Best Houston Super Neighborhoods for People New to Area Based on Their Preferences

IBM Capstone Project

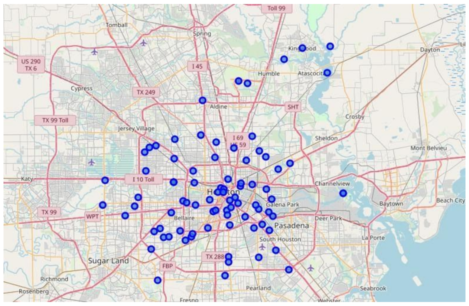
1 Problem Introduction

Moving by itself is stressful, but even more so when moving to a city about which you know very little. Everyone has their preferences of what they would like nearby to where they live, and while some people prefer parks and greenery, others would rather have restaurants and bustling nightlife. My models will first use Foursquare data to cluster the 88 'Super Neighborhoods' of Houston, Texas based on the quantity of over 200 types of venues and then be able to predict which neighborhoods new Houston residents should look at to live based on their preferences and the cluster into which they fall.

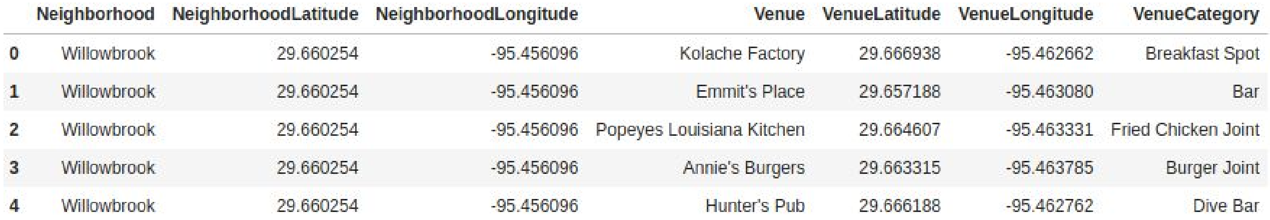
In theory, this program will help people looking to move to Houston focus on specific neighborhoods to live instead of being faced with the daunting task of looking at the entirety of the city.

2 Dataset

To complete this task I will be scraping a list of the 'super neighborhoods' from wikipedia and then using the geopy library on Python to get the latitudes and longitudes of each neighborhood. Nominatim did not have data for every neighborhood, so I googled the remaining neighborhoods and manually placed them into my dataframe. This data resulted in the map of neighborhoods to the right.

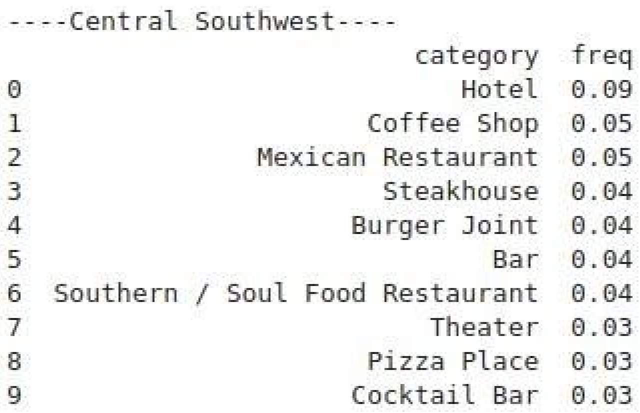


Using the latitudes and longitudes I then utilized the Foursquare API to get a list of venues for each neighborhood. I set the radius for each neighborhood to 1 kilometer since Houston is a pretty spread out city and I think a 1 km walking radius is a decent maximum. I limited the number of venues for each neighborhood to 100 because I didn't want a few neighborhoods with a lot of venues to overpower those which had fewer. This gave me a dataframe of over 2,600 venues which looked like this:



	Neighborhood	NeighborhoodLatitude	NeighborhoodLongitude	Venue	VenueLatitude	VenueLongitude	VenueCategory
0	Willowbrook	29.660254	-95.456096	Kolache Factory	29.666938	-95.462662	Breakfast Spot
1	Willowbrook	29.660254	-95.456096	Emmit's Place	29.657188	-95.463080	Bar
2	Willowbrook	29.660254	-95.456096	Popeyes Louisiana Kitchen	29.664607	-95.463331	Fried Chicken Joint
3	Willowbrook	29.660254	-95.456096	Annie's Burgers	29.663315	-95.463785	Burger Joint
4	Willowbrook	29.660254	-95.456096	Hunter's Pub	29.666188	-95.462762	Dive Bar

From this dataframe I was able to count the venues in each neighborhood, organize them by category, and figure out the most popular type of venues in each individual neighborhood and their frequency as a percentage of the entire amount of venues in the neighborhood (represented as a decimal). An example of the top 10 types of venue in the ‘Super Neighborhood’ Central Southwest is to the right. This is the data I used for modeling.



```
----Central Southwest----
                          category  freq
0                            Hotel  0.09
1                      Coffee Shop  0.05
2               Mexican Restaurant  0.05
3                       Steakhouse  0.04
4                     Burger Joint  0.04
5                              Bar  0.04
6  Southern / Soul Food Restaurant  0.04
7                          Theater  0.03
8                      Pizza Place  0.03
9                      Cocktail Bar  0.03
```