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.