

# Finding Denver-like Neighborhoods in Houston

Jesse Havens

7/26/2020

## **Background:**

When moving from one city to another, it can be hard to know what neighborhoods to look at to buy or rent a home. My wife and I recently went house hunting in Houston, Texas and found it challenging to translate our preferences to our realtor. This is also not ideal for the realtor because they are working for free until their client decides to buy or rent a place. If data science could help translate the neighborhoods from one city to another, that may speed up and simplify the house hunting experience.

## **Business Problem**

For this project I will attempt to ease the pain of moving to a new city by building a clustering algorithm using foursquare to compare neighborhoods from two cities. In this case I will look at Denver, Colorado neighborhoods versus Houston, Texas and cluster local venue frequency to match similar neighborhoods in each city. These two cities are very different, and I am curious to see if a clustering algorithm could have saved my wife and I a lot of pain and heartache looking for a place to buy or rent in Houston.

## **Data Sources:**

I will be scraping a wikipedia page ([Houston neighborhoods](#)) for neighborhoods in Houston. The Denver neighborhoods are available for download in a csv file from this website: [Denver neighborhoods](#). With the neighborhoods in hand, I will use foursquare to gather local venues in each neighborhood then cluster the Denver and Houston neighborhoods together. The clusters for popular neighborhoods will be cross-referenced with neighborhood grades and descriptions from [niche.com](#).

## **Data Cleaning:**

In both the Wikipedia page with Houston neighborhoods and the spreadsheet with Denver neighborhoods there were unnecessary columns that I dropped from the dataframes. There were also different column names used for neighborhoods so I changed the column names to be consistent.

In the Wikipedia table with Houston neighborhoods there were multiple neighborhoods occasionally lumped together in a single entry and also information in parenthesis that was not part of the neighborhood name. To address the multiple neighborhood listings, I split the entries based on a backslash and appended the new neighborhood to the end of the dataframe. The total neighborhood entries went from 88 to 113 with this process. For the information in parenthesis I set up a regular expression to remove any text within parenthesis.

One popular neighborhood in Houston 'Houston Heights' was labeled 'Greater Heights' and was not being found with Nominatim, so I relabeled that neighborhood 'Houston Heights'.

Additional columns with the state and city were added for later use with Nominatim, then the two tables were merged into one dataframe.

With the large number of neighborhoods in both cities, I did not want to manually adjust any latitude/longitude pairs but there were some values that would be returned from Nominatim that were well outside the city limits and even the state. When sending an address to Nominatim I included the city and state, which vastly improved the accuracy of the returned neighborhoods but there were still outliers. To filter out the invalid locations I set up a radius calculation from the city center and filtered any returned values that were beyond 50 kilometers. The radius filter does not guarantee the accuracy of the returned locations, but it does keep locations well outside the city limits from entering the clustering analysis.

With the above cleaning processes most of the recognized neighborhoods were located from each city and all the data were in a single dataframe ready for analysis.

### **Methodology:**

The Foursquare API was used to find the first 100 venues within a 2500m (~1.5 miles) radius around the lat/lon pair determined from Nominatim. The query may have some overlapping areas but for the purpose of selecting a neighborhood to live in, I thought a 1.5 mile radius was a reasonable distance to consider.

One-hot encoding was applied to the returned venues and consisted of 366 unique venue types. The venues were then grouped by neighborhood and the frequency of each venue was calculated for each neighborhood. The top ten venue types were added to our neighborhood dataframe to help interpret the clustering results.

The K-Means clustering algorithm was applied to the venue frequency data to determine ten clusters. With the large number of neighborhoods being analyzed, I wasn't interested in clusters with a single neighborhood, so I cycled through the K-Means clustering algorithm dropping single clusters at a time if there was only one neighborhood. A histogram of the final neighborhood count per cluster is shown in Figure 1. The clustering results were then merged back in with the master dataframe containing the top venues for each neighborhood.

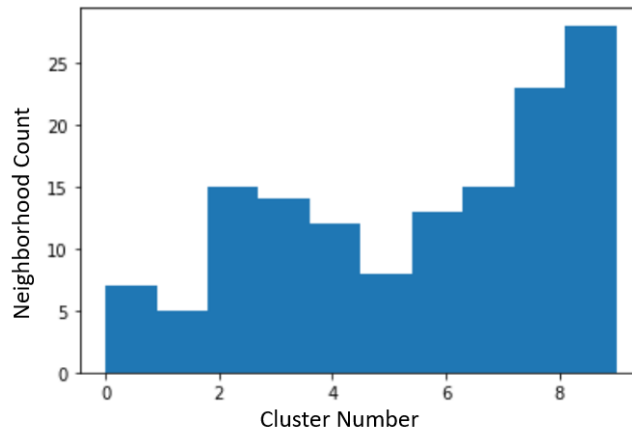












Fig. 1: Neighborhood count per cluster.

By analyzing the common venues in each neighborhood and looking at Niche.com grades, I have relabeled each cluster. The cluster labels can be found in Table 1 with the cluster number, cities containing that cluster, and the map markers of each cluster. There are only two clusters that are unique to each city. Maps of Denver and Houston with the clusters are shown in Figures 2 and 3.

**Table 1: Relabeled K-Means clusters with map markers**

Cluster #	Cluster Label	Cities with Cluster	Symbol
0	Party central	Denver	
1	The outskirts	Denver, Houston	
2	Functional city	Denver, Houston	
3	Dense suburbia	Denver, Houston	
4	Hoppin Metropolis	Denver, Houston	
5	Chain station	Houston	
6	Young professionals suburbia	Denver, Houston	
7	The urban-suburb desirables	Denver, Houston	
8	Discount central	Denver, Houston	
9	Sophisticated living	Denver, Houston	

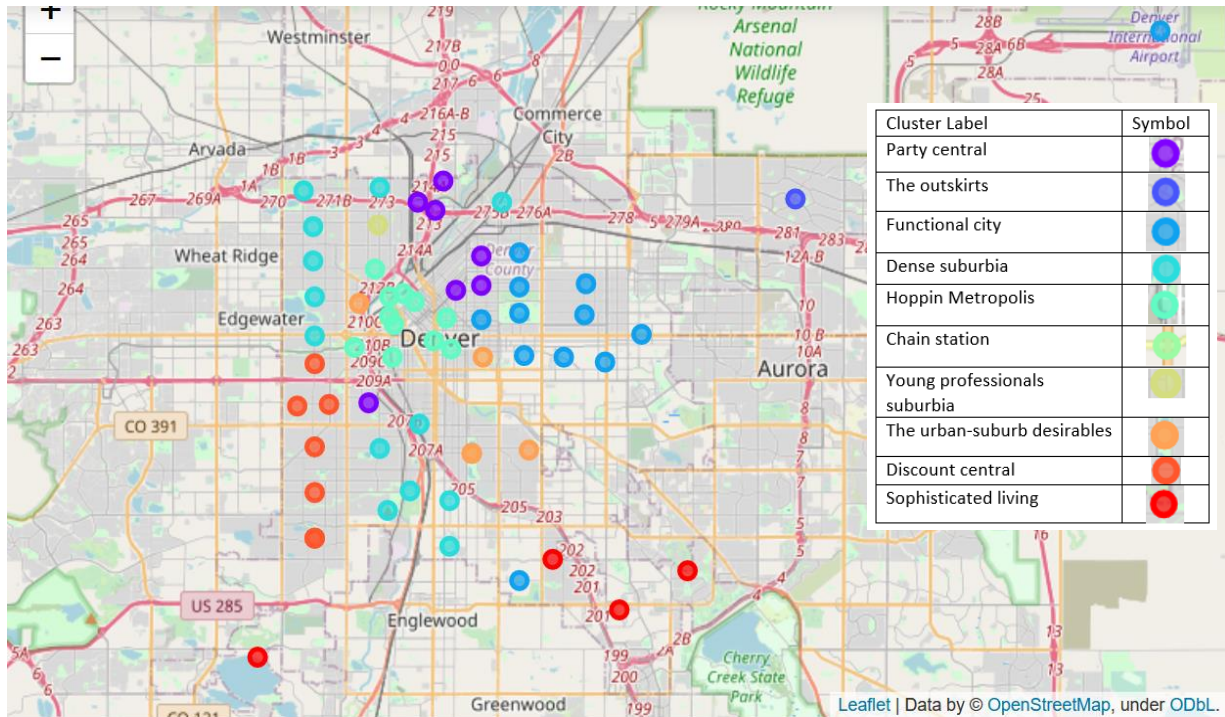


Fig. 2: Denver, Colorado neighborhood clustering results.

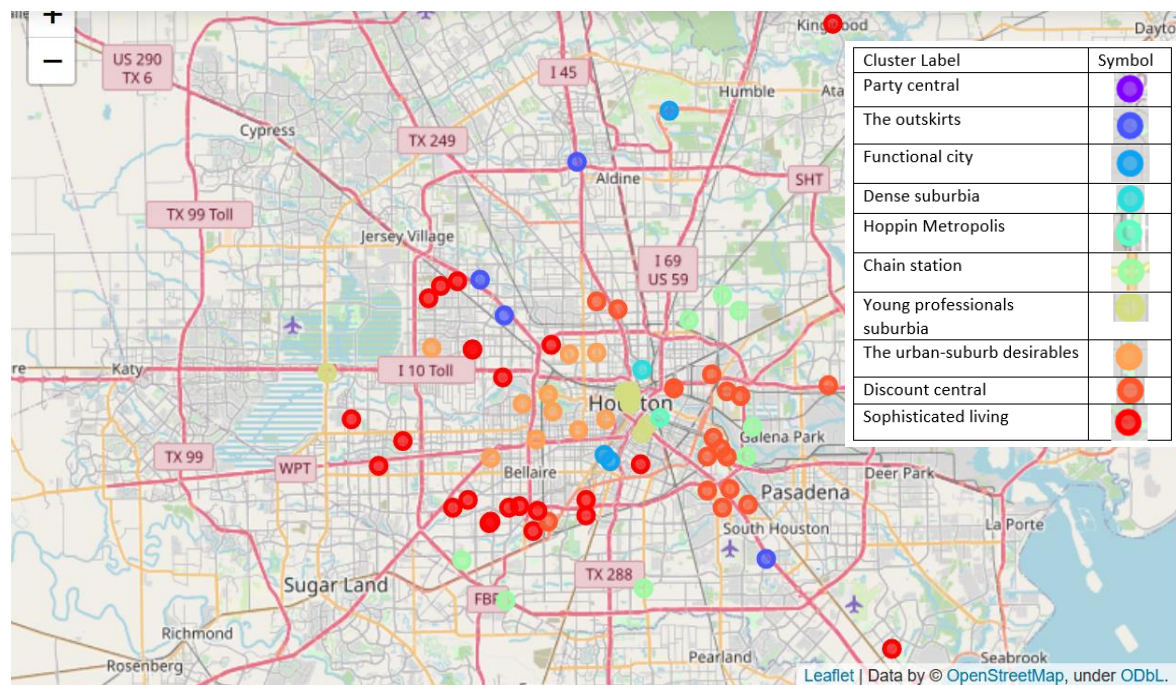


Fig. 3: Houston, Texas neighborhood clustering results.

To dig a little deeper into the results I set up a correlation matrix for each cluster and a function to sort correlations for a given neighborhood. The correlation ranking allows us to understand some of the larger clusters a little better. An example use case can be illustrated for the *Functional city* cluster. In Figure 4 we have the correlation matrix showing higher correlation in lighter colors. One of the things that jumped out at me when I looked at this cluster was the inclusion of Denver and Houston airports. To make the relationships a little clearer we can use the sorted correlation coefficients with Denver International Airport (DIA) as the reference neighborhood. The sorted results are shown in Figure 5. Now we can see that the two airports are fairly well correlated and have poor correlation to virtually all the other neighborhoods.

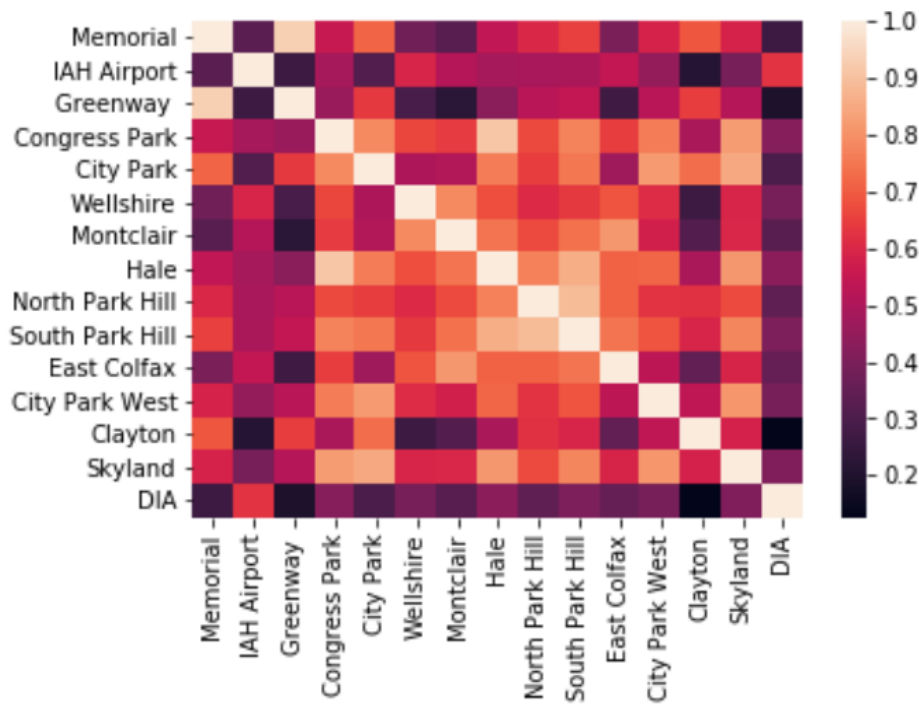


Fig. 5: Correlation matrix for the *Functional city* cluster.

DIA	1.000000
IAH Airport	0.626347
Hale	0.434300
Congress Park	0.422771
Skyland	0.407625
South Park Hill	0.405003
City Park West	0.392576
Wellshire	0.389791
East Colfax	0.356741
North Park Hill	0.343700
Montclair	0.328056
City Park	0.299562
Memorial	0.265690
Greenway	0.191915
Clayton	0.127073

Fig. 6: The *Functional city* cluster sorted by correlation to Denver International Airport (DIA).

If we instead sort the cluster by City Park, which is the neighborhood containing the Denver Zoo, we see much better correlation to most of the neighborhoods and the airports are at the bottom of the list. To analyze the results without getting overwhelmed with data, I will search for these representative neighborhoods to discuss in more detail instead of trying to cover each and every neighborhood. For the *Functional city* cluster, City Park is a better reference neighborhood than DIA.

City Park	1.000000
Skyland	0.850018
City Park West	0.819380
Congress Park	0.787442
Hale	0.761553
South Park Hill	0.751682
Clayton	0.732115
Memorial	0.715121
North Park Hill	0.646600
Greenway	0.639222
Montclair	0.509029
Wellshire	0.501430
East Colfax	0.470753
IAH Airport	0.314079
DIA	0.299562

Fig. 7: The *Functional city* cluster sorted by correlation to City Park.

## **Results:**

For each cluster we have the Folium maps, top venues, correlation matrices, and Niche.com grades. I will step through each cluster and summarize the clustering results within the context of the aforementioned information.



## Party Central –

This is one of two clusters that only appear in one city. The clear distinguishing factor for this cluster looking at the top venues is the breweries and bars. The Niche.com score of Cole, our representative neighborhood, shows that this is an area with great nightlife, but poor crime rates, schools, and housing.

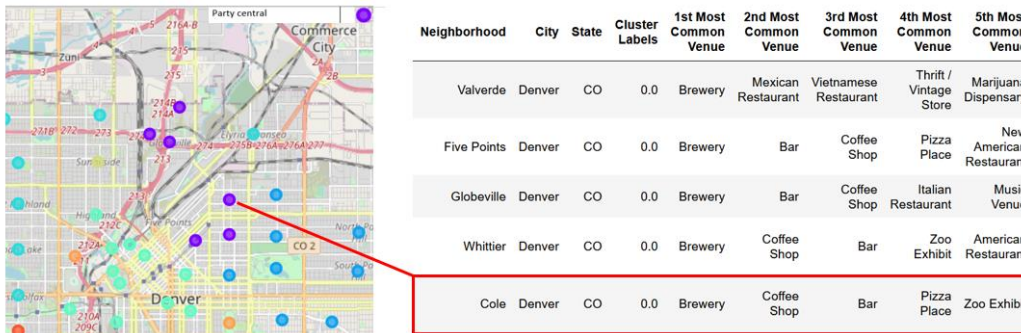


Fig. 8: Left - Folium map zoomed in to the Party Central neighborhoods. Right – top venues for each neighborhood. Red box and arrow indicate the representative neighborhood and location on the map.

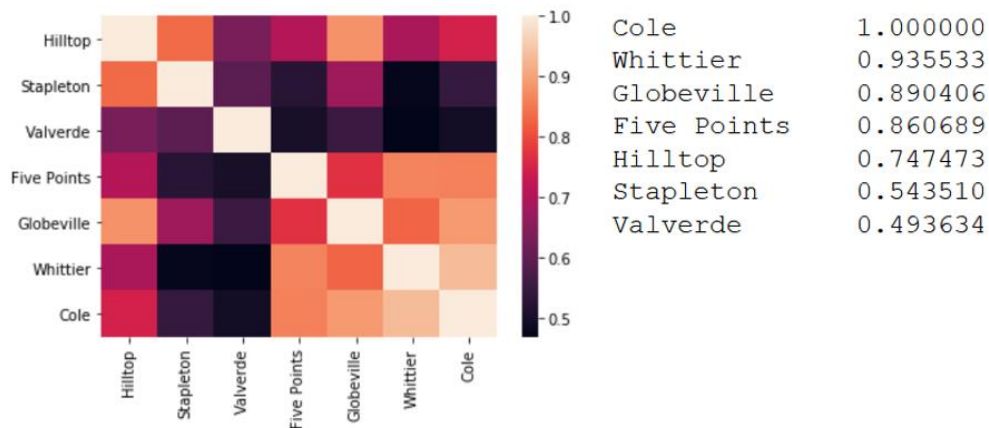


Fig. 9: Left – cross-correlation plot for all neighborhoods. Right – sorted correlation table for the Cole neighborhood. The correlation map shows that Stapleton and Valverde do not cluster well with this group.

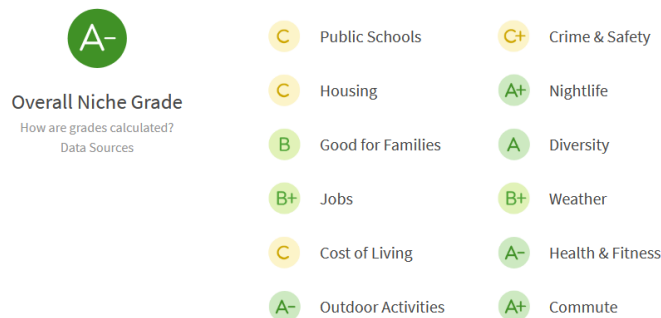


Fig. 10: Niche.com grades for the Cole neighborhood.

The Outskirts –

These neighborhoods tend to have more generic dining options and shops. The Niche.com grades suggest average crime rates, schools, and housing so these are most likely quieter areas that would be good for raising a family. Overall, the correlations are poor, so someone interested in this style of living would need to check out the locations in more detail.

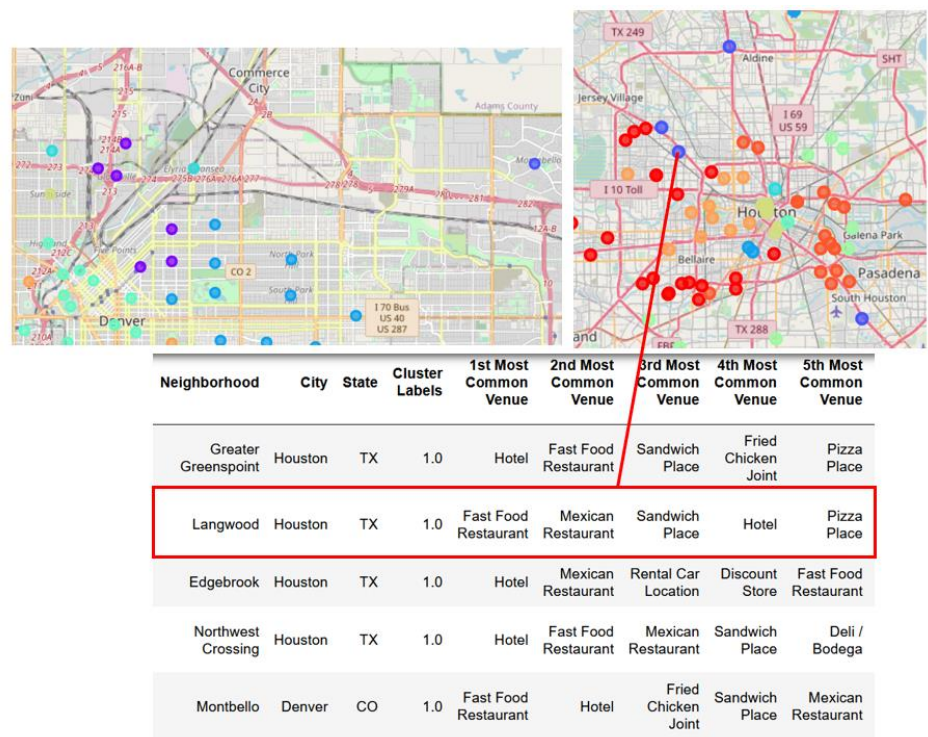


Fig. 11: Top left – Denver map with Montebello shown out to the right. Top right – Houston map with The Outskirt shown to the northwest and southeast of the city center. Bottom – top venues for the outskirts cluster.

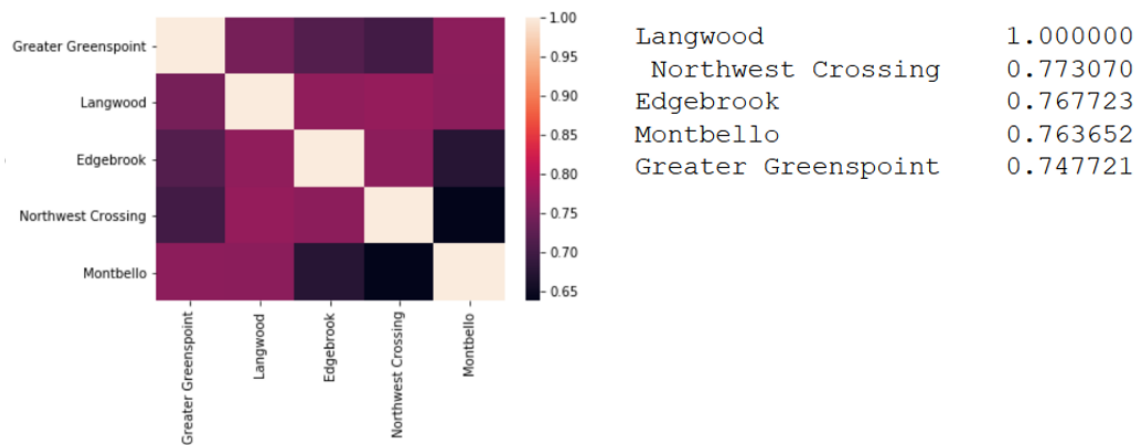


Fig. 12: Left – correlation map for all neighborhoods. Right – Sorted correlations to Langwood in Houston.





Fig. 13: Niche.com grades for the Langwood neighborhood.

### Functional City –

As discussed in the methodology section, airports from each city were clustered in this group but do not correlate well with most of the neighborhoods. For mapping purposes, I have left out the airports so we can see a more interesting view of these neighborhoods. They are both around the city zoo, major transit stations and parks. Niche.com rates City Park and it's closest Houston neighborhood Memorial as two of the best neighborhoods to live in for their respective cities.

Another noteworthy observation is that East Colfax is in this cluster but it has a very poor correlation to City Park, similar to the airports. The crime rate is higher in East Colfax and it is locally known as a seedy area, so it is good that our correlation matrix was able to recognize the poor relationship with some of the more desirable neighborhoods.

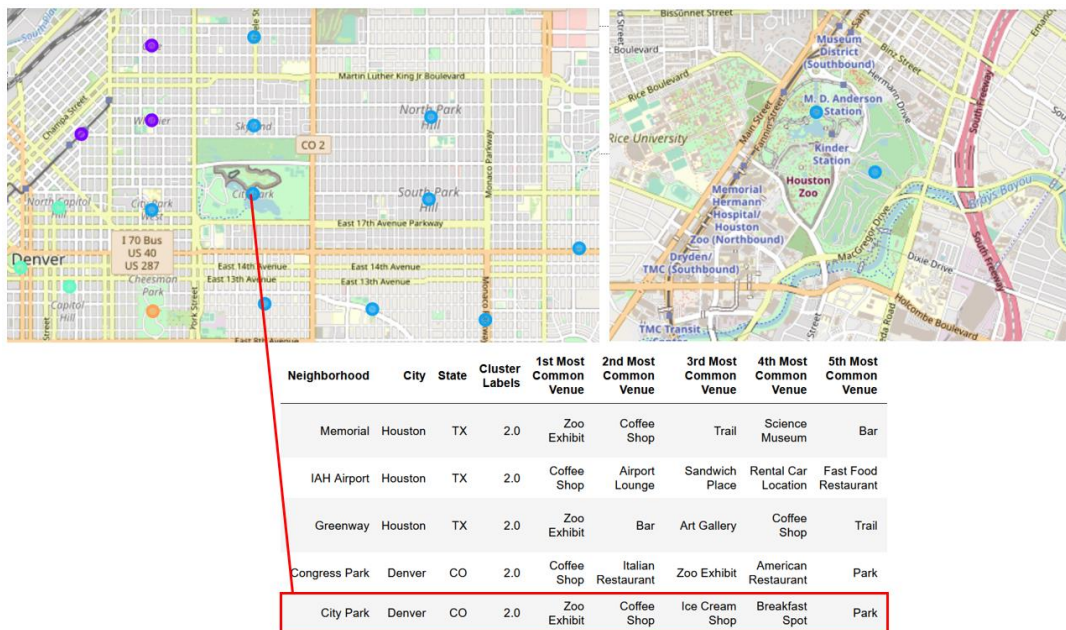


Fig. 14: Top left – Denver map with City Park and surrounding neighborhoods designated by light blue markers. Top right – Houston map zoomed in over the city zoo. Bottom – sampling of top venues for the Functional City cluster.

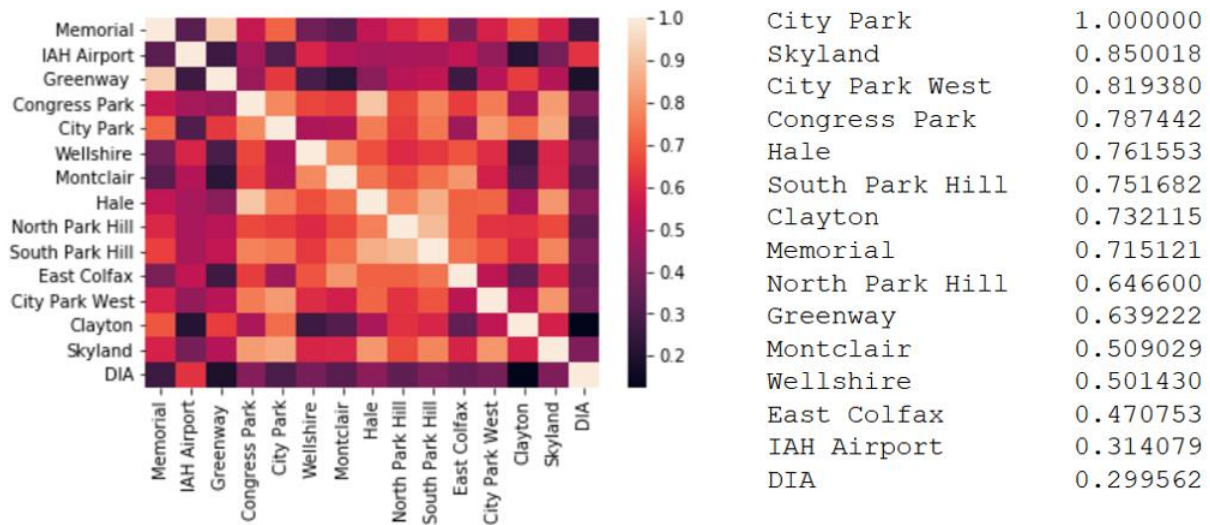


Fig. 15: Left – Correlation map for Functional City. Right – correlations organized by City Park.



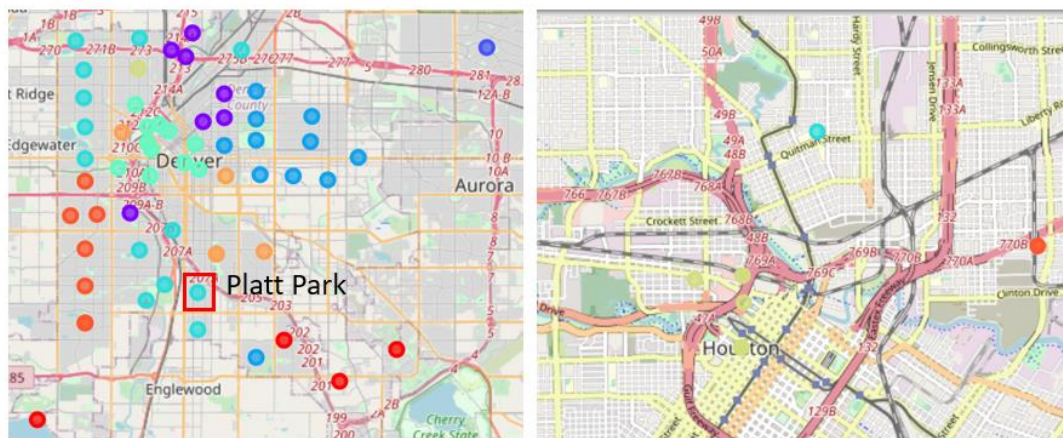
Fig. 16: Niche score for City Park.

## Dense Suburbia –

Next, we have Dense Suburbia, areas just outside the main city hub that tend to have more suburban-style venues but still have a vibrant city atmosphere. This cluster has one of our favorite neighborhoods in Denver, Platt Park. We lived in Rosedale just south of Platt Park for six years and contemplated buying a house in the area. The only Houston neighborhood that clustered in this group was Northside. Since Platt Park is one of my favorite locations on the planet, I have included a little more information on this cluster.

By looking at the correlation map, Northside does not correlate well with the other neighborhoods. In fact, when ordering by Platt Park correlation, Northside is last on the list. To investigate further I have ordered by Northside in Figure (). Chaffee Park and Ruby Hill correlate the strongest with Northside. I am not familiar with that side of town so I pulled up the Niche.com grades for Ruby Hill and Northside shown in Figure(). The Niche.com grades are significantly lower than Platt Park, and Northside is much

lower than Ruby Hill. Unfortunately, it looks like there are not very comparable neighborhoods to Platt Park in Houston.



Neighborhood	City	State	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Northside	Houston	TX	3.0	Mexican Restaurant	Cocktail Bar	Coffee Shop	Bar	Park
West Colfax	Denver	CO	3.0	Coffee Shop	Mexican Restaurant	Brewery	Pizza Place	Park
West Highland	Denver	CO	3.0	Coffee Shop	Mexican Restaurant	Pizza Place	Brewery	Breakfast Spot
Sloan Lake	Denver	CO	3.0	Coffee Shop	Brewery	Mexican Restaurant	Pizza Place	Italian Restaurant
Berkeley	Denver	CO	3.0	Pizza Place	Brewery	Coffee Shop	Breakfast Spot	Park

Fig. 17: Top left, Denver map. Top right, Houston map. Bottom – head of dataframe for Dense Suburbia neighborhoods.

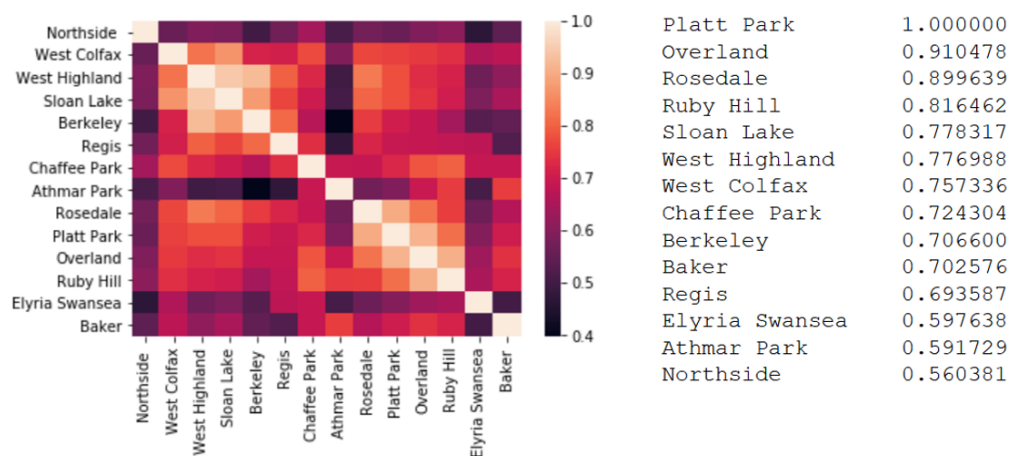


Fig. 18: Left – correlation map for Dense Suburbia. Right – correlations organized by Platt Park.



Fig. 18: Platt Park grades from Niche.com

Northside	1.000000
Chaffee Park	0.640464
Ruby Hill	0.611136
West Highland	0.592292
Overland	0.590408
Sloan Lake	0.583638
Rosedale	0.575434
Regis	0.571966
Platt Park	0.560381
West Colfax	0.557235
Baker	0.543951
Athmar Park	0.513600
Berkeley	0.502775
Elyria Swansea	0.467771

Fig. 19: Correlation scores organized by Northside.

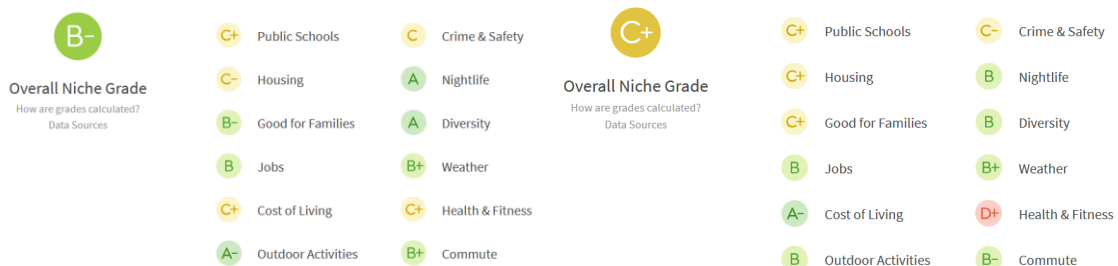


Fig. 20: Left – Ruby Hill Niche.com scores. Right – Northside Niche.com scores.

**Discussion:**

The ability of a K-Means clustering algorithm used on venue frequency to group similar neighborhoods in Denver and Houston worked surprisingly well. The algorithm was able to indicate that there are large portions of Houston with neighborhood vibes that are not found in Denver, predominantly areas with a large amount of fast food restaurants and discount stores.

Overall someone moving from Denver to Houston would likely be attracted to the Urban-suburban Desirables and Sophisticated Living areas and that is well in line with my personal experience searching for a place in Houston. We ended up renting a place in the Heights, and the feel is similar to the areas indicated by the clustering algorithm.

One thing that would probably improve these results would be to do some unsupervised clustering as well, some of the clusters contain clearly distinguished groups such as the airports in the Functional City cluster. But overall the results are well in line with my experiences in both cities.

**Conclusion:**

A K-Means clustering algorithm was used to cluster similar neighborhoods in Denver and Houston. The results were compared with Niche.com grades for the neighborhoods and cross-referenced with the author's own personal experience searching for places to live in Houston. The clustering algorithm performed well, and located the eventual area that the author rented a house in.

The results could possibly be improved by including some unsupervised clustering to see if there are clear outliers in some of the determined clusters. Overall the performance was very good and would have saved the author time and heartache when trying to move from Denver to Houston.