Where to establish a Start-up Company

Applied Data Science Capstone by IBM/Coursera – “The Battle of Neighbourhoods”

Author: Henrique M. L. Pereira

Table Of Contents:

1. Introduction

This is the final part of Coursera Capstone Project. In this notebook I'll describe the problem at hand, explain why it is important, and explain how and where I’ll retrieve the data needed to accomplish my objective. I am asked to use location data to explore geographical locations (New York, Toronto or another city of my choice) using Foursquare location data, to be creative and find a possible problem that can be solved with this approach.

Everyone has faced or will face, sometime of their life, a difficult decision to make: i.e. get or not get married, move away from your parent's house to your own place, have or not have kids, enrol or not in that awesome course, etc...

The situation I present might as well be a one of particular difficulty, mostly due to the challenge it may present. Let's imagine that you, as an entrepreneur, creative professional, and keen on producing innovative products, decide (along with some good and daring colleagues) to start a Start-up company. You are offered some help to establish an office in only 5 cities in Portugal, all of them away from your hometown and your colleagues' hometown.

Choosing Portugal as a country where to setup and establish start-up companies is no new thing to all that need these kinds of things to be done in a well feed investment environment. There are some facts that support this, mainly the annual Web Summit conference (previously held in Dublin), Lisbon's start-up community is thriving - with healthy support from the Portuguese government (1). Moreover, a key driver to the rise of desirability of Portugal as a good tech hub is the record number of foreign companies using

Portugal as a more convenient and affordable way to assemble their digital platforms with quality and low cost services for specific projects, never forgetting the increasing numbers returning to Portugal to establish their tech hub (2).

1.1 Problem Definition

The 5 cities are major hubs in Portugal, diverse in all the ways, financial, economic, artistic, academic, full of opportunities, etc. The 5 cities were selected using data from the Nomad List and are classified as good cities to establish a start-up. These are 'LISBOA', 'PORTO', 'COIMBRA', 'BRAGA', 'AVEIRO'.

The Start-up, and its work, may be defined by the following keywords:

- Innovative;
- Culturally impacting;
- Environment friendly;
- Information Technology;
- Game changing;
- Daring designs and approaches;

The founders of the start-up have the following concerns and needs (with no particular order):

- Affordability of office and living;
- Multicultural environment;
- Near essential commodities;
- Near places to relax and meet people
- Always wanted to live in a buzzing city but without stressing situations;

1.2 Objective

The main objective is, with Data Science and Machine Learning, to advise on the best city and neighbourhood to set up said Start-up company based on the factors above specified, as well all the data from these cities that will be retrieved dynamically.

2. Data

Every data used in the project will be either stored in the resources folder of the GitHub repository associated with this project or be available in the internet. I'll use the data related to the cities of Portugal from official Portuguese Government open source information (available at https://dados.gov.pt/). I'll also use data scrapped from the "Nomad List" (https://nomadlist.com/) to rank the cities so I can put similarities and dissimilarities in perspective when comparing selected neighbourhoods.

First, the data scrapped from Nomad List website, had numerous problems concerning data cleaning, so I used regular expressions to obtain a clean dataset, and then rescaled it. This data is intended to further help on selecting the best place for the start-up. I'm expecting to have matches between neighbourhoods of all cities, so this is the best objective way to select an appropriate neighbourhood or at least the top best ones. I dropped features related to real time data and others that were not important to help the start-up.

Second, I downloaded Portuguese cities neighbourhood's and location data, then used the *geocoder* package to run the addresses found to retrieve the corresponding coordinates to be used in the next part. The features selected were the city, neighbourhood, its latitude and longitude

Finally, I used the Foursquare Explore API to retrieve every venue for each neighbourhood in a radius of 500 meters. I retrieved the parent categories as well for each venue. The features selected were all the venues categories, and parent categories.

All this data was merged together into a single Dataframe.

3. Methodology

For this problem I will first make a data exploratory analysis in order to find hidden patterns among the data, and check if all selected cities have venues that in some way are very different one from another.

After exploring the data for each city, I'll apply 2 machine learning techniques (unsupervised) provided by SciKit-Learn (3), firstly K-Means, and OPTICS, in this order. My intention is to identify clusters found by K-Means and validate them (or not) with the capability of finding outliers provided by OPTICS (in a similar fashion as DBSCAN)

After selecting the best suited neighbourhoods with the appropriate venues for our start-up, I´ll use the data from Nomad List to rank these a little bit further to narrow down the eligible neighbourhoods so that the clients can have an easy task selecting the best one to establish their start-up.

3.1 Data Exploration

Now that we have all the data gathered and cleaned, we can start by exploring it and try to find patterns among all cities.

I found in the dataframe 139 unique Venue categories and a total of 8 unique Parent Venue categories.

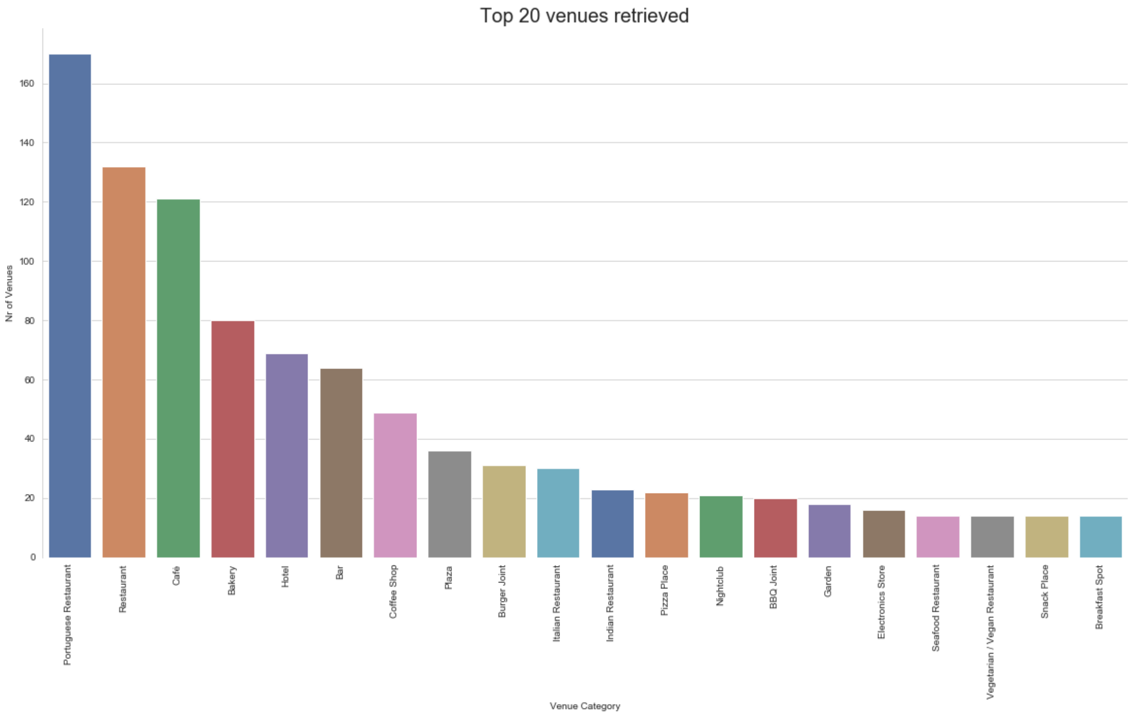


Figure 1 – Top 20 Venues from all Neighbourhoods

Apart from some categories not food related, every other category is food related. The distribution of the Parent Venue categories is worth to be visualised:

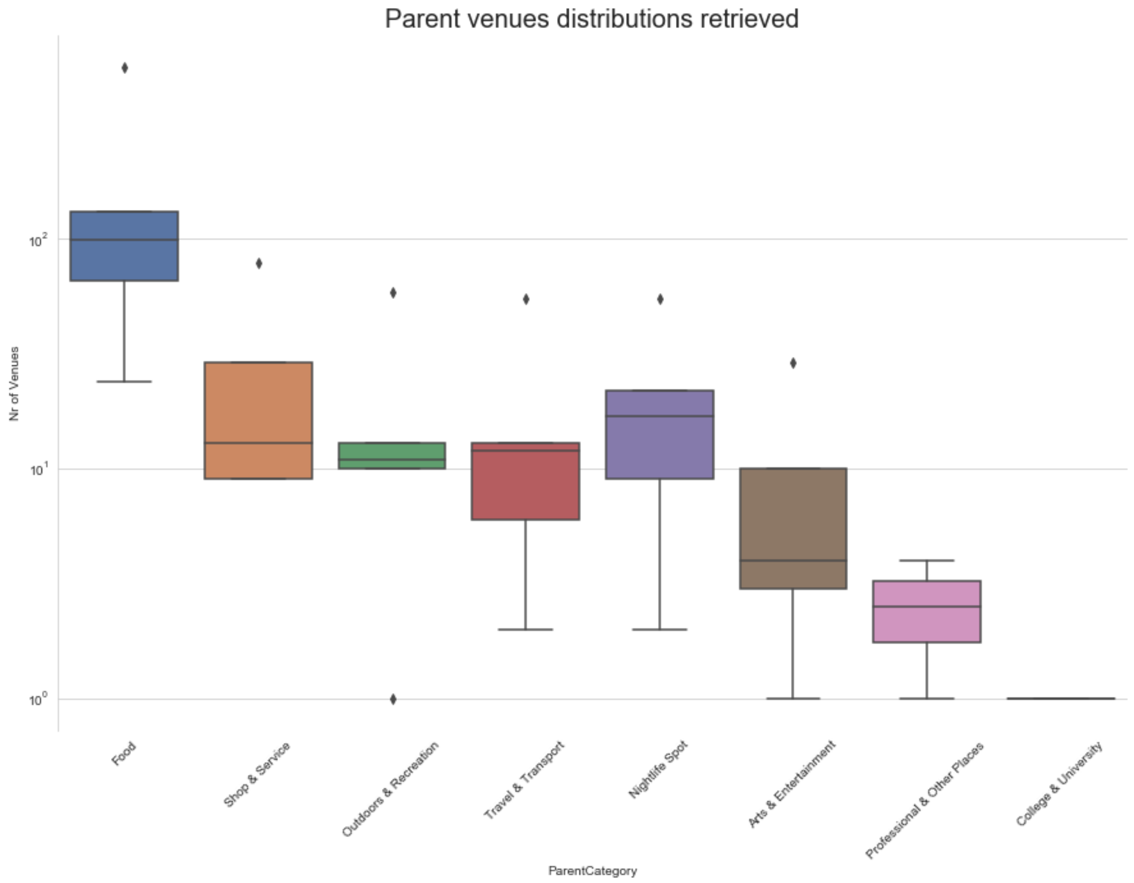


Figure 2 – Box plot distributions of the Parent Categories from all Neighbourhoods

It will be simpler to work only with the Parent categories, in order to prevent the 'curse of dimensionality' (which would be the case when using the original categories) when applying unsupervised learning. I'll ignore the probable outliers for now.

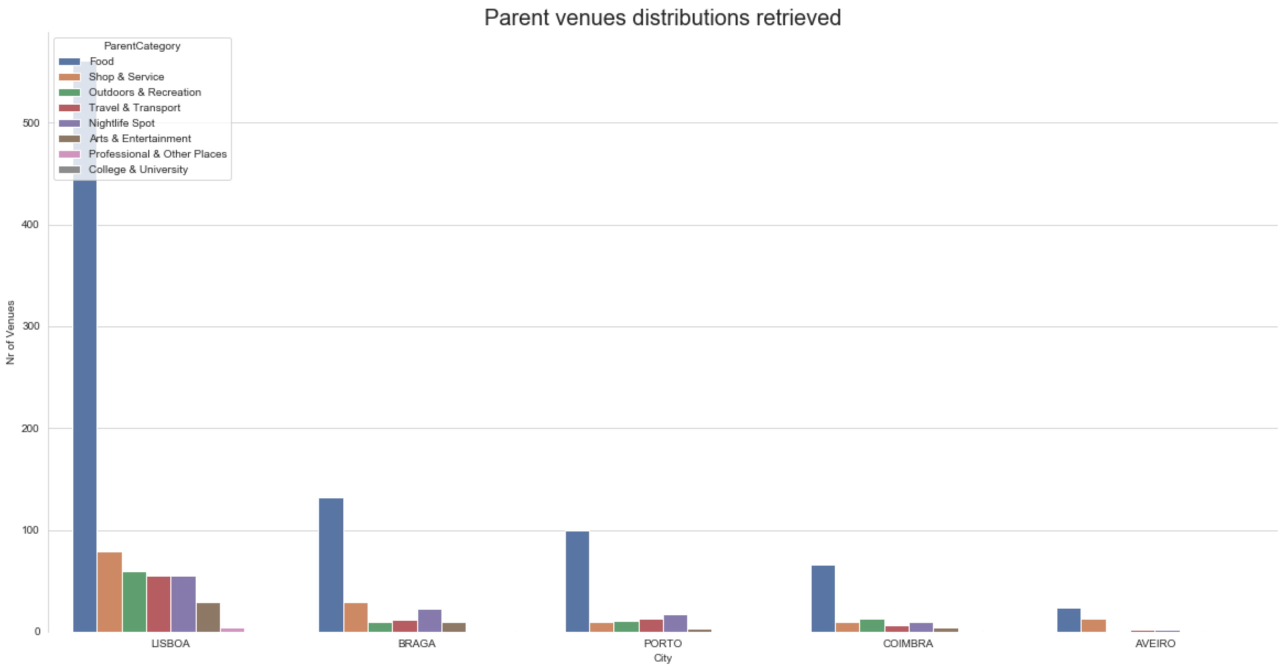


Figure 3 – Number of Venues retrieved by city

It looks like every city is a bit different from each other. Food Venues are the mode for each of them, only changing the proportion of other venues. Also, the amount of venues retrieved for the city of Lisbon is significantly higher than the other cities, regardless of their real dimension. This might be an indicator that there are more venues registered in the city of Lisbon that in the other 4 cities.

3.2 K-Means method

In order to achieve good clustering, first I need to find the optimal number of clusters to do that. In this stage I'm going to try to identify the right number of clusters based on the results of the K-means inertia graph (also kown as elbow method).

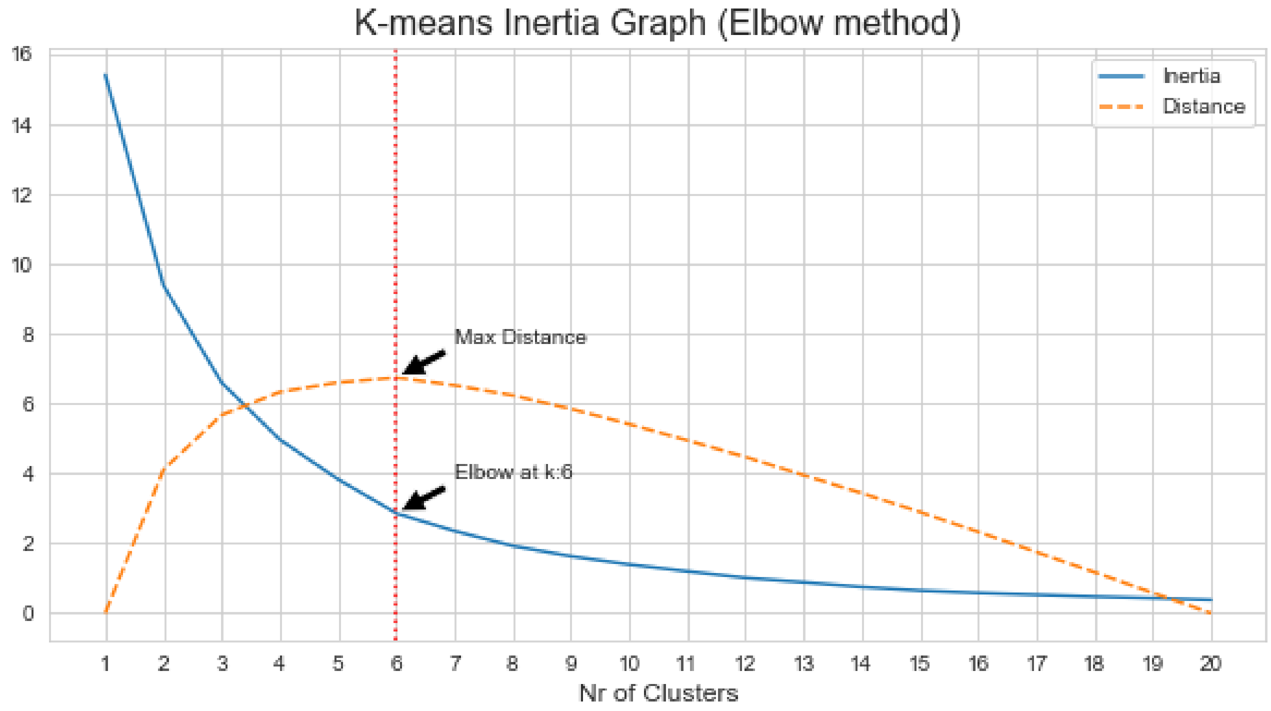


Figure 4 - Elbow Method for K means

The optimal number of clusters is not visually very clear but measuring the distance of each cluster to a line between the first cluster inertia to the last cluster inertia, we can mathematically find the best number of clusters. So the best number of Clusters is 6.

Now I need to profile each cluster (from the best number of clusters found before) to better understand the data.

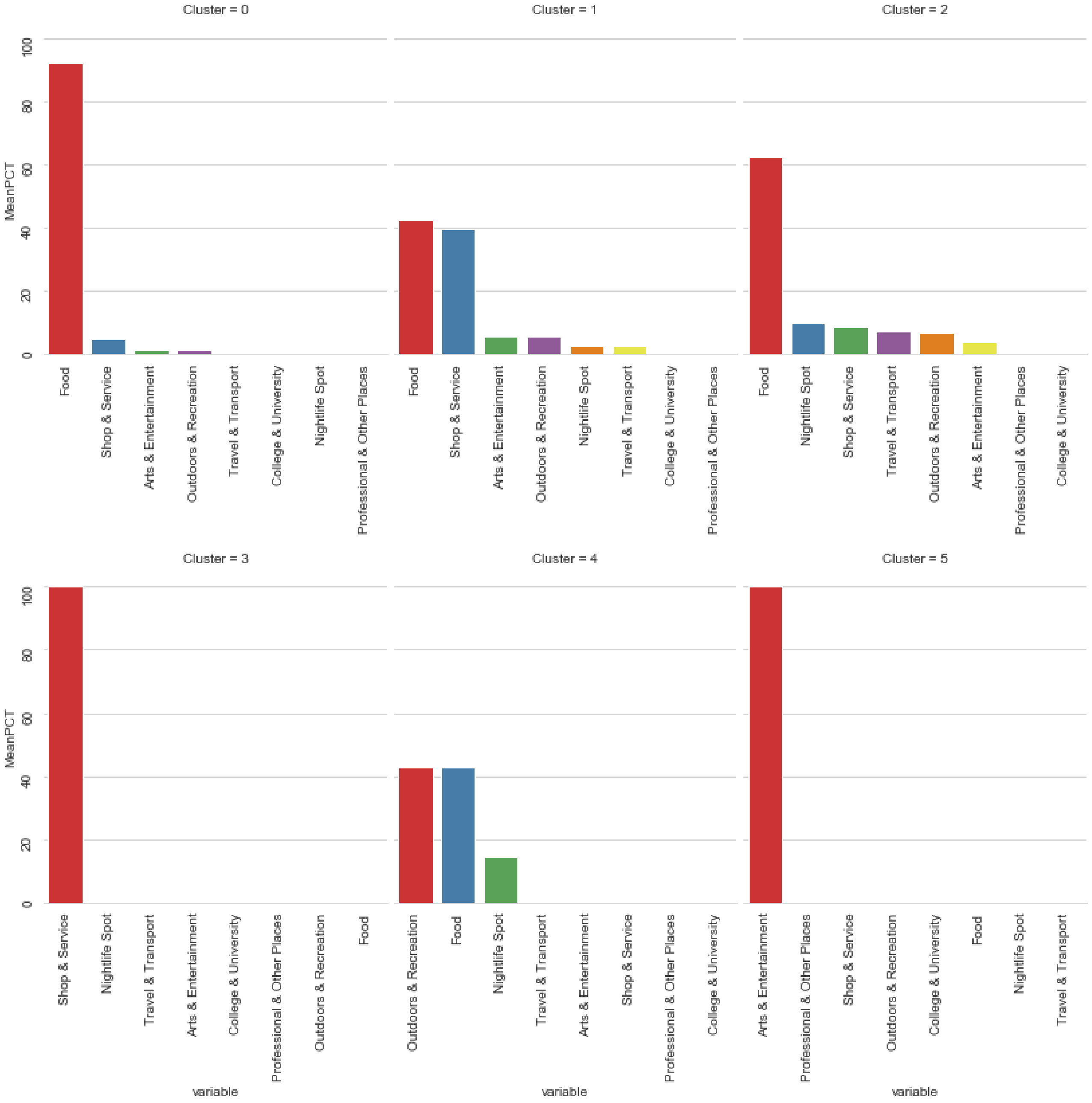


Figure 5 - Profile of each cluster for the K means Method

The classification/naming of these clusters seems very subjective without any other data at the level of information of neighbourhoods, but we can still see different patterns along the clusters.

But instead of trying to name a cluster that can be different over time, my aim is to provide the best neighbourhoods to our start-up based on what it's important to its founders. So, in order to do that, I selected only the neighbourhoods that had values for the following important categories:

- 'Arts & Entertainment',
- 'Outdoors & Recreation',
- 'Professional & Other Places',
- 'Shop & Service',
- 'Travel & Transport'

Applying these restrictions, the target cluster based on the categories ['Arts & Entertainment', 'Outdoors & Recreation', 'Professional & Other Places', 'Shop & Service', 'Travel & Transport'], was cluster number 2.

Plotting these neighbourhoods in a map for each city I'll have the following figures:

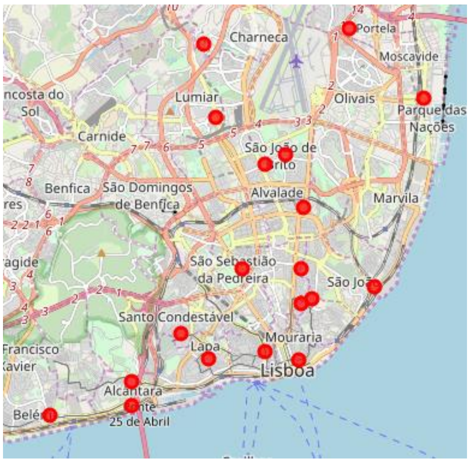


Figure 6 - City of Lisbon selected Neighbourhoods

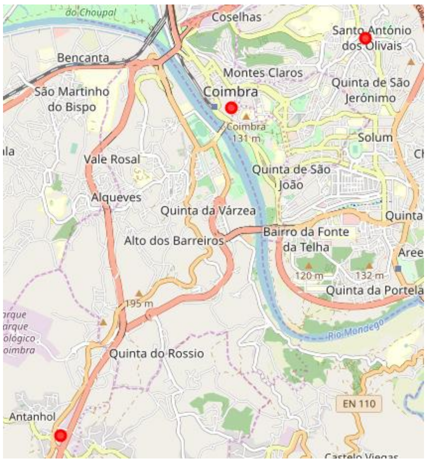


Figure 7 - City of Coimbra selected Neighbourhoods

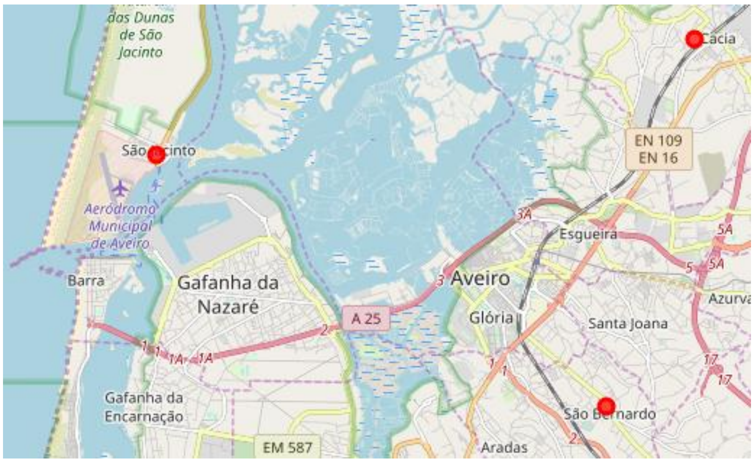


Figure 8 - City of Aveiro selected Neighbourhoods

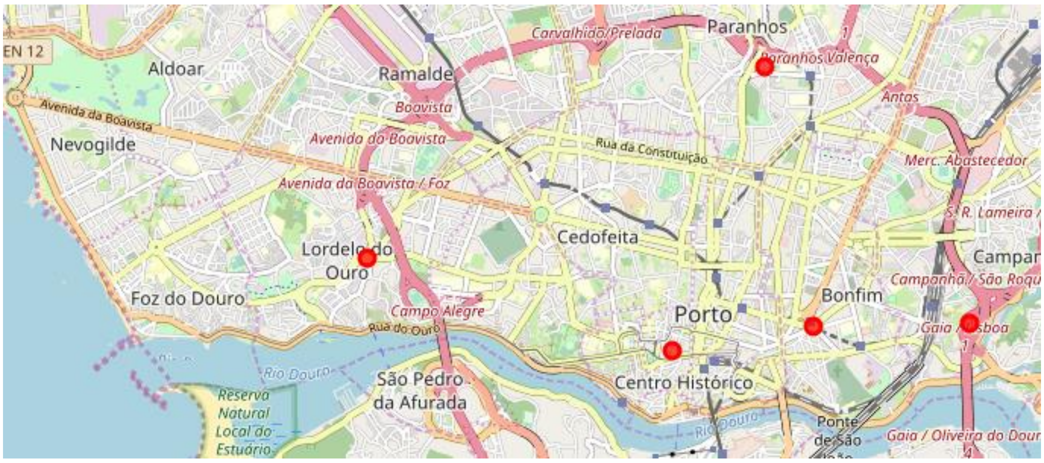


Figure 9 - City of Porto selected Neighbourhoods

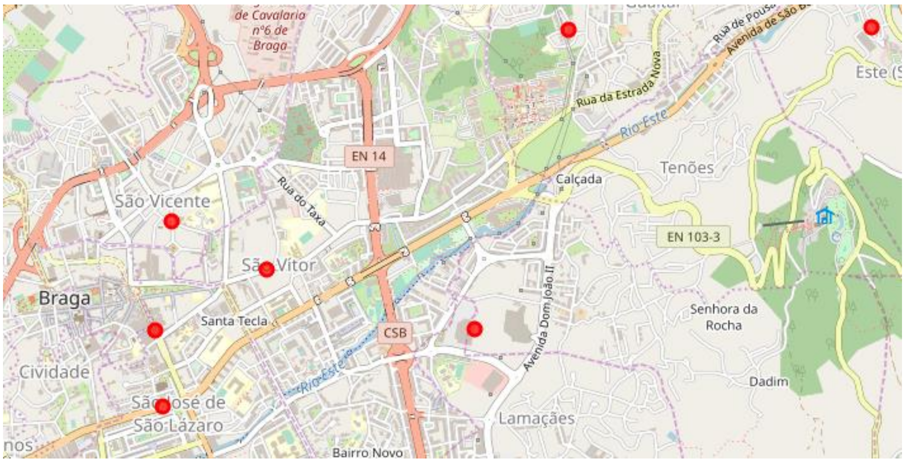


Figure 10 - City of Braga selected Neighbourhoods

Although we found a target cluster of neighbourhoods, the elbow method didn't show any abrupt change in inertia, also I can't, by this method, rule out possible neighbourhoods that might be outliers, thus influencing the overall results.

3.3 OPTICS method

With OPTICS, an advanced algorithm that follows the same principles that of DBSCAN, will identify the right number of clusters and identify any outliers based on every epsilon value.

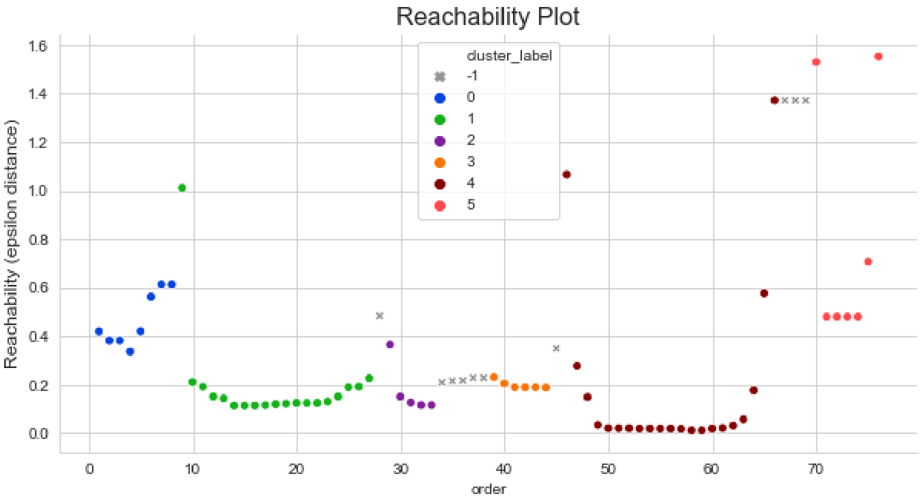


Figure 11 - Reachability Plot for identification of Clusters

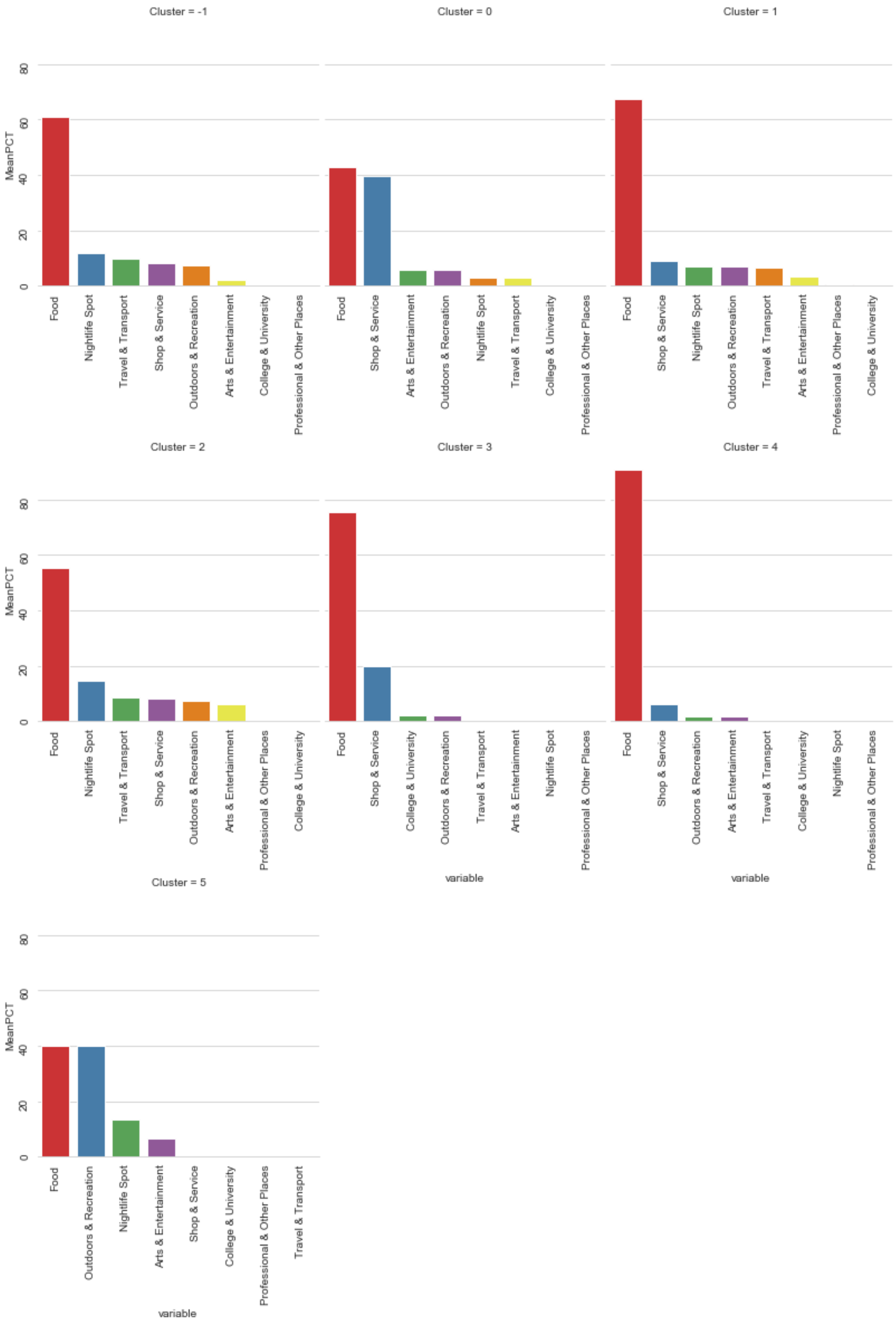


Figure 12 - Profile of each cluster for the OPTICS Method

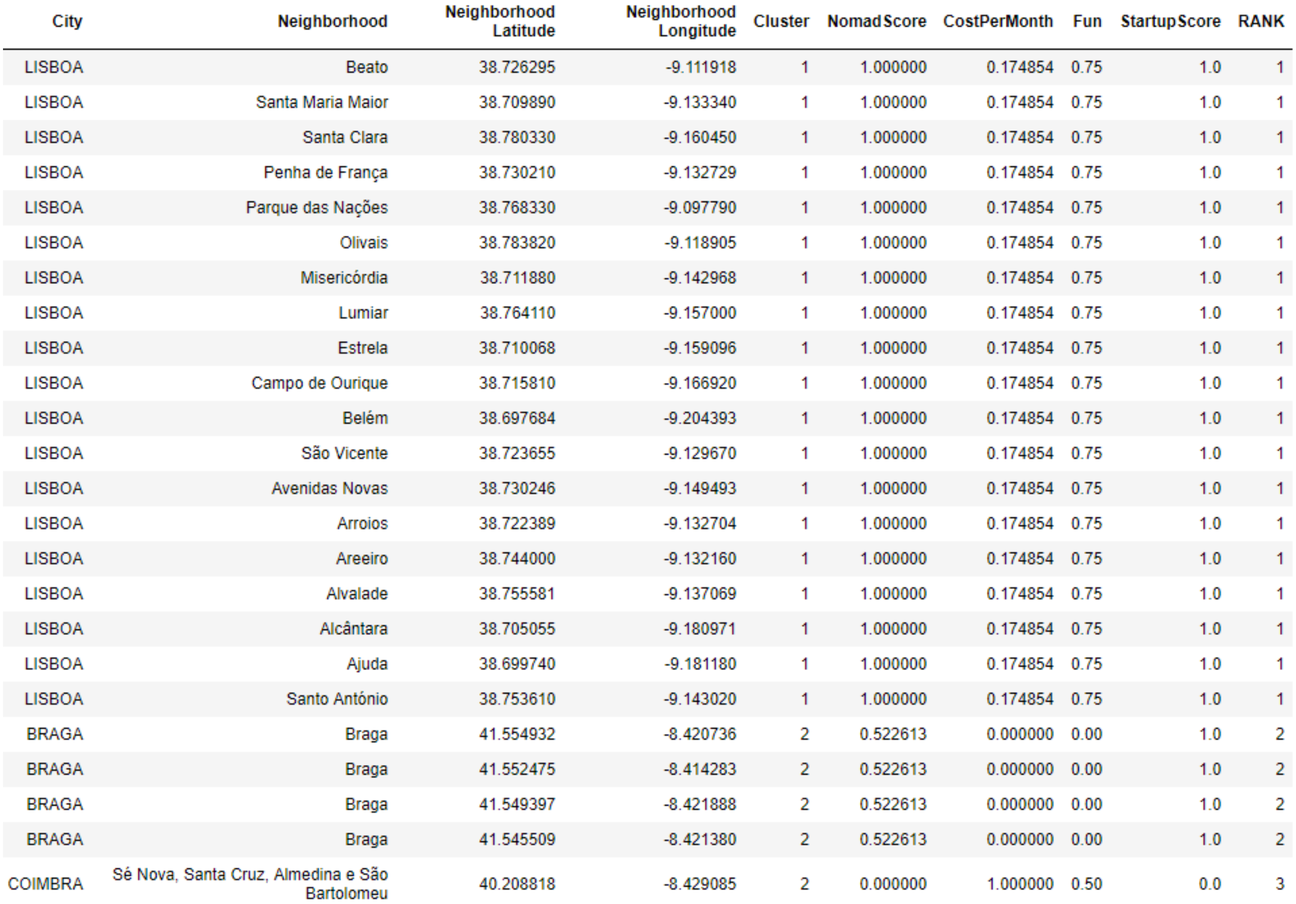
The clusters identified show us the following:

- Cluster nr -1: outliers (will not be used to make any decision)
- Clusters that do not fulfil the interests of the start-up company
- Target Clusters based on categories of interest to the start-up company

Following the same rationale stated in the previous method, one can say that the target clusters based on the categories ['Arts & Entertainment', 'Outdoors & Recreation', 'Professional & Other Places', 'Shop & Service', 'Travel & Transport'], are cluster nr 1 and 2.

Considering the target clusters chosen before, as the most probable neighbourhoods to establish a start-up, we can rank these (to a maximum of rank 3) using the Nomad List dataset. This ranking gives us the following table:

Table 1 - List of the most probable Neighbourhoods to suggest to the Client



City	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Cluster	NomadScore	CostPerMonth	Fun	StartupScore	RANK
LISBOA	Beato	38.726295	-9.111918	1	1.000000	0.174854	0.75	1.0	1
LISBOA	Santa Maria Maior	38.709890	-9.133340	1	1.000000	0.174854	0.75	1.0	1
LISBOA	Santa Clara	38.780330	-9.160450	1	1.000000	0.174854	0.75	1.0	1
LISBOA	Penha de França	38.730210	-9.132729	1	1.000000	0.174854	0.75	1.0	1
LISBOA	Parque das Nações	38.768330	-9.097790	1	1.000000	0.174854	0.75	1.0	1
LISBOA	Olivais	38.783820	-9.118905	1	1.000000	0.174854	0.75	1.0	1
LISBOA	Misericórdia	38.711880	-9.142968	1	1.000000	0.174854	0.75	1.0	1
LISBOA	Lumiar	38.764110	-9.157000	1	1.000000	0.174854	0.75	1.0	1
LISBOA	Estrela	38.710068	-9.159096	1	1.000000	0.174854	0.75	1.0	1
LISBOA	Campo de Ourique	38.715810	-9.166920	1	1.000000	0.174854	0.75	1.0	1
LISBOA	Belém	38.697684	-9.204393	1	1.000000	0.174854	0.75	1.0	1
LISBOA	São Vicente	38.723655	-9.129670	1	1.000000	0.174854	0.75	1.0	1
LISBOA	Avenidas Novas	38.730246	-9.149493	1	1.000000	0.174854	0.75	1.0	1
LISBOA	Arroios	38.722389	-9.132704	1	1.000000	0.174854	0.75	1.0	1
LISBOA	Areeiro	38.744000	-9.132160	1	1.000000	0.174854	0.75	1.0	1
LISBOA	Alvalade	38.755581	-9.137069	1	1.000000	0.174854	0.75	1.0	1
LISBOA	Alcântara	38.705055	-9.180971	1	1.000000	0.174854	0.75	1.0	1
LISBOA	Ajuda	38.699740	-9.181180	1	1.000000	0.174854	0.75	1.0	1
LISBOA	Santo António	38.753610	-9.143020	1	1.000000	0.174854	0.75	1.0	1
BRAGA	Braga	41.554932	-8.420736	2	0.522613	0.000000	0.00	1.0	2
BRAGA	Braga	41.552475	-8.414283	2	0.522613	0.000000	0.00	1.0	2
BRAGA	Braga	41.549397	-8.421888	2	0.522613	0.000000	0.00	1.0	2
BRAGA	Braga	41.545509	-8.421380	2	0.522613	0.000000	0.00	1.0	2
COIMBRA	Sé Nova, Santa Cruz, Almedina e São Bartolomeu	40.208818	-8.429085	2	0.000000	1.000000	0.50	0.0	3

Plotting these Neighbourhoods in a map show us the following:

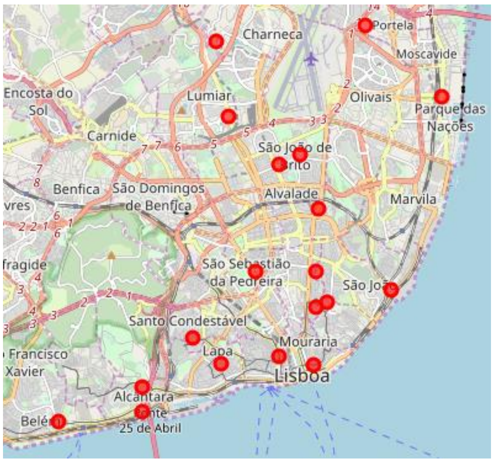


Figure 13 - City of Lisbon selected Neighbourhoods (OPTICS)

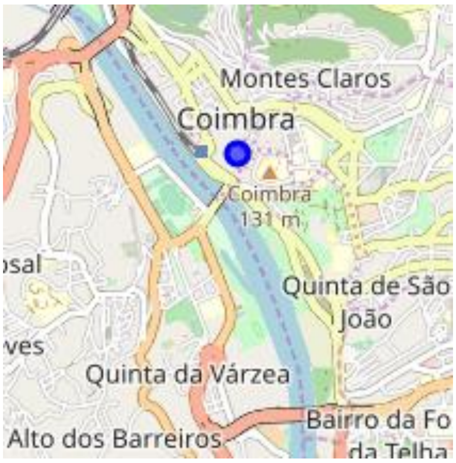


Figure 14 - City of Coimbra selected Neighbourhood (OPTICS)

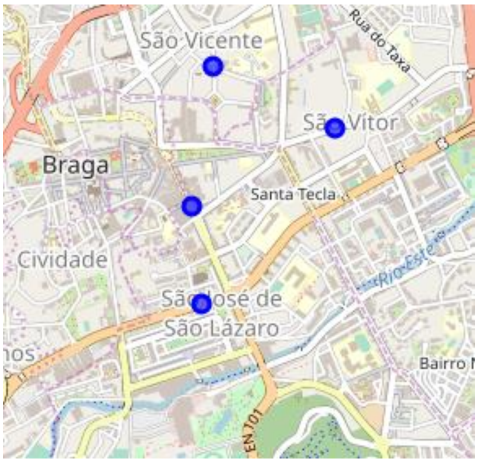


Figure 15 - City of Braga selected Neighbourhood (OPTICS)

The Cluster nº2 is common in the cities of Coimbra and Braga, while the Cluster nº1 is limited to Lisbon. When comparing both clusters it is obvious that the cities of Coimbra and Braga have higher number of Nightlife Spots comparing to Lisbon.

4. Results and Discussion

The results obtained by the exploration of the datasets retrieved and the methodology are very peculiar. First, although not presented in this report and its Jupyter notebook, the time of the day and day of week change the amount and the categories retrieved by the Foursquare Explore Venues API. This fact leads me to advise a workday and during working hours the retrieval of the Venue's Data.

The profile of each city is very different on its venues, one hypothesis I have for further investigation is the fact that smaller cities might have less venues reported to Foursquare, therefore skewing the data to favour the main city (the capital of Portugal - Lisbon). This might not be a surprise if we check the world most popular cities with data on Foursquare.

Furthermore, the amount of venue unique categories is huge comparing to its parent categories. This fact leads me to consider the following scenarios:

- If we consider every category, we might force k-means or other unsupervised learning technique into the 'Curse of Dimensionality' and identifying dissimilar neighbourhoods into the same cluster.
- If we consider only the parent categories, we might get a very good performing unsupervised learning technique but at the cost of losing information.

Without any other data, like the real population for each neighbourhood, the price of housing and commerce per square meter, the overall condition of the neighbourhood in various dimensions, etc., might render this analysis dull and very broad. We might identify a neighbourhood with all the required venues to a start-up thrive, but those venues might have low ratings. The ratings for all the venues were not retrieved because we need to have an advanced account of Foursquare Developer (paid), thus not the scope of this data science project.

After the run of K-Means we found clusters that were easy to tell them apart but somewhat difficult to give them a 'name'. This is also a symptom of the lack of information at the neighbourhood level. The running of the OPTICS algorithm gave the ability to rule out neighbourhoods that are very different form the others, so clustered as outliers. This helps on finding neighbourhood that are similar in a realistic way, even with the lack of further information.

Finally, the use of the Nomad List data, which is at the city level (not neighbourhood) was an appropriate tool to rank the selected neighbourhoods (those which had the venues and characteristics required for the establishment of the start-up) into a small list. This list of neighbourhoods would be presented to the clients so they can make the best-informed decision. Most neighbourhoods presented were located at Lisbon, Porto and Aveiro.

5. Conclusion

The purpose of this project was to broadly identify neighbourhoods for a start-up company to establish in Lisbon, Porto, Braga, Aveiro or Coimbra cities of Portugal. In order to make this identification we collected the needs and desires of that start-up to match the best neighbourhoods with the appropriated venues retrieved from the Foursquare Database. There were two approaches to the clustering of the neighbourhoods and the one being OPTICS because of its capability of identifying outliers. In the end we present a list of candidate neighbourhoods to our clients to establish.

Although the data is very dynamic, the decision selection the best location will be made by the client taking into consideration additional factors like attractiveness of each location, levels of noise or proximity to major roads, real estate availability and it's prices, social and economic dynamics of every neighbourhood. The data retrieved in this project with the previous additional data could be presented as a list of variables from which the Client could chose the most important, feed it to our OPTICS model, ranked by Nomad List and, finally, the presentation of a more refined and short list of the best neighbourhood to settle the Start-up

References

1. **Farmbrough, Heather.** Lisbon 2018: Why Startups Are Booming In The Portuguese Capital. *Forbes.* [Online] 28 de Feb de 2018. https://www.forbes.com/sites/heatherfarmbrough/2018/02/28/all-roads-lead-to-lisbon-why-startups-are-booming-in-the-portuguese-capital/#5399ed1177ea.

2. **Ruivo, Daniel de Castro.** Why Portugal Is the New Land of Opportunity for Tech Startups. *Entrepreneur.* [Online] 21 de Mai de 2018. https://www.entrepreneur.com/article/307526.

3. **Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duche.** Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research.* 2011, Vol. 12, 2825-2830.