# Which Neighborhoods are Good Choices for new Mexican Restaurants at Houston

## 1. Introduction

### Brief Background

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 |

### Business Problem

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 |

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 wit 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.

## 3 Methodology

We will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project.

We cluster the neighbourhoods into 5 clusters, as shown in following figure.

**Cluster Neighborhoods**

In [512]:
```
# set number of clusters
kclusters = 5

#original
#kl_clustering = kl_gym.drop(["Neighborhoods"], 1)

#change to:
kl_clustering = kl_r.drop(["Neighborhoods"], 1)

print(kl_clustering.shape)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(kl_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

```
(88, 11)
```

Out[512]: `array([3, 0, 2, 4, 0, 0, 0, 1, 0, 0])`

In [513]:
```
# create a new dataframe that includes the cluster as well as the top 10 venues for each neighborhood.
kl_merged = kl_r.copy()

# add clustering labels
kl_merged["Cluster Labels"] = kmeans.labels_
```

Wen then generate the cluster labels and merge it back to original data. Then, we can perform further analysis based on the clustering results, as given in next sections.

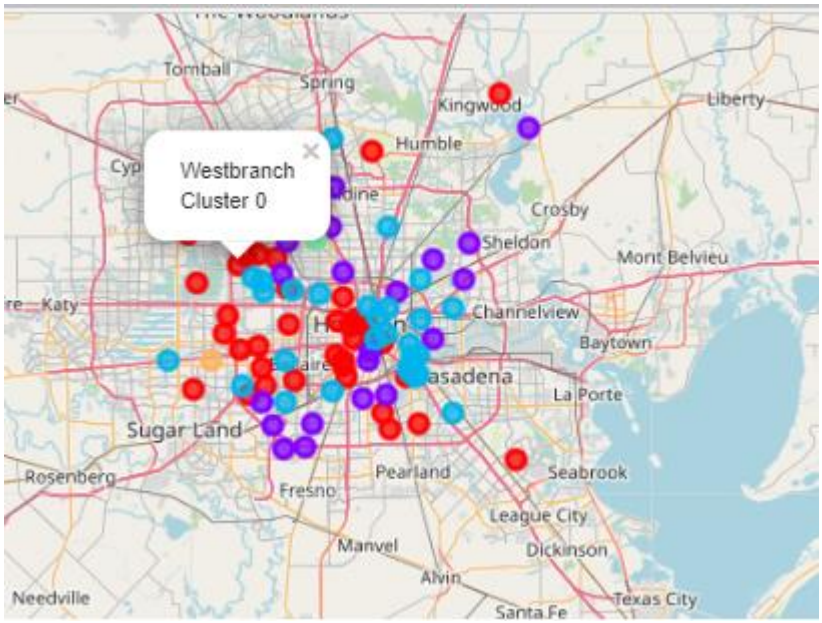
In [518]:
```
# sort the results by Cluster Labels
print(kl_merged.shape)
kl_merged.sort_values(["Cluster Labels"], inplace=True)
kl_merged
```

```
, 15)
```

| Neighborhood | Mexican Restaurant | Pizza Place | Fast Food Restaurant | Sandwich Place | Fried Chicken Joint | Burger Joint | American Restaurant | Vietnamese Restaurant | Chinese Restaurant | Seafood Restaurant | Italian Restaurant | Cluster Labels | Latitude | Longitude |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Willowbrook | 0.040000 | 0.000000 | 0.020000 | 0.010000 | 0.010000 | 0.030000 | 0.010000 | 0.010000 | 0.010000 | 0.010000 | 0.010000 | 0 | 29.952400 | -95.544630 |
| IAH Airport | 0.033898 | 0.050847 | 0.033898 | 0.067797 | 0.016949 | 0.000000 | 0.033898 | 0.000000 | 0.000000 | 0.016949 | 0.016949 | 0 | 29.987890 | -95.349480 |
| Meyerland | 0.040000 | 0.030000 | 0.040000 | 0.030000 | 0.000000 | 0.040000 | 0.010000 | 0.000000 | 0.020000 | 0.000000 | 0.000000 | 0 | 29.684530 | -95.467460 |
| Memorial | 0.060000 | 0.010000 | 0.000000 | 0.010000 | 0.000000 | 0.030000 | 0.050000 | 0.000000 | 0.000000 | 0.030000 | 0.020000 | 0 | 29.772630 | -95.570920 |
| Fourth Ward | 0.020000 | 0.020000 | 0.000000 | 0.010000 | 0.010000 | 0.020000 | 0.030000 | 0.020000 | 0.000000 | 0.000000 | 0.030000 | 0 | 29.757620 | -95.384490 |
| Golfcrest | 0.047619 | 0.047619 | 0.031746 | 0.015873 | 0.047619 | 0.015873 | 0.000000 | 0.015873 | 0.015873 | 0.015873 | 0.000000 | 0 | 29.691230 | -95.298790 |
| Greater Eastwood | 0.036585 | 0.048780 | 0.048780 | 0.036585 | 0.036585 | 0.012195 | 0.024390 | 0.012195 | 0.024390 | 0.000000 | 0.012195 | 0 | 29.735990 | -95.334980 |
| Medical Center | 0.020000 | 0.020000 | 0.000000 | 0.010000 | 0.010000 | 0.040000 | 0.020000 | 0.000000 | 0.010000 | 0.000000 | 0.000000 | 0 | 29.711790 | -95.393150 |

## 4 Results

The results from the k-means clustering show that we can categorize the neighborhoods into 5. We plot the result as following figure.




We further calculated the average value of different restaurant in following table, based on the cluster labels.

```
In [521]: kl_merged.groupby('Cluster Labels').mean()
```

Out[521]:

| Cluster Labels | Mexican Restaurant | Pizza Place | Fast Food Restaurant | Sandwich Place | Fried Chicken Joint | Burger Joint | American Restaurant | Vietnamese Restaurant | Chinese Restaurant | Seafood Restaurant | Italian Restaurant | Latitude | Longitu |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.030222 | 0.025456 | 0.016077 | 0.024366 | 0.010695 | 0.016468 | 0.015643 | 0.008315 | 0.015032 | 0.009439 | 0.012502 | 29.759653 | -95.437 |
| 1 | 0.008773 | 0.049978 | 0.100218 | 0.050279 | 0.053853 | 0.021471 | 0.010871 | 0.001826 | 0.015379 | 0.009865 | 0.000560 | 29.839513 | -95.370 |
| 2 | 0.097117 | 0.043596 | 0.039028 | 0.037571 | 0.035934 | 0.021157 | 0.013641 | 0.012542 | 0.008913 | 0.014417 | 0.007123 | 29.755647 | -95.375 |
| 3 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.142857 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.142857 | 0.000000 | 29.870470 | -95.435 |
| 4 | 0.000000 | 0.015625 | 0.031250 | 0.015625 | 0.031250 | 0.000000 | 0.000000 | 0.265625 | 0.000000 | 0.000000 | 0.000000 | 29.710880 | -95.595 |

• Cluster 0: Neighbourhoods with moderate concentration of Mexican restaurants.

• Cluster 1: Neighbourhoods with low concentration of Mexican restaurants.

• Cluster 2: Neighbourhoods with high concentration of Mexican restaurants.

We also note that cluster 3 and cluster 4 only have 1 observations. And both of these two clusters have 0 concentration of Mexican restaurant.

## 5 Discussion

As observations noted from the results in previous section, the Mexican restaurants are more concentrated at East and South-East Houston. The concentrate level of Mexican restaurant at West Houston are moderate. At the same time, there are less Mexican restaurants at North and East-North Houston.

Based on the findings, we may say that we should open a new Mexican restaurant at cluster 1, or North and East-North Houston.

At the same time, we note that the concentrate level of Pizza place, fast food restaurants, Burger Joint, are higher, and the concentrate level of Italian restaurant is lower, compared to other clusters.

Therefore, we may take cluster 1 as one potential choice but we also need to take further considerations based on other factors such as population density, rents, traffic, among others.

## 6 Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 5 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders. To answer the business question that was raised in the introduction section, the answer proposed by this project is:

The neighborhoods in cluster 1 are the most preferred locations to open a new Mexican restaurant. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new Mexican restaurant.