BATTLE OF WORLD CAPITALS!!!

CAPSTONE PROJECT, COURSERA-IBM SPECIALIZATION
ANIBAL J GUERRA S,
MEDELLÍN, COLOMBIA

INTRODUCTION

- What factors are really needed for a city to achieve the perfect balance that makes it attractive, providing enough quality of life to consider it city as a well-developed city.

- Are there any categories of venue that can be related as a sign of having a balanced development?

BUSSINESS PROBLEM

- For public servers in the government (i.e. major, governors) having clarity about such factors would give them clarity about the decisions that should be made to keep the city on the road to development. Two questions seem obvious: where on that road is the city right now (what does the city "have") and what should we "have" as a city in order to be considered as developed city in the future.

- I would use the Foursquare data related to all of the venues that the capitals of the world have right now, and see if through data science I get to stratify (or cluster) those cities to see if there are specific venues that are characteristic of the cities development in each class.

Hypothesis: It is possible to discriminate groups of cities according to the amount of certain types of venues in them. If so, is it possible to find out which venues are characteristics of well-developed cities.

DATA SOURCE

- More than 200 capitals of the world and along with their geo-localization data.

- Up to 100 venues per city as returned by a foursquare request.

Country		Capital	Latitude	Longitude
Afghanistan		Kabul	34.28N	69.11E
Albania		Tirane	41.18N	19.49E
Algeria		Algiers	36.42N	03.08E
American Samoa		Pago Pago	14.16S	170.43W
Andorra		Andorra la Vella	42.31N	01.32E
Angola		Luanda	08.50S	13.15E

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Regent Park , Harbourfront	43.65426	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery
1	Regent Park , Harbourfront	43.65426	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop
2	Regent Park , Harbourfront	43.65426	-79.360636	Cooper Koo Family YMCA	43.653249	-79.358008	Distribution Center
3	Regent Park , Harbourfront	43.65426	-79.360636	Body Blitz Spa East	43.654735	-79.359874	Spa
4	Regent Park , Harbourfront	43.65426	-79.360636	Morning Glory Cafe	43.653947	-79.361149	Breakfast Spot

FOURSQUARE

METHODOLOGY

- Web scrape from website page for a list of capitals of the world along with their geo-localization data.

- Explore and pre-process geo localization data.

- Use Foursquare API to find the venues corresponding to each capital.

- Clean encode and normalize venues data.

- Apply k-means clustering, find the optimal k value and visualize data using Folium maps

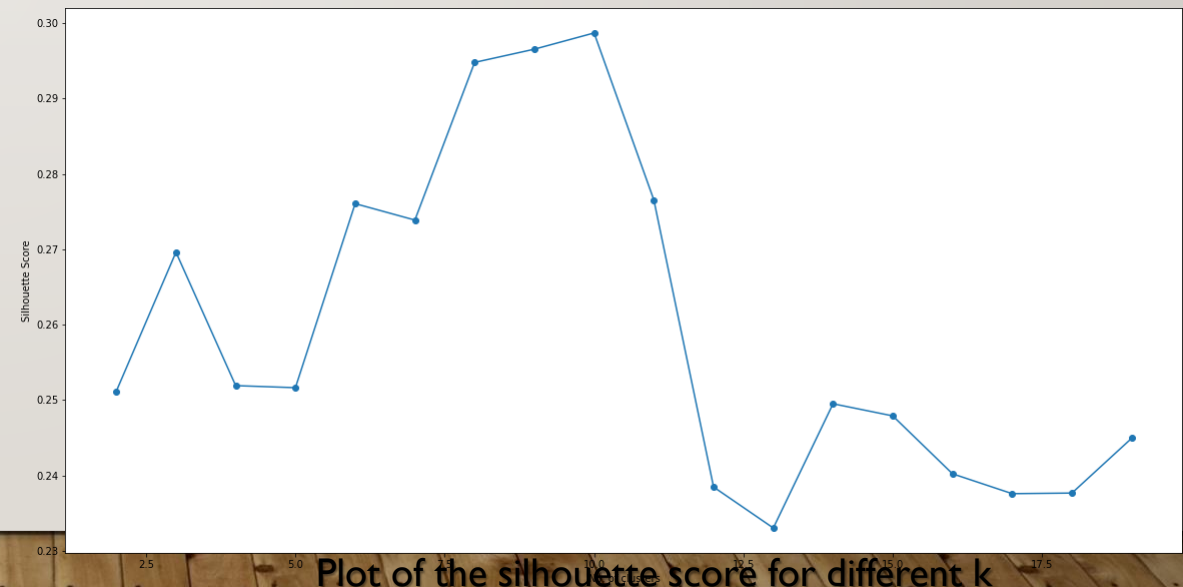- Analyze each cluster to find if any matches the criteria of well-developed cities.

MACHINE-LEARNING CLUSTERING

With selected and normalized input columns:

- Apply K-means clustering algorithm

- Iterate using different k to optimize the k value through the silhouette score metric

- Plot the map with the optimal cluster value.

	Airport	Culture	Food	Spectacle	Stores	Education	Tourism	Leisure	Lodging	Industry	Transport	Health
0	0.384252	-0.351402	-0.87272	-0.859855	-0.790625	-0.210187	-0.234738	-0.952632	-0.344169	-0.289318	-0.417542	-0.308614
1	-0.491485	-0.351402	2.02342	-0.859855	-0.222303	-0.210187	-0.234738	0.162922	1.31009	-0.289318	-0.922717	-0.308614
2	0.384252	0.525985	0.152995	-0.859855	-0.601184	-0.210187	-0.234738	-0.673743	0.574865	-0.289318	-0.417542	-0.308614
3	-0.491485	-0.790096	-1.35541	-0.859855	-0.980066	-0.210187	-0.234738	-1.09208	-1.0794	-0.289318	-0.922717	-0.308614
4	-0.491485	-0.351402	1.54073	0.59719	-0.411744	-0.210187	-0.234738	1.20875	0.758671	-0.289318	0.0876324	-0.308614

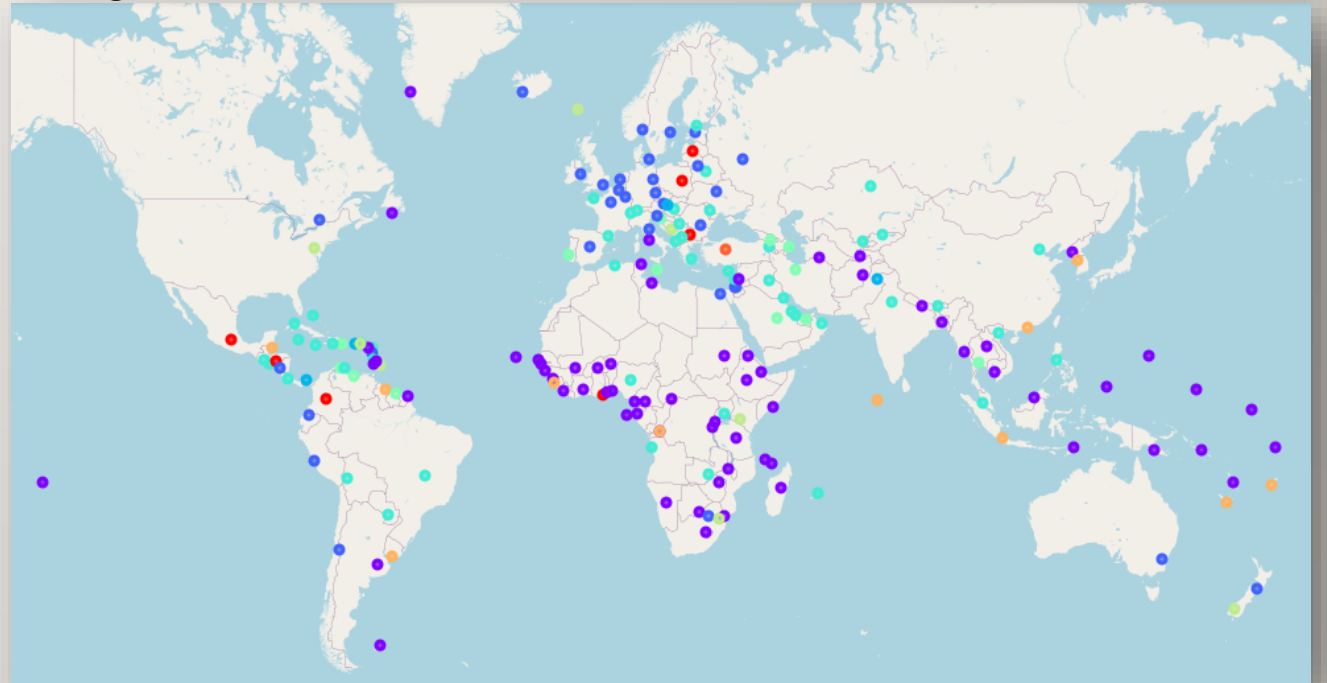Example of final input data

Plot of the silhouette score for different k

RESULTS

- With k=8, this was the resulting clustering :

Cluster number 2 matched the concept of well-developed cities.

Top1	Top2	Top3	Shortcapital
Food	Leisure	Stores	Canberra
Leisure	Food	Stores	Vienna
Food	Leisure	Stores	Brussels
Food	Stores	Leisure	Ottawa
Food	Stores	Leisure	Santiago
Food	Leisure	Lodging	Prague
Food	Airport	Culture	Copenhage
Food	Stores	Leisure	Quito
Food	Leisure	Stores	Cairo
Food	Lodging	Leisure	Tallinn
Leisure	Stores	Food	Paris
Food	Leisure	Stores	Berlin

That cluster was dominated in more than 90% by food, leisure, and stores venues.

CONCLUSIONS

I have built a model to cluster world cities according to the most common venues in them.

It seems that, according to the Foursquare dataset, those venue categories are characteristics of a well-developed economies.

There is space for improvement in the model, using more data and normalizing input data respect to population size.

Nowadays, the commerce and service venues have become an important part of a city development, and according to the results presented here their growth could be related to the definition of a healthy economy.