**Examining popularity of Chinese restaurants in the Houston area**

In Houston, My friends and I are looking to open a Chinese healthy breakfast bar, and the main goal of the project is to find the optimal location.

**Background and Vision**

I am concerned that obesity is an increasing trend in the US. Obesity prevalence in Texas is greater than the national average. Obesity and its complications bring enormous health and economic burden to our community. Obesity is generally caused by eating too much fat and sugars and moving too little. Traditional Chinese breakfast foods are usually low in sugar and fat. In order to promote and advocate a healthy lifestyle, I want to open a Chinese healthy breakfast bar in Houston, my hometown. I will focus on developing low sugar snacks and soy milk.

The main goal of the project is to find the optimal location. Since Houston is an immigrant city and has a diverse culture, people here should be open to healthy changes in lifestyles. Houston’s neighborhoods are very diverse and each has a unique character. Therefore, it is expected that different areas will have different levels of demand for Chinese Breakfast Foods. Using Foursquare data, I would like to identify the neighborhoods in which Chinese foods are popular.

**Data sources and description**

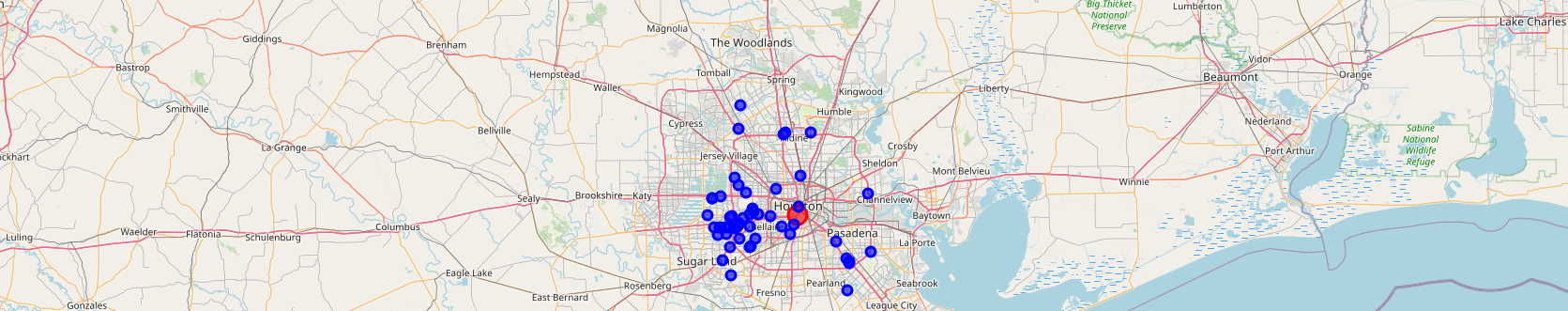
I will explore the foursquare API to extract the data for all Chinese restaurants within a 6 mile radius around my home address. The responses and reviews for each restaurant in different locations will be collected. The responses contain the ratings, the price tiers, and the number of likes. The response data will be integrated with latitude and longitude coordinates values to cluster the current Chinese restaurants by using unsupervised k-means methods. Each cluster would theoretically group together restaurants by geographic proximity and various metrics of performance. By analyzing the average characteristics of each cluster, I hope to be able to pinpoint the area with most potential.

**Methodology section**

K-means clustering is one of the popular unsupervised machine learning algorithms.The objective of K-means is simple: group similar data points together and discover underlying patterns. A cluster refers to a collection of data points aggregated together because of certain similarities. The foursquare API was explored to extract the data for 50 Chinese restaurants within a 6 mile radius of my home address. The responses and reviews for each restaurant in different locations will be collected. The responses contain the ratings, the price tiers, and the number of likes. Missing data were replaced with the average values. The response data will be integrated with latitude and longitude coordinates values to cluster the current Chinese restaurants by using unsupervised k-means methods. I experimented with multiple K values and finally decided to set k=5 to break up large clusters. We then analyzed the average characteristics of each cluster, in order to determine the geographic regions with highest like and rating metrics.

**Results**

Map1: Distribution of the Chinese restaurants in Houston area



Map2: K-means clustering showing 5 clusters

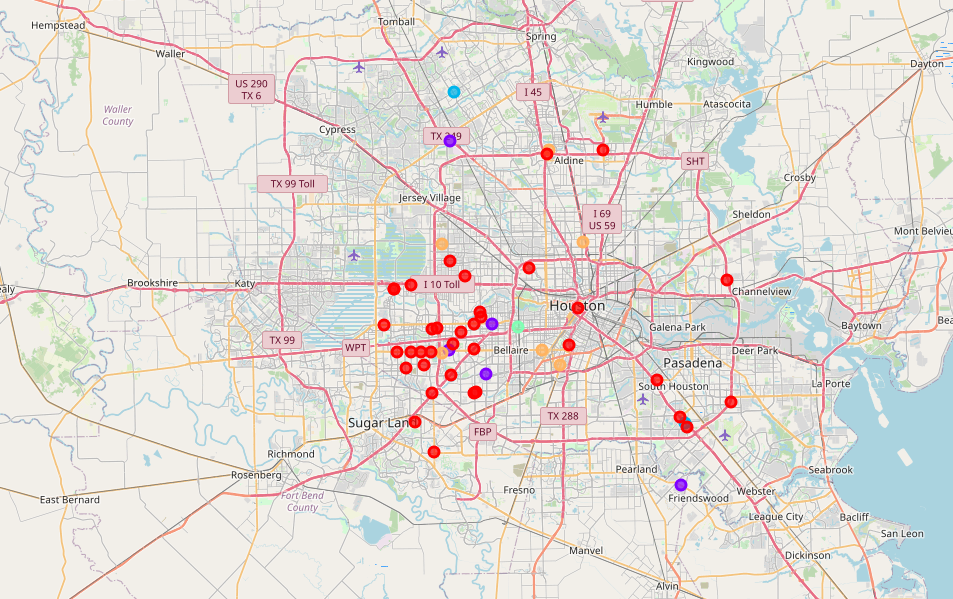


Table 1: The characters of the five clusters

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| cluster | color | size | Latitude | Longitude | Location | Rating | price tier | Likes |
| 0 | Red | 36 | 29.723964 | -95.486482 | Chinatown | 5 | 1.08 | 0.36 |
| 1 | Purple | 5 | 29.723879 | -95.458081 |  | 7.04 | 1.4 | 14 |
| 2 | Blue | 2 | 29.815982 | -95.37841 |  | 7.75 | 1.5 | 21.5 |
| 3 | Green | 1 | 29.732873 | -95.449335 |  | 7.6 | 1 | 29 |
| 4 | Orange | 6 | 29.784847 | -95.446935 | Chinatown | 5.91 | 1.16 | 4.5 |
| total |  | 50 | 29.756309 | -95.4438486 |  | 6.66 | 1.228 | 13.872 |

**Discussion**

1. The numbers of Restaurants are limited because of API day search limits

2. Clusters not strictly limited by geographical proximity because we put several other performance factors into consideration.

3. Chinatown is not the most highly liked or rated cluster.

**Conclusion**

The data suggest that Houstonians have high interest in Chinese food options across the entire city. However, demand seems to be concentrated in specific neighborhoods, such as Midtown and southeast/northwest suburbs. Although the conclusion of the project is that any location has the potential to succeed, and that we should prioritize the quality of food rather than picking an ideal location.

Links to my Notebook on my Github repository, showing my code.

With map view

<https://nbviewer.jupyter.org/github/guoweilleexh123/data-science-capstone-project/blob/master/final%20project-breakfast%20bar%20%281%29.ipynb>

Github

https://github.com/guoweilleexh123/data-science-capstone-project/blob/master/final%20project-breakfast%20bar%20(1).ipynb