Battle of World Capitals!!!

Capstone Project, Coursera-IBM Specialization

Aníbal J Guerra S, ajguerras@gmail.com

Medellín, Colombia

In this project we are meant to put a lot the knowledge we have gotten through the whole IBM data science specialization. My work is intended to exhibit the application of a methodology, the usage of the right tools, data treatment, tools for visualization and to provide a final analysis that answer the initial questions.

1. Introduction

I remember the days when playing the simulation game SIM City ® I could be the hero of my very own city as I designed and created a beautiful and bustling metropolis. Through a long chain of apparently simple decisions, like creating schools, libraries, hospitals, entertainment places, etc.; you could make your city a good and attractive place to live, so it would get larger and more intricate. That could be defined as success. On the opposite case, people in your town would start leaving and the whole system finally would fail. A perfect balance which sounds very hard to find.

I grew up in Latin America, in a country that we were told that we were "on the road to become a first world country". But it always seemed like a never-ending process. After traveling a lot, and visiting very organized cities with a high standards for living, I started to ask myself: what factors really needed for a city to achieve the perfect balance that makes it attractive, providing enough quality of life to consider such city as a well-developed city.

2. Problem Description

Far beyond my youthful meditations, finding the answers to such questions is a very relevant 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.

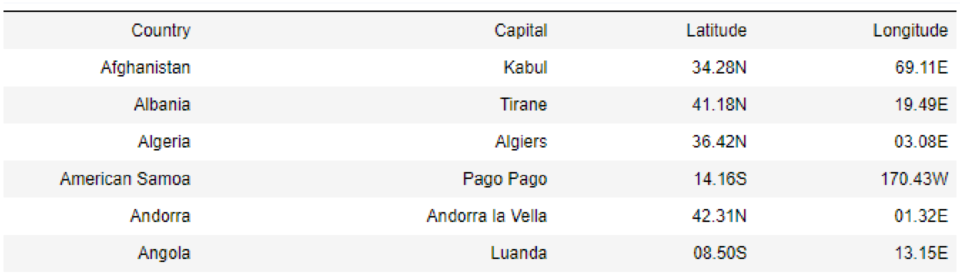
As I am not an expert in such extent, I find those answers hard to find. But for sure data science has the answers. I would use the Foursquare data related to all of the venues that the capitals of the world have right now, and see if through data science I get to stratify (or cluster) those cities to 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, I will have to figure out if such groups correspond to an already known classification (developed, non-developed , or on the development road).

3. Data Description

a. Capitals of the world and their geo-localization data.

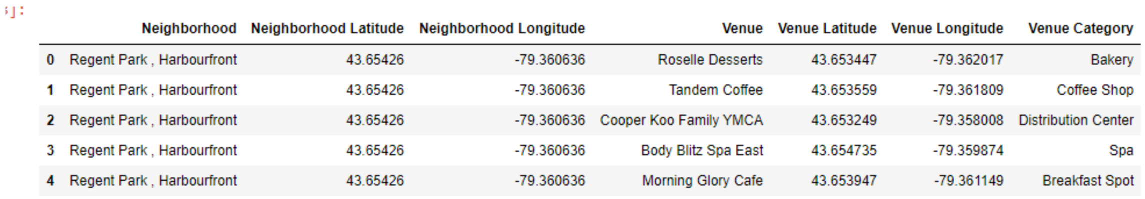
The following figure is an example of the input needed. For every capital of the world, the respective geo-localization coordinates and the country they belong to.



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

b. List of existing venues per each capital.

From Foursquare database we can retrieve relevant information similar to this one:



	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

A similar table per capital. Besides the usual preprocessing, I plan to perform a lot of transformation to make such table useful for the requirements.

The decision model would be based on the amount of venues from certain categories per city. So, prior to counting the venues, I plan to apply some kind of filter that create classes inside the venue's dataset. For example: Food Services (Coffee, Restaurants, Breakfast places), health services (Hospital, Pharmacies), lodging services (hostels, hotels), etc.