Battle of the Neighborhoods

Kathryn Haske

September 15, 2019

Introduction

Which metropolitan areas are most similar? A business planning to expand into new countries or regions may want to know which metropolitan areas are similar to the areas in which they already have existing branches. There are different ways to classify similar. In this project, I will classify similar based on the types of venues foursquare returns for top picks in the metro area.

Data

I used data from Wikipedia for the 100 most populated metropolitan areas.

	Metropolitan	Country	Continent
0	Tokyo	Japan	Asia
1	Delhi	India	Asia
2	Shanghai	China	Asia
3	Jakarta	Indonesia	Asia
4	Seoul	South Korea	Asia

I used geopy geocoders Nominatim to get the latitude and longitude for each metro area.

	Metropolitan	Country	Continent	Latitude	Longitude
0	Tokyo	Japan	Asia	35.6828	139.759
1	Delhi	India	Asia	28.6517	77.2219
2	Shanghai	China	Asia	31.2323	121.469
3	Jakarta	Indonesia	Asia	-6.17539	106.827
4	Seoul	South Korea	Asia	37.5667	126.978

I used the Foursquare API to explore the areas 'Top Picks' and store the venue categories returned. The venue categories will be processed and used to classify the metro areas into clusters based on similarity. The first dataframe contains the 10 most popular venue categories for each location.

	Metropolitan	Country	Continent	Latitude	Longitude	Category 1	Category 2	Category 3	Category 4	Category 5	Category 6	Category 7	Category 8
0	Tokyo	Japan	Asia	35.682839	139.759455	Café	Japanese Restaurant	Italian Restaurant	Sushi Restaurant	Lounge	Garden	French Restaurant	Dessert Shop
1	Delhi	India	Asia	28.651718	77.221939	Indian Restaurant	Bar	Café	Flea Market	Lounge	Asian Restaurant	Ice Cream Shop	South Indian Restaurant
2	Mexico City	Mexico	North America	19.432601	-99.133342	Mexican Restaurant	Art Museum	Museum	Arts & Crafts Store	Bar	Ice Cream Shop	Boutique	Bakery
3	São Paulo	Brazil	South America	-23.550651	-46.633382	Brazilian Restaurant	Japanese Restaurant	Café	Bakery	Snack Place	Bookstore	Cosmetics Shop	Asian Restaurant
4	Lagos	Nigeria	Africa	6.455057	3.394179	Lounge	African Restaurant	Bar	Café	Shopping Mall	Pizza Place	Hotel	Art Gallery

There were four metro areas for which the latitude and longitude were not located. I updated these names and was able to obtain the coordinates for these areas.

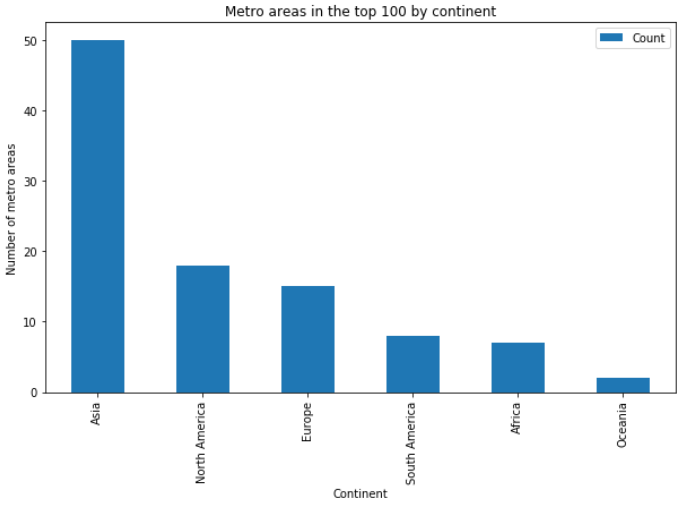
	Metropolitan	Country	Continent	Latitude	Longitude
15	Keihanshin (Kyoto-Osaka-Kobe)	Japan	Asia	NaN	NaN
47	Washington, D.C. - Baltimore	United States	North America	NaN	NaN
53	San Francisco-San Jose-Oakland	United States	North America	NaN	NaN
65	Hong Kong	CHN	Asia	NaN	NaN

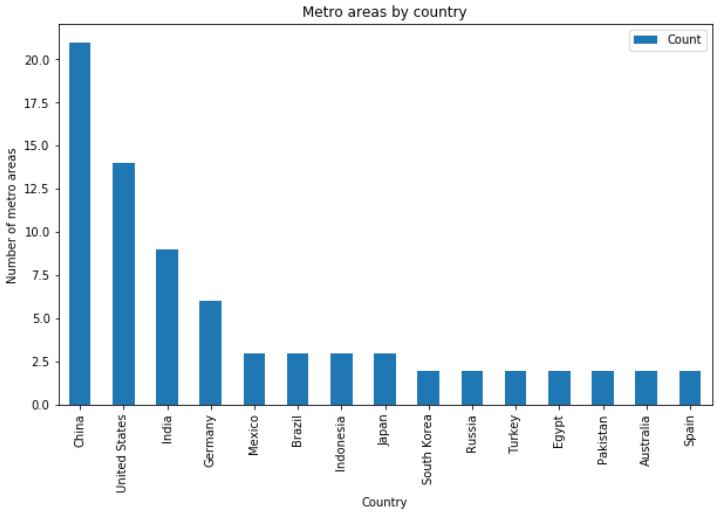
Methodology

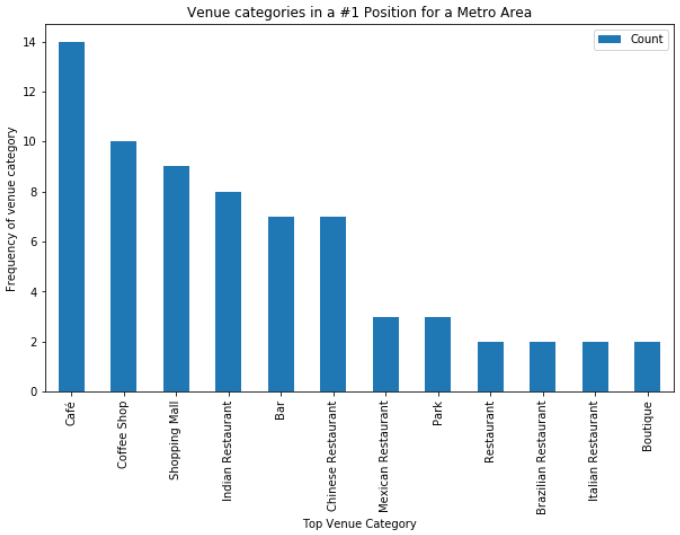
Exploratory Data Analysis

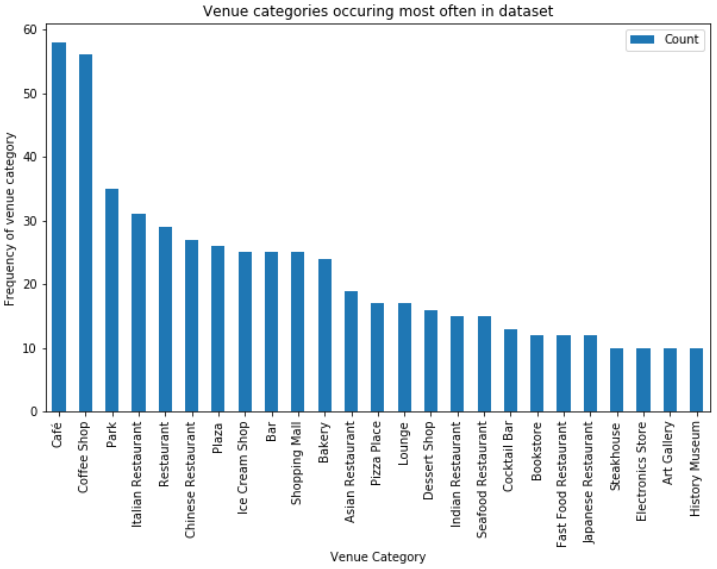
The data consisted of 100 metropolitan areas along with country, continent, latitude, and longitude.

There were 392 unique venue categories returned by Foursquare. I analyzed the 10 most frequently occurring categories for each metro area.









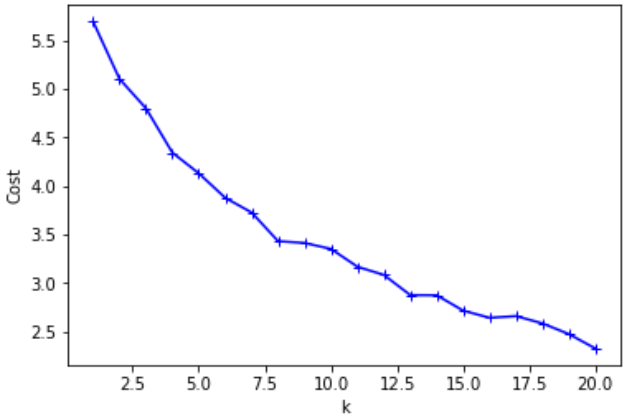
Machine Learning

A second dataframe contains all categories returned along with the ratio of categories for each location. This dataframe will be used for clustering.

	Metro	ATM	Acai House	Accessories Store	Afghan Restaurant	African Restaurant	Alsatian Restaurant	American Restaurant	Amphitheater	Antique Shop	...	Whisky Bar	Wine Bar	Wine Shop	Wings Joint
0	Ahmedabad	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.000000	0.0
1	Alexandria	0.0	0.0	0.0	0.0	0.0	0.0	0.010309	0.0	0.0	...	0.0	0.0	0.000000	0.0
2	Ankara	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.000000	0.0
3	Atlanta	0.0	0.0	0.0	0.0	0.0	0.0	0.019608	0.0	0.0	...	0.0	0.0	0.019608	0.0
4	Bandung	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	...	0.0	0.0	0.000000	0.0

5 rows × 393 columns

I used KMeans to cluster the metro areas into groups. There was not obvious best k (number of clusters) to use, so based on cluster size, I chose k=16.



cat	size
4	21
14	16
1	12
9	9
5	8
13	8
15	7
7	6
12	5
10	2
0	1
2	1
3	1
6	1
8	1
11	1

Cluster Labels and number of metro areas in each cluster

Results

The metro areas were clustered into 16 different clusters using the KMeans method.

Cluster 0

	Metropolitan	Country	Continent
30	Kinshasa	Democratic Republic of the Congo	Africa

Cluster 1

	Metropolitan	Country	Continent
10	Chengdu	China	Asia
12	Shanghai	China	Asia
23	Jakarta	Indonesia	Asia
35	Rhine-Ruhr	Germany	Europe
48	Chūkyō (Nagoya)	Japan	Asia
50	Bandung	Indonesia	Asia
52	Randstad	Netherlands	Europe
53	Busan	South Korea	Asia
67	Manila	Philippines	Asia
77	Berlin/Brandenburg	Germany	Europe
83	Singapore	Singapore	Asia
87	Caracas	Venezuela	South America

Cluster 2

	Metropolitan	Country	Continent
16	Tianjin	China	Asia

Cluster 3

	Metropolitan	Country	Continent
63	Nanchang	China	Asia

Cluster 4

	Metropolitan	Country	Continent
15	Moscow	Russia	Europe
21	London	United Kingdom	Europe
27	Paris	France	Europe
34	Seoul	South Korea	Asia
41	Chicago	United States	North America
42	Washington, D.C.	United States	North America
49	San Francisco	United States	North America
51	Boston	United States	North America
54	Milan	Italy	Europe
58	Riyadh	Saudi Arabia	Asia
60	Wenzhou	China	Asia
62	Hong Kong	China	Asia
65	Philadelphia	United States	North America
70	Santiago	Chile	South America
72	Madrid	Spain	Europe
74	Toronto	Canada	North America
76	Saint Petersburg	Russia	Europe
78	New York City	United States	North America
84	Barcelona	Spain	Europe
91	Hamburg	Germany	Europe
96	Seattle	United States	North America

Cluster 5

	Metropolitan	Country	Continent
1	Delhi	India	Asia
5	Mumbai	India	Asia
28	Bangalore	India	Asia
38	Chennai	India	Asia
57	Hyderabad	India	Asia
61	Pune	India	Asia
75	Ahmedabad	India	Asia
88	Dubai-Sharjah-Ajman	United Arab Emirates	Asia

Cluster 6

	Metropolitan	Country	Continent
20	Tehran	Iran	Asia

Cluster 7

	Metropolitan	Country	Continent
8	Wuhan	China	Asia
37	Jinan	China	Asia
39	Harbin	China	Asia
43	Zhengzhou	China	Asia
59	Shenyang	China	Asia
81	Kanpur	India	Asia

Cluster 8

	Metropolitan	Country	Continent
26	Xi'an	China	Asia

Cluster 9

	Metropolitan	Country	Continent
0	Tokyo	Japan	Asia
3	São Paulo	Brazil	South America
7	Kyoto	Japan	Asia
31	Rio de Janeiro	Brazil	South America
68	Taipei–Keelung	Taiwan	Asia
80	Belo Horizonte	Brazil	South America
82	Frankfurt Rhine-Main	Germany	Europe
85	Ankara	Turkey	Asia
97	Melbourne	Australia	Oceania

Cluster 10

	Metropolitan	Country	Continent
14	Karachi	Pakistan	Asia
19	Kolkata	India	Asia

Cluster 11

	Metropolitan	Country	Continent
64	Yangon	Myanmar	Asia

Cluster 12

	Metropolitan	Country	Continent
22	Hangzhou	China	Asia
24	Surabaya	Indonesia	Asia
29	Changzhou	China	Asia
32	Shantou	China	Asia
56	Beijing	China	Asia

Cluster 13

	Metropolitan	Country	Continent
6	Cairo	Egypt	Africa
11	Dhaka	Bangladesh	Asia
17	Istanbul	Turkey	Europe
47	Bogotá	Colombia	South America
79	Munich	Germany	Europe
92	Sydney	Australia	Oceania
94	Alexandria	Egypt	Africa
99	Khartoum	Sudan	Africa

Cluster 14

	Metropolitan	Country	Continent
2	Mexico City	Mexico	North America
4	Lagos	Nigeria	Africa
9	Los Angeles	United States	North America
25	Buenos Aires	Argentina	South America
36	Lahore	Pakistan	Asia
40	Lima	Peru	South America
55	Dallas-Fort Worth	United States	North America
66	Houston	United States	North America
69	Miami	United States	North America
71	Atlanta	United States	North America
73	Luanda	Angola	Africa
86	Detroit	United States	North America
90	Stuttgart	Germany	Europe
93	Guadalajara	Mexico	North America
95	Phoenix	United States	North America
98	Monterrey	Mexico	North America

Cluster 15

	Metropolitan	Country	Continent
13	Chongqing	China	Asia
18	Bangkok	Thailand	Asia
33	Nanjing	China	Asia
44	Johannesburg	South Africa	Africa
45	Guangzhou	China	Asia
46	Qingdao	China	Asia
89	Shenzhen	China	Asia

Discussion

Using Foursquare venue category data, I grouped the 100 most populated metropolitan areas into 16 clusters. This resulted in 10 clusters with two or more metro areas and 6 clusters with a single metro area, the outliers. Most of the outlier metro areas had fewer than ten venue categories returned by Foursquare. Of the remaining 10 clusters, four contained only metro areas in Asia. Clusters 1, 3, 9, 13, and 14 contained a mixture of metro areas from different continents. Geographic areas tended to contain only a few clusters with the exception of China. China had 21 metro areas in the list.

The results may be improved by including additional data including economic and demographic data. The number of features could be condensed and reduced before clustering. KMeans analysis did not result in an optimal number of clusters to use. A different clustering algorithm such as Agglomerative clustering or DBScan may work better for this dataset.

Conclusion

This study was an attempt to answer the question “Which metropolitan areas are most similar?” This question would be of interest to a business planning to expand among others. The clustering algorithm resulted in nine clusters that contained 5 or more metro areas. The clusters are shown on the interactive map with each cluster having a different colored marker. Further research may include additional data and different clustering algorithms.

Appendix

Notebooks:

Part 1: https://nbviewer.jupyter.org/github/KathrynDH/IBMCapstoneFinalProject/blob/master/Final%20Project%20Get%20Data.ipynb

Part 2: https://nbviewer.jupyter.org/github/KathrynDH/IBMCapstoneFinalProject/blob/master/Final%20Project%20-%20Get%20FourSquare%20Data.ipynb

Part 3: https://nbviewer.jupyter.org/github/KathrynDH/IBMCapstoneFinalProject/blob/master/Explore%20Location%20Data.ipynb

Part 4: https://nbviewer.jupyter.org/github/KathrynDH/IBMCapstoneFinalProject/blob/master/Cluster%20Metro%20Areas.ipynb

Part 5: https://nbviewer.jupyter.org/github/KathrynDH/IBMCapstoneFinalProject/blob/master/Explore%20the%20Clusters.ipynb

Map

https://kathryndh.github.io/clustermap.html

Course Reference:

https://www.coursera.org/learn/applied-data-science-capstone

Data sources:

https://en.wikipedia.org/wiki/List_of_metropolitan_areas_by_population

https://developer.foursquare.com/