Finding a Suitable Location for a New Indian/Asian Restaurant within Houston or Dallas

Kheya Banerjee

December 22, 2020

1. Introduction

1.1 Background

According to reports [1], Dallas, TX and Houston, TX have the 7th and 8th largest Indian immigrant population in US. The other 6 cities are:

1. New York, NY
2. Chicago, IL
3. Washington, DC
4. Los Angeles, CA
5. San Francisco, CA
6. San Jose, CA

1.2 Problem

Although the Indian immigration population in Dallas and Houston is large, the number of Indian restaurants is few. Hence, there is always demand for new authentic Indian dining restaurants. Due to cultural and culinary proximity, we have used Asian restaurants along with Indian.

1.3 Interest

It is extremely important to choose a suitable location for a new restaurant as this will help the business to thrive.

The goal of this project is to determine a location for a new Indian/Asian restaurant within Dallas or Houston area.

2. Data Acquisition and Cleaning

2.1 Data Sources

Various data sources are used in this project. They are:

1. The center latitude and longitude of Dallas, TX and Houston, TX are obtained by the help of Nominatim package which is a child package of geopy.geocoders. Although these are approximate address, it is quite accurate for our purpose.
2. To determine the boundaries of Dallas and Houston city, we have downloaded the list of all zip codes with the respective latitude and longitude values from the website [2].

3. The restaurant location and categories are downloaded from foursquare website [3] as .json file. The total number of venues in food category are:

Dallas	457
Houston	572

2.2 Data Cleaning

Some preprocessing and cleaning of the zip code data is needed as the csv file had only 1 column with all the values separated by ';'. The processing is done as follows:

1. The dataframe is converted to a list.
2. The column labels are also converted to list and split to make column labels for new dataframe
3. Every element(row) is split and separated into columns. Values are written into a new dataframe.

The restaurant information from Foursquare [3] also needed some selection as only the required fields are downloaded and stored into a dataframe. The fields saved in dataframe are:

1. Zip code
2. Zip code Latitude
3. Zip code Longitude
4. Venue name
5. Venue Latitude
6. Venue Longitude
7. Venue Category
8. Venue Category ID

4. For clustering, some preprocessing of data is needed as k-means algorithm does not accept categorical variables. The dataframe needed to be converted to numpy array.

3. Exploratory Data Analysis

3.1 Mapping the zip codes on the city map

The zip codes are mapped on the Dallas and Houston city map using Folium package [4]. The maps are shown below:

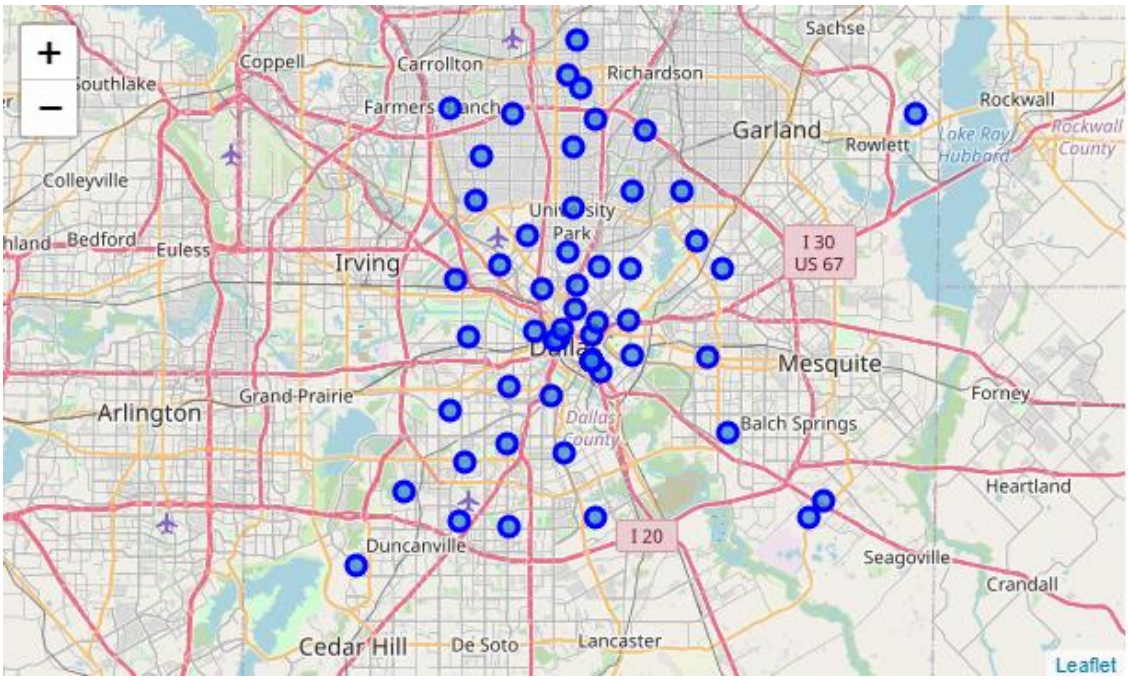

Figure 1 - Dallas city map with all the zip code locations as blue dots

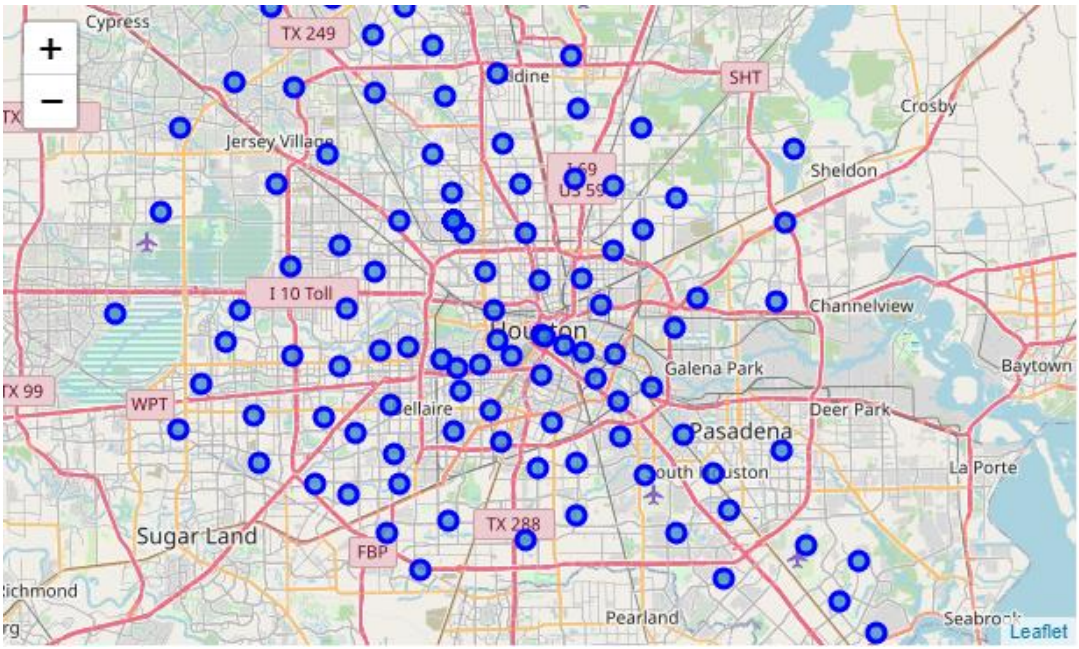

Figure 2 - Houston city map with all the zip code locations as blue dots

3.2 Restaurant Location and Category

The restaurants nearby the previously shown zip codes are downloaded from Foursquare [3]. We have got the following information:

1. Venue name
2. Venue Latitude
3. Venue Longitude
4. Venue Category
5. Venue Category ID

As we are interested in Indian and Asian restaurants, we have gathered a list of "Category ID" from [5]. The restaurants of interest are examined, and their numbers are counted. We got:

	Dallas	**Houston**
Indian	1	9
Chinese	7	17
Japanese	13	11
Other Asian	15	28

These values are also plotted as bar chart.

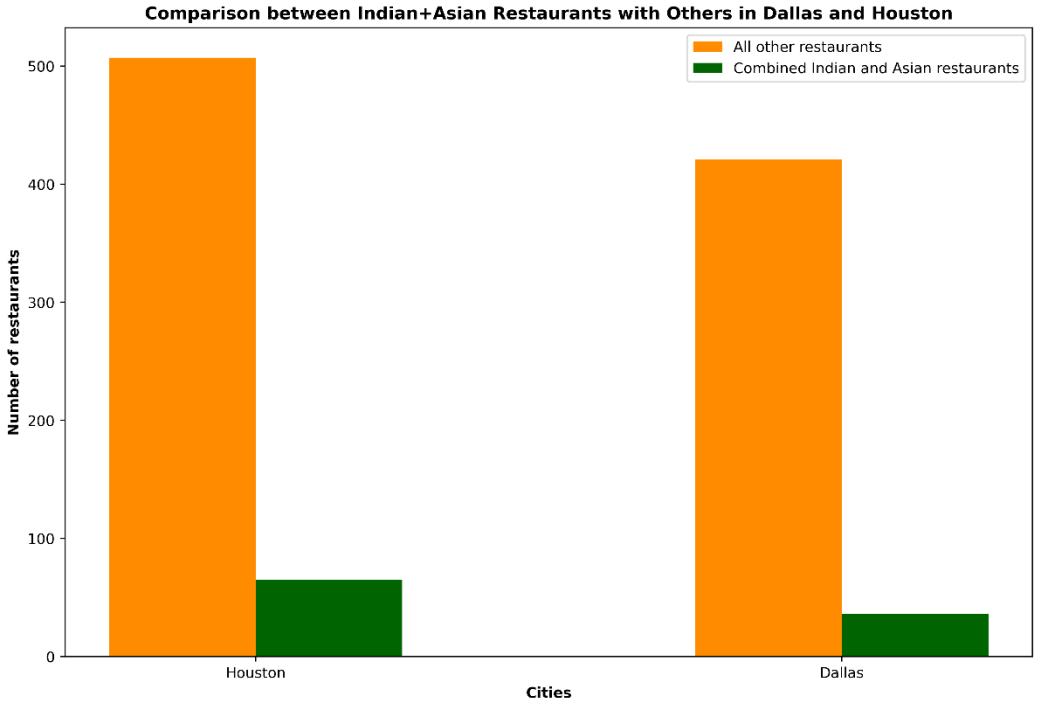

Figure 1 - Comparison between all restaurant categories

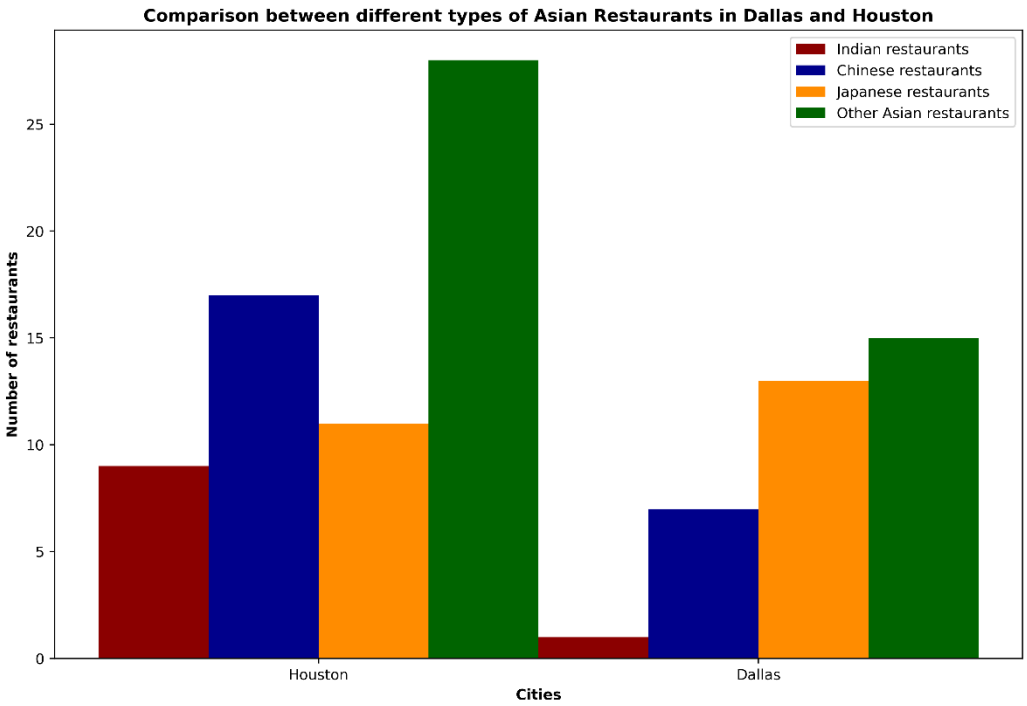

Figure 2 - Comparison between Asian restaurant categories

3.3 Direction of Further Analysis

It is clearly visible from the above graph that Houston has more Asian and Indian restaurants than Dallas. Hence, from now on, we are going to concentrate our analysis on Houston.

The goal is to find a suitable location for a new Indian or Asian restaurant in Houston.

3.4 Mapping the Restaurant location on the city map

The restaurants are placed as shown below:

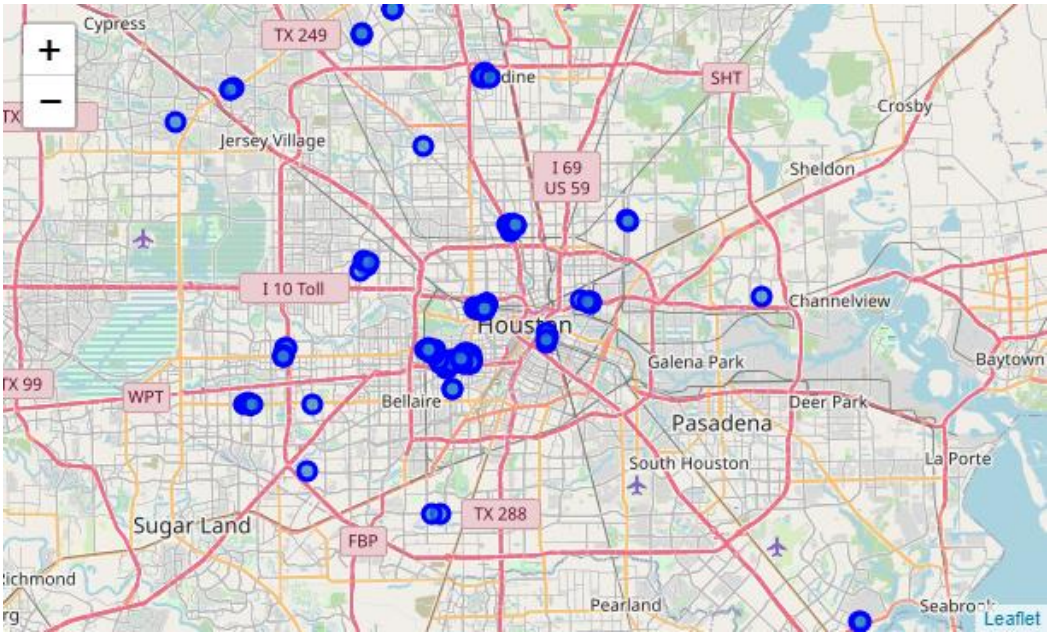

Figure 3 - Restaurant locations in Houston

As the map shows, most of the restaurant locations are concentrated around some region or cluster. Hence, we will cluster the locations to form a more simplified map. We will use k-mean algorithm to accomplish this.

4. Data Clustering

For clustering, the latitude and longitude values of restaurant locations will be used. Some preprocessing of data is needed as k-means algorithm does not accept categorical variables. The dataframe also needed to be converted to numpy array.

4.1 Obtaining the K value for Kmeans clustering

Elbow method is used for this purpose. The idea is to vary the K values and calculate the inertia of the model. The plot is as follows:

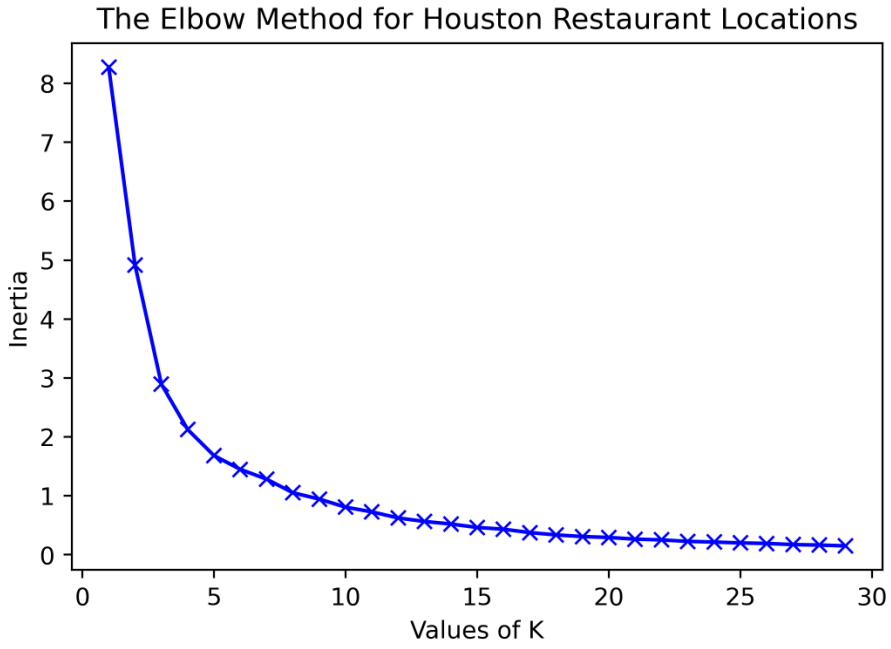

Figure 4 - Elbow method to determine K value

Based on the map (Figure 3 - Restaurant locations in Houston) and elbow method, we have decided to use **K=10** for our analysis.

4.2 Clustering with K-Means

The restaurant locations are clustered into 10 different clusters. The idea is to choose the cluster which is a popular spot for restaurants but with very few Indian or Asian restaurants. So, we chose the top 5 clusters based on the total restaurant numbers and explored each cluster individually. The number of Indian and Asian restaurants for top 5 clusters are as below:

Cluster #	Cluster Label	# Indian Restaurants	# Chinese Restaurants	# Japanese Restaurants	# Other Asian Restaurants
1	5	6	4	7	8
2	0	1	3	1	3
3	3	0	3	2	2
4	7	0	1	0	0
5	6	1	1	0	4

5. Conclusions

Based on our analysis, Cluster 4 (Clus_km = 7) is the most suitable location for opening a new Indian/Asian restaurant. This location is near the Heights area, which is in north of Houston. This cluster does have very few Indian/Asian restaurants (only 1 Chinese restaurant). However, there are many popular restaurants in this location. This means this place is popular for its food venues. Location wise, this is the best location in Houston for a new Indian/Asian restaurant.

6. Future Improvement

There are many factors needed to be considered when opening a new business. For this case, the factors can be:

1. Asian population near that area
2. Property pricing
3. Tax and other expenditures
4. Crime rate

Including these factors can help us understand about the possibility of success much better.

Also, Foursquare has a limit of 50 venues per location. This means it only provides 50 most popular food places for a single location. There may be many other Indian/Asian restaurants that are not being included in the analysis.

7. References:

[1] https://en.wikipedia.org/wiki/Indian_Americans
[2] https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/table/
[3] www.foursquare.com
[4] https://python-visualization.github.io/folium/
[5] https://developer.foursquare.com/docs/build-with-foursquare/categories/