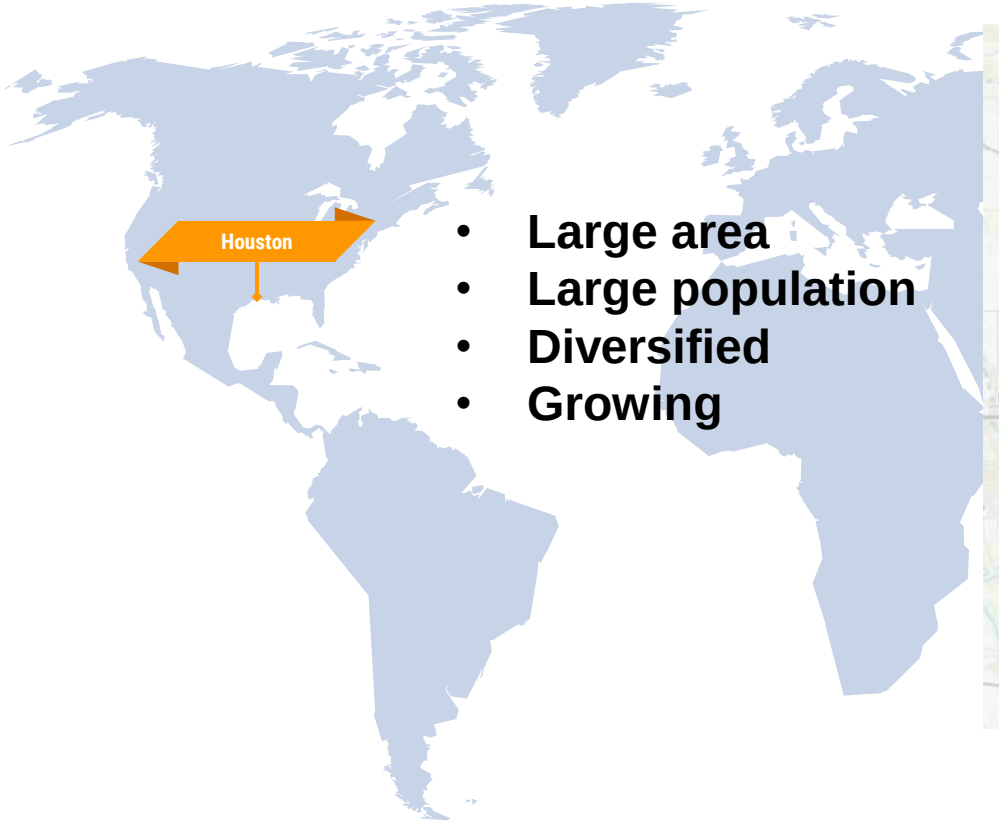
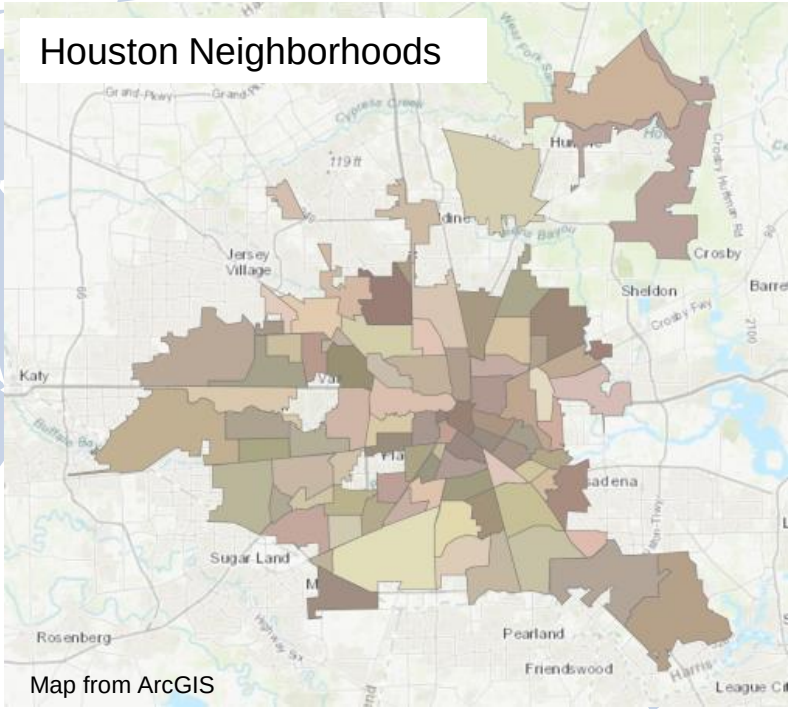
Use clustering to find a best neighborhood in the city of Houston for a new Chinese restaurant

City of Houston



- **Large area**
- **Large population**
- **Diversified**
- **Growing**

Houston Neighborhoods



Map from ArcGIS

Input needed for this study

Asian Population

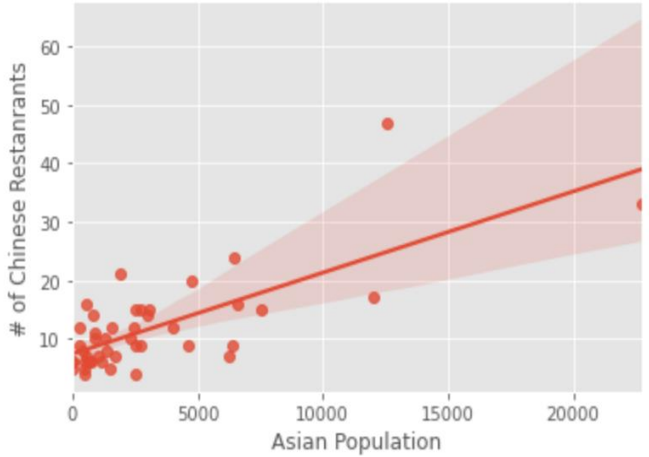
Assume Asian prefers Chinese food

Median Household Income

Larger income suggests more buying power

of Chinese restaurant

Competition

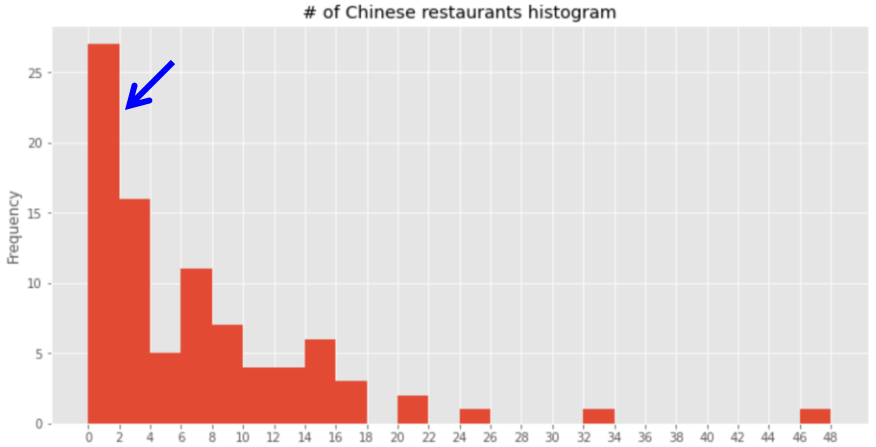


Data source

- Neighborhood information scraped from houstontx.gov
- Location of Neighborhood from google earth, ArcGIS
- Venue information from Foursquare

Data selection

A large number of neighborhoods have very little Chinese restaurants. We'll consider it as risky and drop neighborhoods with less than 3 Chinese restaurants



Data normalization

Input Data Range

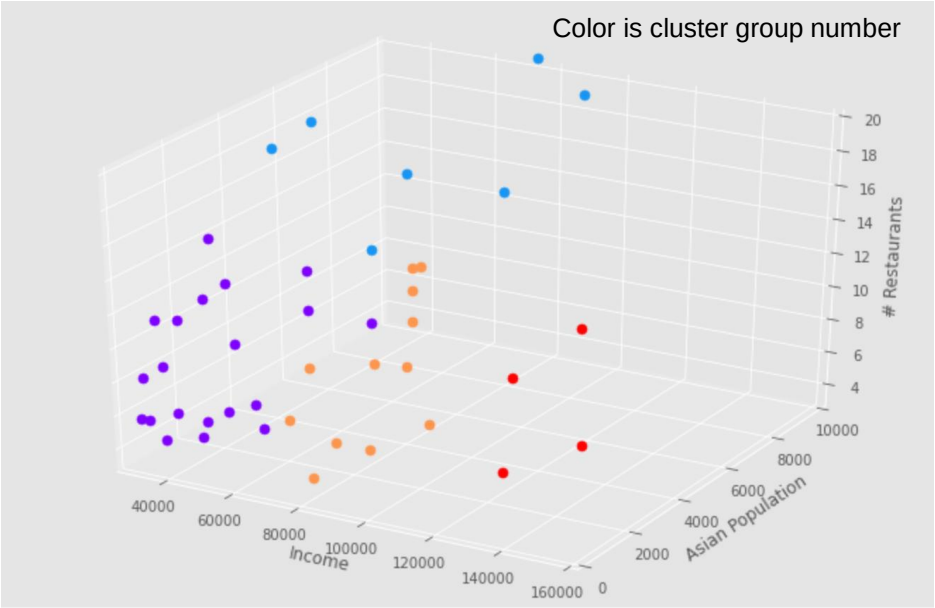
	Pop_a	Income	Count
count	45.000000	45.000000	45.000000
mean	3050.733333	67349.777778	11.666667
std	4174.597157	30303.732516	7.931525
min	0.000000	29124.000000	4.000000
25%	606.000000	42928.000000	6.000000
50%	1573.000000	58305.000000	9.000000
75%	3064.000000	90626.000000	15.000000
max	22723.000000	152092.000000	47.000000

```
[[  2483  82869      9]
 [  1338 152092      8]
 [ 22723  42928     33]
 [  6280  51510      7]
 [   606  37879      6]]
```

```
[[-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]]
```

K-means clustering

- Unsupervised learning
- Group data into K clusters and discover underlying patterns.



Clustering result

Characters of each cluster

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

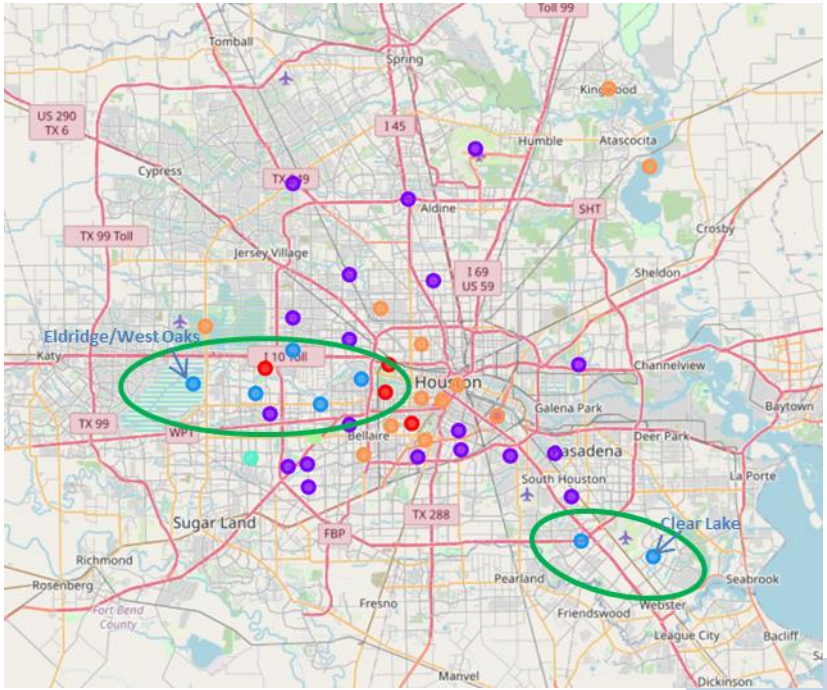
- Cluster 0, 1, 5: Low target customers. Not recommended.
- Cluster 3, 4: High target customers, low income and high competition. Could be a reasonable choice if opening a low-cost, highly competitive restaurant.
- Cluster 2 has reasonable number of target customers, spending power. Mid level of existing Chinese restaurants suggests that there is demand and the competition is not too severe.
- Cluster 2 is recommended

Conclusion

Characters of each cluster

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

Recommend cluster 2 for reasonable number of target group, buying power and reasonable competitions.
Clear Lake and Eldridge/West Oaks to be more specific.



Limitation and Discussion for Future Research

- Target group can be fine tuned.
- Could be more specific about the restaurant style (sub-branch of Chinese restaurant)
- Median household Income does not full reflect spending power for restaurants.
- Additional factors can help further improve the result:
  - Rent cost
  - More recent data
  - More data points from the post few years
  - Impact of COVID-19?

References

- 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