Coursera Capstone Project

Restaurant preference

Agenda

Introduction and Business value

Data and methodology

Result and conclution

Introduction and Business value

There are many restaurants in US. As one of the most international countries in the world, restaurants in US are very diversified. There are great potential opportunities in restaurant industry.

Restaurant owners are interest in below topics:

1. What is most popular food in US?
2. Is there any trend in food preference among cities in US?
3. If I'm considering expanding my chain to another city, what is good choice for me?

Data

Where is data from？

- Information of major US cities https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population.
- Foursquare location data

How to handle data?

- Data cleaning -> data in same format
- Normalize data -> data in same scale
- Data filtering ->only keep relavant data

	City	2019estimate	2016 population density	lantitude	longitude	American Restaurant	Bar
0	New York	1.000000	1.000000	40.6635	73.9387	0.000000	0.000000
1	Los Angeles	0.477350	0.295253	34.0194	118.4108	0.086957	0.000000
2	Chicago	0.323142	0.416637	41.8376	87.6818	0.074074	0.000000
3	Houston	0.278316	0.122166	29.7866	95.3909	0.052632	0.052632
4	Phoenix	0.201635	0.104648	33.5722	112.0901	0.200000	0.000000

K-means clustering and correlation

Correlation

$$\rho_{X,Y} = \frac{\mathcal{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

K-means clustering

- Initialize cluster centroid
- Assign data to corresponding centroid
- Update centroid

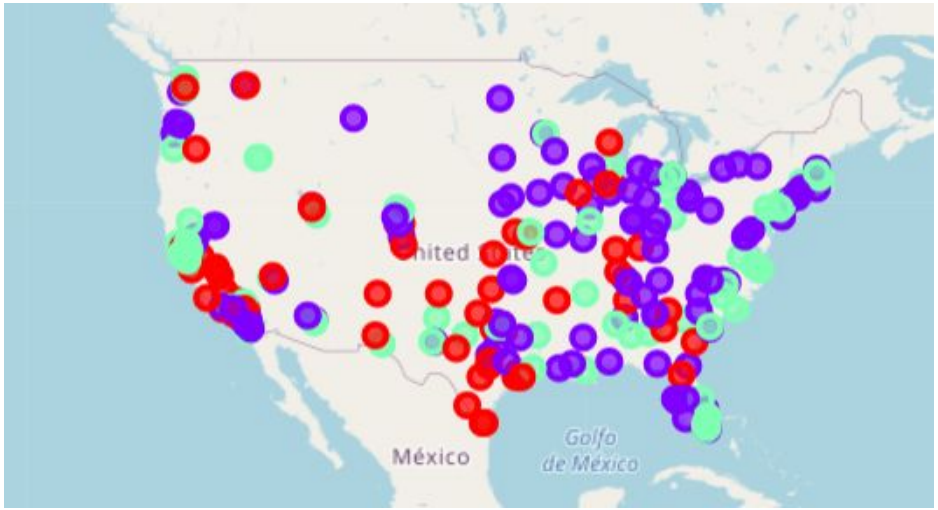
Results

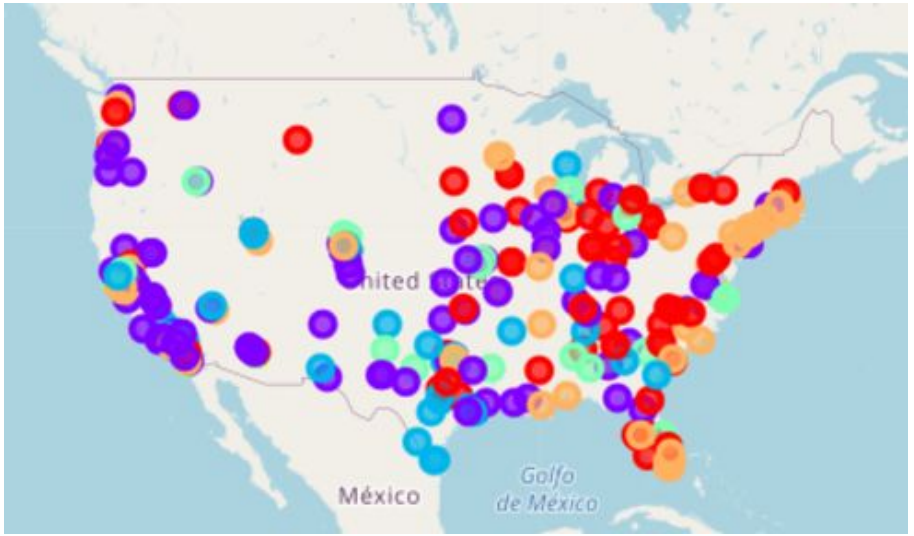
Correlation between restaurant categories

	Mexican Restaurant	Chinese Restaurant	American Restaurant	Italian Restaurant	Thai Restaurant	Vietnamese Restaurant	Japanese Restaurant	Pizza Place	Indian Restaurant	Seafood Restaurant	Bar	Sushi Restaurant
Mexican Restaurant	1.000000	-0.309054	-0.269092	-0.233103	-0.150031	-0.120604	-0.121465	-0.260768	-0.126263	-0.112146	-0.156490	0.014794
Chinese Restaurant	-0.309054	1.000000	-0.264826	-0.189202	-0.120468	-0.030998	-0.152485	0.075209	-0.127642	-0.011854	-0.096827	-0.089321
American Restaurant	-0.269092	-0.264826	1.000000	0.004084	-0.004291	-0.120740	0.058407	-0.108653	-0.051484	-0.120967	0.053275	-0.080084
Italian Restaurant	-0.233103	-0.189202	0.004084	1.000000	0.024596	-0.157952	-0.020287	0.028273	-0.070676	-0.047519	-0.083204	-0.072576
Thai Restaurant	-0.150031	-0.120468	-0.004291	0.024596	1.000000	0.017922	0.237825	-0.062025	0.003348	-0.056393	0.000127	-0.013346
Vietnamese Restaurant	-0.120604	-0.030998	-0.120740	-0.157952	0.017922	1.000000	-0.081756	-0.107052	0.035497	0.096621	0.010563	-0.045085
Japanese Restaurant	-0.121465	-0.152485	0.058407	-0.020287	0.237825	-0.081756	1.000000	-0.046815	0.046411	-0.100998	-0.042725	0.034917
Pizza Place	-0.260768	0.075209	-0.108653	0.028273	-0.062025	-0.107052	-0.046815	1.000000	0.040533	-0.073359	-0.032501	0.011225
Indian Restaurant	-0.126263	-0.127642	-0.051484	-0.070676	0.003348	0.035497	0.046411	0.040533	1.000000	0.058139	-0.056013	-0.020546
Seafood Restaurant	-0.112146	-0.011854	-0.120967	-0.047519	-0.056393	0.096621	-0.100998	-0.073359	0.058139	1.000000	-0.025251	-0.090539
Bar	-0.156490	-0.096827	0.053275	-0.083204	0.000127	0.010563	-0.042725	-0.032501	-0.056013	-0.025251	1.000000	-0.079702
Sushi Restaurant	0.014794	-0.089321	-0.080084	-0.072576	-0.013346	-0.045085	0.034917	0.011225	-0.020546	-0.090539	-0.079702	1.000000
Caribbean Restaurant	-0.245829	0.060640	-0.097263	-0.085834	-0.075796	-0.084579	-0.096245	0.089202	-0.069726	-0.027037	0.019129	-0.077490
Breakfast Spot	-0.056721	-0.086416	-0.016878	-0.013030	-0.023171	-0.055523	-0.039783	-0.026596	-0.000402	-0.056756	0.057384	-0.109884
Latin American Restaurant	-0.146549	-0.020053	-0.086118	-0.035365	-0.061499	-0.081233	-0.027347	0.228618	-0.075727	-0.005401	-0.024894	0.000590

Results

Clustering result





Conclusion

1. Mexican, Chinese and American restaurants are most popular restaurant categories.
2. Japanese and Thai, American and Bar, Italian and Pizza show positive correlation.
3. Cluster of cities is related with locations although location data is not used in clustering.
4. If k=3, cities can be dividled into Mexican dominanted, American dominated and diversified.
5. If k=5, one of the group is dominated by Mexican food with limited space for other restaurants. Another group is dominated by Chinese food. With rest of clusters slightly prefer American, Chinese food and very diversified.