Preference of restaurants of US major cities

1. Introduction

Some of the restaurants are operating locally. The owner of the restaurants might considering expanding their business to other cities. The restaurant might be very successful in one city and not very popular in the other city. The restaurant owner would like to know which cities are good targets to expand the business.

For example, a pizza restaurant currently operating in city A in US is very successful and would like to find a target city to open a new chain. Our project's goal is to cluster major cities of US based on preference of restaurant categories. For example, some cities have more fast food restaurant so residents there prefer to eat more fast food. Some cities are more international cities so there a a lot of international food. Owner of this pizza restaurant can get a list of cities similar to city A and consider open a restaurant there.

Extra information like popular restaurant categories, trend of restaurant numbers in cities will be provided.

Target group of interest will be people in restaurant industry. They can use the information to improve their business.

2. Data

2.1 Raw data

In order to clustering preference of restaurants in different cities in US, first step is to get a list of major cities and their location. I get data from this wiki page https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population.

It can be seen the page contains essential information like city name and location. Some extra information like state, population are also included.

2.2 Data cleaning and transformation

Columns of city name, density , population, latitude and longitude are kept from table. Population, and population density are very important in determining whether opening a restaurant in a city is a good idea or not.

Data cleaning is done to extract latitude, longitude from one columns to two separate columns. All numerical data are transferred to float format.

Population and density data are normalized to make cluster better performed. It can be seen that NY, the largest city have density/population value of 1 in final table. Other cities have value between 0 and 1.

2.3 Utilizing fourSquare to get restaurant information

Location data and foursquare are used to get restaurant information in each city. Command to search for "restaurant" related venues within 5k of center of city is used. We get all related restaurant info and only keep category and restaurant names. Returned results contains venue data. Only venue category is kept and a list of restaurant categories in each city is get.

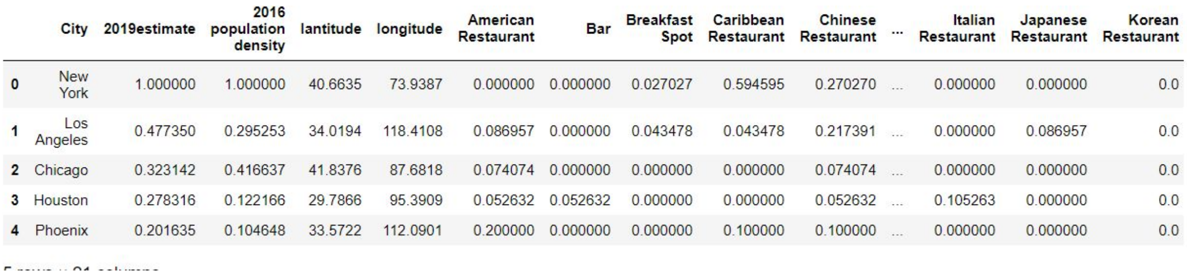
For example in notebook, it can be seen that in city Abilene, Chinese food is very popular while in city Akron Italian food is very popular.

It can be seen from result table that some irrelevant venue categories like “kitchen supply” is included. Those information is unrelated to our study of restaurant preference. It can also be seen some ambiguous category like “food” and “dinner” are included. Only top restaurants categories with frequency larger than 100 are kept and we make sure those are clear and detailed categories. Other venue rows are deleted. Mexican restaurant, Chinese restarant, American restaurant are example of restaurant categoris kept.

2.4 Finailze data for clustering

Total number of restaurants we have for each city is different. It is not good idea to use restaurant numbers directly for clustering, they are not in same scale. We will get frequancy of restaurant categories in each city to better perform clustering.

Restaurant categories and city data are merged together. Features used for clustering are frequency of restaurant categories, population, density. Location data is excluded for clustering since we focus more on preference and city characteristic instead of location. Final data table can be seen in below image.



	City	2019estimate	2016 population density	lantitude	longitude	American Restaurant	Bar	Breakfast Spot	Caribbean Restaurant	Chinese Restaurant	...	Italian Restaurant	Japanese Restaurant	Korean Restaurant
0	New York	1.000000	1.000000	40.6635	73.9387	0.000000	0.000000	0.027027	0.594595	0.270270	...	0.000000	0.000000	0.0
1	Los Angeles	0.477350	0.295253	34.0194	118.4108	0.086957	0.000000	0.043478	0.043478	0.217391	...	0.000000	0.086957	0.0
2	Chicago	0.323142	0.416637	41.8376	87.6818	0.074074	0.000000	0.000000	0.000000	0.074074	...	0.000000	0.000000	0.0
3	Houston	0.278316	0.122166	29.7866	95.3909	0.052632	0.052632	0.000000	0.000000	0.052632	...	0.105263	0.000000	0.0
4	Phoenix	0.201635	0.104648	33.5722	112.0901	0.200000	0.000000	0.000000	0.100000	0.100000	...	0.000000	0.000000	0.0

5 rows × 21 columns

3. Methodology

3.1 Correlation data

A correlation matrix is derived from frequency of restaurant categories of different cities to find if there is any trend between different restaurant categories.

It is similar to applying regression with one restaurant category as x and the other restaurant category as y. It can be seen if there is any trend between restaurant categories.

Formula of correlation is

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

3.2 Clustering using k means method

K means methodology is performed for clustering. The data contains over 15 dimensions of features and k means methodology is good for large data set.

We are not sure how many clusters should be used. Two k values 3 and 5 will all be tried.

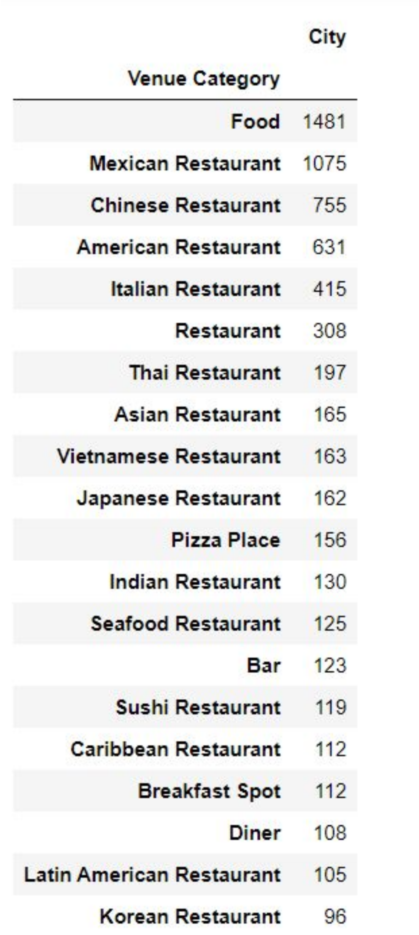
The basic concept for k means method is first randomly choose centroid of clusters and assign the data to different clusters according to distance. Centroids are updated with mean of data in the group. The process is repeated untill converge.

“Sklearn” library is used to perform k means clustering in code.

4. Results

4.1 Most popular restaurants.

By calculating number of restaurants in each category, top 15 restaurant categories are shown below. It can be seen that Mexican, Chinese, American and Italian restaurants are most popular restaurants.

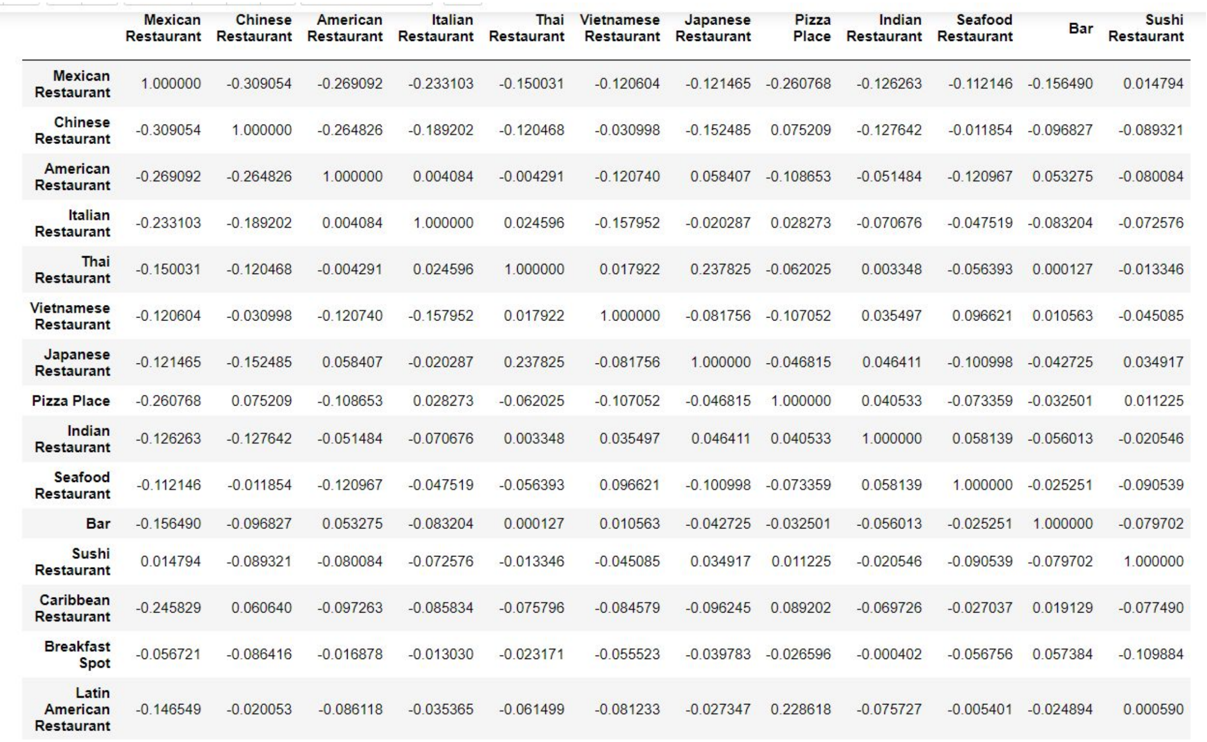


	City
Venue Category	
Food	1481
Mexican Restaurant	1075
Chinese Restaurant	755
American Restaurant	631
Italian Restaurant	415
Restaurant	308
Thai Restaurant	197
Asian Restaurant	165
Vietnamese Restaurant	163
Japanese Restaurant	162
Pizza Place	156
Indian Restaurant	130
Seafood Restaurant	125
Bar	123
Sushi Restaurant	119
Caribbean Restaurant	112
Breakfast Spot	112
Diner	108
Latin American Restaurant	105
Korean Restaurant	96

4.2 Trends in restaurants

Correlation matrix of restaurant frequancy in different cities can be seen below. Since in each city, sum of frequency of restaurants is 1, generally we see negative correlation between restaurants.

It is interesting to see that Japanese restaurant and Thai and Japanese restaurants are positively correlated. American restaurants and bar are positively correlated. Italian restaurants and pizza place are positively correlated.

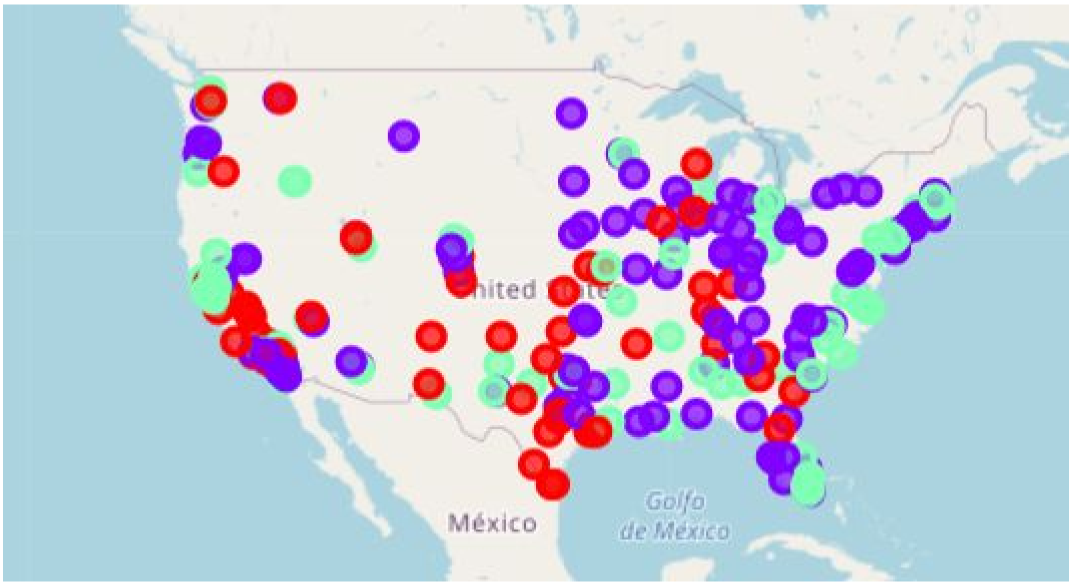


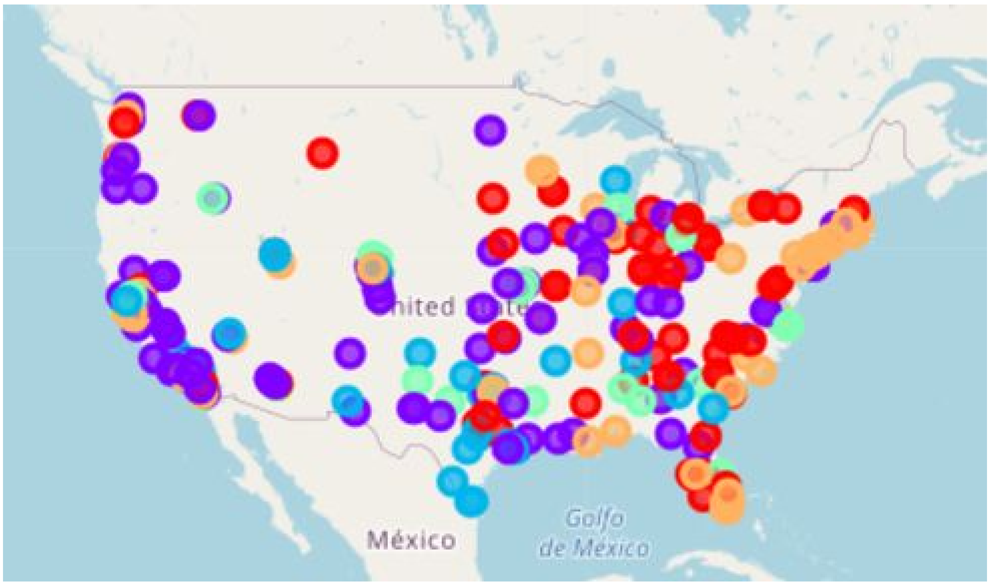
	Mexican Restaurant	Chinese Restaurant	American Restaurant	Italian Restaurant	Thai Restaurant	Vietnamese Restaurant	Japanese Restaurant	Pizza Place	Indian Restaurant	Seafood Restaurant	Bar	Sushi Restaurant
Mexican Restaurant	1.000000	-0.309054	-0.269092	-0.233103	-0.150031	-0.120604	-0.121465	-0.260768	-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

4.3 Clustering result

Below are clustering result with k=3 and k=5.

Location data is actually not a feature in clustering but it can be seen that 0 cluster appear mostly in middle south area while number cluster 3 appear most in eastern coast for k=3. 0 and 1 categories appear most commonly in eastern coast and number 3 and 4 category appears most frequently in western coast. It can be seen people’s preference of restaurants are correlated with geography locations.

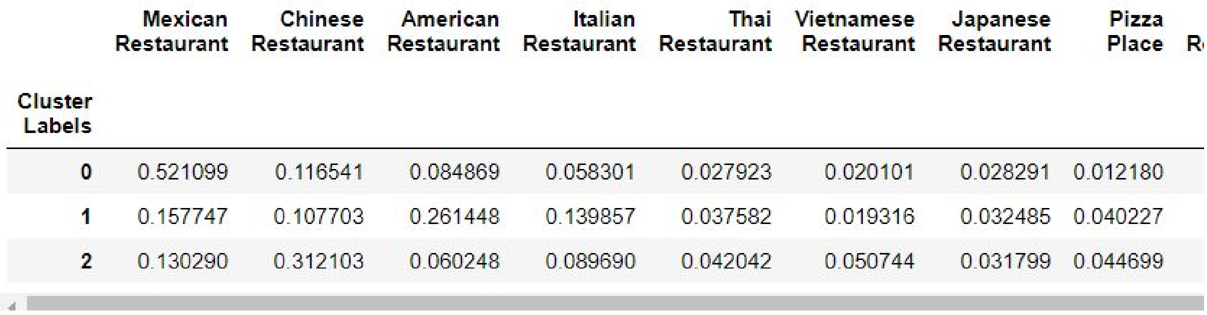




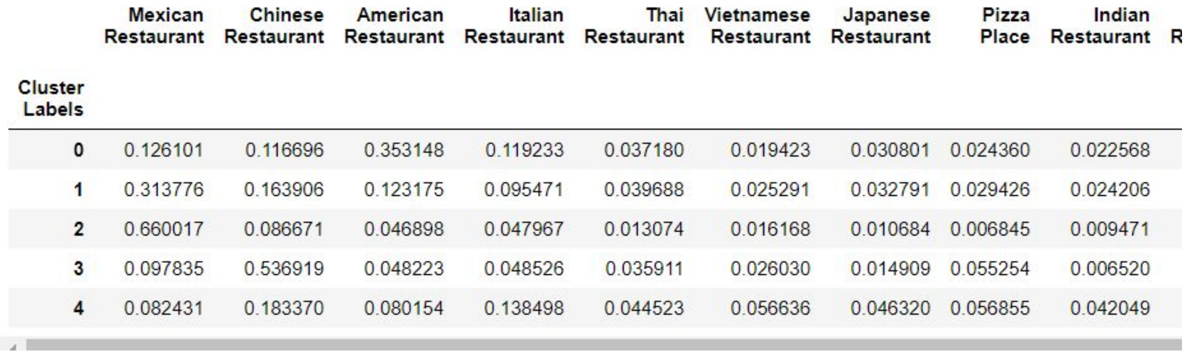
Looking at average frequency group by cluster result, result can seen in table below.

It can be seen that for k=3, Mexican food is very popular in cluster 1. Cluster 2 cities have prefer more American restaurant but overall have diverse preference Cluster 3 cities prefer Chinese restaurant most.

For k=5, cluster 0 prefer American food mostly but have relatively diverse preference. Cluster 1 prefer Mexican food most but also have a relatively diverse preference. Cluster 2 is big fun for Mexican food with little room for other type of food to operate. Cluster 3 prefer Chinese restaurant most and have very low frequency for rest of categories.Cluster 4 has most diverse preference with relatively average frequency for each restaurant type.



Cluster Labels	Mexican Restaurant	Chinese Restaurant	American Restaurant	Italian Restaurant	Thai Restaurant	Vietnamese Restaurant	Japanese Restaurant	Pizza Place	R
0	0.521099	0.116541	0.084869	0.058301	0.027923	0.020101	0.028291	0.012180	
1	0.157747	0.107703	0.261448	0.139857	0.037582	0.019316	0.032485	0.040227	
2	0.130290	0.312103	0.060248	0.089690	0.042042	0.050744	0.031799	0.044699	



Cluster Labels	Mexican Restaurant	Chinese Restaurant	American Restaurant	Italian Restaurant	Thai Restaurant	Vietnamese Restaurant	Japanese Restaurant	Pizza Place	Indian Restaurant	R
0	0.126101	0.116696	0.353148	0.119233	0.037180	0.019423	0.030801	0.024360	0.022568	
1	0.313776	0.163906	0.123175	0.095471	0.039688	0.025291	0.032791	0.029426	0.024206	
2	0.660017	0.086671	0.046898	0.047967	0.013074	0.016168	0.010684	0.006845	0.009471	
3	0.097835	0.536919	0.048223	0.048526	0.035911	0.026030	0.014909	0.055254	0.006520	
4	0.082431	0.183370	0.080154	0.138498	0.044523	0.056636	0.046320	0.056855	0.042049	

5. Discussion

   It is not very surprising that we can find some “restaurant bundles”. American restaurants and bars are positively correlated probably because some type of Americans enjoy steak and wine together. It is interesting to see that Japanese restaurants and Thai restaurants have highest correlation. I guess they represent certain type of “Asian food”. Chinese restaurant is not in the “Asian food” bundle probably because there are many American Chinese food, it is very different from traditional Chinese food.

   From k=5 cluster, we found cluster 2 is huge fan for Mexico food and cluster 3 prefer Chinese food most. For cities lie in those clusters, it is hard to start a restaurant in different category. It is also not very good idea to open Mexican or Chinese in those cities because of fierce competition.

   Clustering result can be seen in “cluster_data.csv”. Restaurant oweners can check in detail which cluster each city belongs to. Most cities in Western coast have divergence taste. Cities in Eastern coast slightly prefer more Mexican food. If average price is included in clustering, more detailed results can be found.

6. Conclusion

   In this report, we discussed trend of restaurant preference among major US cities and cluster the cities according to their preference using K means method. Result shows location is relavant to customer’s preference.

   Restaurant owners can refer to cluster result to choose potential cities to open a new restaurant in US.