Predict Location of A New Asian Restaurant in Southwest Houston, TX

Applied Data Science Capstone by IBM/Coursera

Deming Wang
Sugar Land, TX

***Abstract*—This project aims at analyzing restaurant distribution in west Houston and targets to stakeholders who are interested in opening a new Asian restaurant in Southwest Houston – Sugar Land and Missouri City.**

Keywords—data science, K-Means, uncertainty analysis, Foursquare API, Python

I. Introduction

Sugar Land and Missouri City are two cities located in the southwestern part of the Houston-The woodlands-Sugar Land metropolitan area. Located about 19 miles (31 km) southwest of downtown Houston, both cities are populous suburban municipality centered around the junction of Texas State Highway 6 and U.S. Route 59.

Due to the culture diversity and increasing population these years, demand for authentic Asian food is expected to remain strong in near future. Currently, Houston China town (near Highway 8 & Bellair Blivd) and Katy Asian Town (near Highway 10 & Highway 99) are two most popular places for Asian food courts in Houston area. Unfortunately, there is no similar place in Sugar Land/Missouri City area. In this project, we will try to see how all restaurants are distributed in west Houston and find out the best possible locations for a new Asian restaurant in Sugar Land/Missouri City area.

II. DAta

Based on definition of our problem, the following factors are key answers to our problem:

- All existing restaurants in west Houston
- Locations of these restaurants
- Food type (category) of all restaurants: for example, Asian restaurant, Mexican restaurant, Cafe, etc.

It is possible to get all restaurants for each neighborhood, but many neighborhoods are in irregular shapes and it is hard to use neighborhood center to search restaurants. To have better coverage, many small cycles will be created in selected area, then restaurants will be searched via Foursquare API based on center of each cycle. The following data will be needed to extract/generate the required restaurant information:

- coordinates of the center in both cities using folium map.
- generate small areas in folium map
- retrieve all restaurants data (name, category, and coordinates) for each area using Foursquare API

A. Data Sources

In this project, all restaurants information will be downloaded using Foursquare API. Restaurant data from Foursquare include restaurant name, latitude, longitude, category.

B. Data acquistion

To be able to retrieve data from Foursquare API, 1534 points are generated around the center point (Latitude: 29.6589382 Longitude: -95.7276974). To accurately calculate distance between two points, the geodetic coordinates (latitude and longitude) is converted to projected coordinates (x and y) using WGS84 datum when generating points. Later, x and y are converted back to latitude and longitude for API calling.

Now, the coordinates of centers of areas to be evaluated, equally spaced (distance from every point to its neighbors is exactly the same). The following map shows all points and coverage of each cycle.

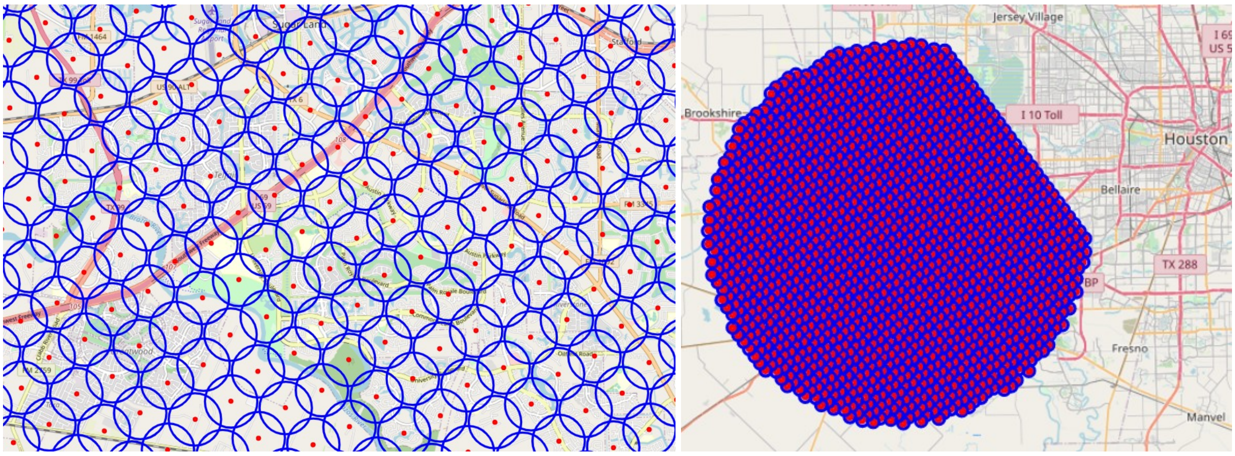


Fig. 1. 1534 points generated for west houston area.

According to Foursquare API documentation (https://developer.foursquare.com/docs/build-with-foursquare/categories/), all restaurant categories are listed under category - Food with id 4d4b7105d754a06374d81259. 3702 restaurants are downloaded from Foursquare.

The following map shows the distribution of 3702 restaurants in west Houston area.

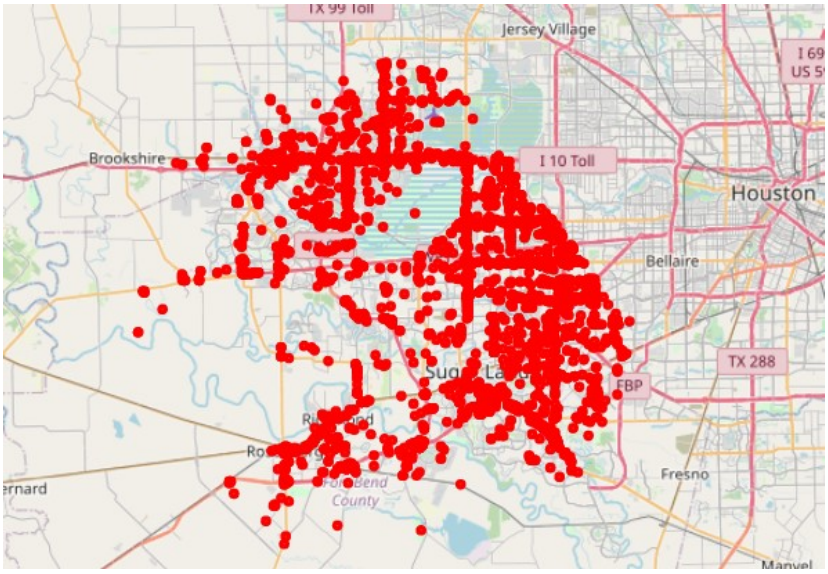


Fig. 2. Example of a figure caption. (*figure caption*)

C. *Data cleaning*

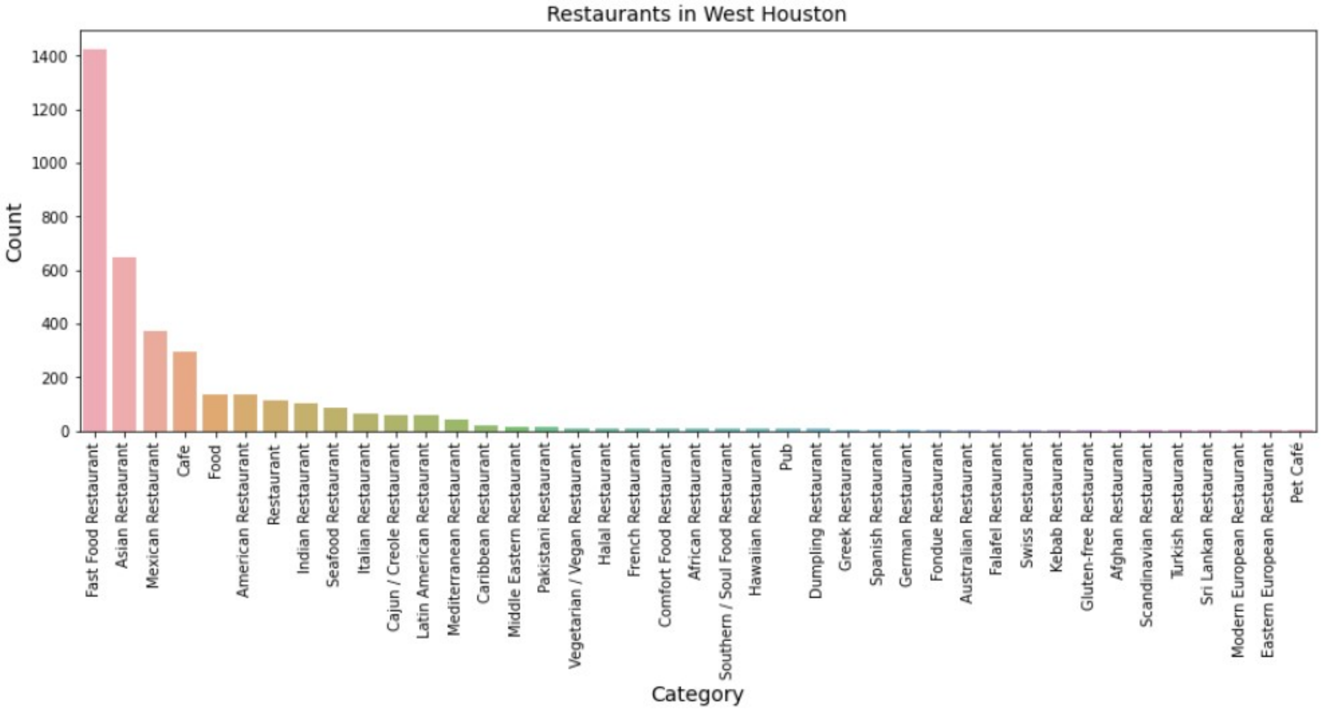
Foursquare API contains a list of food categories and some categories are very detailed, which may make cluster algorithm inefficient. To improve it, the following categories are grouped to general categories including food, fast food restaurant, pub and café.

TABLE I. RESTAURANT CATEGORY MAPPING

Original value	***New value***
[Empty]	Food
Sandwich Place Fried Chicken Joint Pizza Place Fish & Chips Shop Mac & Cheese Joint Breakfast Spot Donut Shop Food Truck Bakery Burger Joint Bistro Wings Joint Bagel Shop	Fast Food Restaurant

Original value	***New value***
Snack Place Hot Dog Joint	
Irish Pub Gastropub	Pub
Creperie BBQ Joint Buffet Café Deli / Bodega Steakhouse Salad Place Food Court Soup Place	Cafe

As expected, the flowing chart shows the distribution based on restaurant categories. Unquestionably, “Fast Food Restaurant”,” Asian Restaurant” and “Mexican Restaurant” are top 3 restaurants in west Houston.



Now, all restaurants in west Houston area are downloaded and cleaned. Each restaurant has location and category information. This concludes the data preparation phase, and these data will be used later for advanced analysis.

III. Methodology

A. K-Means test

In this project, our efforts will be dedicated on detecting areas in Sugar Land/Missouri City that have high Aisan restaurant density and similar restaurant distribution as Houston China town and Katy Asian town, which have the most successful Asian restaurant business.

The whole area will be divided into N x N grids, and K-Means is used to create clusters of restaurant categories.

The following chart shows a quick test with N=18, and run K-Means up to 30 clusters. Unfortunately, there is no clear trend to identify a optimal K value from sum squared of distance using elbow method. So silhouette score is selected to identify the optimal K. Although K=8 has the highest score, but the score is only about 0.31, which is too small to be accepted.

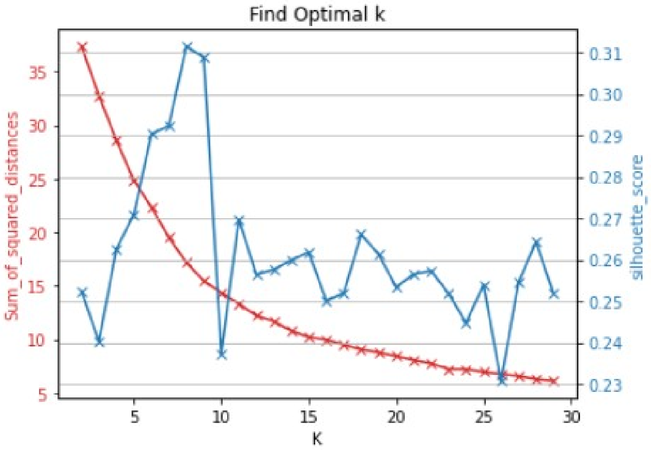


Fig. 3. Sum of suqared distance and silhouette scores for 18 x 18 grid and clusters from 2 to 30 in K-Means clustering

B. Uncertainty analysis

To find the optimal K, we'll run an uncertainty analysis for grid size and generate scenarios for N from 10 to 70. The following plot shows there are no clear elbow method can be used to identify optimal K, but silhouette score line get higher as N increases.

We would like to choose a N value with high silhouette score and keep K as small as possible. From the plot, N=65 and K=20 seems acceptable.

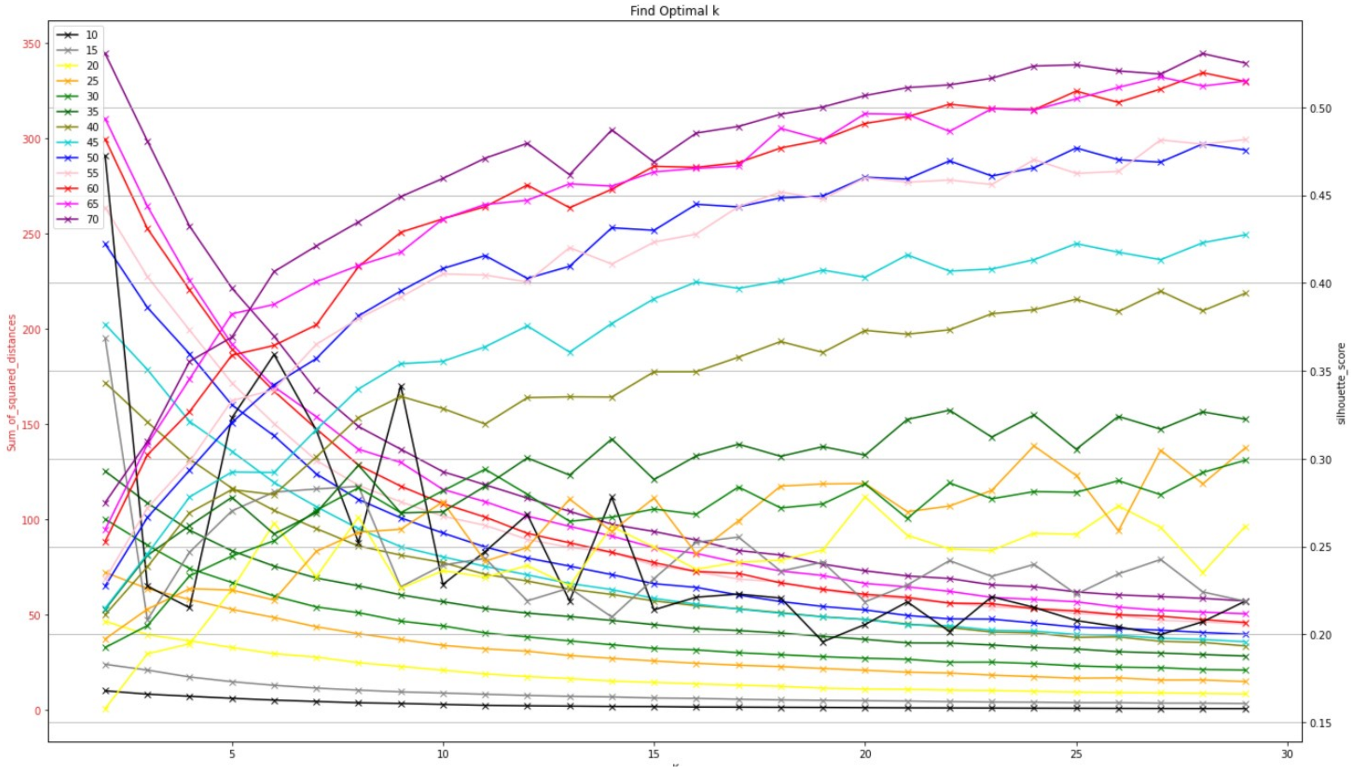


Fig. 4. K-Means score for N (10 to 70)

From above plot, silhouette score for K=20 is close to 0.5, and this will be used as the best clustering parameter.

C. Clustering analysis

Run K-Means model for N=65, and select K=20 as the final parameter for clustering restaurants in west Houston. Finally, 20 clusters are generated.

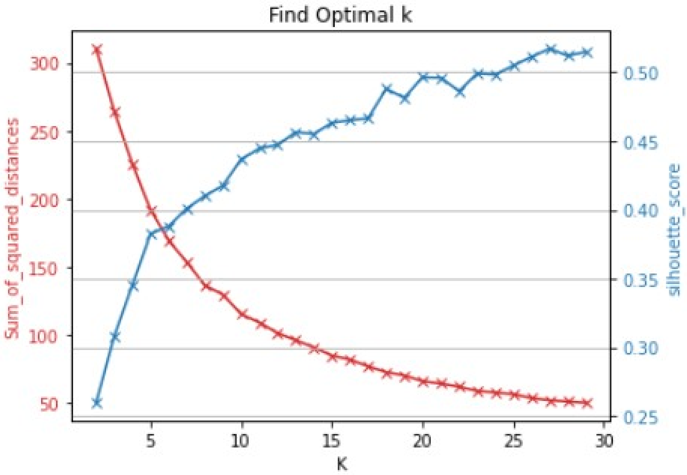


Fig. 5. K-Means clustering score for grid 65x65

IV. Results

A. Restaurants overview

After clustering restaurants in 65x65 grid, all restaurants from 20 clusters are plotted in figure 6. Most popular restaurants are located along highway 59 and 6 in Sugar Land and Missouri city.

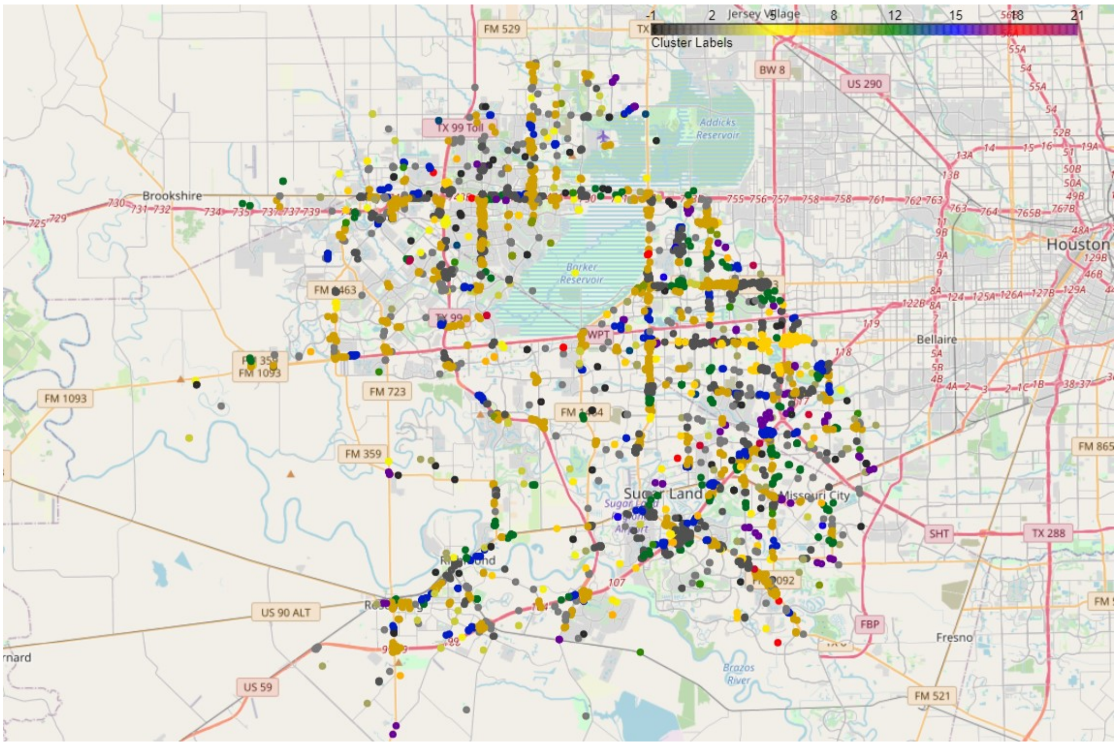


Fig. 6. Example of a figure caption. (*figure caption*)

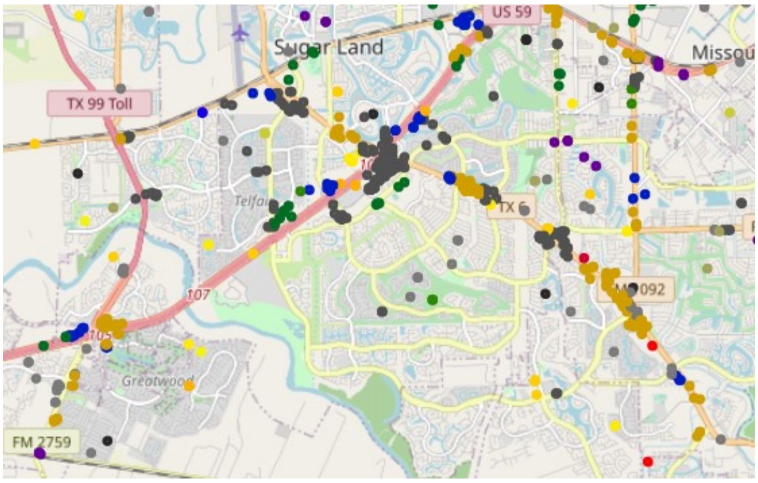


Fig. 7. Restaurants in Sugar Land and Misourri City

B. *Most common restaurants in clusters*

To have a better view about Asian restaurants in the whole area, we'll group these restaurants based on cluster label and create a 5-most-common restaurants in each cluster. Table II shows top 5 most common restaurants in 20 clusters.

TABLE II. TOP 5 RESTAURANT CATEGORY IN CLUSTERS

Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Cafe	Vegetarian / Vegan Restaurant	Eastern European Restaurant	Gluten-free Restaurant	German Restaurant
1	Asian Restaurant	Fast Food Restaurant	Mexican Restaurant	Cafe	American Restaurant
2	Fast Food Restaurant	Mexican Restaurant	Cafe	Pakistani Restaurant	American Restaurant
3	Food	Vegetarian / Vegan Restaurant	Halal Restaurant	Gluten-free Restaurant	German Restaurant
4	Mexican Restaurant	Italian Restaurant	**Asian Restaurant**	Food	Fast Food Restaurant
5	American Restaurant	Fast Food Restaurant	Mexican Restaurant	Spanish Restaurant	Cafe
6	Indian Restaurant	Fast Food Restaurant	Middle Eastern Restaurant	**Asian Restaurant**	Cajun / Creole Restaurant
7	Asian Restaurant	Fast Food Restaurant	Cajun / Creole Restaurant	Food	Cafe
8	Cajun / Creole Restaurant	Asian Restaurant	Fast Food Restaurant	Vegetarian / Vegan Restaurant	Eastern European Restaurant
9	Fast Food Restaurant	Asian Restaurant	Mexican Restaurant	Cafe	American Restaurant
10	Italian Restaurant	Restaurant	American Restaurant	Indian Restaurant	Fast Food Restaurant
11	Restaurant	Indian Restaurant	American Restaurant	Vegetarian / Vegan Restaurant	Eastern European Restaurant
12	Cafe	Fast Food Restaurant	Restaurant	Food	Caribbean Restaurant
13	Caribbean Restaurant	Vegetarian / Vegan Restaurant	Eastern European Restaurant	Gluten-free Restaurant	German Restaurant
14	Mexican Restaurant	Fast Food Restaurant	**Asian Restaurant**	Cafe	American Restaurant
15	Middle Eastern Restaurant	Vegetarian / Vegan Restaurant	Eastern European Restaurant	Gluten-free Restaurant	German Restaurant
16	Food	Fast Food Restaurant	Mexican Restaurant	Seafood Restaurant	American Restaurant
17	Latin American Restaurant	Fast Food Restaurant	Vegetarian / Vegan Restaurant	Eastern European Restaurant	Gluten-free Restaurant
18	Seafood Restaurant	Vegetarian / Vegan Restaurant	Dumpling Restaurant	Gluten-free Restaurant	German Restaurant
19	Mediterranean Restaurant	Afghan Restaurant	Eastern European Restaurant	Gluten-free Restaurant	German Restaurant

As we can see, Asian restaurants are popular in west Houston area, especially in cluster 1, 7(1st Most Common Venue), 8 and 9(2nd Most Common Venue).

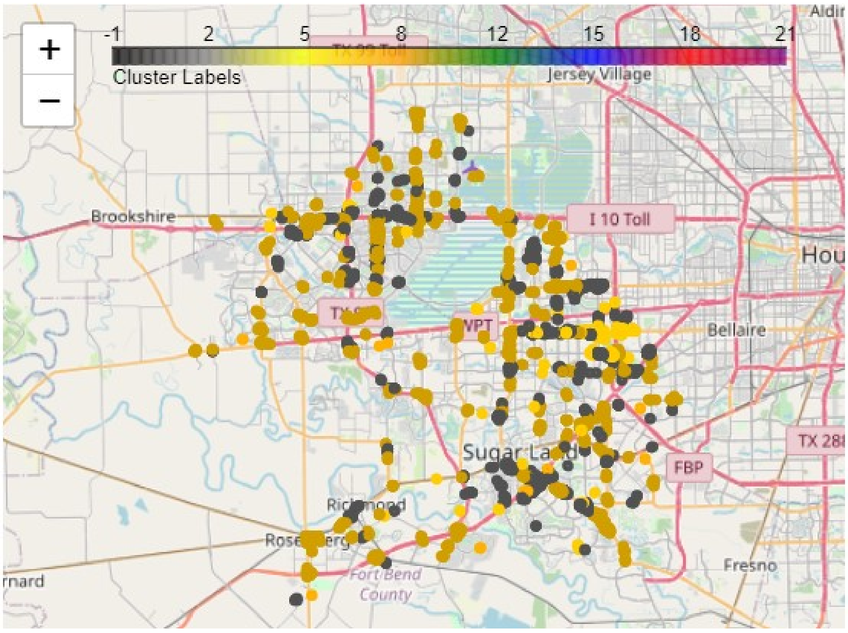


Fig. 8. Example of a figure caption. (*figure caption*)

Figure 8 Restaurants distribution from cluster 1, 7, 8 and 9.

V. DISCUSSTION

To have better understanding of restaurants distribution and density, bar charts and pie charts are created to explore cluster 1,7,8 and 9 respectively.

A. Restaurants in cluster 1

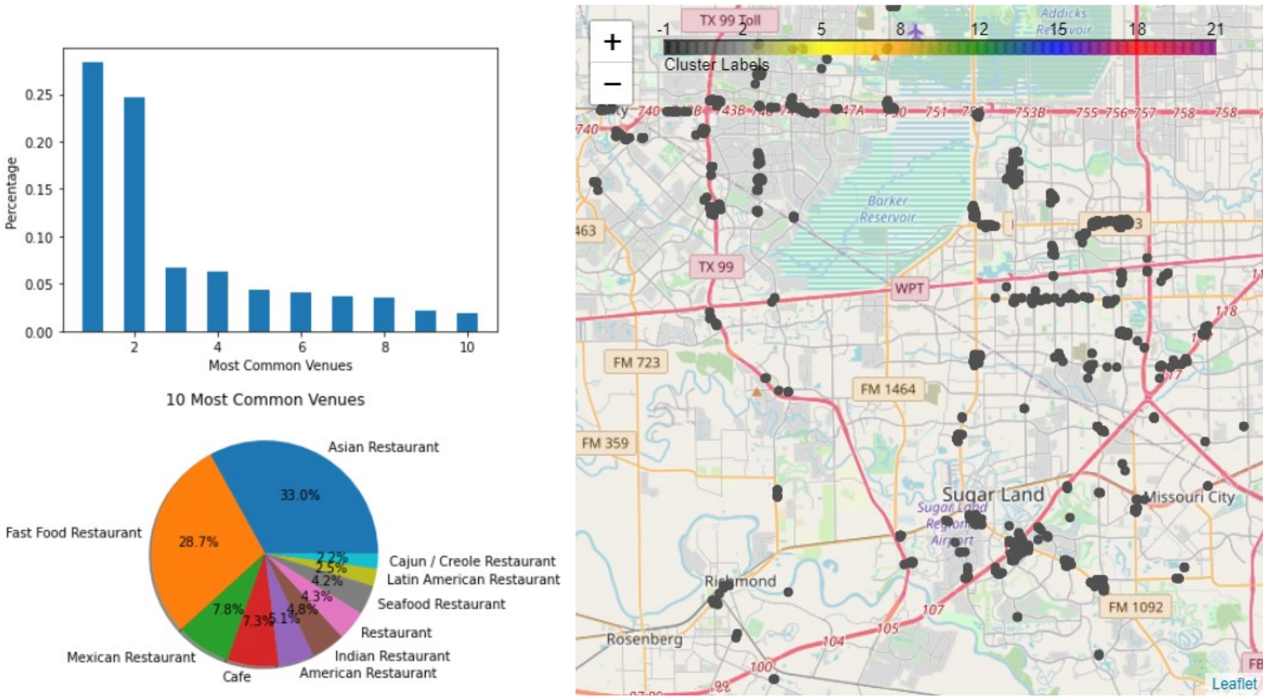


Fig. 9. Restaurants in cluster 1

From above bar chart and map, it is clear that this category is a popular category with 33% Asian restaurants and 28.7% Fast food restaurant, and it's widely spreaded out in west Houston. Most restaurants in Sugar Land/Missouri City area are located near the intersection of Highway 59 and 6.

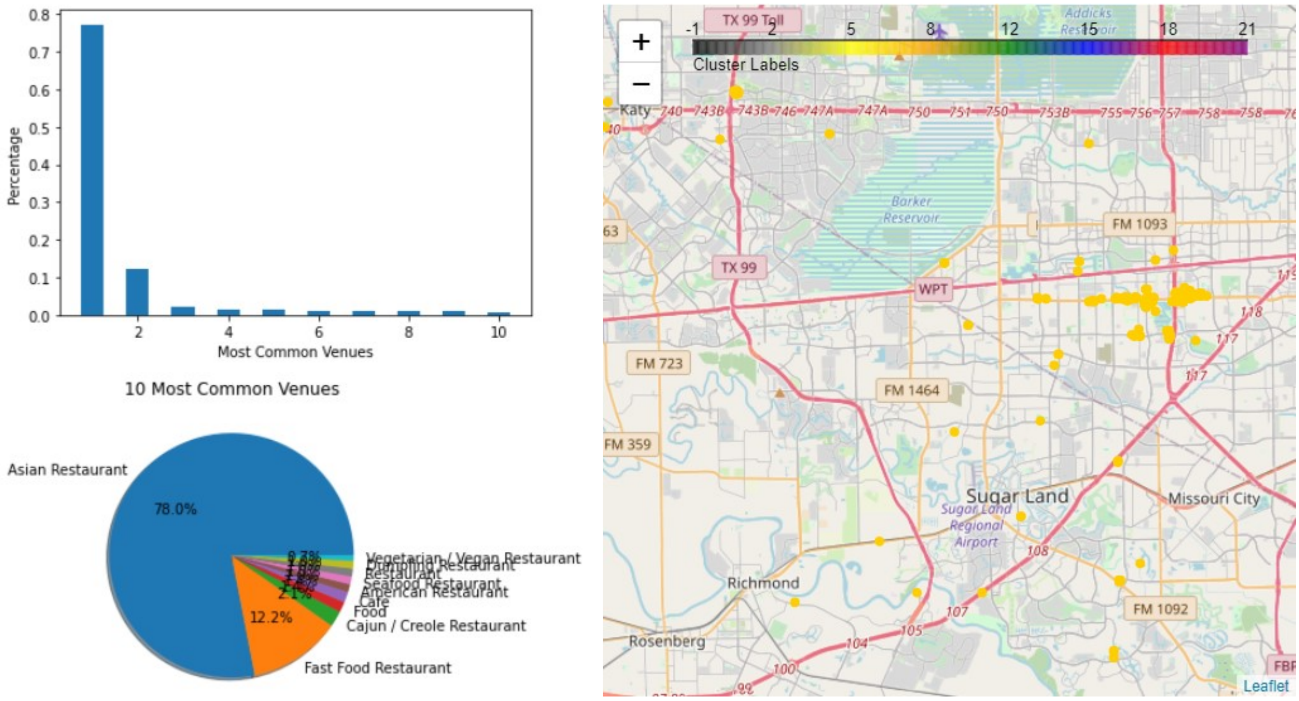


Fig. 10. Restaurants in cluster 7

For cluster 7, 78% restaurants are Asian restaurants. These are hot spots for Asian food. Both Houston China town and Katy Asian town fall into this category. In Sugar Land, there are 5 Asian restaurants in the plaza near Hight 6 and Austin Parkway. This is probably the best location for future Asian restaurant with the best traffic.

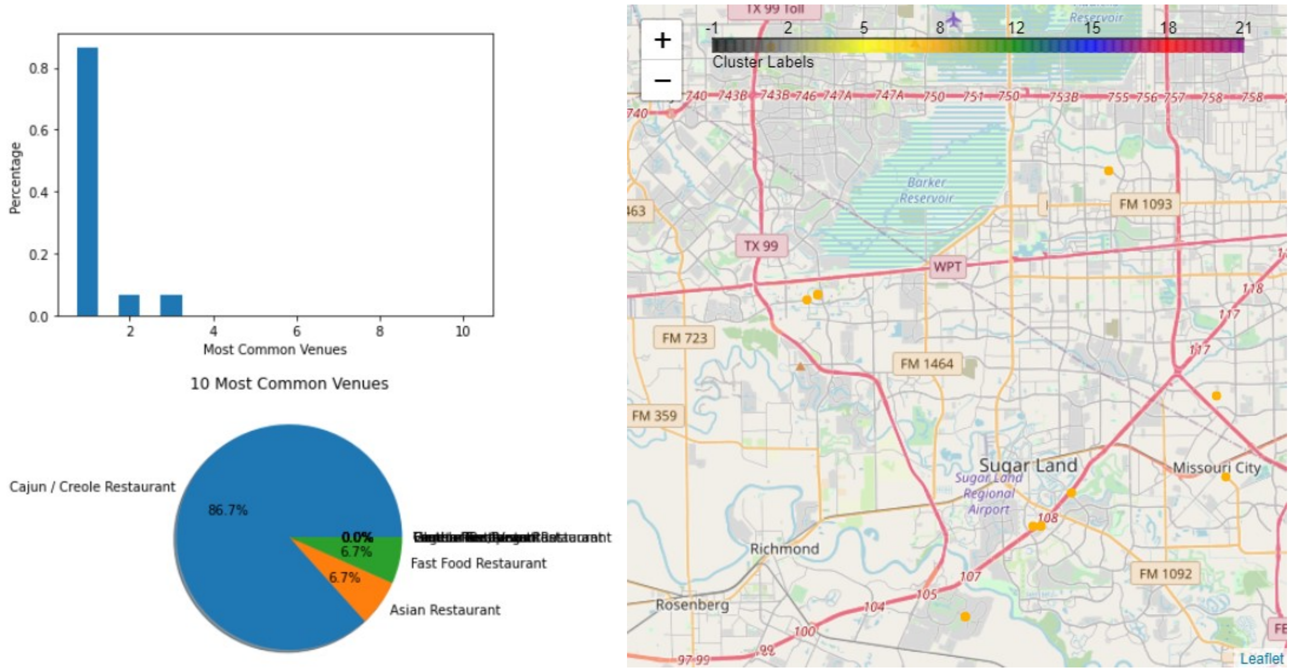


Fig. 11. Restaurants in cluster 8

Category 8 contains 6.7% Asian restaurants, but not many restaurants are observed from map.

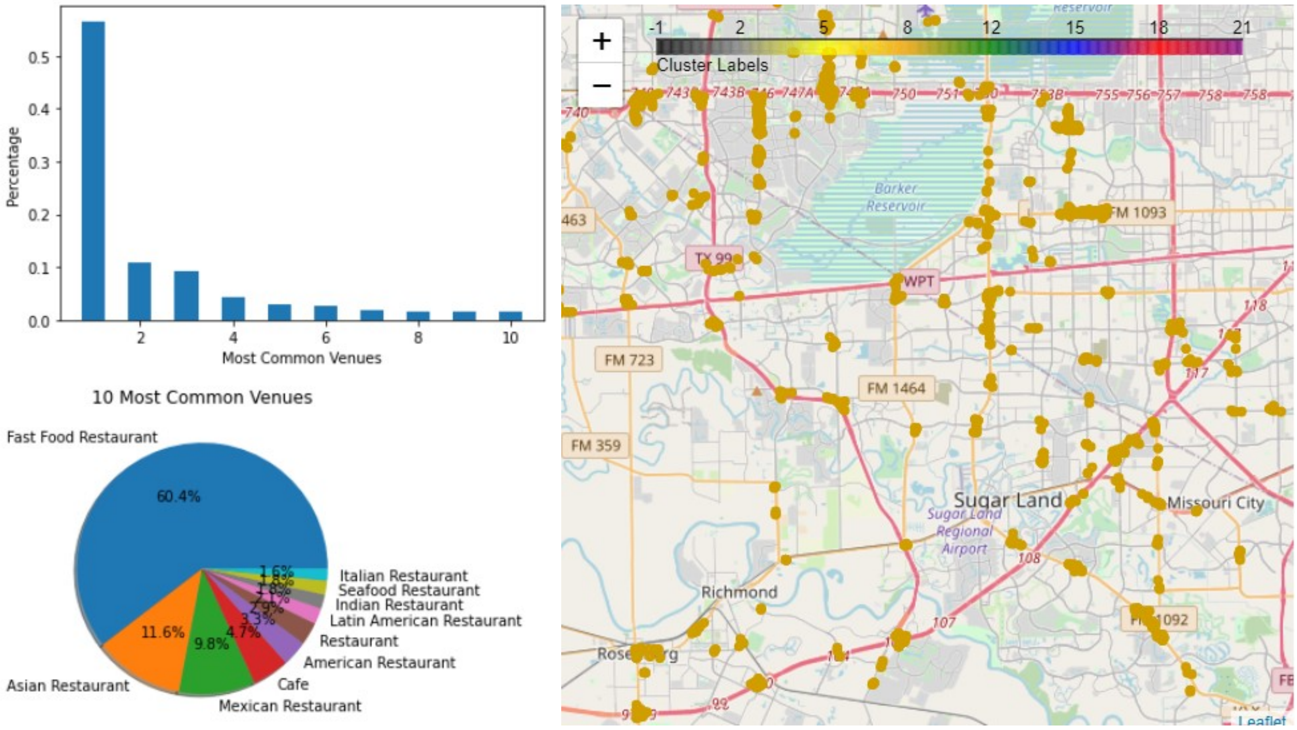


Fig. 12. Restaurants in cluster 9

Category 9 is very popular also. There are 60.4% Fast food restaurants and 11.6% Asian restaurants. All these restaurants are mainly located along highways.

VI. Conclustion

For a new business, the most important factor is choosing a right place to start the business. This is more crucial and a very big challenge for restaurant business because there are some many restaurants in every city. This project tries to use data science algorithm (K-Means) to analyze current restaurant distribution in West Houston area, and find out potential locations for a new Asian restaurants in Sugar Land and Missouri City.

After analyzing all the clusters with Asian restaurants in 1st and 2 most-common-venues, and considering the percentage/number of restaurants in each cluster, cluster 1,7,and 9 seem to be the best clusters for a new Asian restaurant in Sugar Land/Missouri area. Most of restaurants in these clusters are opened along highway 59 and 6. Especially, cluster 7 has the highest percentage of Asian restaurants and also covers Houston China town, Katy Asian Town and a plaza with 5 Asian restaurants near Highway 6 and Austin Pkwy in Sugar Land. Therefore that plaza or a nearby location in Sugar Land might be a perfect place for opening a similar Asian restaurant as restaurants in Houston China town and Katy Asin Town. Cluster 1 and 9 areas are also suitable for a new Asian restaurant business and more challenges will be expected with high percentage of fast-food restaurants.

References

[1] Sugar Land, Wikipedia https://en.wikipedia.org/wiki/Sugar_Land,_Texas

[2] Missouri City, Wikipedia https://en.wikipedia.org/wiki/Missouri_City,_Texas

[3] Foursquare Venue Category Hierarchy https://developer.foursquare.com/docs/build-with-foursquare/categories/

[4] Selecting the number of clusters with silhouette analysis on KMeans clustering https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

Project Link in Github

[1] Project implentation using python: https://github.com/wdmhouston/CapestoneProject_PredictingLocation/blob/main/CapestoneProject_PredictingLocation.ipynb

[2] Project report: https://github.com/wdmhouston/CapestoneProject_PredictingLocation/blob/main/CapestoneProject_PredictingLocationon_Report.pdf

[3] Project presentation: https://github.com/wdmhouston/CapestoneProject_PredictingLocation/blob/main/CapestoneProject_PredictingLocationon_Presentation.pdf