The Battle of Neighborhoods: Buenos Aires data – Airbnb

Author: Máximo La Pietra – maximo.lapietra@ing.austral.edu.ar

1) Introduction/Business Understanding:

1.1 Background

Buenos Aires is the capital and largest city of Argentina. The city is located on the western shore of the estuary of the Río de la Plata, on the South American continent's southeastern coast. Buenos Aires' quality of life was ranked 91st in the world in 2018, being one of the best in Latin America.In 2012, it was the most visited city in South America, and the second-most visited city of Latin America (behind Mexico City).

It is known for its preserved European architecture and rich cultural life. Buenos Aires is a multicultural city, being home to multiple ethnic and religious groups.

Several languages are spoken in the city in addition to Spanish, contributing to its culture and the dialect spoken in the city and in some other parts of the country. This is because since the 19th century the city, and the country in general, has been a major recipient of millions of immigrants from all over the world, making it a melting pot where several ethnic groups live together and being considered one of the most diverse cities of the Americas.

According to the World Travel & Tourism Council, tourism has been growing in the Argentine capital since 2002. In a survey by the travel and tourism publication Travel + Leisure Magazine in 2008, visitors voted Buenos Aires the second most desirable city to visit after Florence, Italy. In 2008, an estimated 2.5 million visitors visited the city.

As an Argentinian future engineer who is extremely interested in travelling and knowing new cultures, I believe that Buenos Aires is a truly wonderful location for international tourists to visit.

1.2 Problem Description

Global tourism activity has shown sustained growth in recent years decades, both in terms of the movement of people and the volume of foreign exchange. From 1990 to the present, the number of international travelers has grown on average 4.1% annually, estimating that in 2015 it would be around 1,180 million, tripling the number of international tourists since then.

In terms of annual generation of income from international tourism, the figures are significant as they reach USD 1,350 billion. Average annual growth stands at 6.7% for the 25-year period, which implies that the volume of international tourism spending was increased by a figure of six.

Moreover, various estimates indicate that globally tourism explains 9% of global GDP, generates 1 in 11 jobs, is responsible for 6% of total exports, and 30% of service exports.

A similar scenario to the international one has been verified in Argentina where the activity tourism has recorded significant growth over the last 25 years, with the arrival of numerous investments and the creation of jobs, where in particular, the City of Buenos Aires (CABA) has positioned itself as one of the most important tourist destinations within the region and the country.

Thus, tourism has come to present significant relevance within the economic sphere of the City, reason why it is extremely important to be able to quantify the economic impact this activity generates.

As a consequence, the objective of this document is to estimate the impact of the offer of the activities related to tourism, and the distribution of the venues, hotels/hostels and important places of the capital city of Argentina.

Airbnb is a booming industry with the latest rise in tourism worldwide. Over 20% of the total tourists worldwide are aged between 19-25. And around 80% of them prefer to spend less on accommodation by opting for hostels. This means the demand for hostels and other accommodations is only going to grow further and more people would want to invest in this platform.

- How should a new businessperson decide where to open a particular accommodation?
- What factors should he look at before investing?
- Which neighborhood venues affect a user's rating for *location* of hostel?

At the same time, it is difficult for a traveler, especially first-timers, to select a hostel from among many options. Hostel reviews are subjective and differ from person-to-person and one cannot solely depend on them to make a decision. It is especially important to consider other aspects like price and neighborhood, which can greatly influence one's experience of the city/country.
I will try to answer the following questions:

- 1) What is the determining factor / price of a publication?
- 2) What is the percentage of occupation of the apartments?
- 3) What is the average income of a host at AirBnB in Buenos Aires?
- 4) Is there seasonality in reviews?
- 5) Which neighborhood / neighborhoods dominate the AirBnB platform in BsAs?
- 6) How does proximity to transportation affect hostel rating?

1.3 Target Audience

This project will be useful for two groups of audience:

1. Travelers: Help them make an informed decision while choosing a particular accommodation by providing an in-depth analysis of hostels and their neighborhood.

2. Businessperson: Provide useful information and models which can help them where to open their first/next business related to tourism and hosting travelers.

2) Analytic Approach

I will be taking two approaches in the project.

Firstly, I will use exploratory data analysis (EDA) to uncover hidden properties of data and provide useful insights to the reader, both future travelers and investors.

Secondly, I will use prescriptive analytics to help a businessperson decide a location for new hostel. I will use *clustering* (KMeans).

3) Data Requirements

Following are the datasets used in the project along with the reasons for choosing them:

1. [Airbnb Buenos Aires listings] (http://insideairbnb.com/get-the-data.html): This dataset was downloaded from the Airbnb webpage. This is the core dataset with which I'll work.

2. [Airbnb Buenos Aires reviews] (http://insideairbnb.com/get-the-data.html): Also, from the Airbnb webpage, in order to analyze the reviews in Buenos Aires.

3. [Foursquare API] (https://developer.foursquare.com/docs/api): This API will help me get the venues around the hostel which I will use for EDA and clustering.

4.[BuenosAiresNeighborhoods](http://cdn.buenosaires.gob.ar/datosabiertos/datasets/barrios/barrios.geojson): Thisgeojson file will help me get the location of the neighborhoods which I will use for EDA and clustering.

Firstly, I will use the list of Airbnb posts from listings dataset and use Foursquare API to get venues around the accommodations. I will then use EDA to explore the neighborhood and how it affects the price of the hostel or hotel. I will also use the reviews dataset to analyze seasonality and tendencies regarding tourism in Buenos Aires.

Secondly, I will combine the above data with the neighborhood geolocation and develop clustering models to better understand the ideal place to be a host, or the cheaper and most enjoyable place to stay.

4) Methodology

4.1 Feature Extraction

In order to obtain the features, One Hot Encoding is used in terms of categories. Each category represents a venue. After, each category is transformed into a binary classification, with the values being 1 or 0. Moreover, the data is grouped by the neighborhood they are obtained from, using the mean of each neighborhood feature to do calculations. This provides us a venue for each row and each column will contain the frequency of a certain category

4.2 Exploratory Data Analysis

- Analyzing graphs containing the most common venues in Buenos Aires.

- Checking for tendencies and patterns.

- Evaluating the occupancy rates and income statistics.

- Evaluating the influence of being a superhost.

4.3 Unsupervised Machine Learning

A clustering algorithm is implemented in order to obtain similar neighborhood in both cities. In this case, K-Means is used due to its simplicity and its similarity approach to found patterns.

K-Means:

- K-Means is a clustering algorithm. This algorithm search clusters within the data and the main objective function is to minimize the data dispersion for each cluster. Thus, each group found represents a set of data with a pattern inside the multi-dimensional features.

The number of clusters must be determined before running the algorithm, as it is one of the inputs. Therefore, in order to determine the optimal number of clusters, the elbow method will be utilized. This method consists of a chart that compares the mean square error vs number of clusters is done and the elbow is selected.

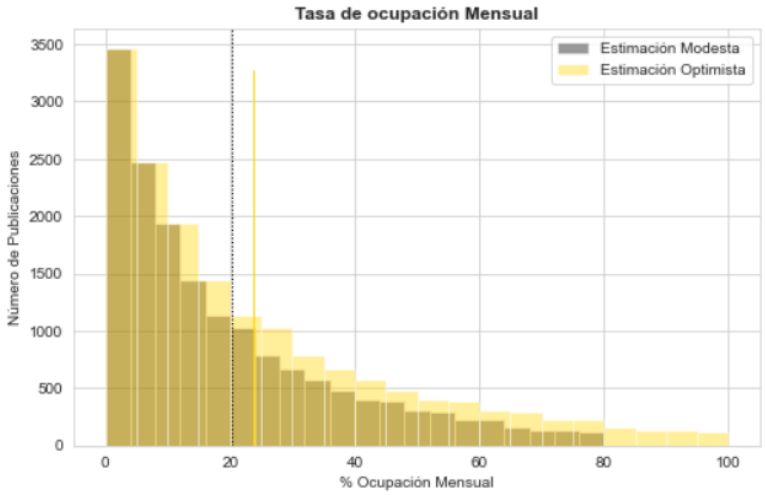
5) Data Preparation / Pre – Processing

Explained in the notebook. The cleaned datasets are available locally and can be sent by email if requested.

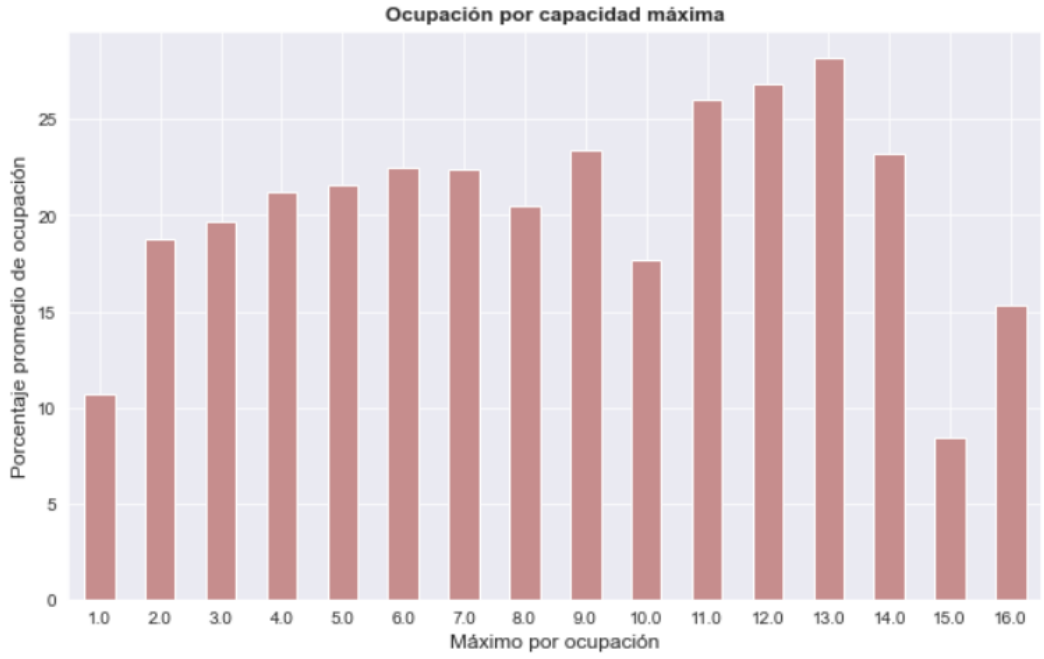
6) Analysis

6.2 Exploratory Data Analysis

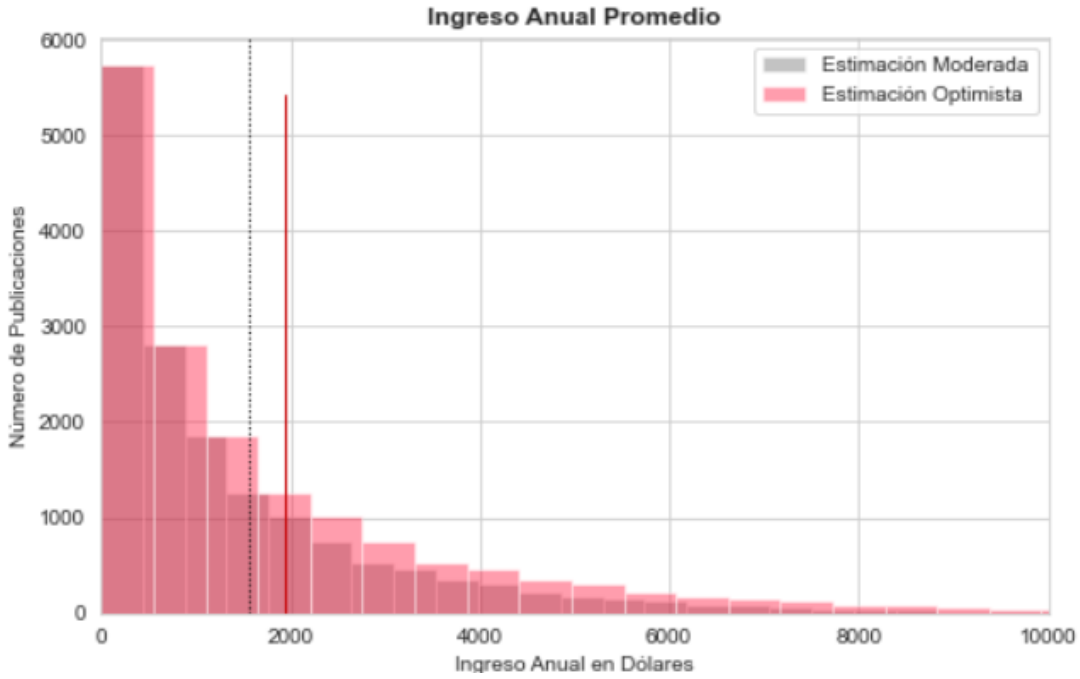
Firstly, a monthly occupancy rate was calculated. Airbnb official information has been retrieved from their webpage. Its distribution looks like this:



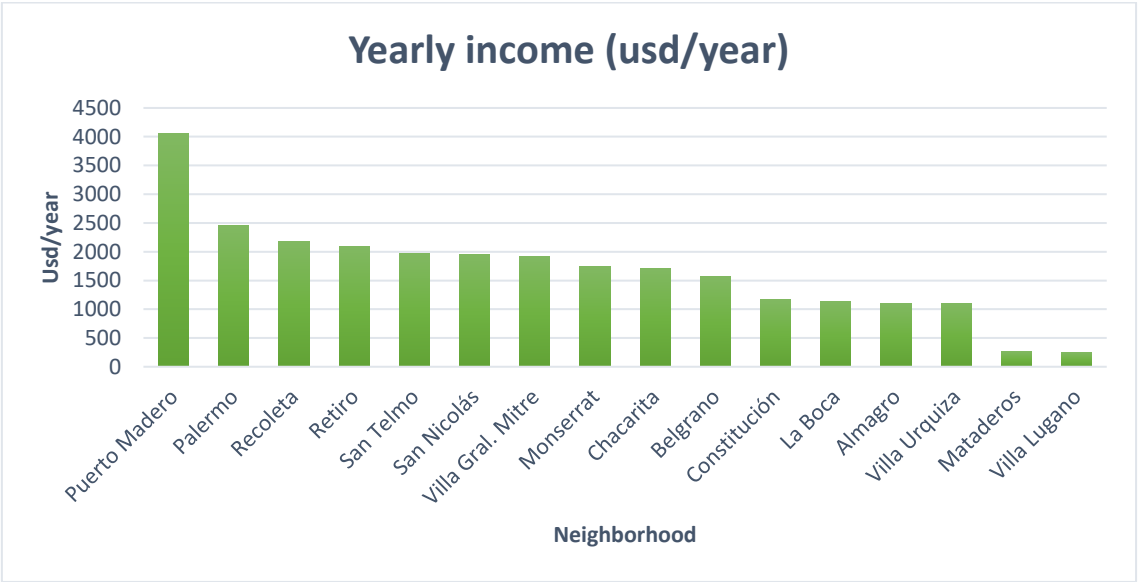
By the distribution of maximum guests allowed:



The average annual income is distributed like this:

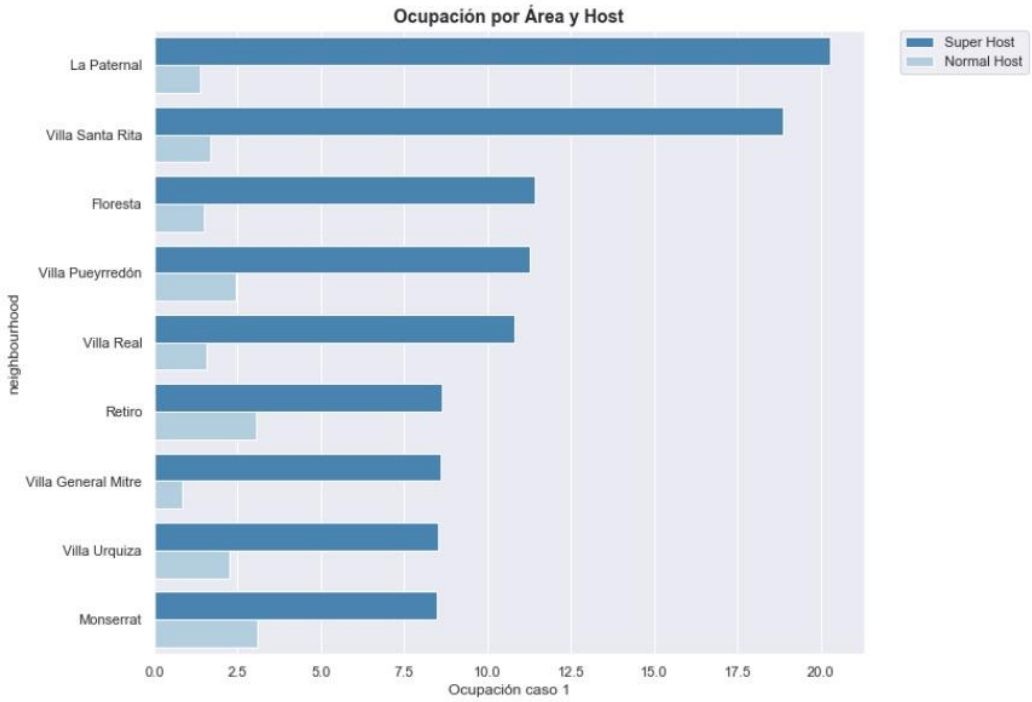


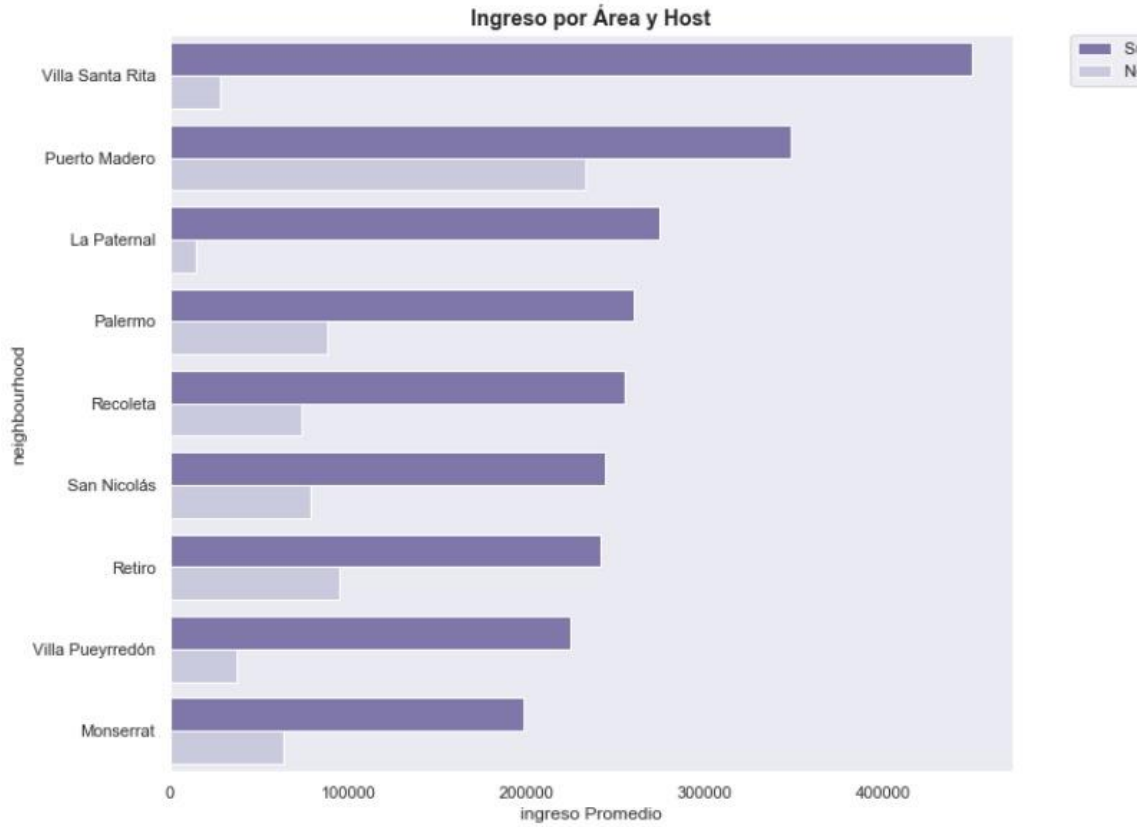
The main neighborhoods, order by income are:



By analyzing room types:

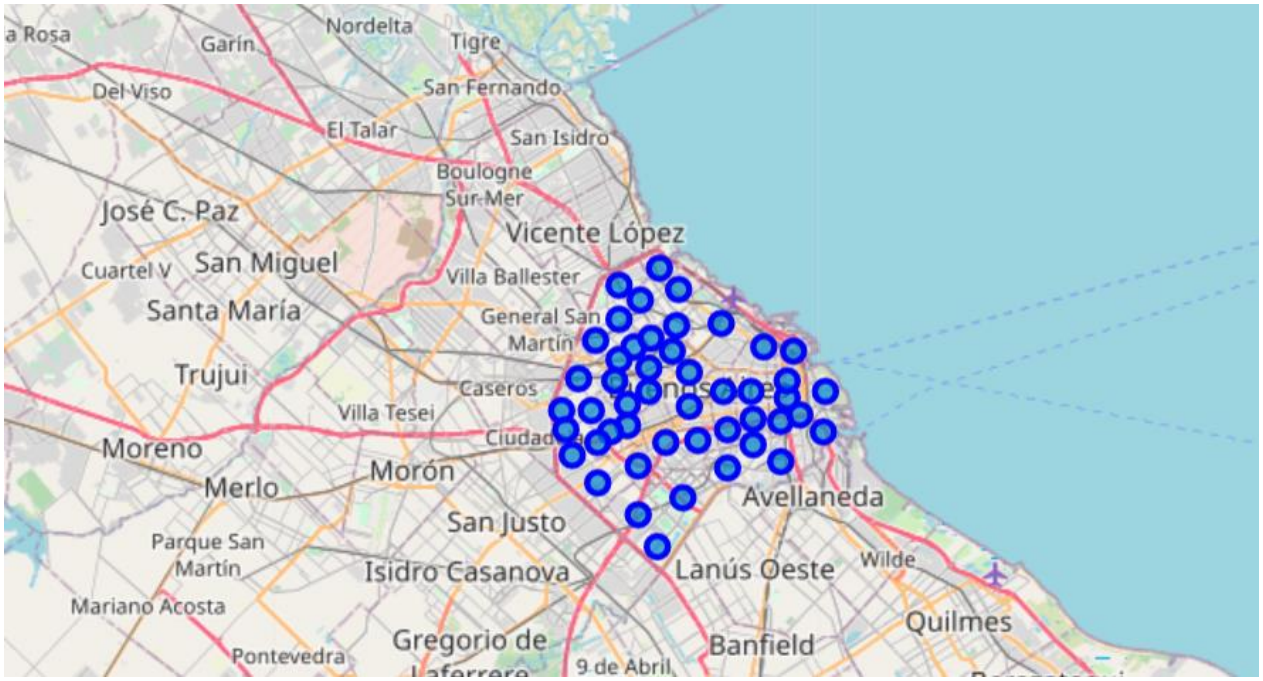
Being a superhost represents lots of different benefits, wich can be expressed in the next graphics:



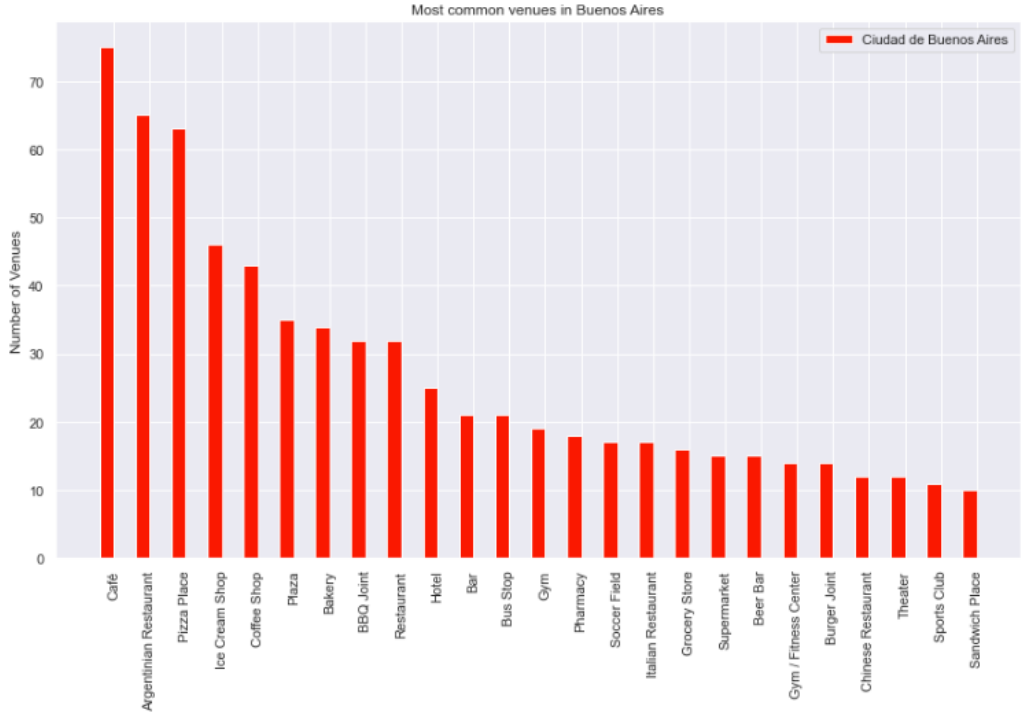


Seasonality trends are shaped like the next distribution:

Analyzing the neighborhoods, inside the map of Buenos Aires:

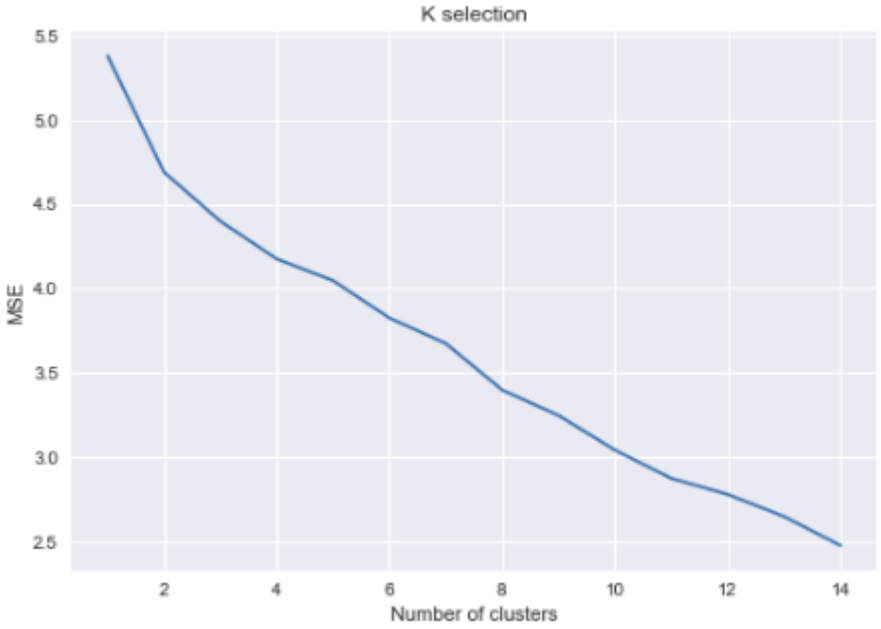


The most common venues in Buenos Aires are:



6.2 Clustering

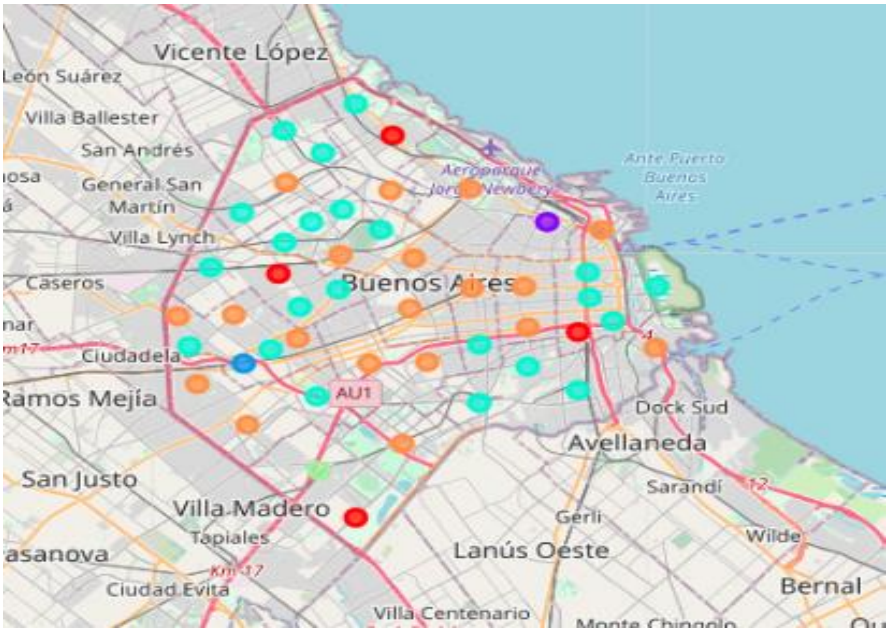
Moving forward, the clustering algorithm is implemented. The mean squared error (MSE) is plotted vs the number of clusters in order to obtain the optimal number of clusters. The number of clusters start with the value of 1 increasing until value of 15. The graph looks like this:



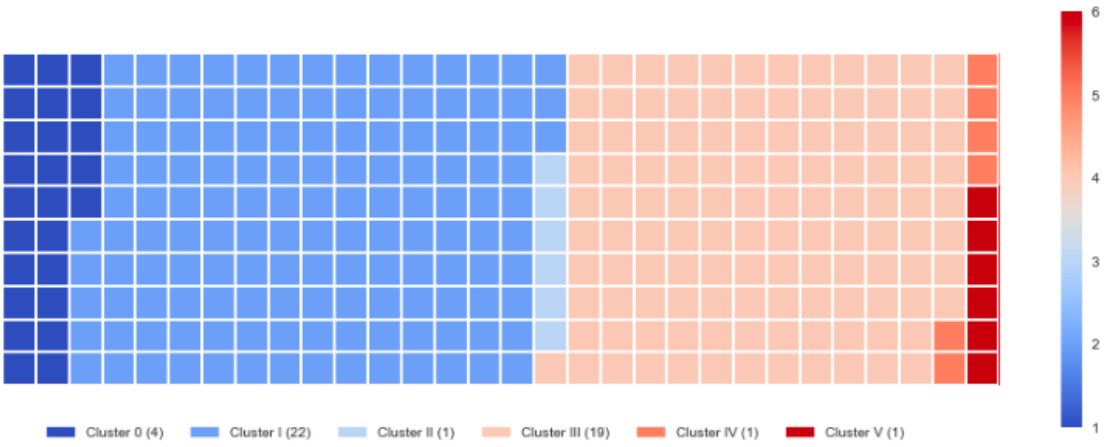
As expected, the MSE will decrease as the number of clusters increases. It is possible to see that the elbow is shown in k=6.

Once the number of clusters is selected, the clustering algorithm is repeated through samples and each neighborhood is labeled according to the clusters found.

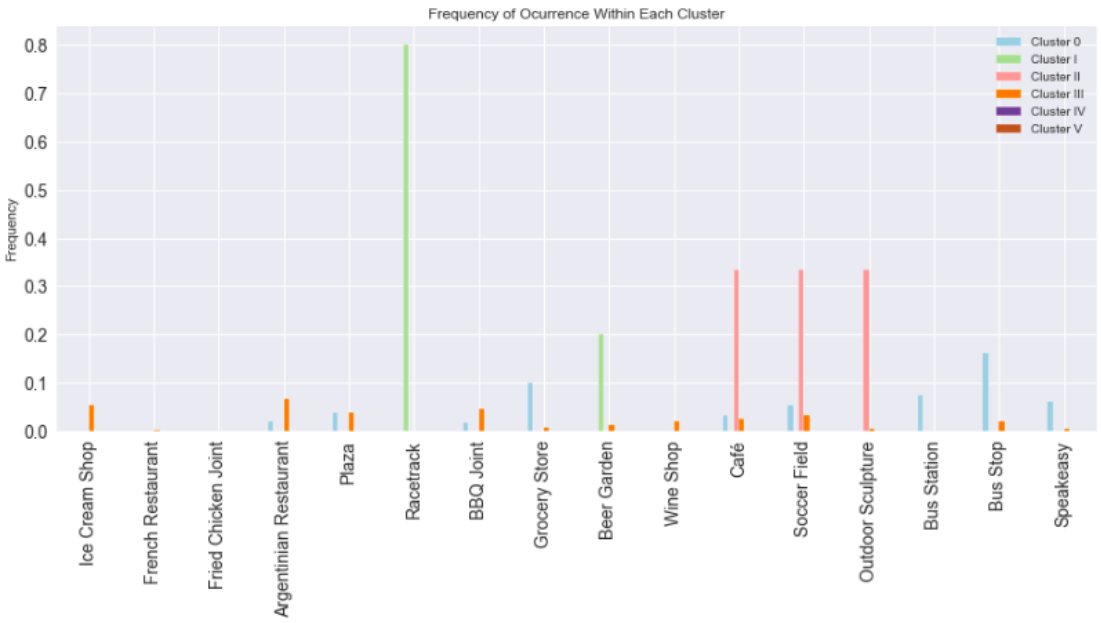
For visualization purposes, the geographical data is again plotted but representing the different clusters (each cluster has one particular color).



Moreover, the proportion of neighborhoods assigned to each cluster can be depicted in a waffle chart. It is represented by this image (there are two big clusters):

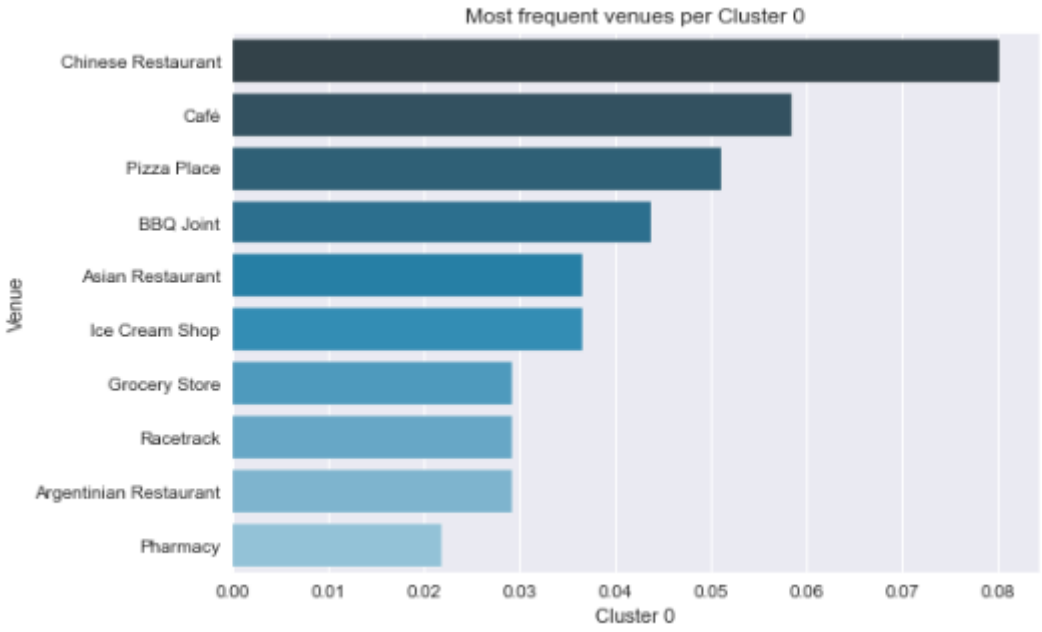


For further analysis, the following bar chart is employed in order to explore the insights within the clusters. The graph below portrays the features with higher frequency in the centroids.

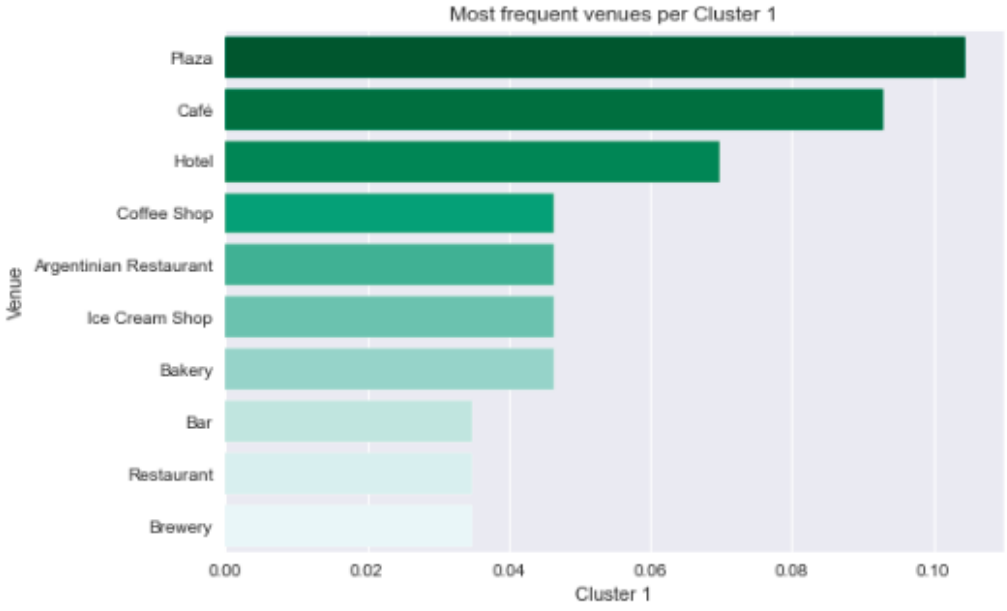


For in depth study, here are the top 10 venues per cluster:

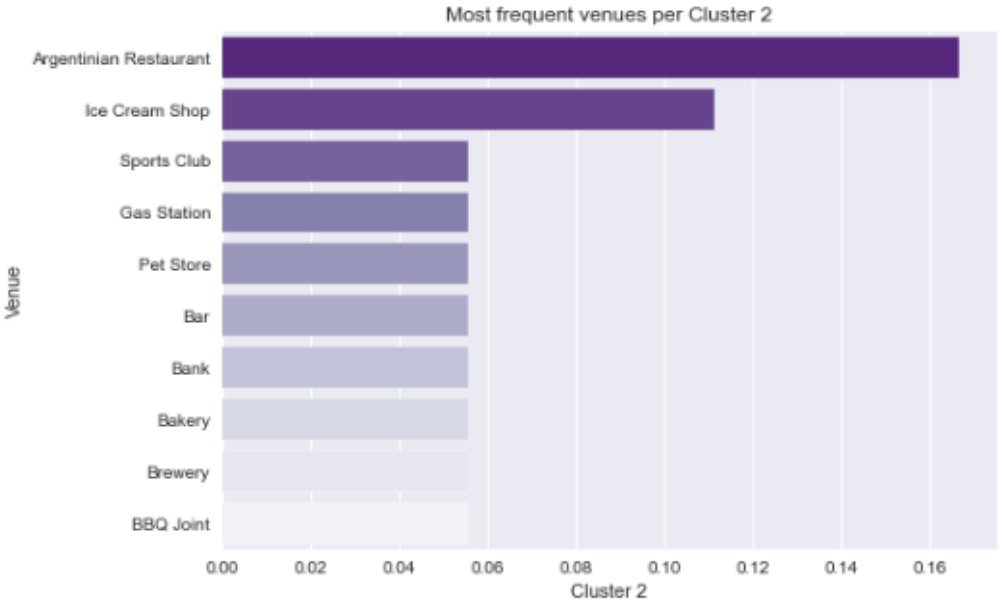
- Cluster 0


Most frequent venues per Cluster 0
Chinese Restaurant
Café
Pizza Place
BBQ Joint
Asian Restaurant
Ice Cream Shop
Grocery Store
Racetrack
Argentinian Restaurant
Pharmacy
Venue
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
Cluster 0

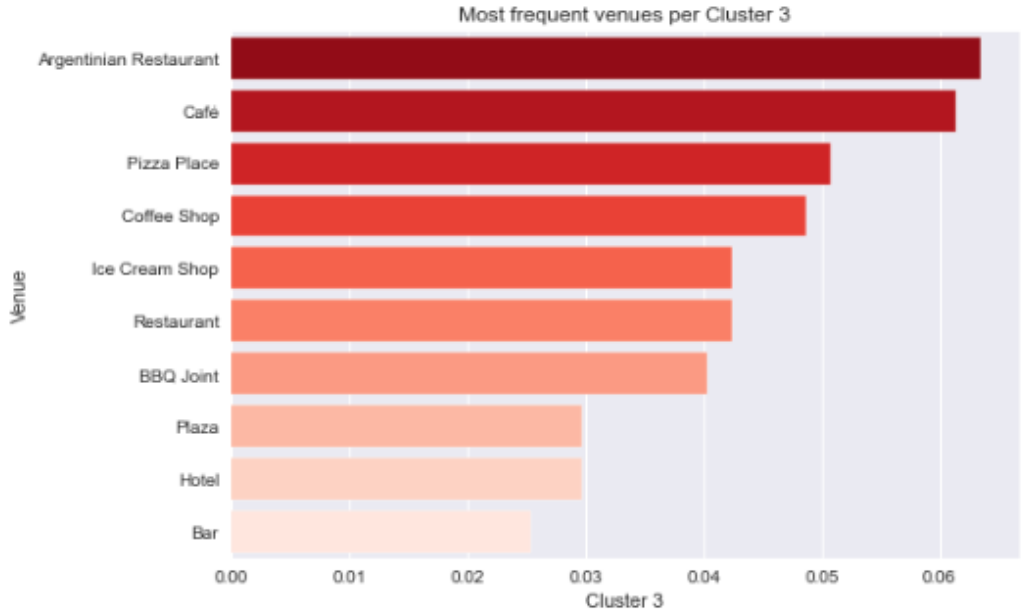
- Cluster 1


Most frequent venues per Cluster 1
Plaza
Café
Hotel
Coffee Shop
Argentinian Restaurant
Ice Cream Shop
Bakery
Bar
Restaurant
Brewery
Venue
0.00
0.02
0.04
0.06
0.08
0.10
Cluster 1

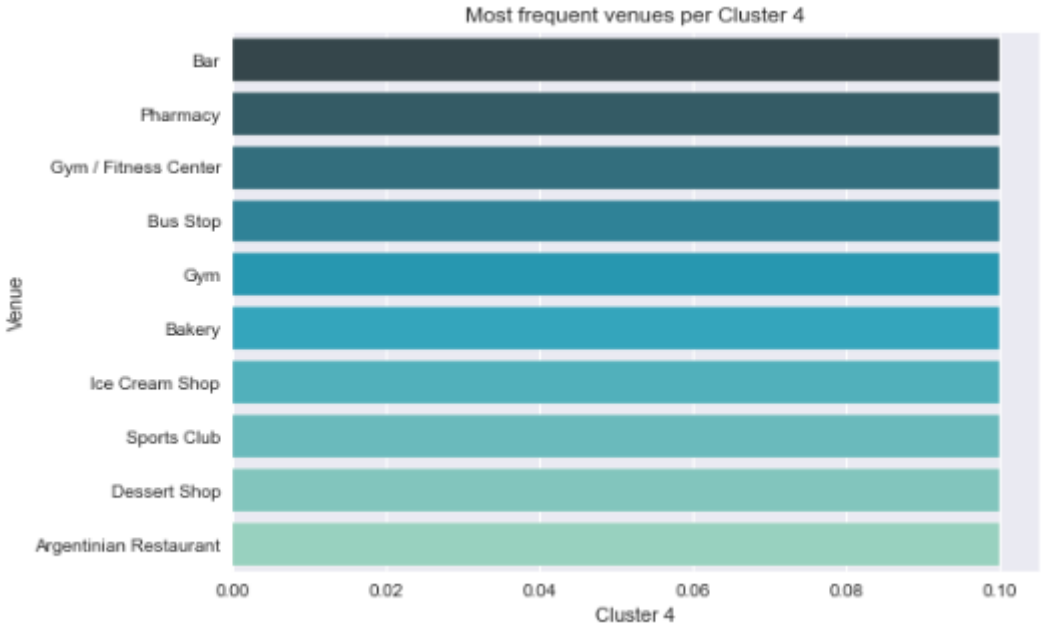
- Cluster 2



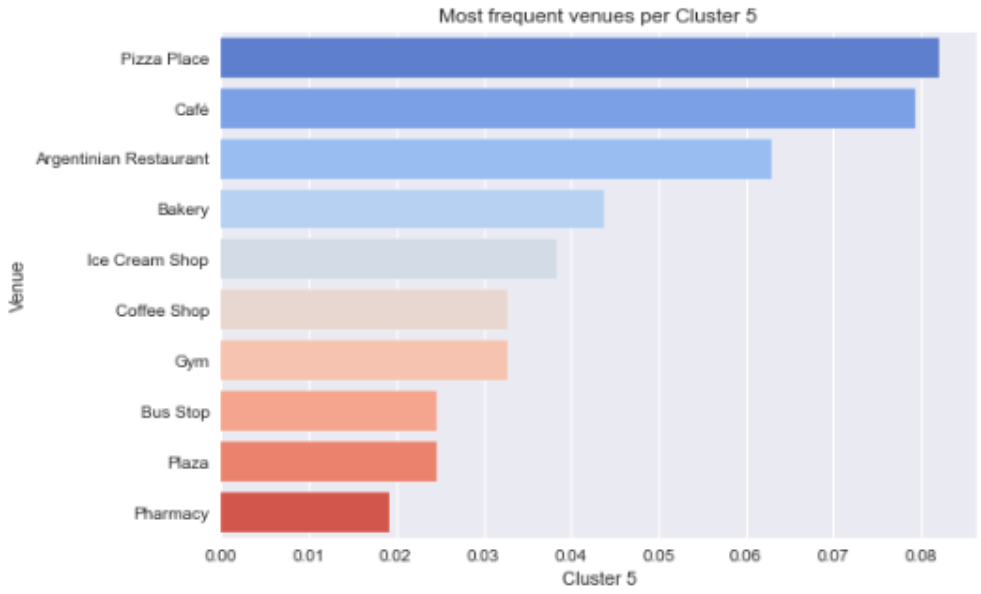
- Cluster 3



- Cluster 4



- Cluster 5



7) Conclusion

We got a glimpse of the tourism scene in Buenos Aires and were able to find out some interesting insights which might be useful to travellers as well as people with business interests. Let's summarize our findings:

- The price of a post is higher in Palermo, Puerto Madero and Recoleta.
- Palermo has de widest offer of accomodations, and venues nearby.
- Hotel rooms are the most expensive places to stay.
- Being a SuperHost is a simple way of increasing profits.
- The starting price of hostels does not vary much depending on its distance from the city center.
- We can group Buenos Aires neighborhoods into 6 clusters.
- Cafes and Argentinian Restaurants are the most common venues (in clusters 2 and 3 especially).
- There is seasonality in Argentina: peaks of tourism in March, July and November.
- 1666 Usd is the approximate annual income for most of the posts in Airbnb.

There are many things which I have assumed while making the above claims since we were working with limited. I'll try to expand the dataset for a more comprehensive study.

This analysis is not perfect due to the limitations of using Foursquare API without a pro account, limiting the calls one can do. However, this project, is useful for differentiating neighborhoods within a city and finding similar ones in another.