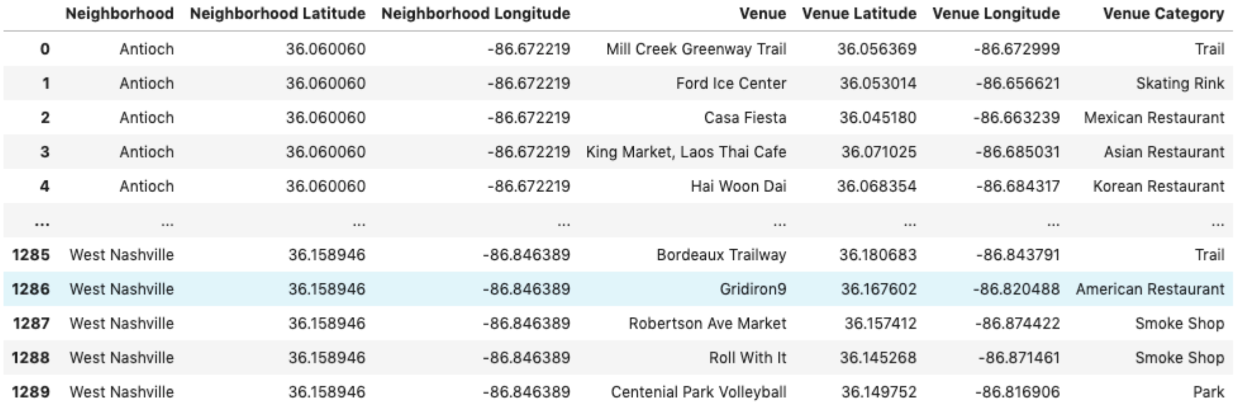
Methodology

In order to be able to cluster neighbors in Nashville and New York together, we first need to get the latitude and longitude of each neighbor in Nashville using the geo library. The following plots shows the part of the result.



	Neighborhood	Latitude	Longitude
0	Antioch	36.060060	-86.672219
1	Bellevue	36.064782	-86.939446
2	Donelson	36.162557	-86.669997
3	East Nashville	36.172556	-86.759721
4	Germantown	36.279498	-86.873611
5	Green Hills	36.103670	-86.816666

After getting the coordinate of each neighbor, we now can use the four square API to get the nearby venues of each neighbors. After some work, we found nearly 1300 venues in Nashville.
And around 200 unique kind of venues.
Some sample are shown in the following figure

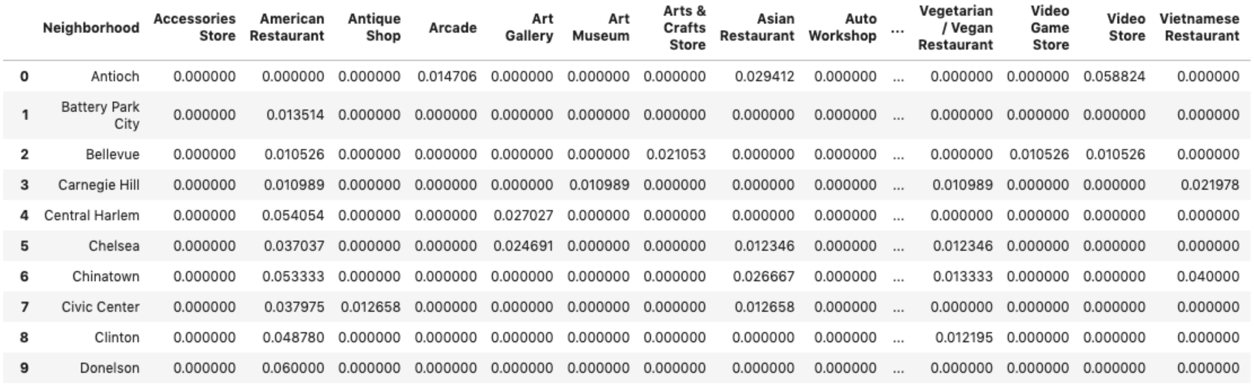


	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Antioch	36.060060	-86.672219	Mill Creek Greenway Trail	36.056369	-86.672999	Trail
1	Antioch	36.060060	-86.672219	Ford Ice Center	36.053014	-86.656621	Skating Rink
2	Antioch	36.060060	-86.672219	Casa Fiesta	36.045180	-86.663239	Mexican Restaurant
3	Antioch	36.060060	-86.672219	King Market, Laos Thai Cafe	36.071025	-86.685031	Asian Restaurant
4	Antioch	36.060060	-86.672219	Hai Woon Dai	36.068354	-86.684317	Korean Restaurant
...	...	...	...	...	...	...	...
1285	West Nashville	36.158946	-86.846389	Bordeaux Trailway	36.180683	-86.843791	Trail
1286	West Nashville	36.158946	-86.846389	Gridiron9	36.167602	-86.820488	American Restaurant
1287	West Nashville	36.158946	-86.846389	Robertson Ave Market	36.157412	-86.874422	Smoke Shop
1288	West Nashville	36.158946	-86.846389	Roll With It	36.145268	-86.871461	Smoke Shop
1289	West Nashville	36.158946	-86.846389	Centenial Park Volleyball	36.149752	-86.816906	Park

We did the same for the New York Manhattan data and
found there are around 3300 venues and 330 unique kind of venues.

Next, we found there are around 166 common venues in Nashville and Manhattan, so we removed the rows in both dataset whose venues are not in the common venues.

Then we one-hot encoded the venues and then calculated the frequency of each venue appears in each neighbor, the results is shown in following figure.

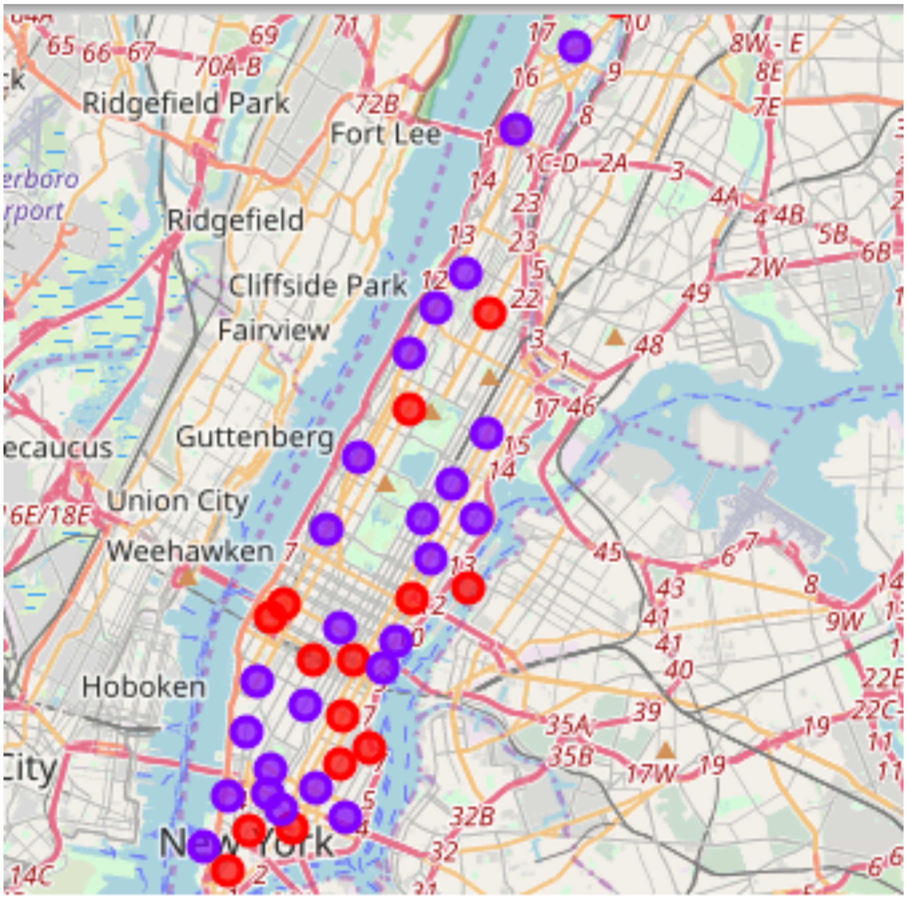


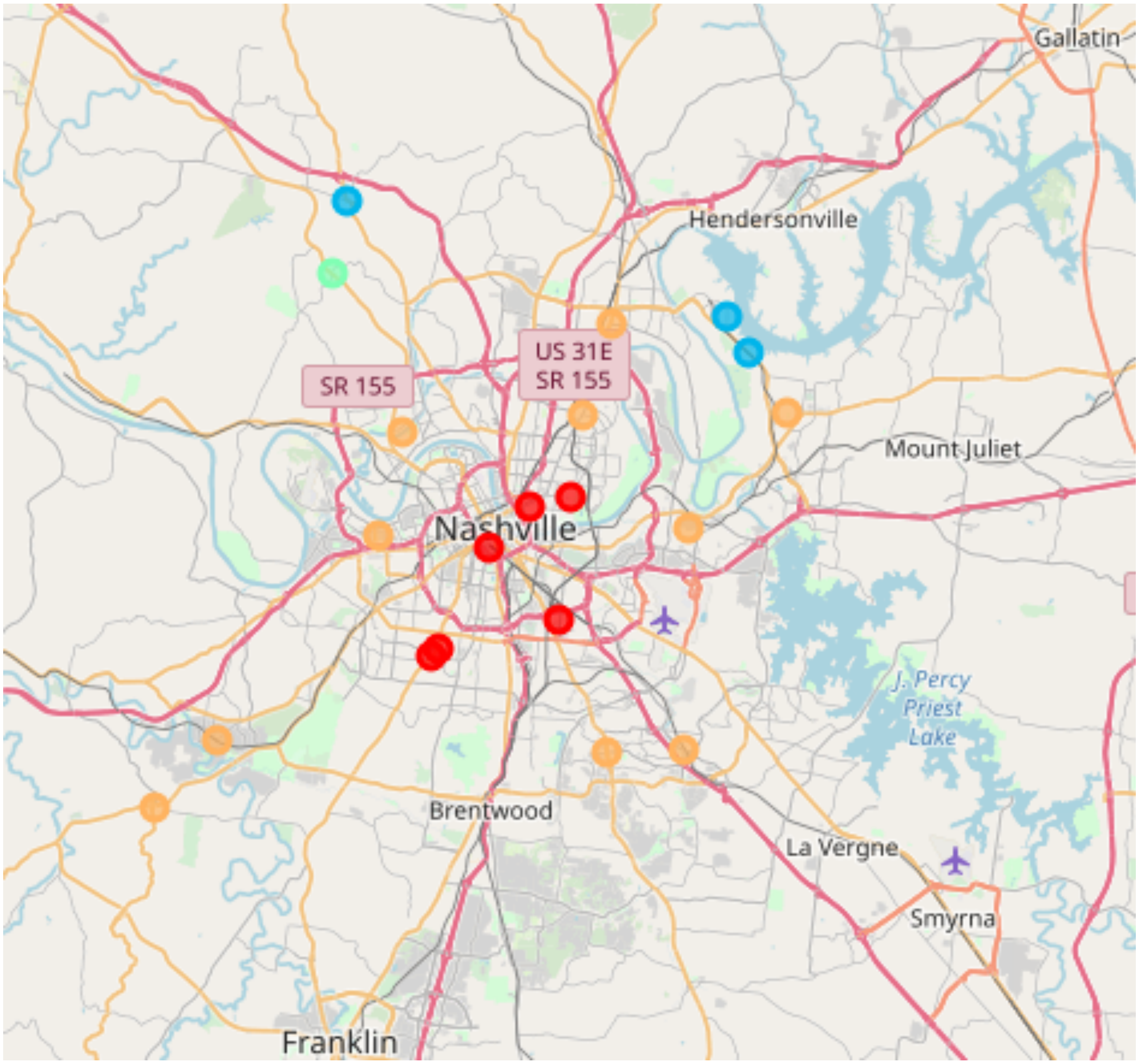
	Neighborhood	Accessories Store	American Restaurant	Antique Shop	Arcade	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Auto Workshop	...	Vegetarian / Vegan Restaurant	Video Game Store	Video Store	Vietnamese Restaurant
0	Antioch	0.000000	0.000000	0.000000	0.014706	0.000000	0.000000	0.000000	0.029412	0.000000	...	0.000000	0.000000	0.058824	0.000000
1	Battery Park City	0.000000	0.013514	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000
2	Bellevue	0.000000	0.010526	0.000000	0.000000	0.000000	0.000000	0.021053	0.000000	0.000000	...	0.000000	0.010526	0.010526	0.000000
3	Carnegie Hill	0.000000	0.010989	0.000000	0.000000	0.000000	0.010989	0.000000	0.000000	0.000000	...	0.010989	0.000000	0.000000	0.021978
4	Central Harlem	0.000000	0.054054	0.000000	0.000000	0.027027	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000
5	Chelsea	0.000000	0.037037	0.000000	0.000000	0.024691	0.000000	0.000000	0.012346	0.000000	...	0.012346	0.000000	0.000000	0.000000
6	Chinatown	0.000000	0.053333	0.000000	0.000000	0.000000	0.000000	0.000000	0.026667	0.000000	...	0.013333	0.000000	0.000000	0.040000
7	Civic Center	0.000000	0.037975	0.012658	0.000000	0.000000	0.000000	0.000000	0.012658	0.000000	...	0.000000	0.000000	0.000000	0.000000
8	Clinton	0.000000	0.048780	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.012195	0.000000	0.000000	0.000000
9	Donelson	0.000000	0.060000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.000000

Then we used this dataset as the input to the k-means cluster algorithms.

Results

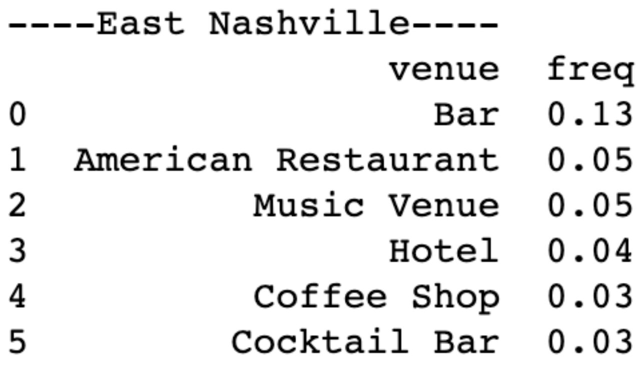
To better see the results, we plot two figures of the Nashville area and Manhattan area, with the same color means the same category.



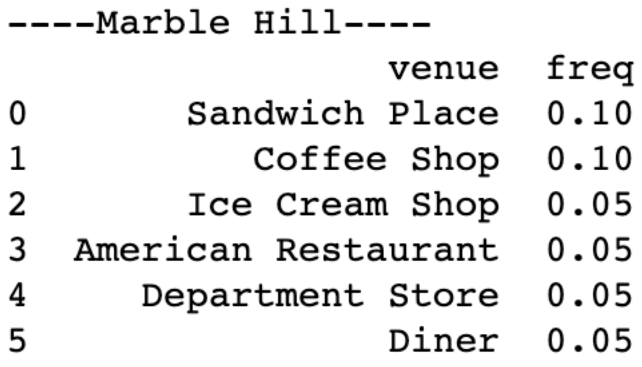


Discussion

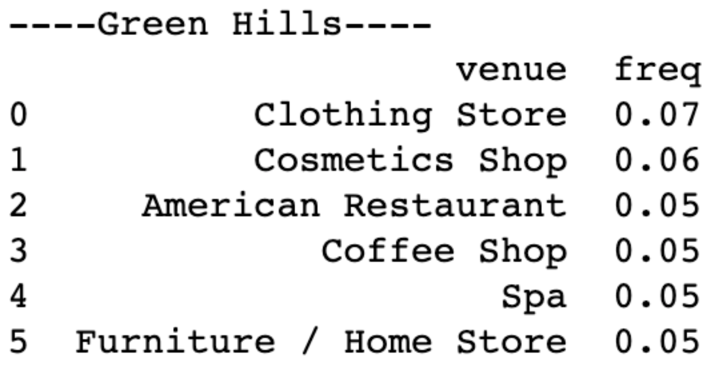
From the above two figures, we see that there are some neighbors in Downtown Nashville are in the same cluster as some of the neighbors in Manhattan. We looked at some examples, such as the East Nashville, Green Hills in Nashville area and Marble Hills in Manhattan are in the same cluster, we looked at the most frequent venues nearby, shown in the following three figures.



```
----East Nashville----
                 venue  freq
0                  Bar  0.13
1  American Restaurant  0.05
2          Music Venue  0.05
3                Hotel  0.04
4          Coffee Shop  0.03
5         Cocktail Bar  0.03
```



```
----Marble Hill----
                 venue  freq
0       Sandwich Place  0.10
1          Coffee Shop  0.10
2       Ice Cream Shop  0.05
3  American Restaurant  0.05
4     Department Store  0.05
5                Diner  0.05
```



```
----Green Hills----
                   venue  freq
0         Clothing Store  0.07
1         Cosmetics Shop  0.06
2    American Restaurant  0.05
3            Coffee Shop  0.05
4                    Spa  0.05
5  Furniture / Home Store  0.05
```

we see there are many common venues such as Coffee Shop and American Restaruants. So we can say this method works fine.

Conclusion

We successfully find some common neighbors in Nashville and Manhattan area, with the similar nearby venues.

However, one disadvantage of this study is it only considers the downtown area in New York, which caused the result that there are no corrspondings areas to the urban areas in Nashville. To compensate this, we can study other neighbors of New York city separately. The reason I didn’t consider all the neighbors in New York together is there are too many neighbors in New York and there is a limit on the API calls you can make one day. Another reason is that I am afraid there would be data imbalance if I add all the neighbors in New York. But studying other areas in New York is pretty simple, just change the selection of New York neighbors and it can be done pretty easily.