IBM Data Science specialization Coursera

Applied Data Science Capstone

Final Project:

The Battle of the Neighbourhoods

Title:

«Et voilà!» Restaurant in Buenos Aires

Author: Juan Ignacio Morales Volosín

Table of Contents

Abstract

The best neighbourhood to open a restaurant in Buenos Aires is sought. Georeferenced data and Buenos Aires government datasets are used to gain insight about this problem. All data is processed and results are extracted with the aid of a k-Means clustering algorithm. An interpretation is carried out, with criteria as competition and potential attendance. A final decission is obtained, namely the neighbourhood of Balvanera, presenting favorable conditions.

Introduction

Problem Description

This project is a preliminary study about establishing the better location to open a new restaurant in Buenos Aires, Argentina. To accomplish this, the city will be devided into its traditional neighbourhoods.

As a first criterion in the selection, competition should be taken into consideration. Accordinlgy, the neighbourhoods presenting a lower amount of restaurants will be preferred.

As a second criterion, potential attendance is of interest. In this case, neighbourhoods with a bigger population will have an advantage. However, there is also the factor of public transportation, which helps in letting people living further from a neighbourhood to potentially travel to it and eat at a restaurant there.

The question is, therefore, which neighbourhood has less restaurants, is more populated and better connected? Its answer is crucial for entrepreneurs in the field of gastronomy.

Background

The Autonomous City of Buenos Aires (or simply Buenos Aires) is the capital city of Argentina and also the largest one in this country. This is in the South American continent, with the city placed on the western shore of the estuary Río de la Plata. It is included in the Greater Buenos Aires, which comprises an urban conglomeration with the fourth biggest population in the Americas. The city is an autonomous district, not being a part of the province of the same name which surrounds it by land. Internally, Buenos Aires is subdivided in communes, that serve some administrative purposes and, at the same time, it is traditionally divided in fourty-eight neighbourhoods.

Presenting one of the highest living qualities in South America, a numerous population and brimming with tourism, this multicultural city offers an attractive prospect to anyone interested in the gastronomical business.

Data

Description

Foursquare

Its API will be used to obtain data related to the first criterion in the selection of a neighbourhood to open a restaurant. It will be fed the coordinates of each neighbourhood, returning a json file. From it the relevant information will be extracted, presenting mainly venues, their categories and their coordinates.

ArcGIS

Their API will also be used, feeding it with the name of a location in order to obtain its geographical coordinates. In this study in particular, it will be used mainly to get the coordinates of Buenos Aires and all its neighbourhoods.

Buenos Aires – Datasets

Additionally to the already mentioned sources, datasets provided by the government of the City of Buenos Aires.

Neighbourhoods

https://data.buenosaires.gob.ar/dataset/barrios/archivo/juqdkmgo-191-resource (in Spanish).

In this dataset different fields can be found inside the .csv file, but the site does not give any explanation about them. Upon inspection of these fields, it is possible to gain some understanding and to infer the following:

- WKT: this is a string field and contains polygons in wkt format, which is widely used and, in this dataset, contains the shape of each neighbourhood.
- Barrio: this is another string field, containing, in this case, the name of the neighbourhood.
- Comuna: this is an integer field, containing the number of the commune to which the neighbourhood belongs.
- Perimetro: this is a float field, containing the perimeter of the neighbourhood, measured in meters.
- Area: this is a float field, containing the area of the neighbourhood, measured in square meters.

Population

https://data.buenosaires.gob.ar/dataset/estructura-demografica/archivo/c44be985-8d7f-4aa4-972e-a7f8f0b796dc (in Spanish)

The explanation given about the fields in this csv file is quite straight-forward: “BARRIO” is a string containing the name of the neighbourhood and “POBLACION” is an integer with the population of said neighbourhood. Unfortunately, this is a dataset with information from 2010, but it’s the best made available by the City’s government.

Bus Stops

https://data.buenosaires.gob.ar/dataset/colectivos-gtfs (in Spanish).

This dataset includes a number of files compressed in a .zip file. The one of interest for this study is “stops.txt”, which internally is actually a .csv file. The site gives no explanation about the dataset, but the fields “stop_lat”, “stop_lon” (both floats) will prove to be the latitude and longitude coordinates of each bus stop. It must be noticed that the listed stops belong not only to the City of Buenos Aires, but also to Buenos Aires Province. This will have to be considered when this set is cleaned and prepared.

Train Stations

https://data.buenosaires.gob.ar/dataset/estaciones-ferrocarril/archivo/juqdkmgo-1021-resource (in Spanish).

In this case, the site explains the dataset, presented in .csv format. For this study, the fields “long”, “lat” (both floats) and “barrio” (string) contain the relevant information: longitude, latitude and neighbourhood,

correspondingly. As in the previous case, the stations belong to both the City and the Province. However, those belonging to the Province have no listed neighbourhood, which simplifies the cleaning and preparation.

Subway Stations

https://data.buenosaires.gob.ar/dataset/subte-estaciones/archivo/juqdkmgo-1992-resource (in Spanish).

Once again, the site does not explain this dataset, which is comprised of a .csv file. From the name of its fields and their inspection it can be inferred that "long" and "lat" (both floats), are the geographical coordinates of each subway station. This is the relevant information to extract for this study. All subway stations are located in the City, as it is the only place in Argentina with such a transportation system.

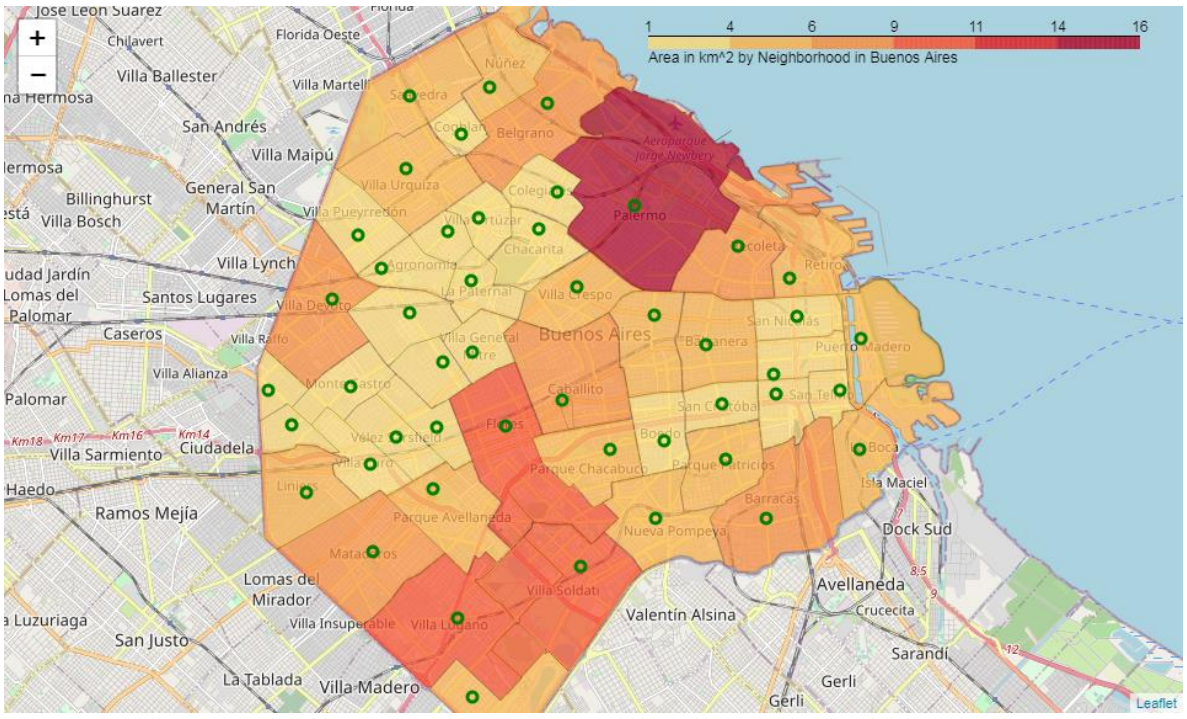
Tram Stations

https://data.buenosaires.gob.ar/dataset/premetro/archivo/juqdkmgo-1651-resource (in Spanish).

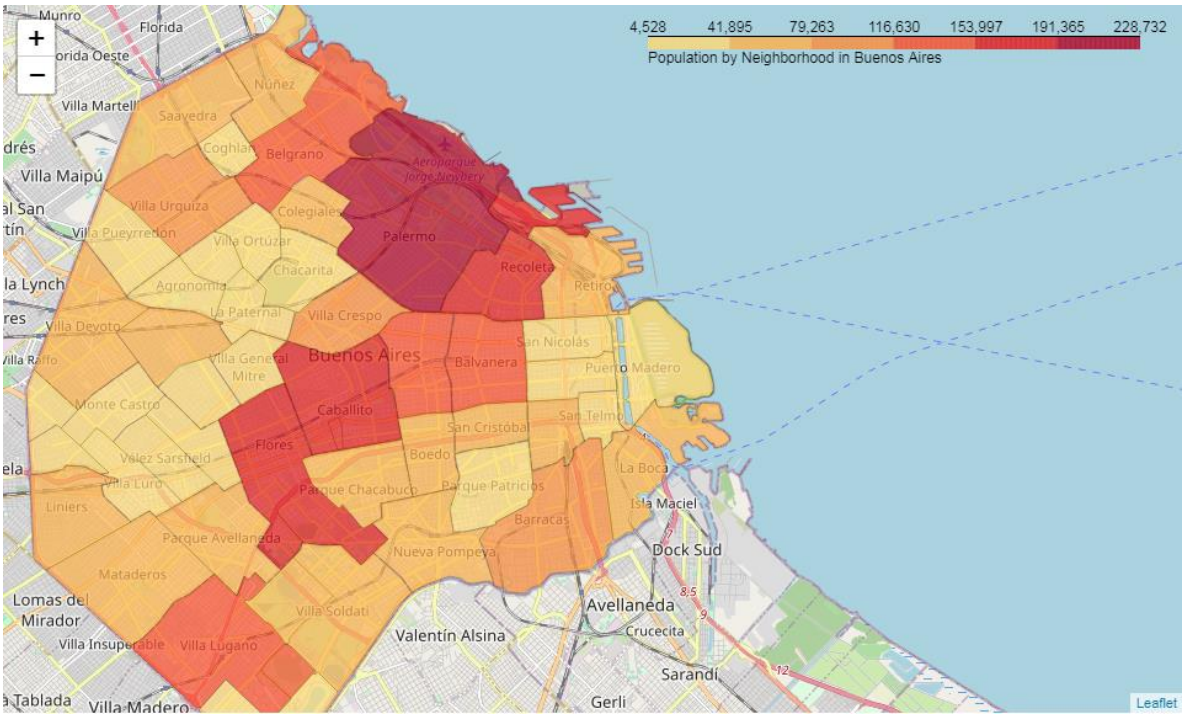
No explanation can be found in the site about this dataset. Analogously to previous cases, upon inspection of the .csv file, the fields "long" and "lat" (both floats) can be recognized and the geographical coordinates of the tram stations. These are the only relevant fields for this study. Also, all tram stations are found in the City only.

Usage

The starting point will be the Neighbourhood dataset, because after some preparation of this set, the names of the neighbourhoods will be fed to the ArcGIS API in order to obtain the coordinates of each of them. These coordinates will be appended to this dataset.



Neighbourhoods colored by size and their coordinates



Population by neighbourhood

The datasets Bus Stops, Train Stations, Subway Stations and Tram Stations will be cleaned and prepared with the aim of obtaining only the geographical coordinates of each stop/station reported in them. Later, using these datasets and the field WKT of the Neighbourhood ones, each stop/station will be assigned to a neighbourhood. Consequently, all stops/stations outside of the limits of the City of Buenos Aires will be dropped. Finally, each of these datasets related to the public transportation system will be turned into dataframes counting how many stops/stations there are in each neighbourhood.

Bus stops in green; train stations in blue; subway stations in yellow; tram stations in orange

Finally, all the dataframes will be combined in a single dataframe in which each row will represent a neighbourhood containing all the prepared information described up to this point and allowing the analysis which ultimately will help answer the question stated in the initial parts of this study.

Methodology

Up to this point in the study, the data has been understood, cleaned and prepared. Different datasets have been used and combined, along with data obtained from georeferential APIs. This will be used for the analysis that will come in the following sections.

To accomplish this, there are factors to be taken into consideration: **competition** and **potential attendance**, the latter influenced by **population** and how well connected is the neighbourhood, i. e. the total **number of stops or stations** that it has. Also, as the neighbourhoods differ in size, it is more appropriate to analyze these factors as area densities. That is, dividing the values of *Restaurants*, *Population* and *Total Transp* in each row, by the value of *Area* in that same row.

Then, a new dataframe will be created with the name of the neighbourhood, its geographical coordinates, its restaurant density, its population density and its transportation density.

Because each feature of interest has very disparate values (some much bigger than the others), the data will be normalized using the Min-Man method.

Results and Discussion

From this comprehensive dataframe and the normalized data extracted from it, it is possible to proceed and gain true understanding of the data, aiming to answer the business problem.

In order to extract better insight out of all the information gathered up to this point, the k-Means clustering algorithm will be used to find similar neighbourhoods and make the final analysis simpler. This will hopefully give a better notion of the neighbourhoods which better suite the problem in study.

As a first approach, before a finer analysis of each cluster, the mean values for each of them are presented. These will be sorted placing in the first rows those clusters where lower competition is expected.

The clustering results will be evaluated taking into consideration that in the first place, the lowest possible level of competition (red bars) is of interest. Secondly, the highest possible level of population density (blue bars). Finally, in the third place, the highest possible level of transport density, or transportation access (purple bars). With this in mind, the graphic of grouped bars is presented as follows.

Restaurant densities; Population densities; Transport Densities

Now, the exploration of each cluster in the order of appearance in the graphic, serving as a further explanation to it.

Cluster 7: This appear to be a cluster with very little competition, a medium level of population density, but on the lower side about how well connected it is.

Cluster 5: This appears to be a cluster with very little competition, againt, but with a low level of both population density and transportation access.

Cluster 3: This small cluster appears to have very little competition, again, but a very low level of both population density and transportation access.

Cluster 4: This is another small cluster offering very little competition, once more, but a very high level of population density and a high level of transportation access. Particularly, the neighbourhood of **Balvanera** seems to be the more adequate so far and, because of the densities shown, it will be difficult to find a better one.

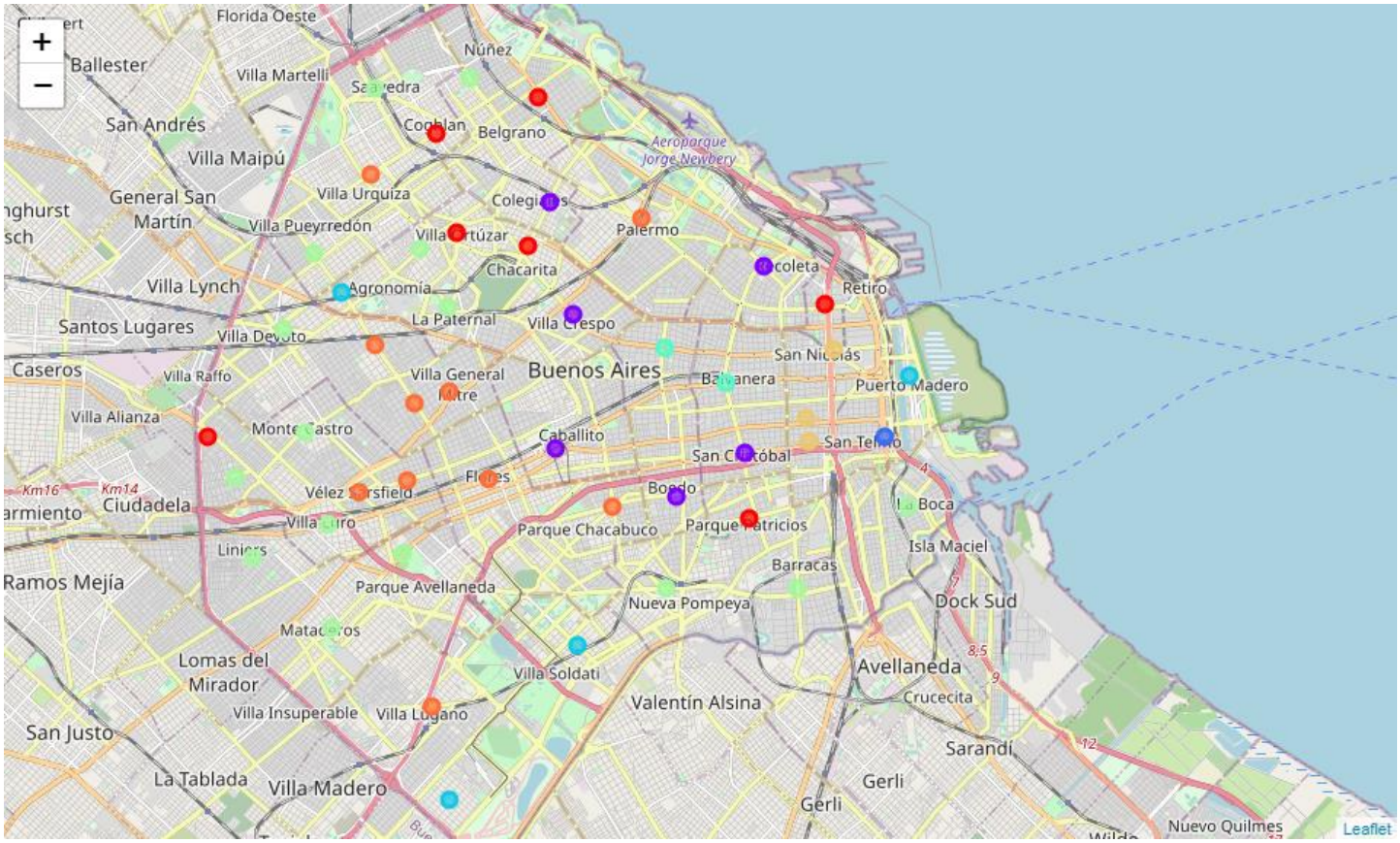
Cluster 1: This cluster shows a low level of competition, a high level of population density and a medium level of transportation access.

Cluster 0: This cluster shows a low level of competition, combined with a medium to low level of both population density and transportation access.

Cluster 6: This small cluster shows a low level of competition, a medium of population density and very high transportation access.

Cluster 2: This small cluster simply is the neighbourhood of San Telmo. It shows a the highest level of competition, a medium of population density and high transportation access.

To bring this section to a close, a final visualization of the neighbourhoods with their markers colored accordingly to the cluster they belong to.



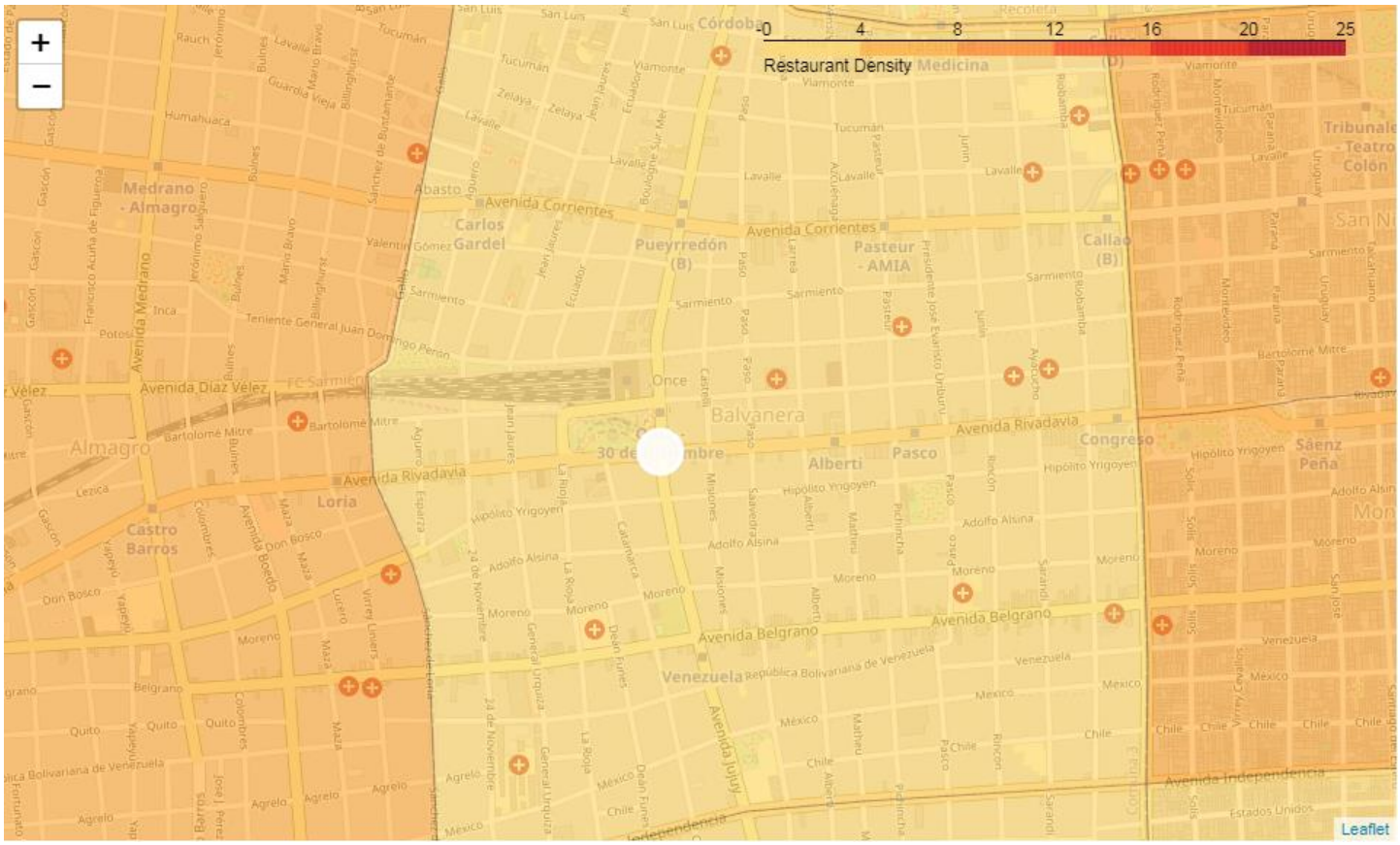
Cluster 0: red; Cluster 1: purple; Cluster 2: dark blue; Cluster 3: blue; Cluster 4: cyan; Cluster 5: green; Cluster 6: yellow; Cluster 7: orange

Conclussion

This division into eight clusters served as a filter, useful in realizing the general properties of the neighbourhoods in each of them. As seen, the level of competition is low in a number of them, but the other criteria fail. On the other extreme of the spectrum, when the seconday criteria (popupation and transport density) become both higher, the competition also increases. This left **Cluster** 4 as a promising candidate, with reasonably little competition and high values for the remaining features. When inspected closely, Cluster 4 truly had the best combination of features, especially considering the neighbourhood of **Balvanera**. This presented very little competition, very high density population and high access to public transport, making it ideal in this study.

Therefore, **answering the problem question**, those entrepreneurs looking **for opening a restaurant in Buenos Aires** should do good use of this study and further research what other conditions offers the neighbourhood of **Balvanera**.

Finally, a visualization of the selected neighbourhood.



The selected neighbouhood: Balvanera, and its non-normalized restaurant density

Future Research

For everyone with the intention of undertaking the project of opening a restaurant in Buenos Aires, It would be in their best interest to keep in mind that this is a preliminary study. Other factors might still play an important role in choosing such a location. Some of which could be mentioned now: income of the population living in the neighbourhoods, price of buying or renting the necessary premises, closeness to other kind of venues who could attract clients, tourist affluence, and so on, even criminality rates.

On the other hand, the data used to feed the model used to cluster the neighbourhoods, could have been polished a bit more before the modeling. It would seem that cluster 2, which is the neighbourhood of San Telmo sets a very high bar for restaurant density. This actually makes sense, because that neighbourhood is very visited by tourists, because of its history and picturesque places. In a refined follow-up to this study, it is suggested that outlier cases be handled more efficiently.