Use clustering to find a best neighborhood in the city of Houston for a new Chinese restaurant

By Bing Bai

November 7, 2020

1. Introduction

Houston is located in Southeast Texas near Gulf of Mexico. It is one of the most populous cities in US. Houston is a very diversified and growing city. Its main industries include aerospace and aviation, manufacturing, energy, and life sciences and biotechnology.

Our client is interested in opening a Chinese restaurant in the Houston area. There are more than 80 neighborhoods in Houston. Each one has its own characteristics. Chinese restaurant may be underrepresented in some neighborhood and they desire more Chinese restaurants while some other neighborhoods have little desire or already have enough Chinese restaurants. This study is to find the best neighborhood for the new restaurant to increase the chance of success. In this study, we'll consider Chinese food is one branch of Asian food. We'll check Asian population in the neighborhood as an indicator for demand, number of existing Chinese restaurant to represent competition, and median household income in the neighborhood as buying power.

This report discusses the process used to accomplish our goal, including data acquisition, analysis, preprocessing, clustering, conclusion and discussion. The complete code is available in GitHub.

2. Data

3 sets of data are acquired for this study.

1. I obtain the neighborhood information from houstontx.gov. The website has links to PDF files, one for each neighborhood. I did the web scraping using BeautifulSoup and regular expression to get a list of neighborhoods and their corresponding PDF file.

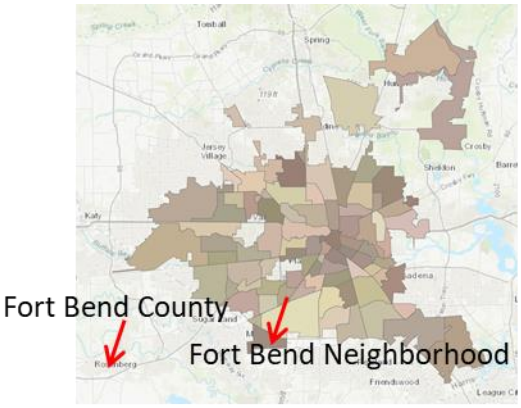
   Then use the table extraction method in Camelot to read the PDF files. The PDF file has both 2000 and 2017 data; we'll only use the 2017 data and extract 'Total Population', 'Population density', 'Ethnicity' and 'Median Household Income'. For Ethnicity, only keep the Asians percentage.

   One item in the PDF file, Median Household Income, has very long name and is very close to the next cell. It creates problems if we try to read in the whole table. I resolved this issue by breaking the table into 2 tables using this line during the read-in process. All the needed information is read in, reformatted and combined afterward.

   After read in all the information needed from the PDF files, I compute neighborhood area by dividing Population with Population density. Then get an estimated radius by assuming all neighborhoods are circle shaped. This will be used later for map display and venue search. We also compute Asian population by multiple Population with Asian percentage. Then discard the extra information, only need Income, radius and Asian population, save to a dataframe.

	Neighborhood	Income	Radius	Pop_a
0	Acres Home	41358	2717	258
1	Addicks Park Ten	82869	4395	2483
2	Afton Oaks / River Oaks	152092	1725	1338
3	Alief	42928	3487	22723
4	Astrodome	51510	1618	6280

2. I extracted the latitude/longitude manually using google earth and cross checking with ArcGIS, the information was saved as an excel file. Some locations have same names. For example, search for Fort Bend returns Fort Bend county location instead of Fort Bend Houston neighborhood. By cross checking the neighborhood map on ArcGIS, it's been correctly updated.

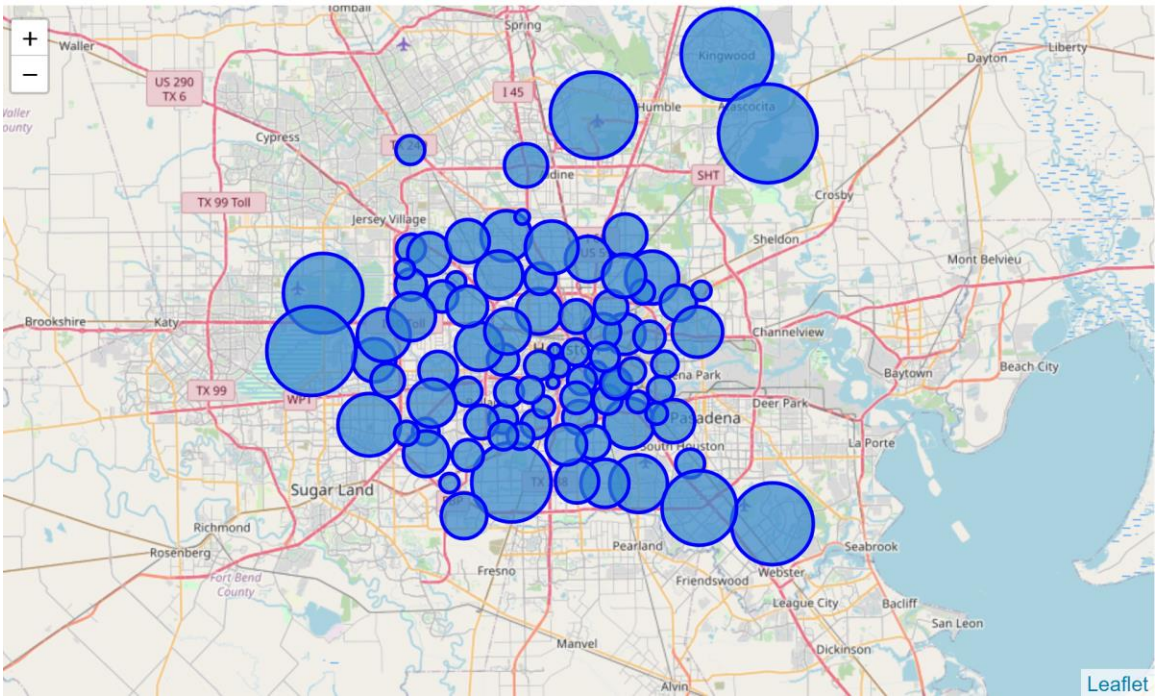


Houston Neighborhood map from ArcGIS

After QC, combine Latitude/Longitude into the original neighborhood dataframe.

	Neighborhood	Income	Radius	Pop_a	Latitude	Longitude
0	Acres Home	41358	2717	258	29.870842	-95.436447
1	Addicks Park Ten	82869	4395	2483	29.814219	-95.645506
2	Afton Oaks / River Oaks	152092	1725	1338	29.748383	-95.440128
3	Alief	42928	3487	22723	29.683789	-95.592861
4	Astrodome	51510	1618	6280	29.684442	-95.403142

Then use folium to display the neighborhoods on Houston map. The size of the circles is determined by the estimated radius.



3. With latitude/longitude, we can extract the existing Chinese restaurant for each neighborhood using Foursquare. Then count the number of Chinese restaurants and add it to the dataframe. The following table showed the final dataframe which contains all the information needed for this study.

	Neighborhood	Income	Radius	Pop_a	Latitude	Longitude	Count
0	Acres Home	41358	2717	258	29.870842	-95.436447	2
1	Addicks Park Ten	82869	4395	2483	29.814219	-95.645506	9
2	Afton Oaks / River Oaks	152092	1725	1338	29.748383	-95.440128	8
3	Alief	42928	3487	22723	29.683789	-95.592861	33
4	Astrodome	51510	1618	6280	29.684442	-95.403142	7

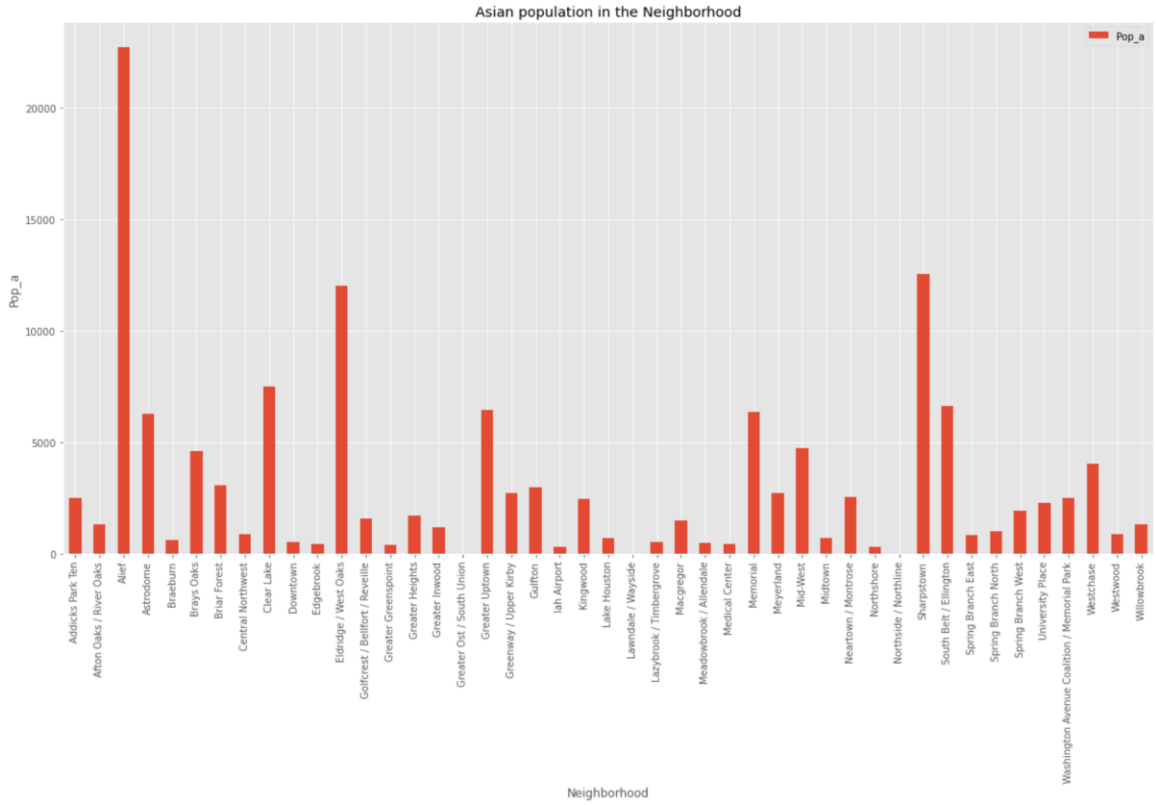
3. Methodology

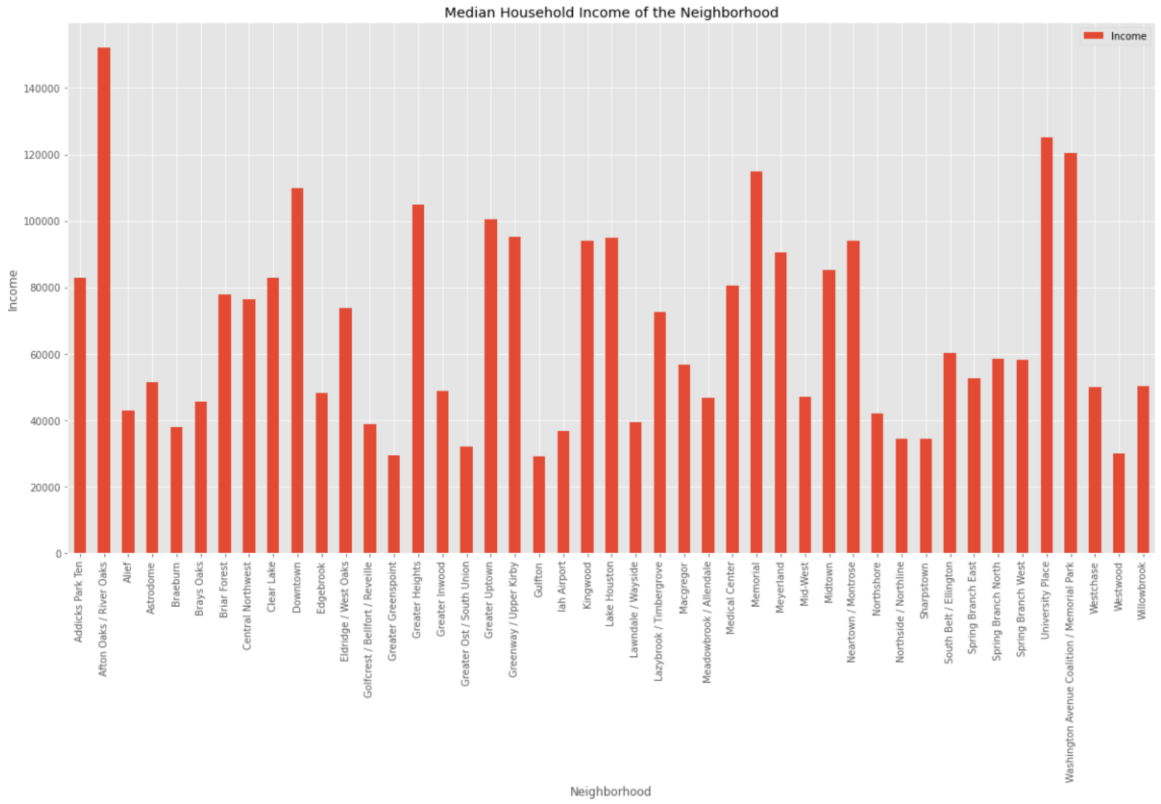
We will mainly use the following data for our analysis: Number of Asian population in each neighborhood, Median household income, and Number of Chinese restaurants. Our hypothesis is that more Asian population will increase the demand for more Chinese restaurants. More average house income suggests stronger buying power and number of existing Chinese restaurants in the neighborhood indicates the competition level.

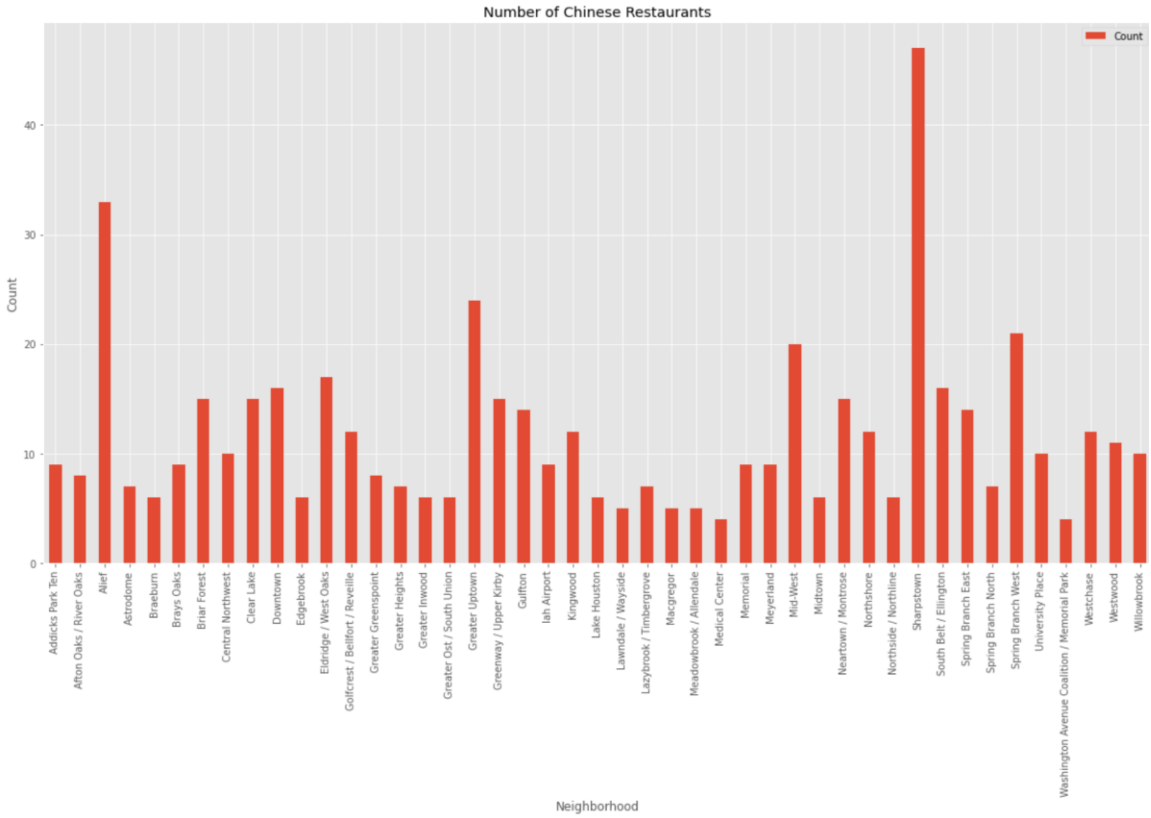
Next, we will generate clusters (use k-means clustering) of the neighborhood to find the ideal neighborhood for opening a new Chinese restaurant.

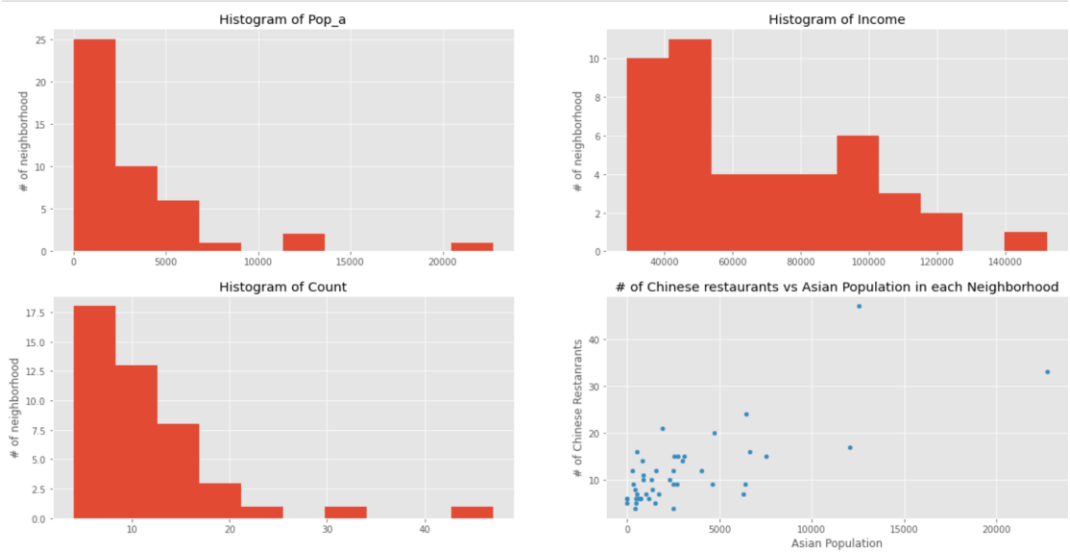
4. Analysis

A large number of neighborhoods have very little Chinese restaurants. We'll consider that it is risky to open a Chinese restaurant there. Let's drop neighborhoods with less than 3 Chinese restaurants. Then we performed some QC to better understand the input data. Histogram QC shows the distribution of the input data. We can also observe a loose correlation between Asian Population and Number of Chinese Restaurants.




Median Household Income of the Neighborhood
Income
Neighborhood
Addicks Park Ten
Afton Oaks / River Oaks
Alief
Astrodome
Braeburn
Brays Oaks
Briar Forest
Central Northwest
Clear Lake
Downtown
Edgebrook
Eldridge / West Oaks
Golfcrest / Bellfort / Reveille
Greater Greenspoint
Greater Heights
Greater Inwood
Greater Ost / South Union
Greater Uptown
Greenway / Upper Kirby
Gulfton
Iah Airport
Kingwood
Lake Houston
Lawndale / Wayside
Lazybrook / Timbergrove
Macgregor
Meadowbrook / Allendale
Medical Center
Memorial
Meyerland
Mid-West
Midtown
Neartown / Montrose
Northshore
Northside / Northline
Sharpstown
South Belt / Ellington
Spring Branch East
Spring Branch North
Spring Branch West
University Place
Washington Avenue Coalition / Memorial Park
Westchase
Westwood
Willowbrook


Number of Chinese Restaurants
Count
Neighborhood
Addicks Park Ten
Afton Oaks / River Oaks
Alief
Astrodome
Braeburn
Brays Oaks
Briar Forest
Central Northwest
Clear Lake
Downtown
Edgebrook
Eldridge / West Oaks
Golfcrest / Bellfort / Reveille
Greater Greenspoint
Greater Heights
Greater Inwood
Greater Ost / South Union
Greater Uptown
Greenway / Upper Kirby
Gulfton
Iah Airport
Kingwood
Lake Houston
Lawndale / Wayside
Lazybrook / Timbergrove
Macgregor
Meadowbrook / Allendale
Medical Center
Memorial
Meyerland
Mid-West
Midtown
Neartown / Montrose
Northshore
Northside / Northline
Sharpstown
South Belt / Ellington
Spring Branch East
Spring Branch North
Spring Branch West
University Place
Washington Avenue Coalition / Memorial Park
Westchase
Westwood
Willowbrook



Since different type of input data has different range. We'll first normalize each one before clustering.

```
X = df_hou_select.drop(['Neighborhood','Latitude','Longitude'], axis=1)
X= preprocessing.StandardScaler().fit(X).transform(X)
X[0:5]
```

```
array([[-0.13753389,  0.51790934, -0.3400102 ],
       [-0.41491114,  2.8280276 , -0.46751403],
       [ 4.76562372, -0.8150065 ,  2.7200816 ],
       [ 0.78229266, -0.52860696, -0.59501785],
       [-0.59223878, -0.98350233, -0.72252168]])
```

Then we use K-means clustering to divide the neighborhoods into 6 groups. To better understand the characteristic of each group, we categorize each input data into 5 ranges: 'Low', 'Lowmid', 'Mid', 'Midhigh', and 'High'.

5. Results

The neighborhood in Houston was clustered into 6 clusters, and we summarized the characteristics for each cluster here.

Cluster	0	1	2	3	4	5
Asian population	Low	Low	Mid	High	Mid	Low
Income	High	Low	Mid	Low	Low	Mid
# of Chinese restaurants	Low	Low	Mid	High	High	Low

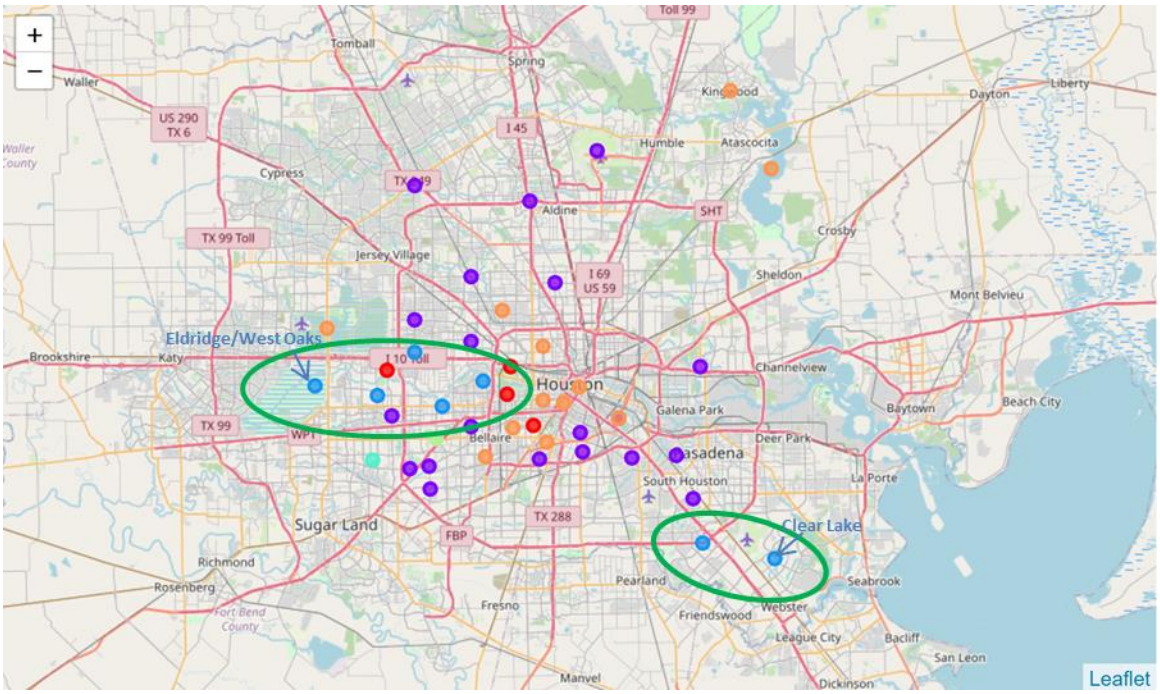
We consider Asian population as the target customer for the new restaurant. Low Asian population (cluster 0, 1, 5) is not ideal. Cluster 3 and 4 are similar, they have reasonable number of target customers and large number of competitions. It kind of supports our assumption that more Asian population will have more demand for Asian food. But clusters 3 and 4 both have low income, which

suggests that the demand is mostly for low-price restaurants. And the competition is high for these neighborhoods. If our stakeholder have authentic food, competitive advantage and relative low cost, cluster 3 (only one neighborhood in this cluster, Alief) is still a reasonable choice for opening a Chinese restaurant.

Cluster 2 Neighborhoods

	Neighborhood	Latitude	Longitude	Pop_a	Income	Count	binned_Pop_a	binned_Income	binned_Count
0	Briar Forest	29.747739	-95.587269	3064	77819	15	low	lowmid	lowmid
1	Clear Lake	29.585614	-95.132725	7509	82744	15	lowmid	mid	lowmid
2	Eldridge / West Oaks	29.756956	-95.659072	12037	73874	17	mid	lowmid	lowmid
3	Greater Uptown	29.761469	-95.467242	6450	100485	24	lowmid	mid	mid
4	Mid-West	29.737519	-95.514211	4725	47138	20	lowmid	low	lowmid
5	South Belt / Ellington	29.601289	-95.216136	6611	60278	16	lowmid	lowmid	lowmid
6	Spring Branch West	29.790917	-95.544683	1920	58305	21	low	lowmid	lowmid

Cluster 2 has reasonable number of target customers, spending power, which is promising. Mid level of existing Chinese restaurants suggests that there is demand and the competition is not too severe. We would recommend cluster 2 to our stakeholder, 2 neighborhoods to be more specific: Clear Lake and Eldridge/West Oaks. The following map shows the clustering; our preferred group (#2) is mostly within the circled area.



6. Conclusion

In this study, I labeled the neighborhoods in Houston into 6 groups using the following criteria: spending power, target population and number of competitors. Best on this clustering, I recommended cluster 2 to the stakeholder, which has reasonable target population and spending power, and the competition is not severe. To be more specific, 2 neighborhoods from this cluster were recommended: Clear Lake and Eldridge/West Oaks.

7. Discussion

The assumption we used in this study may be overly simplified. Here are some limitations of this study:

1. Asian population may not always prefer Chinese food. And some may crave for a specific branch of Chinese food, for example: Szechuan or Hunan restaurant, not other branches. Further study is needed.
2. Use Median household Income as indicator for spending power does not consider other living costs, such as house price etc. And this study does not consider people going to nearby neighborhood for food.
3. Other factors, such as rent, safety, are not considered in this study.

To further improve this study, we can study the demand for the specify branch of Chinese food if we get more detailed information from our stakeholder. Can and rent cost to the study if the data is available. Use more recent data and more data point from past few years, to understand if there is a growing demand or shrinking demand. Moreover, this study improves the chances of success for our stakeholder, but the food quality and services are still the key to attract and maintain loyal customers.

8. References and resources

- Houston neighborhood information: https://www.houstontx.gov/planning/Demographics/super_neighborhoods_2.html
- Neighborhood map on ArcGIS: https://www.arcgis.com/home/webmap/viewer.html?webmap=e87cdc21ac3a43ecb2cdf2c31d75ca8e
- Google earth, Foursquare, Anaconda, GitHub
- Documents and codes for this project: https://github.com/aggiebane/Capstone_Houston_Clustering

Thank you for looking at this study. It is the final project for Coursera class: Applied Data Science Capstone. Any suggest and question is welcomed.