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