CLUSTERING THE WORLD CAPITALS

February 2020

BACKGROUND – Sister Cities

- Sister cities or twin towns is a known concept in today's world where two or more cities form a sort of agreement to promote their ties

- The concept is mainly aimed at increasing the friendship and understanding between the cities that ultimately helps in increasing trade and tourism

Dull, Perth and Kinross, Scotland is twinned with Boring, Oregon, USA.

RESEARCH PROBLEM & INTERESTS

- The objective of this report is to form clusters of capital cities that are closer to each other based on some physical attributes (Venue data in this case).

- This piece of research can be of interest for many different groups like tourists, travelers & international students

A restaurant in Paris

A restaurant in London

DATA SOURCES

For this research item, we will use two data sets.

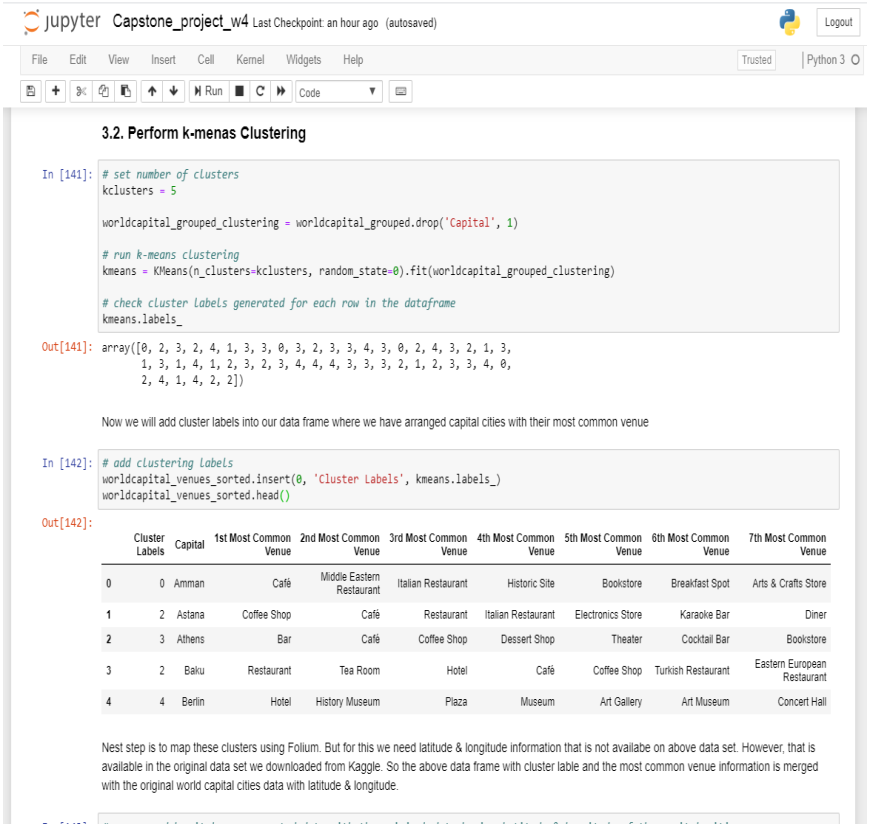
- Country names and their capital cities along with the latitude and longitude information available on this link

- Location data for the selected cities accessed using Foursquare API

Country	Capital	Venue	Venue_Category
Somaliland	Hargeisa	Hiddo - Dhawr	Comfort Food Restaurant
South Georgia and South Sandwich Islands	King Edward Point	جزيرة سندويشة	Beer Bar
South Georgia and South Sandwich Islands	King Edward Point	Bilinmeyen Yer	Racetrack
South Georgia and South Sandwich Islands	King Edward Point	Tang Ke Lek Harbour	Harbor / Marina
French Southern and Antarctic Lands	Port-aux-Français	Book Time	Bookstore

METHODOLOGY

- 50 capital cities were selected for cluster analyses (list of cities given in Appendix)
- Each city has at least 60 venues extracted from Foursquare location data
- One hot encoding was used to transform the categorical variable (venue category) into numeric data
- Scikit-learn was used to fit k-means clustering algorithm with k = 5



jupyter Capstone_project_w4 Last Checkpoint: an hour ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help

3.2. Perform k-menas Clustering

In [141]:
```
# set number of clusters
kclusters = 5

worldcapital_grouped_clustering = worldcapital_grouped.drop('Capital', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(worldcapital_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_
```

Out[141]:
```
array([0, 2, 3, 2, 4, 1, 3, 3, 0, 3, 2, 3, 3, 4, 3, 0, 2, 4, 3, 2, 1, 3,
       1, 3, 1, 4, 1, 2, 3, 2, 3, 4, 4, 4, 3, 3, 3, 2, 1, 2, 3, 3, 4, 0,
       2, 4, 1, 4, 2, 2])
```

Now we will add cluster labels into our data frame where we have arranged capital cities with their most common venue

In [142]:
```
# add clustering labels
worldcapital_venues_sorted.insert(0, 'Cluster Labels', kmeans.labels_)
worldcapital_venues_sorted.head()
```

Out[142]:

	Cluster Labels	Capital	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
0	0	Amman	Café	Middle Eastern Restaurant	Italian Restaurant	Historic Site	Bookstore	Breakfast Spot	Arts & Crafts Store
1	2	Astana	Coffee Shop	Café	Restaurant	Italian Restaurant	Electronics Store	Karaoke Bar	Diner
2	3	Athens	Bar	Café	Coffee Shop	Dessert Shop	Theater	Cocktail Bar	Bookstore
3	2	Baku	Restaurant	Tea Room	Hotel	Café	Coffee Shop	Turkish Restaurant	Eastern European Restaurant
4	4	Berlin	Hotel	History Museum	Plaza	Museum	Art Gallery	Art Museum	Concert Hall

Nest step is to map these clusters using Folium. But for this we need latitude & longitude information that is not availabe on above data set. However, that is available in the original data set we downloaded from Kaggle. So the above data frame with cluster lable and the most common venue information is merged with the original world capital cities data with latitude & longitude.

RESULTS

- Table below shows number of capital cities in each cluster
- The chart shoes the mapping of capital cities where each dot represent one city and different colors represent different clusters

Cluster	City Count
1	4
2	7
3	12
4	17
5	10

DISCUSSION on RESULTS

CLUSTER 1

- There are four capitals in this cluster
- All four cities belong to Muslim countries and three out of four cities speak the same language (Arabic)
- Looking at the location data, all of these cities have Café as 1st most common venue and then coffee shop and restaurants at 2nd and 3rd most common places.

Capital	Continent	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Cairo	Africa	Café	Plaza	Coffee Shop
Tehran	Asia	Café	Persian Restaurant	Coffee Shop
Amman	Asia	Café	Middle Eastern Restaurant	Italian Restaurant
Kuwait City	Asia	Café	Coffee Shop	Middle Eastern Restaurant

DISCUSSION on RESULTS

CLUSTER 2

- Has seven capital cities from five different continents however four of the cities belong to Asian continent.
- All seven cities have Coffee shop as 1st most common venue and then Café and Restaurant as second and third most common places.

Capital	Continent	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Seoul	Asia	Coffee Shop	Café	Korean Restaurant
Bishkek	Asia	Coffee Shop	Café	Asian Restaurant
Ulaanbaatar	Asia	Coffee Shop	Restaurant	Café
Phnom Penh	Asia	Coffee Shop	Chinese Restaurant	Asian Restaurant
Ottawa	Central America	Coffee Shop	Café	Restaurant
Nicosia	Europe	Coffee Shop	Greek Restaurant	Bar
Nairobi	Africa	Coffee Shop	Café	African Restaurant

DISCUSSION on RESULTS

Capital	Continent	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Tirana	Europe	Cocktail Bar	Café	Italian Restaurant
Baku	Europe	Restaurant	Tea Room	Hotel
Manama	Asia	Café	Hotel	Breakfast Spot
Sarajevo	Europe	Café	Restaurant	Hotel
Zagreb	Europe	Café	Restaurant	Bar
Guatemala City	Central America	Café	Pizza Place	Fast Food Restaurant
Astana	Asia	Coffee Shop	Café	Restaurant
Pristina	Europe	Restaurant	Bar	Hotel
Vientiane	Asia	Hotel	Café	Asian Restaurant
Skopje	Europe	Café	Hotel	Bar
Monaco	Europe	Italian Restaurant	French Restaurant	Restaurant
Podgorica	Europe	Café	Hotel	Bar

CLUSTER 3

- There are twelve capital cities in cluster 3 and eight of those are from Europe
- Most of the cities falling under this cluster have Café and Restaurants as 1st and 2nd most common venue

DISCUSSION on RESULTS

Capital	Continent	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Oranjestad	North America	Caribbean Restaurant	Bar	Shopping Mall
Nassau	North America	Boat or Ferry	Seafood Restaurant	Caribbean Restaurant
Minsk	Europe	Cocktail Bar	Bar	Restaurant
Brussels	Europe	Sandwich Place	Hotel	Bar
Sofia	Europe	Coffee Shop	Restaurant	Vegetarian / Vegan Restaurant
Santiago	South America	Bar	Pizza Place	Chinese Restaurant
San Jose	Central America	Bar	Sandwich Place	Hotel
Prague	Europe	Café	Bakery	Vietnamese Restaurant
Copenhagen	Europe	Italian Restaurant	Café	Scandinavian Restaurant
Santo Domingo	North America	Hotel	Pharmacy	Ice Cream Shop
Tallinn	Europe	Café	Asian Restaurant	Park
Helsinki	Europe	Scandinavian Restaurant	Hotel	Sushi Restaurant
Athens	Europe	Bar	Café	Coffee Shop
Budapest	Europe	Clothing Store	Coffee Shop	Hotel
Reykjavik	Europe	Bar	Seafood Restaurant	Hotel
Jakarta	Asia	Chinese Restaurant	Hotel	Noodle House
Kuala Lumpur	Asia	Malay Restaurant	Hotel	Asian Restaurant

CLUSTER 4

- Cluster 4 is the largest cluster with 17 cities mostly from Europe and America
- The most common venues in this cluster are Bar and Restaurants

DISCUSSION on RESULTS

Capital	Continent	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Jerusalem	Asia	Historic Site	Hotel	Restaurant
Vienna	Europe	Café	Hotel	Plaza
Paris	Europe	Hotel	French Restaurant	Plaza
Tbilisi	Europe	Hotel	Caucasian Restaurant	Restaurant
Berlin	Europe	Hotel	History Museum	Plaza
Rome	Europe	Italian Restaurant	Plaza	Ice Cream Shop
Tokyo	Asia	Historic Site	Soba Restaurant	Convenience Store
Saint Helier	Europe	Hotel	Coffee Shop	Pub
Riga	Europe	Restaurant	Eastern European Restaurant	Hotel
Mexico City	Central America	Mexican Restaurant	Museum	Art Museum

CLUSTER 5

Last cluster has 10 capital cities and most common venues are Hotel & Restaurants.

CONCLUSION

- k-mean clustering algorithm does a good job in clustering the world capital cities based on the location data
- Resulting clusters not only make sense based on the available venue information but those are also closer to each other in terms of geography, culture & religion

https://www.colourbox.com/vector/conclusion-vector-15359599

APPENDIX

LIST OF SELECTED 50 CAPITAL CITIES – 1

Sr. #	Capital	Country	No. of Venues
1	Amman	Jordan	100
2	Astana	Kazakhstan	81
3	Athens	Greece	100
4	Baku	Azerbaijan	97
5	Berlin	Germany	100
6	Bishkek	Kyrgyzstan	82
7	Brussels	Belgium	100
8	Budapest	Hungary	100
9	Cairo	Egypt	61
10	Copenhagen	Denmark	100
11	Guatemala City	Guatemala	67
12	Helsinki	Finland	100
13	Jakarta	Indonesia	99

Sr. #	Capital	Country	No. of Venues
14	Jerusalem	Palestine	82
15	Kuala Lumpur	Malaysia	100
16	Kuwait City	Kuwait	100
17	Manama	Bahrain	70
18	Mexico City	Mexico	100
19	Minsk	Belarus	100
20	Monaco	Monaco	88
21	Nairobi	Kenya	74
22	Nassau	Bahamas	63
23	Nicosia	Cyprus	100
24	Oranjestad	Aruba	92
25	Ottawa	Canada	100

LIST OF SELECTED 50 CAPITAL CITIES – 2

Sr. #	Capital	Country	No. of Venues
26	Paris	France	100
27	Phnom Penh	Cambodia	70
28	Podgorica	Montenegro	100
29	Prague	Czech Republic	100
30	Pristina	Kosovo	80
31	Reykjavik	Iceland	100
32	Riga	Latvia	100
33	Rome	Italy	100
34	Saint Helier	Jersey	62
35	San Jose	Costa Rica	100
36	Santiago	Chile	100
37	Santo Domingo	Dominican Republic	71
38	Sarajevo	Bosnia and Herzegovina	76

Sr. #	Capital	Country	No. of Venues
39	Seoul	South Korea	100
40	Skopje	Macedonia	100
41	Sofia	Bulgaria	100
42	Tallinn	Estonia	91
43	Tbilisi	Georgia	61
44	Tehran	Iran	100
45	Tirana	Albania	100
46	Tokyo	Japan	100
47	Ulaanbaatar	Mongolia	100
48	Vienna	Austria	100
49	Vientiane	Laos	86
50	Zagreb	Croatia	100