



Which Neighborhoods are Good Choices for new Mexican Restaurants at Houston



1. Introduction

No matter your definition of Mexican food—be it fajitas and margaritas or mole and Rompope — there’s no better place than Houston to find it. As the home to the country’s third-largest Mexican population, Mexican food is very popular at Houston. As shown in the following figure, Mexican restaurant is the 1st most venues at Houston, ranks higher than Gas Station, Bar, and all other restaurant.

VenueCategory	Neighborhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude
Mexican Restaurant	257	257	257	257	257	257
Pizza Place	184	184	184	184	184	184
Fast Food Restaurant	179	179	179	179	179	179
Sandwich Place	154	154	154	154	154	154
Discount Store	145	145	145	145	145	145
Fried Chicken Joint	125	125	125	125	125	125
Coffee Shop	120	120	120	120	120	120
Gas Station	107	107	107	107	107	107
Burger Joint	96	96	96	96	96	96
Bar	95	95	95	95	95	95

So, if someone want to open a Mexican restaurant at Houston, where should her choose the location?

In this project, we want to help people in exploring better locations for a new Mexican restaurant to make her smart and efficient decision, using data science methodology and machine learning techniques like clustering, and data from wiki and four square.

2. Data

To solve the problem, we will need the following data:

- List of neighborhoods in Houston. This defines the scope of this project which is confined to the city of Houston, Texas. This data is from Wiki “List of neighborhoods in Houston”.

https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Houston

- Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data. This data is got by Geocoder.

- Venue data, particularly data related to Mexican restaurant. We will use this data to perform clustering on the neighborhoods. Th is data is from Four Square API. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data.

This is a project that will make use of many data science skills, including working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium).

In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

Firstly, we need to get the list of neighbourhoods in Houston. Fortunately, the list is available in the Wikipedia page. As shown in following figure and the notebook, there are 88 neighborhoods in our data.

```
In [524]: df1 = pd.read_csv('houston.csv')
          print(df1.shape)
          df1.head(100)
```

(88, 1)

Out[524]:

	Neighborhood
0	Willowbrook
1	Greater Greenspoint
2	Carverdale
3	Fairbanks
4	Greater Inwood
5	...

However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame. After getting the neighborhood data, we get the geographical coordinates, we merge the neighborhoods with the coordinates, as shown in following figure.

```
In [139]: # check the neighborhoods and the coordinates
          print(df1.shape)
          df1.head(100)
```

(88, 3)

Out[139]:

	Neighborhood	Latitude	Longitude
0	Willowbrook	29.952400	-95.544630
1	Greater Greenspoint	29.939670	-95.407480
2	Carverdale	29.849590	-95.542450
3	Fairbanks	29.852730	-95.524190
4	Greater Inwood	29.869770	-95.480440
5	Acres Home	29.870470	-95.435360
6	Hidden Valley	29.888470	-95.414600

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. After applying the Foursquare API to explore the neighborhoods, as shown in following figure.

```
In [152]: # convert the venues list into a new DataFrame
venues_df = pd.DataFrame(venues)

# define the column names
venues_df.columns = ['Neighborhood', 'Latitude', 'Longitude', 'VenueName', 'VenueLatitude', 'VenueLongitude', 'VenueCategory']

print(venues_df.shape)
venues_df.head()
```

(5115, 7)

```
Out[152]:
```

	Neighborhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	Willowbrook	29.9524	-95.54463	Bed Bath & Beyond	29.953517	-95.543865	Furniture / Home Store
1	Willowbrook	29.9524	-95.54463	Babin's Seafood House	29.955088	-95.544452	Seafood Restaurant
2	Willowbrook	29.9524	-95.54463	Costco	29.954658	-95.547697	Warehouse Store
3	Willowbrook	29.9524	-95.54463	buybuy BABY	29.953127	-95.543557	Kids Store
4	Willowbrook	29.9524	-95.54463	Pho An 2	29.956606	-95.543805	Vietnamese Restaurant

Based on the previous data, we conduct data cleaning and analyze the data to make sure the data is ready to be used for clustering.