Coursera Capstone Project for IBM Data Science Professional Qualification

By Meriel O'Conor, November 2019

Question: If you move from LA to Nashville, which neighborhood might you want to live in?

Introduction/Business Problem

I live in Nashville and there are lots of people relocating from LA to Nashville. Some common themes for moving seem to be for more affordable housing, to escape the traffic and enjoy the vibrant music scene.

When you first arrive it can be hard to work out which area to live in. This project aims to identify the most closely paralleled neighborhoods between LA and Nashville in order to suggest where new people might want to look for accommodation. Using Foursquare data about which businesses and amenities are in each neighborhood I hope to find areas which have a similar mixture. Then adding house price data to see how affordable the equivalent neighborhood is.

The target audience for this piece of research would be a journalist/blogger wanting to give people information to help them make choices. The end user would be someone moving from LA to Nashville.

Data

Given our problem we need data on neighborhoods, businesses in each neighborhood and house price data.

The data sources for this project were:

- Foursquare data, for businesses and their type and location https://foursquare.com/, accessed 11/15/19
- Zillow data on house prices, to gauge affluence of the neighborhoods https://www.zillow.com/research/data/, accessed 11/15/19
- The names of neighborhoods in LA were scraped from http://www.laalmanac.com/communications/cm02a90001-90899.php , accessed 11/15/19
- The names of Nashville neighborhoods were extracted from https://nestinginnashville.com/buying-a-home-in-nashville/zip-code-map/ , accessed 11/15/19

The names of neighborhoods in LA and Nashville are just tables where each zip code is paired with a recognizable name for the neighborhood. There are several cleaning stages to undertake before arriving at the final dataframe, visible in the notebook. Here is the head of the LA data frame:

	PostalCode	Neighbourhood
0	90001	Los Angeles (South Los Angeles), Florence-Graham
1	90002	Los Angeles (Southeast Los Angeles, Watts)
2	90003	Los Angeles (South Los Angeles, Southeast Los ...
3	90004	Los Angeles (Hancock Park, Rampart Village, Vi...
4	90005	Los Angeles (Hancock Park, Koreatown, Wilshire...

For the Foursquare data we need the neighborhood to be matched with venues in the given area, with their location, name and category. To achieve that we need to have an account to access the API, add in our client name, secret and version. Then we need latitudes and longitudes of the neighborhoods, which we get using pgeocode. Then a function is created to loop through each neighborhood to extract the venues and this is mapped to a dataframe.

Here is the head of the LA table of venues:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Los Angeles (South Los Angeles), Florence-Graham	33.9731	-118.2479	Bill's Drive In	33.974500	-118.244225	Burger Joint
1	Los Angeles (South Los Angeles), Florence-Graham	33.9731	-118.2479	Mi Lindo Nayarit Mariscos	33.974523	-118.256784	Mexican Restaurant
2	Los Angeles (South Los Angeles), Florence-Graham	33.9731	-118.2479	Avila's El Ranchito	33.978609	-118.230469	Mexican Restaurant
3	Los Angeles (South Los Angeles), Florence-Graham	33.9731	-118.2479	Tom's Jr.	33.989227	-118.247519	Burger Joint
4	Los Angeles (South Los Angeles), Florence-Graham	33.9731	-118.2479	Northgate Gonzalez Markets	33.988665	-118.258117	Grocery Store

For the Zillow data we need to match the neighborhoods to the median house price, Zillow calls it Zhvi. Here is the head of a table the table where neighborhoods are matched to house price:

	Neighbourhood	Zhvi	state_code	PostalCode
0	Beverly Hills	4777200.0	CA	90210
1	Santa Monica	3942400.0	CA	90402
2	Los Angeles (Castellemare, Pacific Highlands, ...	3010200.0	CA	90272
3	Malibu	2950800.0	CA	90265
4	Beverly Hills	2710300.0	CA	90212

Methodology

This analysis was done using Python in Jupyter notebooks. The entire methodology is available in the notebook.

An overview of the major steps of the process is:

1. Data on LA neighborhoods was parsed using BeautifulSoup4 from the aforementioned website
2. Zip codes that were later discovered to have little value, including PO Boxes, where retroactively removed
3. Latitudes and Longitudes of each neighborhood were imported using pgeocode
4. LA neighborhoods were mapped using Folium
5. LA venues were extracted using the Foursquare API
6. The venue information was one hot encoded
7. A dendrogram was produced by creating a linkage matrix using scipy's cluster package
8. The Nashville neighborhood data was extracted from the above website
9. This data was edited in excel and reuploaded as an array which was then converted to a data frame
10. Steps 3-7 were then repeated for the Nashville data
11. The one hot encoded tables for LA and Nashville were joined
12. A combined dendrogram for LA and Nashville was created
13. House price data was extracted from Zillow and saved as a CSV
14. The CSV was uploaded and combined with the geographical data
15. Median house price by neighborhood was plotted

Exploratory Data Analysis

In order to understand the data before drawing any conclusions it was important to map it, to check all the information extracted from websites was accurate. Then it was also necessary to inspect some of the venues to check the Foursquare data was reasonable. I did this by looking at my neighborhood (Sylvan Park / Sylvan Heights / The Nations / Charlotte Park), and confirming they are indeed familiar. Once a combined dendrogram was drawn and house price data incorporated different neighborhood examples were looked at individually and further bar charts and heatmaps were made to explore the findings.

Inferential statistical tests

As this project does not rely on testing if samples are representative of a population – as whole populations were used, no inferential statistics were relevant.

Machine Learning

The Machine Learning aspect of this project is the use of agglomerative hierarchical clustering which is represented as dendrograms. The algorithm merges the closest objects to create clusters, it iterates through adding the next closest object to each cluster, creating a linkage matrix which can be displayed as a dendrogram.

There are different methods for determining the distance between clusters, including: single-linkage, complete-linkage, average-linkage, centroid-linkage etc… In this project the Ward method was used, which minimizes the within-group sum of squares. At the beginning of hierarchical clustering each object is its own cluster and therefore has a sum of squares of zero. When each merge is made the sum of squares increases but the Ward method attempts to minimize it.

In this project the algorithm has no knowledge of the neighborhoods, where they are geographically or the similarity of their names, all it looks at is whether the combination of venues is similar.

The number of clusters is not specified and for the purposes of this analysis it is not helpful or sensible to attempt to draw a line at a particular distance to create clusters as it may falsely separate neighborhoods that aren't that far away from each other. It is better viewed as a continuum with the colors only being used for ease of reading and reference, not as a suggestion of discrete groups.

Results

Firstly, the neighborhoods in both LA and Nashville were plotted on maps using Folium.

Fig.1: Map of LA neighborhoods

Each purple dot represents a neighborhood
175 neighborhoods in total

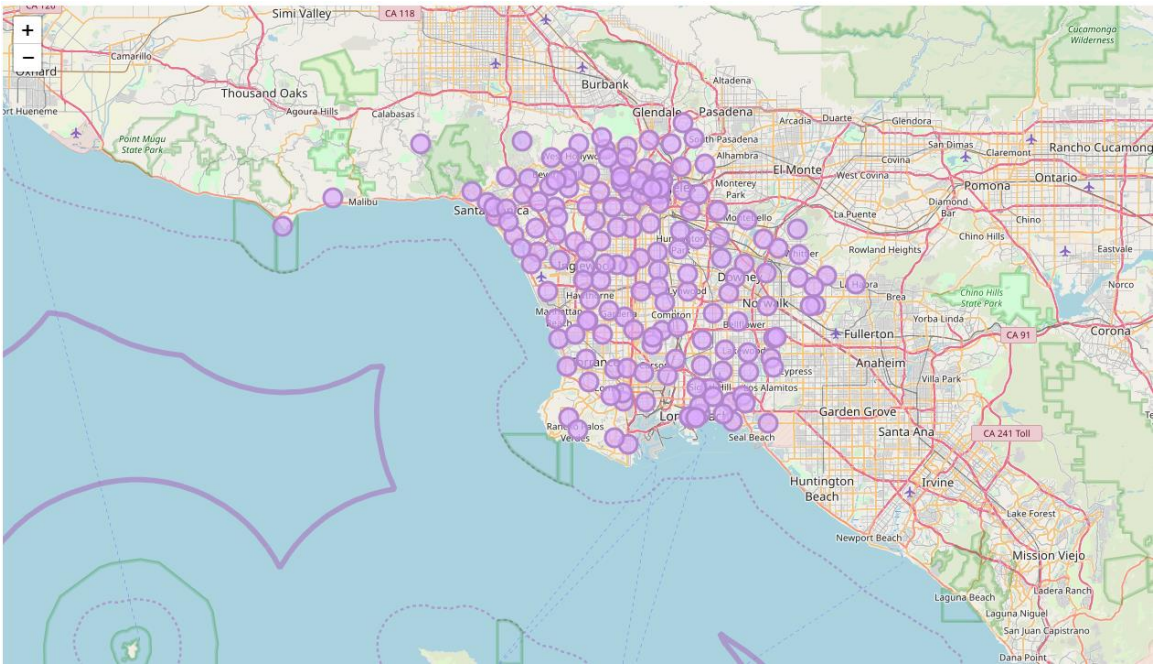
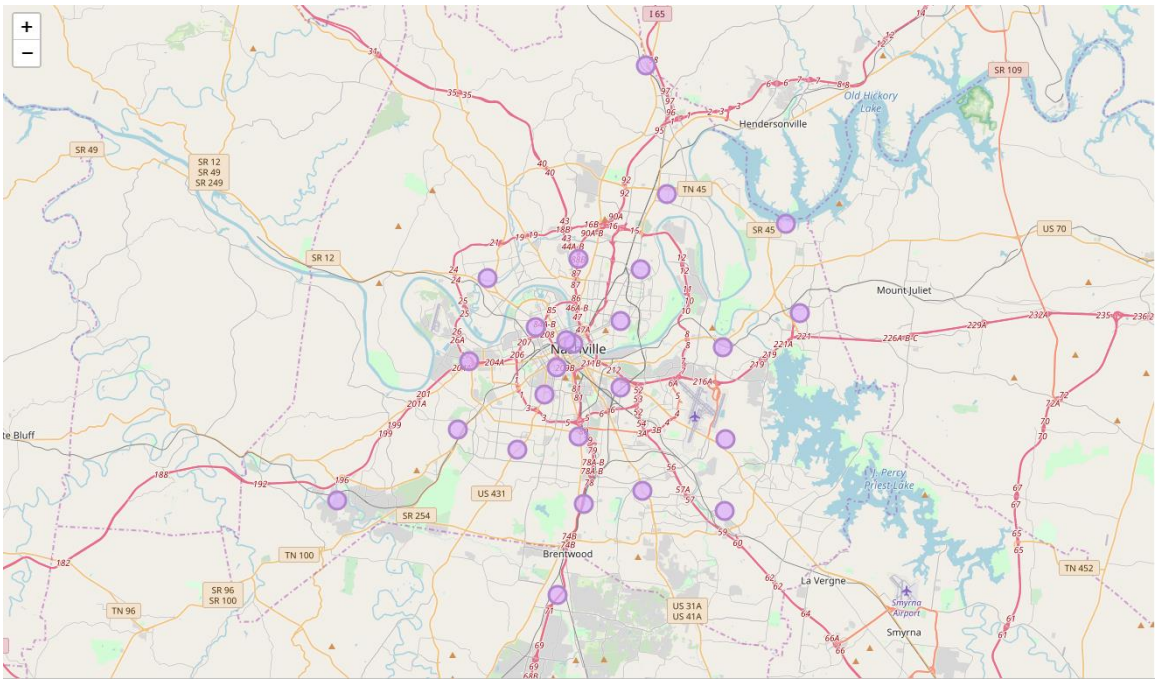


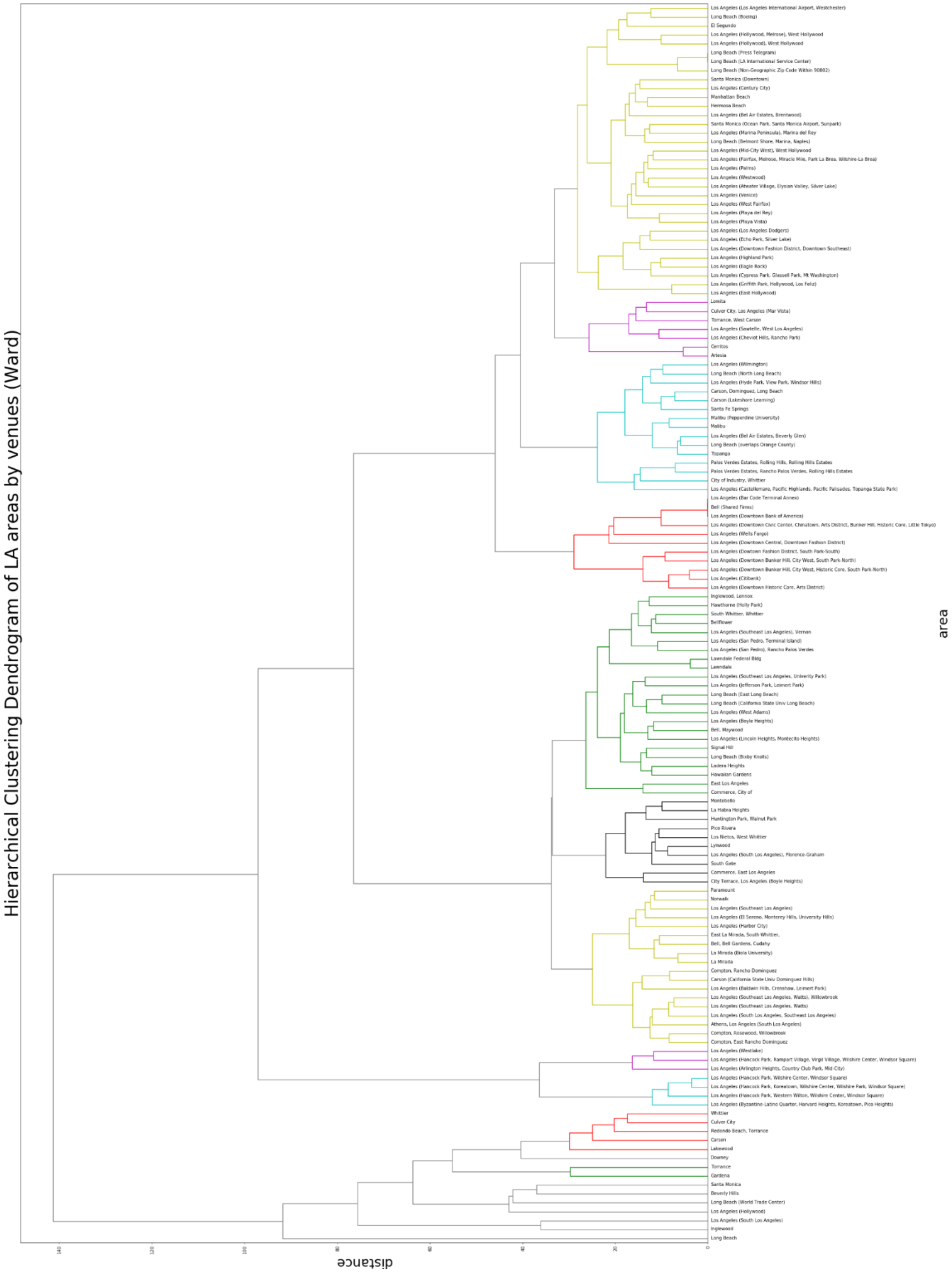
Fig.2: Map of Nashville neighborhoods

Each purple dot represents a neighborhood
24 neighborhoods in total



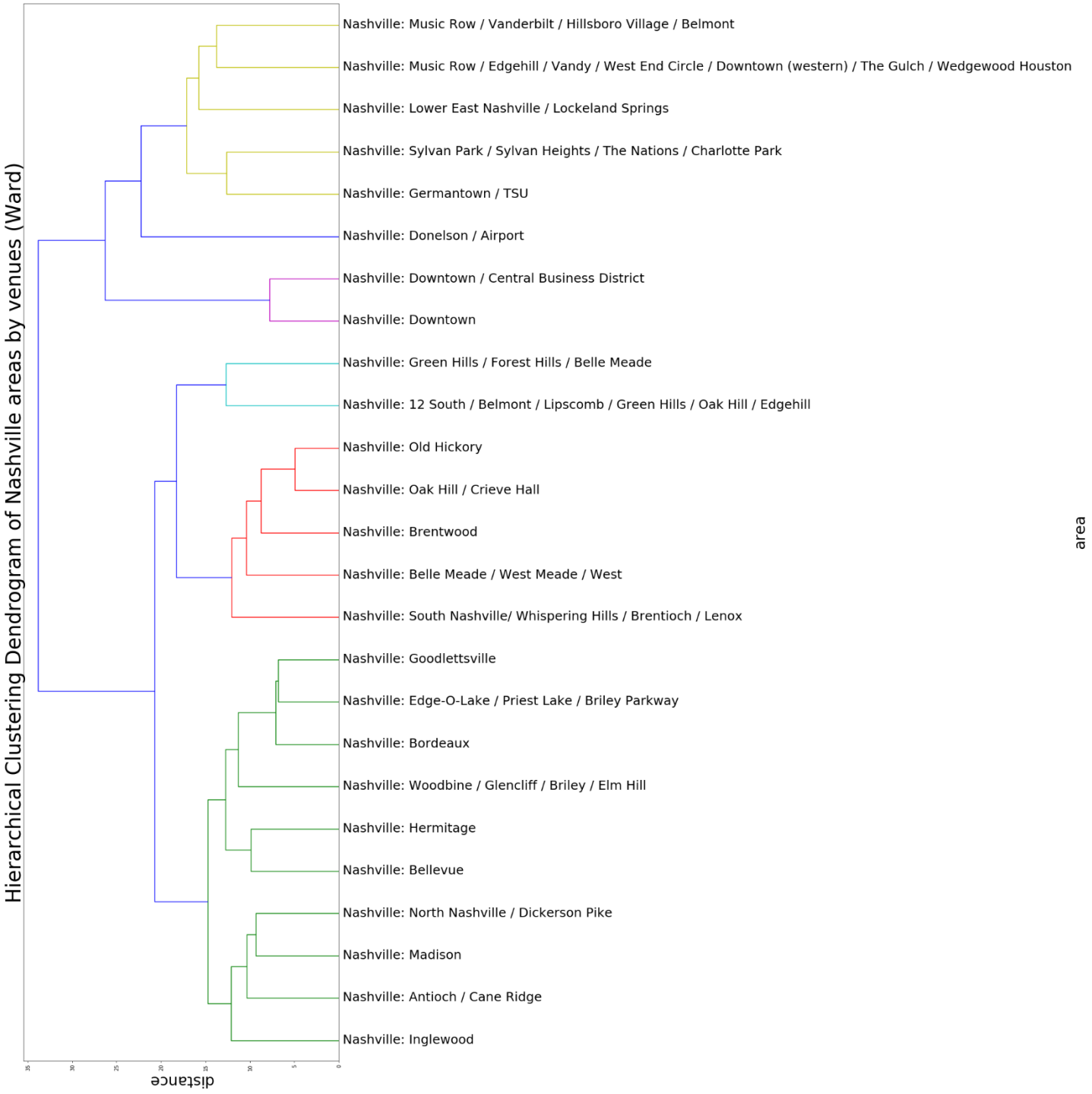
These maps show that there are far fewer neighborhoods in Nashville than LA.

Fig.3: Dendrogram to show similarity of LA neighborhoods



There is clearly a lot of information in the dendrogram. Some notable points include there is a distant cluster of neighborhoods which includes Beverly Hills and Santa Monica. Neighborhoods with similar names seem to cluster together, despite the geographical information not being available to the algorithm; only venue information is supplied.

Fig.4: Dendrogram to show similarity of Nashville neighborhoods



In Fig 5 the Nashville neighborhoods are placed in orange boxes.

Fig.5: Dendrogram to show similarity of LA and Nashville neighborhoods

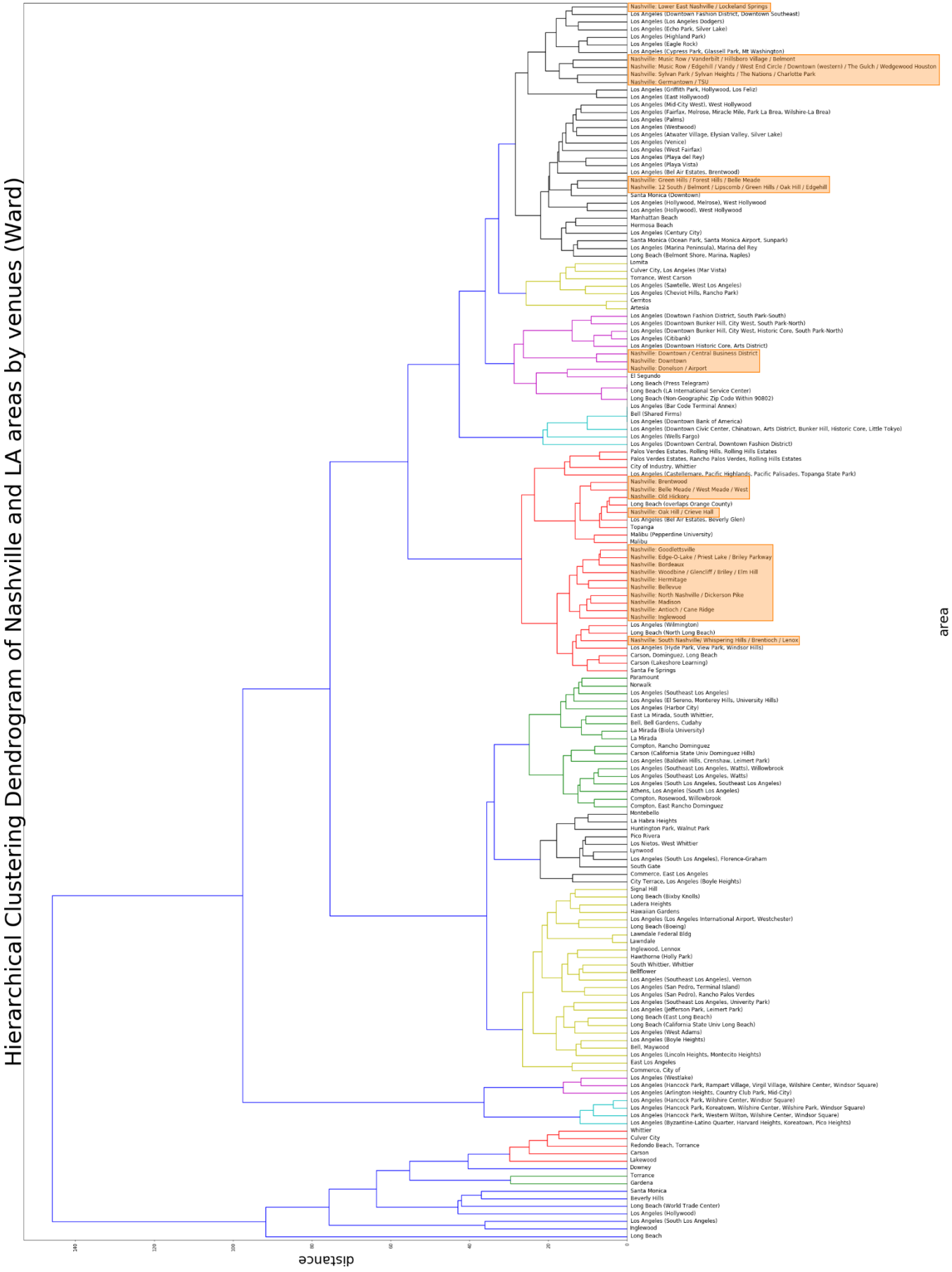
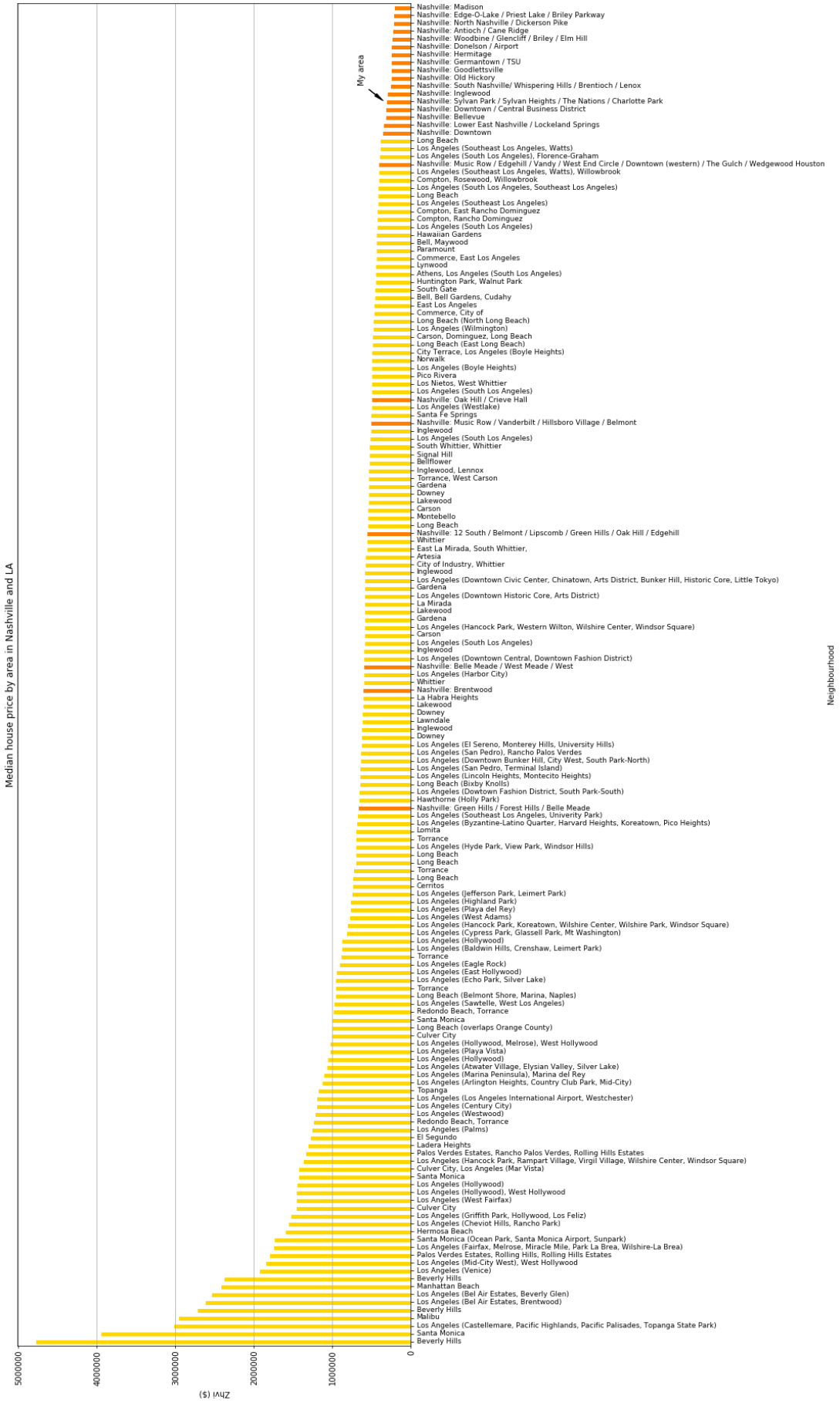


Fig.6: Bar chart of Median house prices by neighborhood



Nashville properties are in TN orange while LA is in yellow. Most of the Nashville properties are at the lower end of the spectrum with six neighborhoods interspersed within the LA price zone.

Fig.7: Bar chart of LA neighborhoods most similar to Green Hills – **mean applies to LA only*

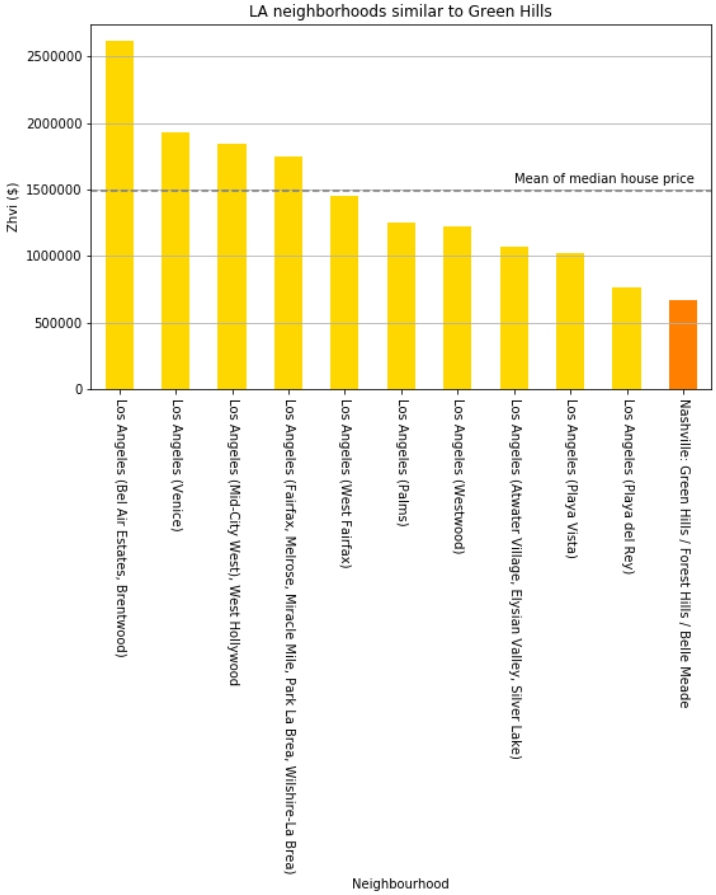


Fig. 8: Bar chart of LA neighborhoods most similar to Sylvan Park / Sylvan Heights / The Nations / Charlotte Park – **mean applies to LA only*

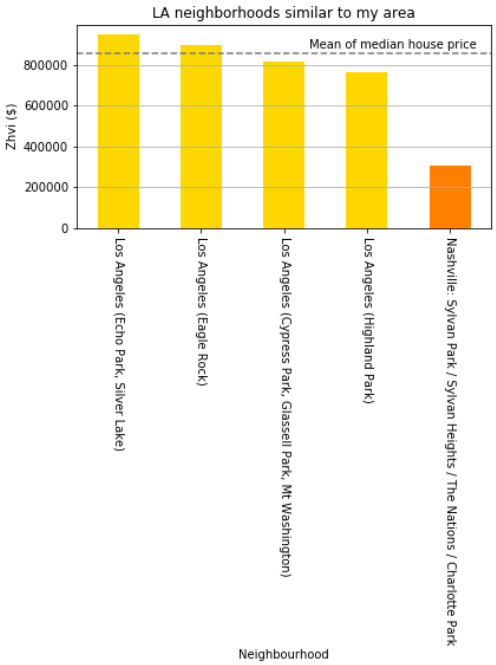


Fig.9: Table of areas where the most common venue is a hotel

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
Beverly Hills	Hotel	Coffee Shop	American Restaurant	Boutique	Italian Restaurant	Park	Sushi Restaurant	Steakhouse	Cafe	Clothing Store
El Segundo	Hotel	Coffee Shop	Sandwich Place	Pizza Place	Airport Lounge	Airport Service	American Restaurant	Mexican Restaurant	Burger Joint	Beach
Los Angeles (Dowtown Fashion District, South Park-South)	Hotel	Coffee Shop	Bar	New American Restaurant	American Restaurant	Sushi Restaurant	Taco Place	Theater	Beer Bar	Mexican Restaurant
Los Angeles (Hollywood)	Hotel	Lounge	Pizza Place	Coffee Shop	Bar	Nightclub	Burger Joint	Gym	Mexican Restaurant	Music Venue
Los Angeles (Hollywood), West Hollywood	Hotel	Gym	Burger Joint	Vegetarian / Vegan Restaurant	Coffee Shop	Trail	Japanese Restaurant	Clothing Store	Comedy Club	Hotel Bar
Los Angeles (Hollywood, Melrose), West Hollywood	Hotel	Gym	Mexican Restaurant	Clothing Store	New American Restaurant	Burger Joint	French Restaurant	Cafe	Vegetarian / Vegan Restaurant	Sushi Restaurant
Nashville: Donelson / Airport	Hotel	Sandwich Place	Fast Food Restaurant	Pizza Place	Gas Station	American Restaurant	Mexican Restaurant	Automotive Shop	Pharmacy	Discount Store
Nashville: Music Row / Edgehill / Vandy / West End Circle / Downtown (western) / The Gulch / Wedgewood Houston	Hotel	Pizza Place	Music Venue	American Restaurant	Mexican Restaurant	Coffee Shop	Burger Joint	Taco Place	Sushi Restaurant	Bar

Fig.10 Heatmap of neighborhoods similar to South Nashville

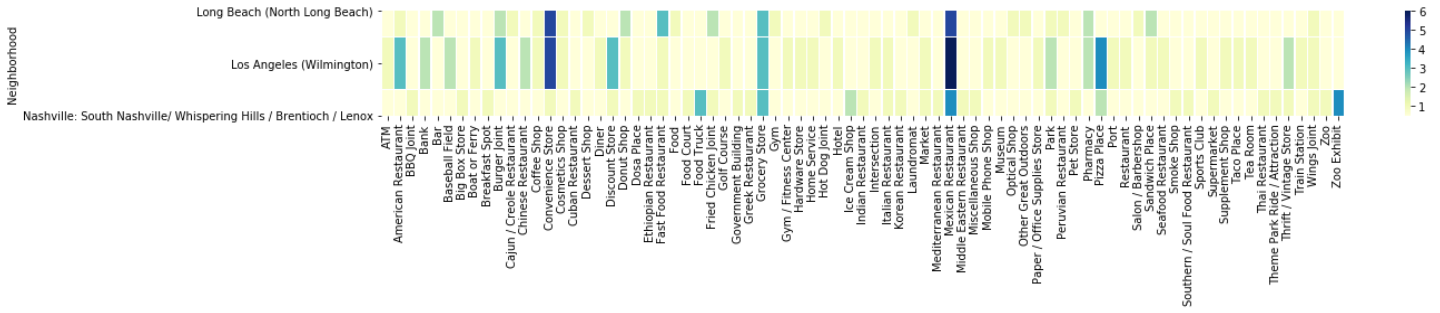


Fig.11 Table of Nashville neighborhoods with at least three music stores/venues

Neighborhood	Venue	Venue Category
Downtown	Live On The Green Music Festival	Music Venue
Downtown	Bourbon Street Blues and Boogie Bar	Music Venue
Downtown	Ryman Auditorium	Music Venue
Downtown	Redneck Riviera	Music Venue
Downtown	The Stage on Broadway	Music Venue
Downtown	B.B. King's Blues Club	Music Venue
Downtown	Ascend Amphitheater	Music Venue
Music Row / Edgehill / Vandy / West End Circle / Downtown (western) / The Gulch / Wedgewood Houston	Music Row	Music Venue
Music Row / Edgehill / Vandy / West End Circle / Downtown (western) / The Gulch / Wedgewood Houston	The Station Inn	Music Venue
Music Row / Edgehill / Vandy / West End Circle / Downtown (western) / The Gulch / Wedgewood Houston	The High Watt	Music Venue
Music Row / Edgehill / Vandy / West End Circle / Downtown (western) / The Gulch / Wedgewood Houston	Exit/In	Music Venue
Music Row / Edgehill / Vandy / West End Circle / Downtown (western) / The Gulch / Wedgewood Houston	The Cannery Ballroom	Music Venue
Music Row / Edgehill / Vandy / West End Circle / Downtown (western) / The Gulch / Wedgewood Houston	Carter Vintage Guitars	Music Store
Music Row / Edgehill / Vandy / West End Circle / Downtown (western) / The Gulch / Wedgewood Houston	Ryman Auditorium	Music Venue
12 South / Belmont / Lipscomb / Green Hills / Oak Hill / Edgehill	Blackbird Studio	Music Venue
12 South / Belmont / Lipscomb / Green Hills / Oak Hill / Edgehill	Guitar Center	Music Store
12 South / Belmont / Lipscomb / Green Hills / Oak Hill / Edgehill	Corner Music	Music Store
12 South / Belmont / Lipscomb / Green Hills / Oak Hill / Edgehill	Fork's Drum Closet	Music Store
Music Row / Vanderbilt / Hillsboro Village / Belmont	Music Row	Music Venue
Music Row / Vanderbilt / Hillsboro Village / Belmont	Musicians Corner @ Centennial Park	Music Venue
Music Row / Vanderbilt / Hillsboro Village / Belmont	Exit/In	Music Venue
Downtown / Central Business District	Live On The Green Music Festival	Music Venue
Downtown / Central Business District	Bourbon Street Blues and Boogie Bar	Music Venue
Downtown / Central Business District	Ryman Auditorium	Music Venue
Downtown / Central Business District	The Stage on Broadway	Music Venue
Downtown / Central Business District	Redneck Riviera	Music Venue
Downtown / Central Business District	B.B. King's Blues Club	Music Venue

Discussion

We can't examine every neighborhood pair so taking a few examples: Green Hills, my area, Beverly Hills, El Segundo and South Nashville.

Green hills is the most expensive area of Nashville (median house price $666,400) (see Fig 6). Its most similar neighborhood in LA is downtown Santa Monica, but we don't have house price data for that zip code. The next nearest places are Bel Air, Playa Vista, Playa del Ray, West Fairfax, Venice, Atwater Village, Westwood, Palms, Fairfax, West Hollywood (respective zip codes: 90049, 90094, 90293, 90035, 90291, 90039, 90024, 90034, 90036, 90048). The mean of their median house prices is $1,491,200 (see Fig 7). So if you moved from any of those areas to Green Hills you'd be getting a similar range of amenities for less than half the price!

My area is Sylvan Park / Sylvan Heights / The Nations / Charlotte Park (37209) with a median house price of $308,000. It is most like six LA areas: Echo Park, LA dodgers, Downtown Fashion District, Cypress Park, Eagle Rock, Highland Park (90026, 90090, 90021, 90065, 90041, 90042). We have Zillow data for four of them (not 90090 and 90021). The mean of their median house prices is $856,775, so 2.8 times more expensive (see Fig 8).

The dendrogram shows the most expensive part of LA – Beverly Hills has no real equivalent in Nashville. With median house prices there of $4.7Mil they are in a league of their own. Beverly Hills' most common amenity are hotels, followed by coffee shops, American restaurants and boutiques (see Fig 9). In Nashville the airport area has hotels as its most common venue but it is followed by sandwich places, fast food and pizza places – clearly a different ambience. Music row also has hotels as its number one venue but music venues dominate. You'd therefore have to accept if you came from Beverly Hills you'd be living in a pretty different neighborhood.

El Segundo is closely related to Donelson/Aiport on the dendrogram. We can see they are both airport areas with similar facilities, so a natural pair.

South Nashville is deemed to be close to Wilmington and North Long Beach (90744, 90805). This seems to be based on them all having lots of Mexican restaurants (see Fig 10).

Lots of people from LA are coming to escape the traffic. One neighborhood that is between some purportedly bad highways is 'Carson, Dominguez, Long Beach' (90810) which is not too distantly related to South Nashville (see Fig 5). So people moving from there might want to consider that area. Nearly every neighborhood in LA has traffic issues so it may not be relevant to highlight many particular examples.

There are 6 postcodes within the LA price range - 37215, 37027, 37205, 37204, 37212, 37220. On the dendrogram Brentwood, Belle Meade and Oak Hill (37027, 37205, 37220) are most like Bel Air, Topanga and Long Beach (90077, 90290, 90740). 12th South and Green Hills (37204, 37215) are closely related and their closest neighborhoods are mentioned above. Music Row (37212) is closely related to my area – 37209 and discussed above.

It is important to note that when interpreting the dendrograms they are most accurate at showing areas that are very similar; neighborhoods appearing far away from each other may just be closer to their cluster rather than very far from a neighboring cluster. It is also worth noting that as venues change in the two cities it can dramatically alter the dendrogram.

Not being familiar myself with LA makes it harder to assess the accuracy of its data but given how the Nashville data looks correct I'm happy to assume the LA data is also representative (see Fig 3 and 4).

Conclusion

It is of course hard to parallel neighborhoods across cities to find their twin. And the sheer ratio of LA to Nashville neighborhoods (7:1) means there is more diversity in LA which is unmatched in Nashville (see Fig 1 and 2). Nevertheless, there are some neighborhoods that have a close degree of affinity, such as LA and Nashville's downtowns, Green Hills and downtown Santa Monica, Lower East Nashville and the 90021 area.

If people are coming for the music scene alone they might like Nashville's Downtown areas, Music Row and 12 south, as those have the most music infrastructure (see Fig 11). These people may be less interested in finding a similar neighborhood to back home.

Most areas of Nashville are considerably more affordable, so if the primary objective is cost there are many options. Nashville also has considerably less traffic so again there are many options for the relocator.

Anyone looking to find the most similar neighborhood between LA and Nashville can use the dendrogram in Fig 5 to try and find their nearest neighborhood and the bar chart in Fig 6 to check the affordability.

Bonus visuals:

LA venue categories word cloud

Nashville venue categories word cloud

LA is notably more dominated by Mexican Restaurants and Coffee Shops. It also has much more presence of Ice Cream. Nashville has a lot of Pizza Places and suspiciously broad ‘Food Restaurants’.