2020

A Guide to choose your Favorite Neighborhood in Houston (Texas)!

CAPSTONE PROJECT

SHOHREH AMINI

1. Introduction

Many people relocate due to their jobs every year. Moving from one place to another is very stressful specifically when you do not have enough information regarding the new place you are going to live in.

When the destination is a large metropolitan city, it has many different neighborhoods to consider as an option to live in. Finding the right location to live is not an easy task since there are many parameters which affect the decision to choose the right neighborhood.

Accessing a tool that allows you to compare different neighborhoods in a large city, can provide valuable insights regarding each area and ultimately helps you in your decision making process.

In this work, the neighborhoods in Houston (Texas) are analyzed. The Greater Houston metropolitan area is the fifth most populous metropolitan area in the US. The diverse industry and affordable housing make this city more attractive to the people who would like to relocate to a place which offers the amenities of a large city with lower cost.

2. Available Data

To complete this work we require several types of data, which must be extracted from different sources.

1. In order to analyze different neighborhoods and suburbs of Houston, we first need to find the geographical data related to Houston suburbs. This data was extracted from the "US zip code latitude and longitude" web page (https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/table/). The location data was extracted for Texas (the filter is only available at state level) and stored in a csv file.
2. List of 26 suburb cities was manually extracted from various pages and, stored in a csv file.
3. Foursquare location data will be used to find various information regarding each suburb of Houston. Then this information will be used to analyze each suburb and make recommendations regarding the characteristic of the neighborhoods.

3. Methodology

3.1 Data Integration

The data from different sources should be integrated in order to have a dataset which includes all the required data for further analysis. In this work, the first set of data includes all Texas zip codes, with their corresponding city and the latitude and longitude. Another set of data includes 26 of the most important Houston suburbs (cities).

Based on the objective of this research, we subset the first dataset for the cities which are included in the second dataset only, since we would like to focus on analyzing the Houston and its suburbs.

3.2 Data Evaluation and Cleaning

After integration and dropping the unnecessary columns, the resulted data frame has 250 rows (data points) and 4 columns (features), which does not include any NULL values. The number of data for each city is evaluated. The highest number of data comes from the main city (Houston,

with 182 data points). Many of the suburbs only include one representative data point and therefore, decision was made to only consider the suburbs with more than 3 data points. At this point we end up with 224 location data in total, which comes from 7 cities (includes Houston and suburbs).

Using ***FOLIUM*** library, all the data points are depicted on the map (Figure 1).

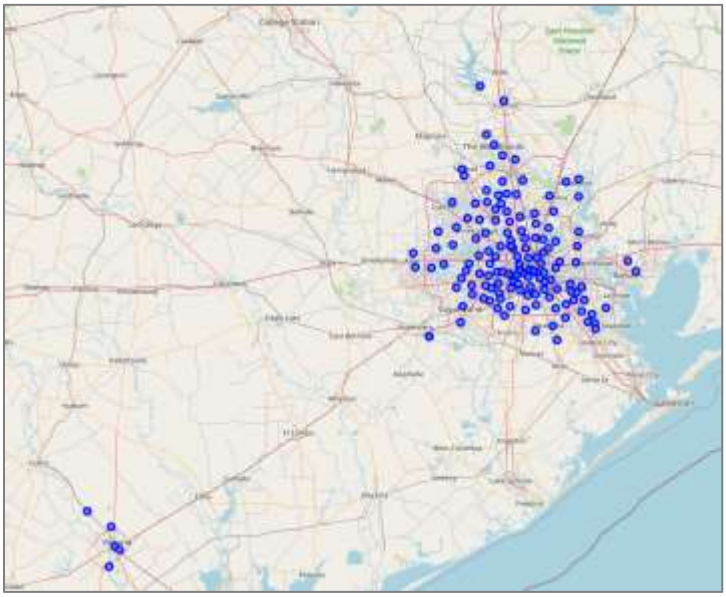

Figure 1. All available data points from Houston and its suburbs

From the map it can be observed that some data points are too faraway from the main city and therefore we would like to remove those from our analysis.
The updated map after removing the *Victoria* suburb, is presented in Figure 2.

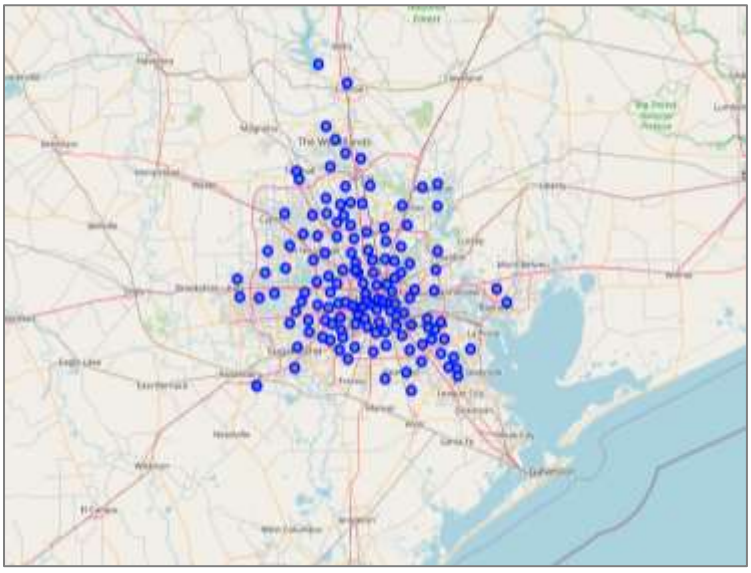

Figure 2. All available data points from Houston and its suburbs Excluding "Victoria"

3.3 Exploring Houston and its Suburbs

First, the *Foursquare* credentials are defined, and then a function is created to extract the venues and category of the venues for all available data points (different zip codes) in different cities.
In the next step, this function is used to create a data frame which includes all the venues which are found at the radius of *1000 meter* from every data point. Other features beside the name of the venue is the latitude, longitude and the venue category.
The resulted data frame includes 7186 data records. Groupby() function is used to find the number of venues for each city. Table below shows the result.

Table 1. Number of Venues resulted from Foursquare for each city

City	Zip_Code
Houston	6290
Humble	252
Katy	164
Pasadena	211
Spring	247
Sugar Land	22

In total *319 unique categories* exist in the entire list of venues. Since the purpose of this study is to generally evaluate the similarities and differences between the cities around Houston, all different types of restaurants and bars are integrated into single categories. For instance, Italian restaurant and French restaurant are both converted to *restaurant*. Consequently, we end up having *254 unique* categories.

In the next step we would like to make the unique venue categories as the input features, while the venue locations (from different zip codes) would be our data points. Then, by taking the mean of the frequency of each venue category, the data will be grouped for each single zip code. This decreases the data points to *217*.
In order to be more focused on the most important venues, we find the *7 most frequent venues* for each zip code and consider them as the feature inputs and based on these features we will evaluate the similarities of different suburbs, through clustering.

K-means is used as an unsupervised clustering method to partition the data into ***6 clusters*** based on similarities of the top venues in the neighborhoods (zip code zones).

4. Result

Since most of the data comes from the city of Houston, we expect that in each cluster we have several data points from this city. The following map (Figure 3) shows the result of the clustering. Different colors represent different clusters. The data point label shows the zip code of the data point as well as the cluster number to which the data belongs to.
The majority of the data falls into either **cluster-0** or **Cluster-1**. The rest of the clusters only include few data points.

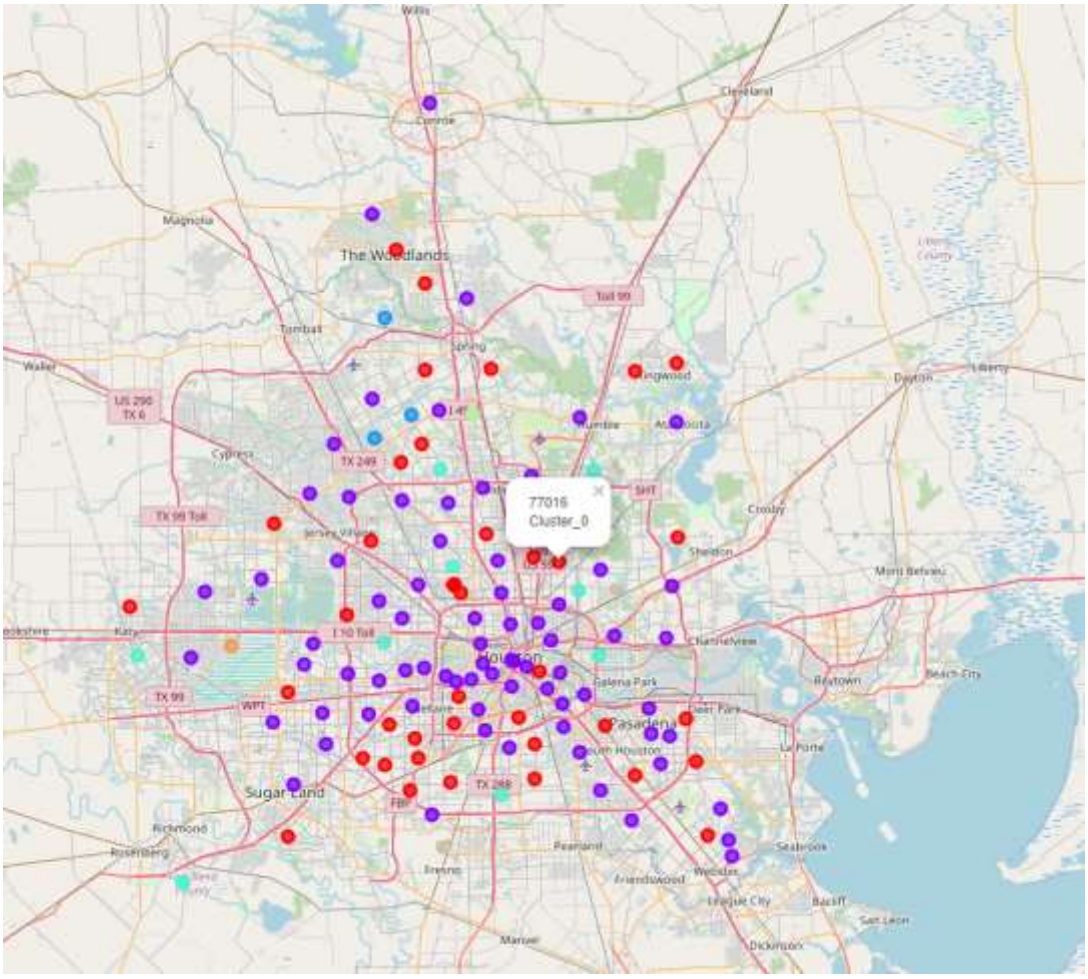

Figure 3. Houston and suburb map with color coded clusters resulted from K-means clustering based on similarities of the venues

In the following section more details regarding different neighborhood is discussed.

5. Discussion

The following observation can be made by studying the characteristics of the data points which are included in each cluster.

Cluster-0: The first cluster, which includes 125 data point is the largest cluster. This cluster is distributed evenly within the Houston and suburb cities. The red dots on the map represents the data points in this cluster. The most frequent categories among all the data records in the first three columns are: *Spa*, *Bar* and *Restaurant* respectively.

Cluster-1: This cluster has 75 data points, which is the second largest cluster among the 6. The data points in this cluster are shown as purple dots on the map. Similar to the first cluster, this one is also evenly distributed within Houston and other cities. The most frequent categories of the first three columns in this cluster are: *Restaurants, Sandwich/pizza place and bar*.

Cluster-2: This cluster includes only 3 data points, which are located towards the North East, and far from the Houston downtown (shown with blue dots on the map). The most dominant categories in this cluster are related to sport and recreation such as Golf course, pool, lake and Yoga studios.

Cluster-3: This cluster is the third largest cluster, with 10 data points and is shown with turquoise dots on the map. The data points are distributed in some zip codes within Houston as well as other suburbs in all direction. The most frequent venue categories for all data set are: *Construction & Landscaping, park, Fried Chicken and burrito place*.

Cluster-4: Only 3 data points fall into this cluster, which is shown with light green color on the map. The data points are located towards the north of Houston. The most frequent categories in this cluster includes: *Home Services, Boat or Ferry, park and Yoga Studio*.

Cluster-5: This cluster includes only one single data point, which is located in the western boundary of Houston and is shown with Orange dot on the map. The most common categories for this zip code are: *Park, Yoga Studio, Disc Golf, Field and Farmers market*.

6. Conclusion

In this study the number of different types of venues in the zip codes within the Houston area, and specifically in the center of the city, is greater compared to the suburbs. However, the distribution of the clusters shows that within almost all the suburbs we still can find good number of similar venues such as restaurant, bars, sport facilities, etc. Therefore, any Houston suburb provides many options for these categories. For more accurate analysis, more data points are required especially from the suburb area.

Finally, this analysis will be more comprehensive if other data such as crime rate, school ranking, house prices, etc. for each individual zip code (neighborhood) can be integrated to the current data set. Analysis of such an integrated data set will generate valuable insight for the people who are new to the city and would like to evaluate different neighborhoods to buy or rent a house or relocate to this area.