Health Care Analysis

IBM Professional Certificate

By Marcos Negre

Table of Contents

Introduction and Business Problem

Introduction

There are different factors which affect the health and lifestyle of people around the world. Related data is going to be used to compare the life expectancy, the investments in healthcare, as well as the cost of life in each country. We are going to contrast the data with most common venues in each capital cities. In this report it is going to be shown another way to use the Foursquare data. We are going to use it to support the clusters which are going to be categorized with information retrieved from well-known websites.

Business Problem

This data is going to be used to categorize and cluster different countries among groups. Consequently, there are going to be some unexpected results and we are going to contrast with real world data from Foursquare. This research may be used to shed light on more specific problems regarding life styles around the globe. The results will be compared with most common venues of each country to know the habits of the population of each country. This way, there will be made some interactive maps which can be used to discuss the outcome of the project.

Methodology

The steps followed along the process are:

- Download and Explore Datasets
- Clean Data
- Explore Countries Selected
- Retrieving Foursquare Data
- DSBC Clustering
- K-Means Clustering
- Results Visualization

There are three datasets we are going to take into account for the project: Wikipidia expenses per capita in health care per country, Numbeo cost of living ranking and Wikipedia life expectancy per country.

Then, the data will be cleaned and merged into a single dataset containing the valuable information. Consecutively, the geolocation of the capital of each country will be retrieved from internet, so it is possible to get the foursquare information for each location. Once we have the Data, it is applied DSBC Clustering and K-Means clustering to compare the results between both.

Finally, some plots are shown to understand the results.

Results

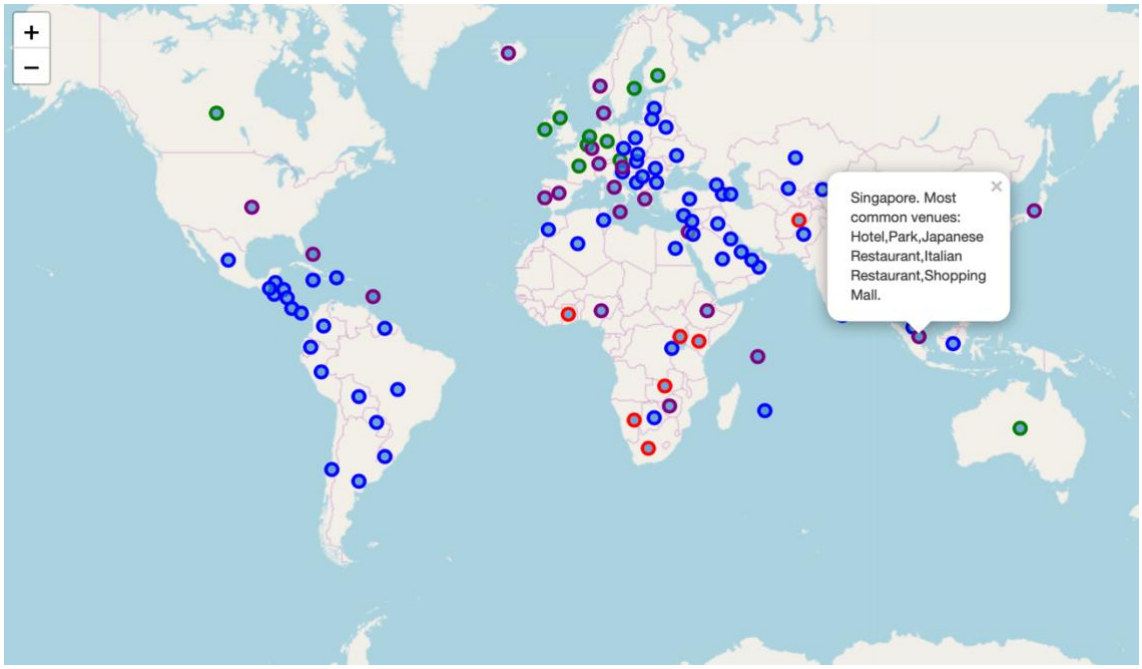
The results are as follows:

Firstly, it is shown the map of the chosen countries. The ones selected must have data for all the categories analyzed here:

Selected countries for healthcare analysis.

DBSCAN

Once the countries are chosen, the data will be normalized and combined to gather results. As an example, for each map we will open the information window for a country, in the first case it will be Singapore. The first clustering method used here is DBSCAN:



DBSCAN clustering.

K-Means clustering

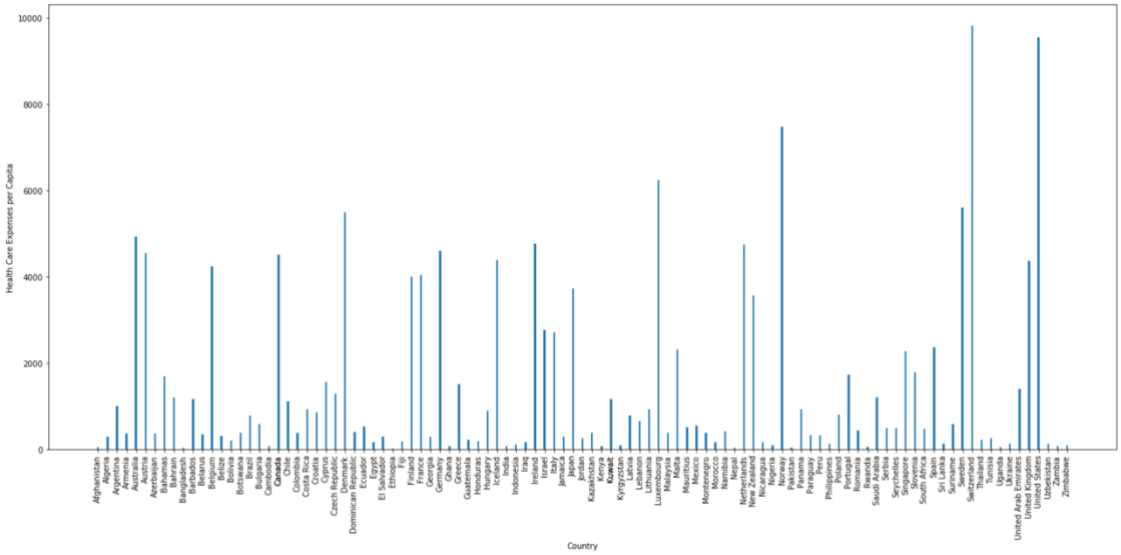
The combination of the data has shown up interesting results, but what about k-means clustering?

K-Means clustering.

At first sight, the results with the K-Means method seem more realistic. Now it is going to be clustered again with each one of the properties, evaluated separately.

Clustering for Health Care Expenses:

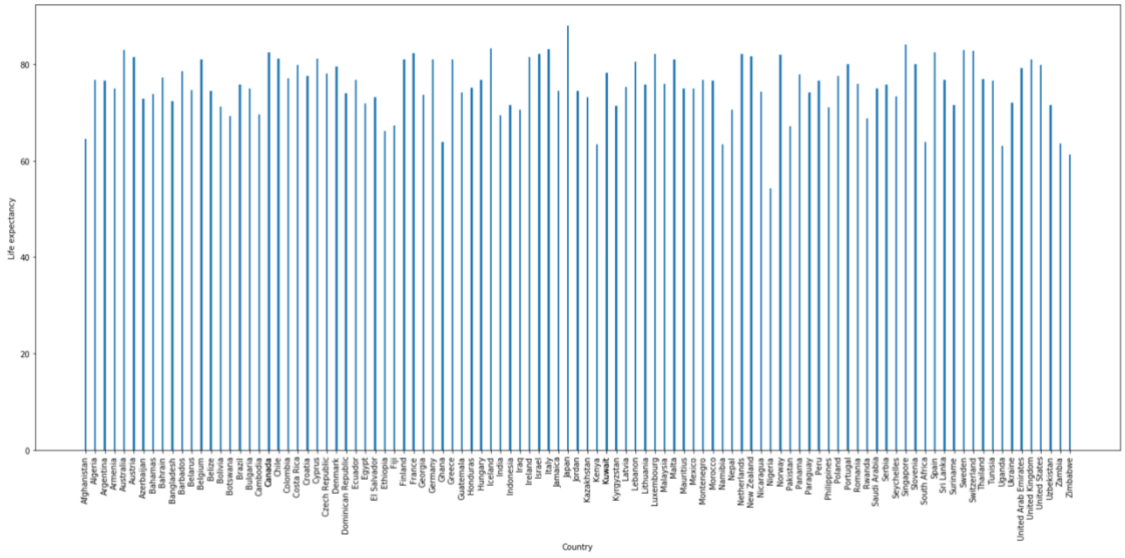
K-Means clustering for healthcare expenses (per person / per year).



Bar plot for healthcare expenses (per person / per year).

Clustering for Life Expectancy:

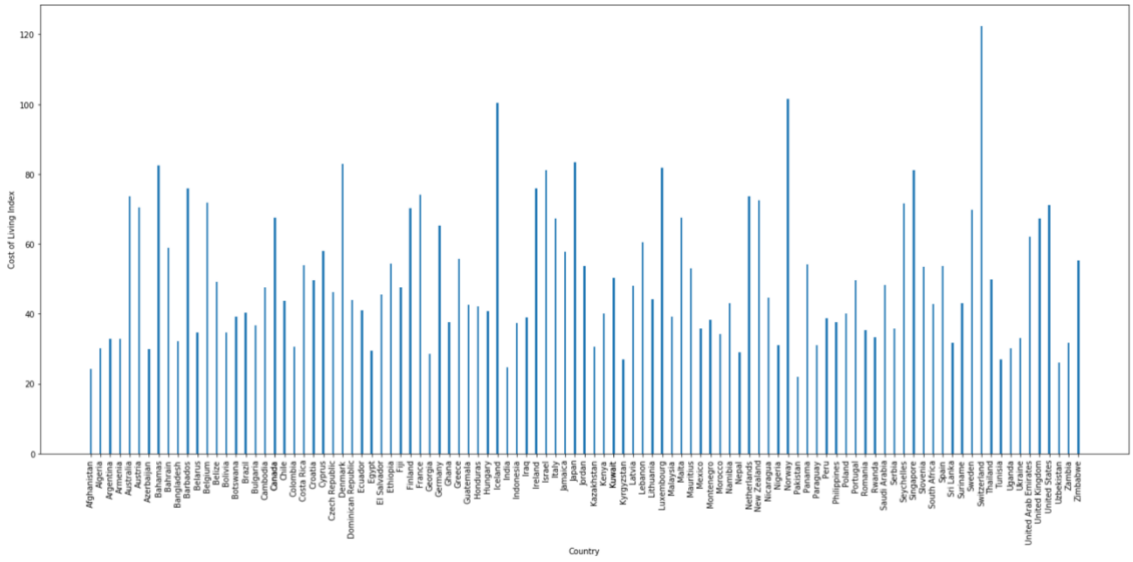
K-Means clustering for life expectancy.



Bar plot for life expectancy.

Clustering for Cost of Life:

K-Means clustering for cost of life ratio.



Bar plot for cost of life ratio.

Discussion

The results show three key points.

Firstly, there are substantial differences between first world countries when it comes to expend money in healthcare services. Nonetheless, that results do not impact the life expectancy and, in fact, are more related to the cost of life of that specific country compared to the years a person lives in average.

The most common places vary a lot from country to country. In USA the most common restaurants by far at the ones catalogued as fast food restaurant and pizza restaurant. On the other hand, countries like Japan have other venues at the top of the most common places. This can be deeper analyzed going country by country on each map.

Thirdly, the foursquare data retrieved for this project is not wide enough to do a more exhaustive research. Since the account used to gather the data is a lite account, it is not good enough to see the complete range of venues at the same time. Apart from that, in order to look for the most common venues, there are some countries which are not very familiarized with foursquare showing an evident lack of information.

Conclusion

In order to sum up, the results might be unexpected for some of the countries which, apparently, seem to have the potential to have the higher life expectancy. Although it is a first approach to the issue and therefore, there are factor that have not been taken into account in this project, the information presented here is a valuable start to discuss this topic and perform further researches.