EU member states capitals comparison

Victor Silva

January 26, 2020

1. Introduction

1.1 Background

For someone who have EU citizenship it is easy to move between all of member states, but with 28 countries to choose one can be confused on where to live or visit. The number of options is so high that, if we consider locations with more than 300,000 inhabitants, one would have to choose between 120 alternatives, according to the list of cities in the European Union by population within city limits on Wikipedia, and if all villages, towns and cities were considered this number surpasses thousands of places.

1.2 Problem

Data might help those who want to move to other countries to choose where to live, based on what venues each city and neighbourhood has to offer. This project aims to group similar capitals of EU countries into clusters and to compare similar neighbourhoods on those clusters to help people who want to live abroad or just meet places similar of those they love.

1.3 Interest

This project will be helpful for anyone who wants to move to an EU member state or want to know places there but don't know where to start.

2. Data acquisition and cleaning

2.1 Data sources

To achieve our objective the model will consider the type of venues existing on each neighbourhood of each place, which will be purchased using Foursquare API, and the cost of living on those cities, based on NUMBEO Cost of Living Index, which is explained below:

> *These indices are relative to New York City (NYC). Which means that for New York City, each index should be 100(%). If another city has, for example, rent index of 120, it means that on an average in that city rents are 20% more expensive than in New York City. If a city has rent index of 70, that means on an average in that city rents are 30% less expensive than in New York City. (NUMBEO)*

Additionally, a CSV dataset was created manually to map all neighbourhoods of the UE members capitals with the city names and neighbourhoods.

Those data will be used in order to cluster cities and neighbourhoods.

2.2 Data cleaning

Data downloaded or scraped from multiple sources were combined into one table. There were 21 subdivisions (4% of the total of 474 subdivisions) that weren't found on Foursquare database, so I decided to not use them in the model of subdivisions.

The limitation of number of requests on Foursquare was another issue, because I couldn't restart the kernel to restart running the code from the start. To surpass that obstacle, I write a code to save a CSV every time the request was successful and to read this CSV when don't.

Another problem I have faced was to get the coordinates with *Geopy*, because sometimes the API didn't work properly. I tried many times to get the latitude and longitude of each neighbourhood, but there was one city that never worked to get the coordinates: Riga, on Latvia. I tried to change the names of the districts to it's original language and to English but I didn't succeed with it, so I decided to export the coordinates I got until there and manually map the boroughs of this city and save into another CSV and reimported it.

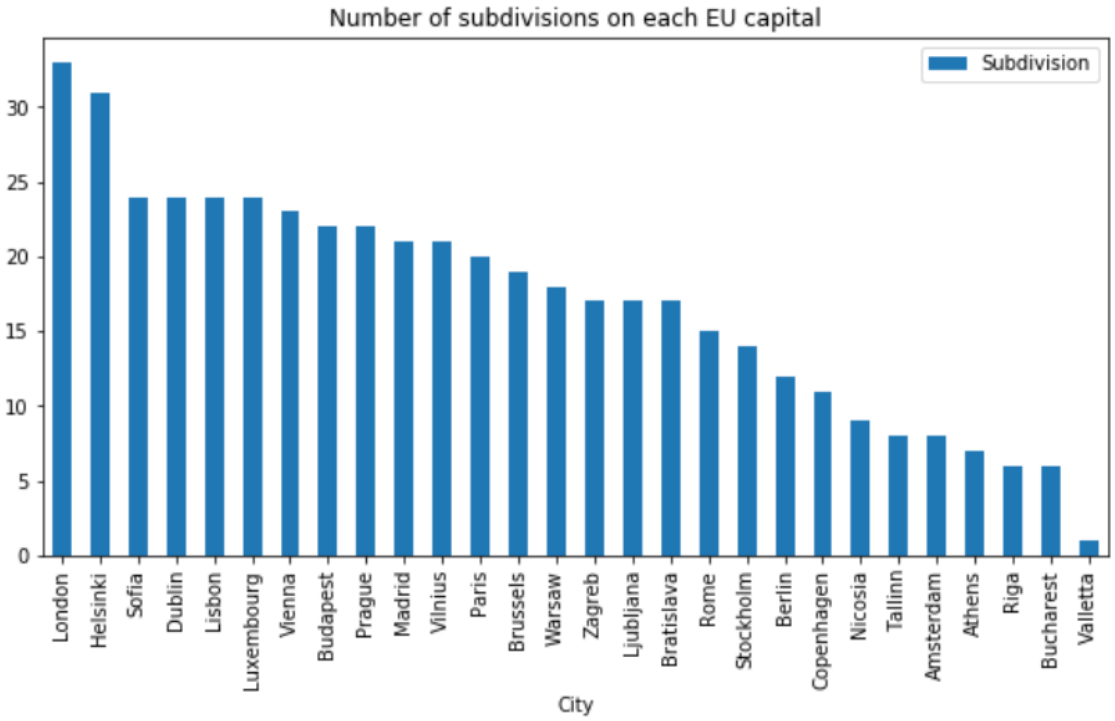
The last issue about data was the missing value of the Cost of Index Ranking of Valleta (Malta) in Numbeo list. For some reason the city wasn't at the ranking list, so I had to enter manually in the Valleta page in Numbeo website and input the value at the dataframe.

3. Exploratory analysis

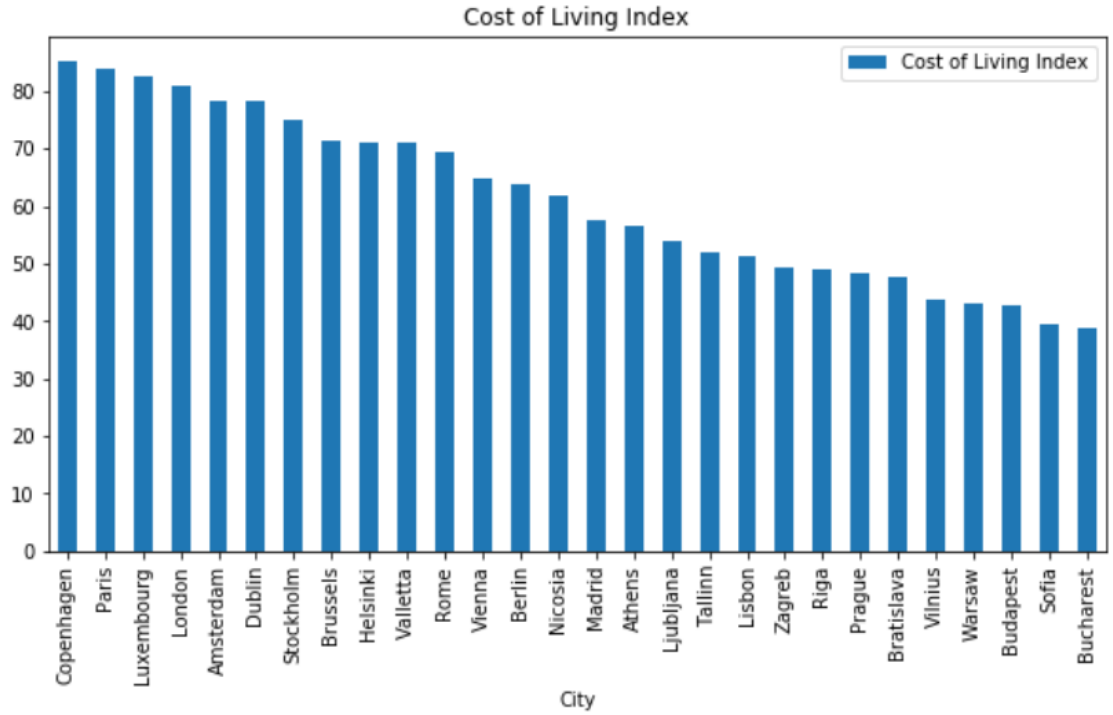
3.1 Understanding data

I worked with the capitals of all 8 European Union member states. The main data used was city names, subdivisions, cost of living on each city and venues on each neighbourhood.

The first thing that I understood was the great variance of city sizes and their subdivisions (some cities like London have boroughs and neighbourhoods while Valletta doesn't have any subdivision).

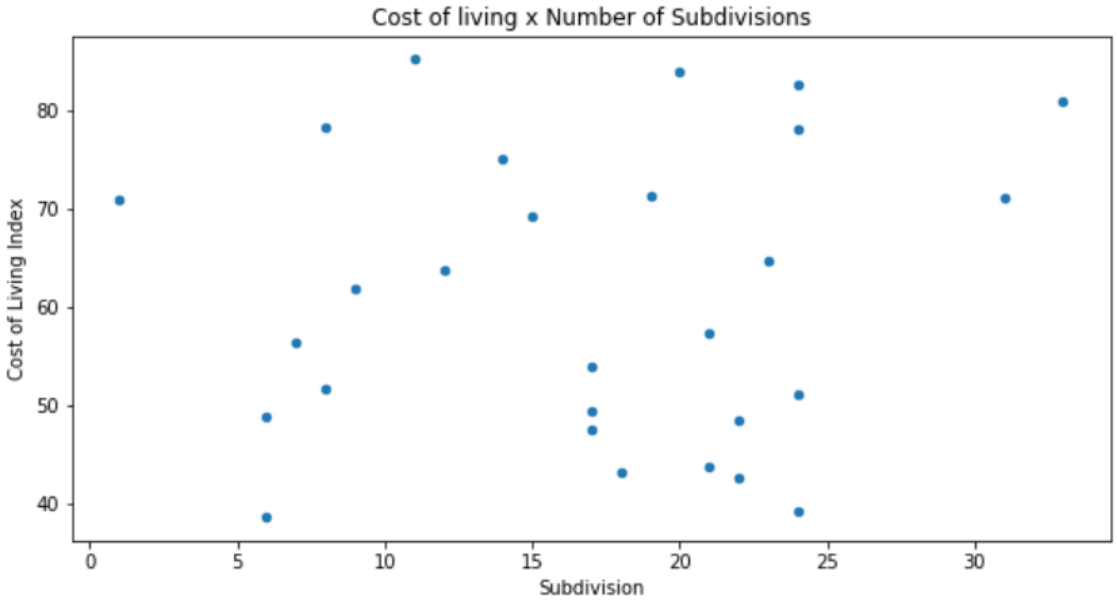


Another variable that has a lot of dispersion is the cost of living on each city. Copenhagen (Denmark), for example, has a Cost of Living Index of 85.25, that is 2.2 times higher than Bucharest (Romania), with 38.64.



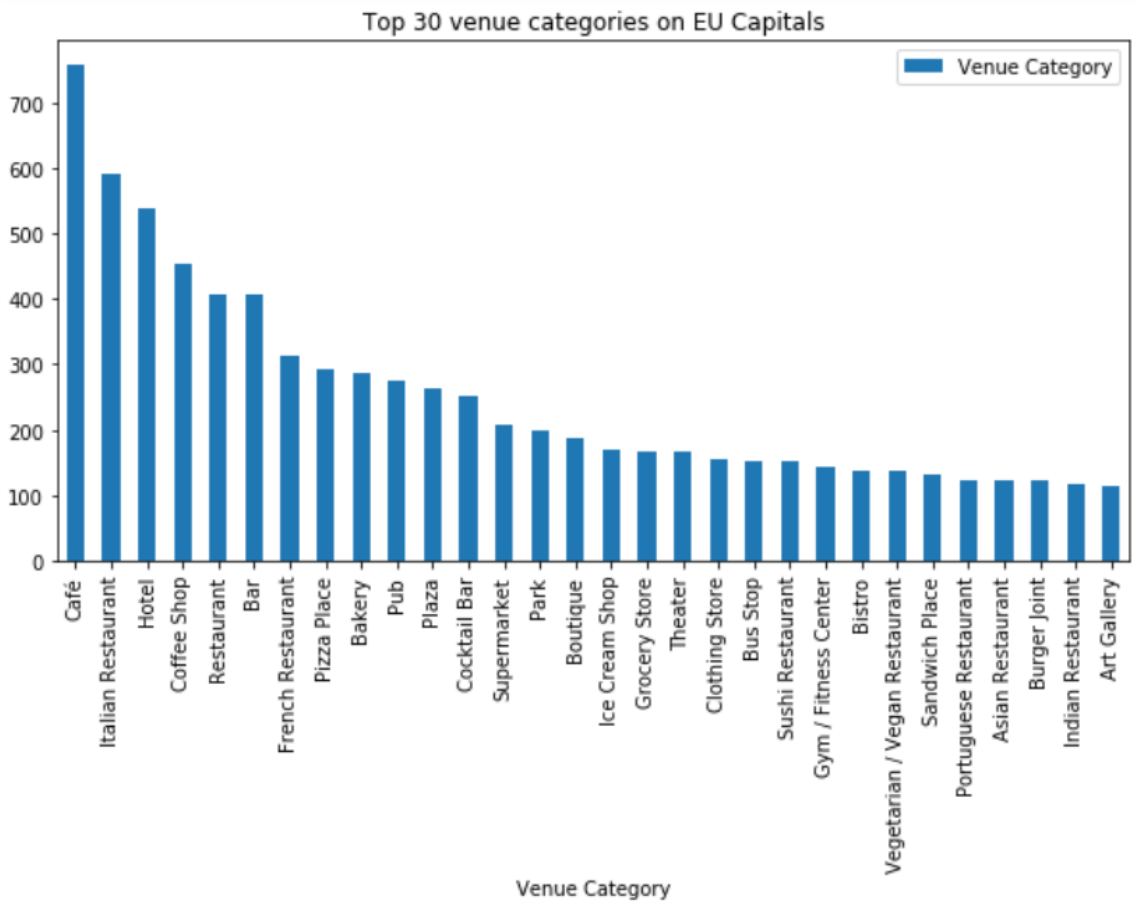
3.2 Relationship between city size and cost of living

And I got curious to see if larger cities (with more subdivisions) has higher cost of living, so I did a scatter plot and there isn't a strong correlation of those two variables.

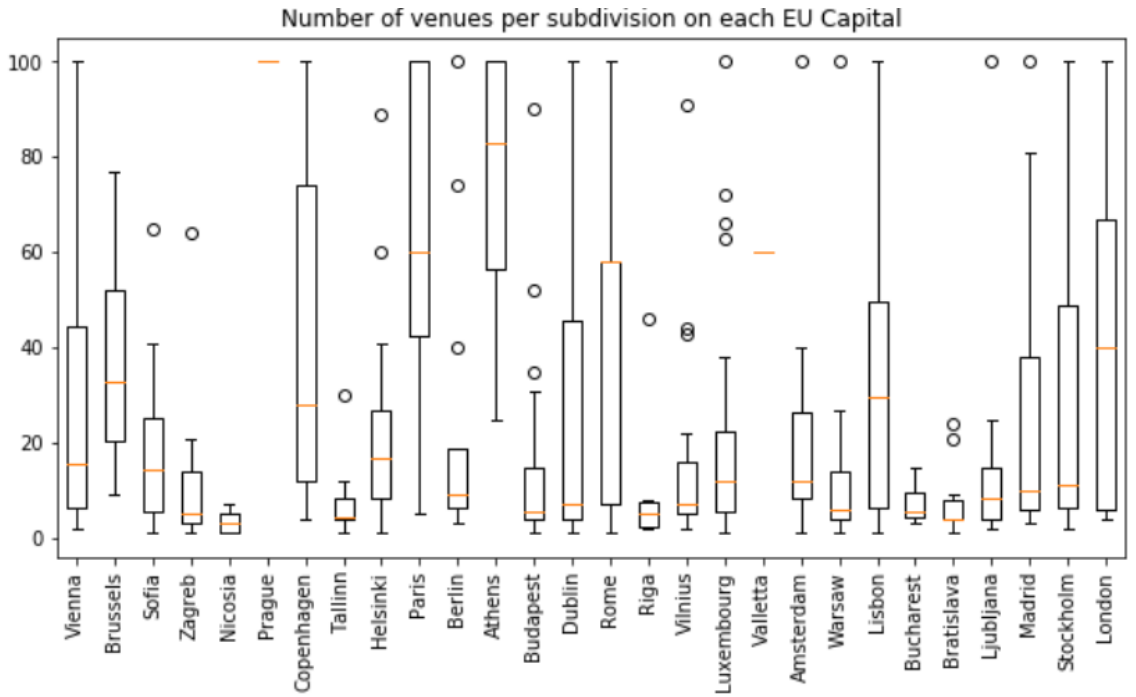


3.3. Venues overview

In a total, there is 480 venue categories on the dataframe, if we select the top 30 of them (that corresponds to 7.540 venues – 54% of the total of 13,853 venues), we can see that there is a lot of places to eat.



The number of venues in each subdivision also variates a lot in each city, as we can see on the following boxplot.



4. Cluster Modelling

In order to better recommend similar cities for those who want to travel or move, I decided to develop a cluster model based on k-means to create 5 cluster of cities and 5 cluster of neighbourhoods and compare them to see which model is best for use.

> K-means clustering is *a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. (WIKIPEDIA)*

The algorithm chooses k random points and calculates the distance of this points to the registers of dataset and then, select another k random points and the process continues. The disposition with the lowest sum of distances of each point to it's mean is the one that will be used.

4.1. Preparing for k-means

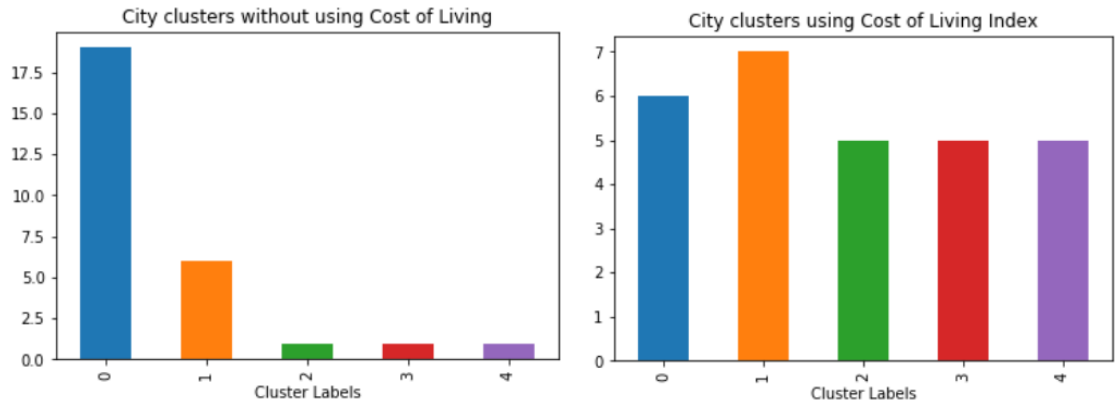
After getting the venues of each subdivision of the cities I prepared two datasets:

1. One showing the top 10 venues classes of each city;
2. One showing the top 10 venues classes of each neighbourhood.

In addition of that data, I also used the normalized cost of living of each city on both datasets, transforming the original values to numbers from 0 to 1.

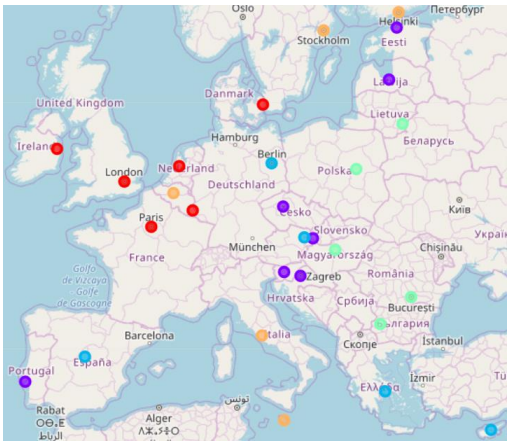
4.2. Results of model

At first, I have applied k-means without using cost of living, just with the top 10, and the clustering result had basically two clusters with more than one city and three others with just one city each. It seemed to me that using the variable it presents more solid results, as we can see as follows:

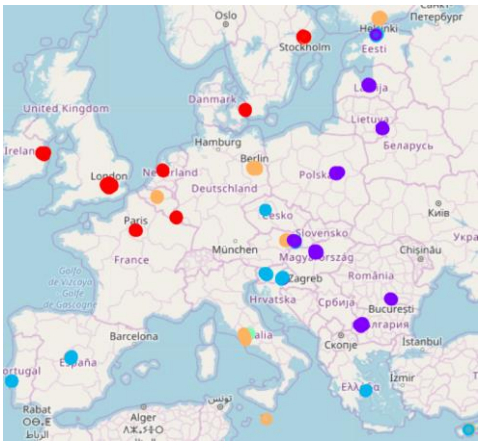


When comparing the maps (or lists) of clusters, it's possible to see that neighbourhood tends to stay at the same cluster as its city belongs to (see the two first maps), but there's always some neighbourhoods that differs from the mean of the city (see the third map).

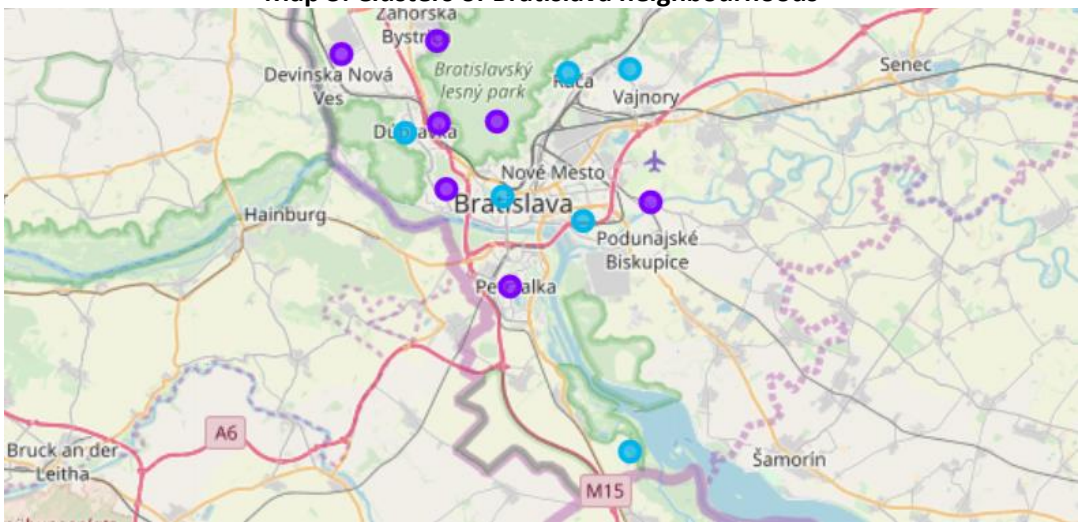
Map 1: Cluster of cities



Map 2: Cluster of neighbourhoods



Map 3: Clusters of Bratislava neighbourhoods



5. Conclusions

This project has showed that data can really help people who wants to choose a place to travel or to move based on their interests. but it appears that it works better recommending cities than neighbourhoods.

On neighbourhood clustering there were groups with more than 100 places to choose and when I tried to increase the value of K, new clusters had just one register, so it's not a matter of the k number. Considering variables like population density, population aging, area, quality of public transportation and other social and demographic indices can enrich the model and, later, the results.

Another point of improvement of the model is how to get the venue information, because there are smaller and bigger neighbourhoods, so using a radius of 500 meters on borough of 8km² is to get less information than necessary. Beside this, some places had 10 or less venues and it can affect the model.

6. Future directions

I recommend improving the model in two ways:

1. Adding new variables to the existing model and try to get more venues to the database using google maps API or other.
2. Create a second model that gets information from user experiences and cluster the users.

7. References

ABOUT COST OF LIVING INDICES AT THIS SITE. NUMBEO. Source: <https://www.numbeo.com/cost-of-living/cpi_explained.jsp>. Accessed January 11, 2020.

CURRENT COST OF LIVING INDEX. Source: <https://www.numbeo.com/cost-of-living/rankings_current.jsp>. Accessed January 11, 2020.

K-MEANS CLUSTERING. Source: <https://en.wikipedia.org/wiki/K-means_clustering>. Accessed January 26, 2020.