THE BATTLE OF NEIGHBORHOODS

IBM Data Science Professional Certificate - Capstone Project Report

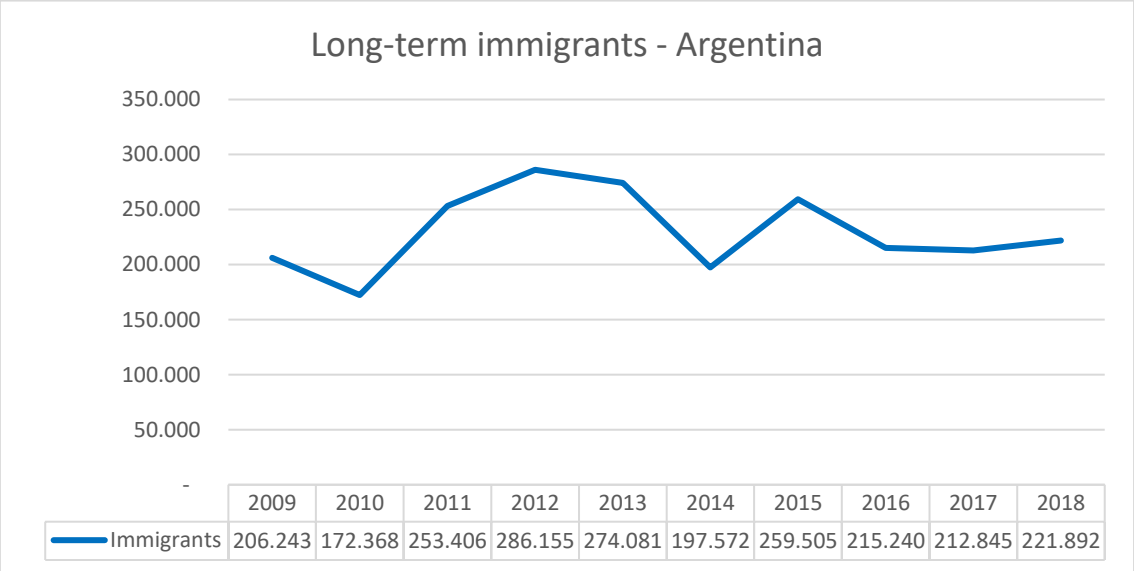
Leonardo Fernández
leohfermamdez@gmail.com

Table of contents

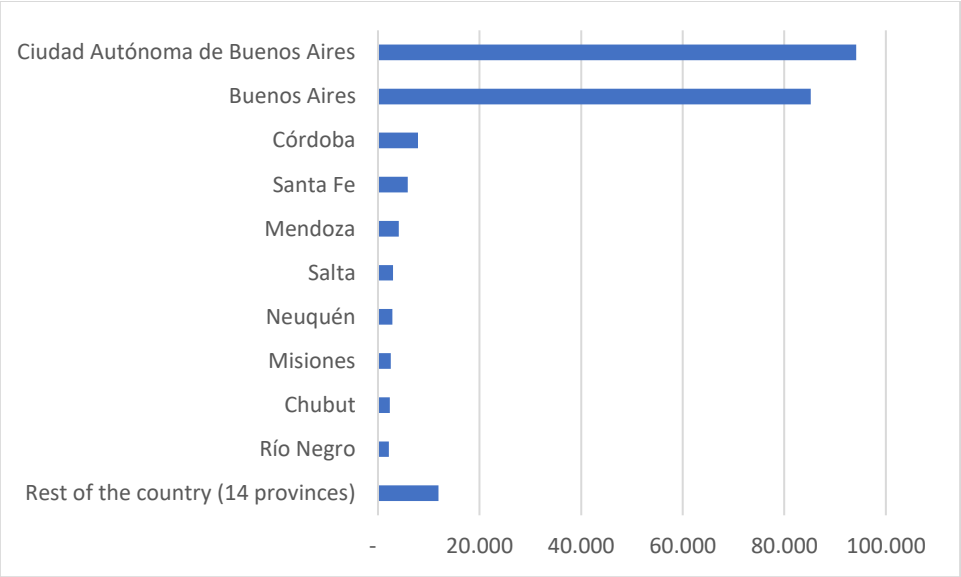
Introduction

Background

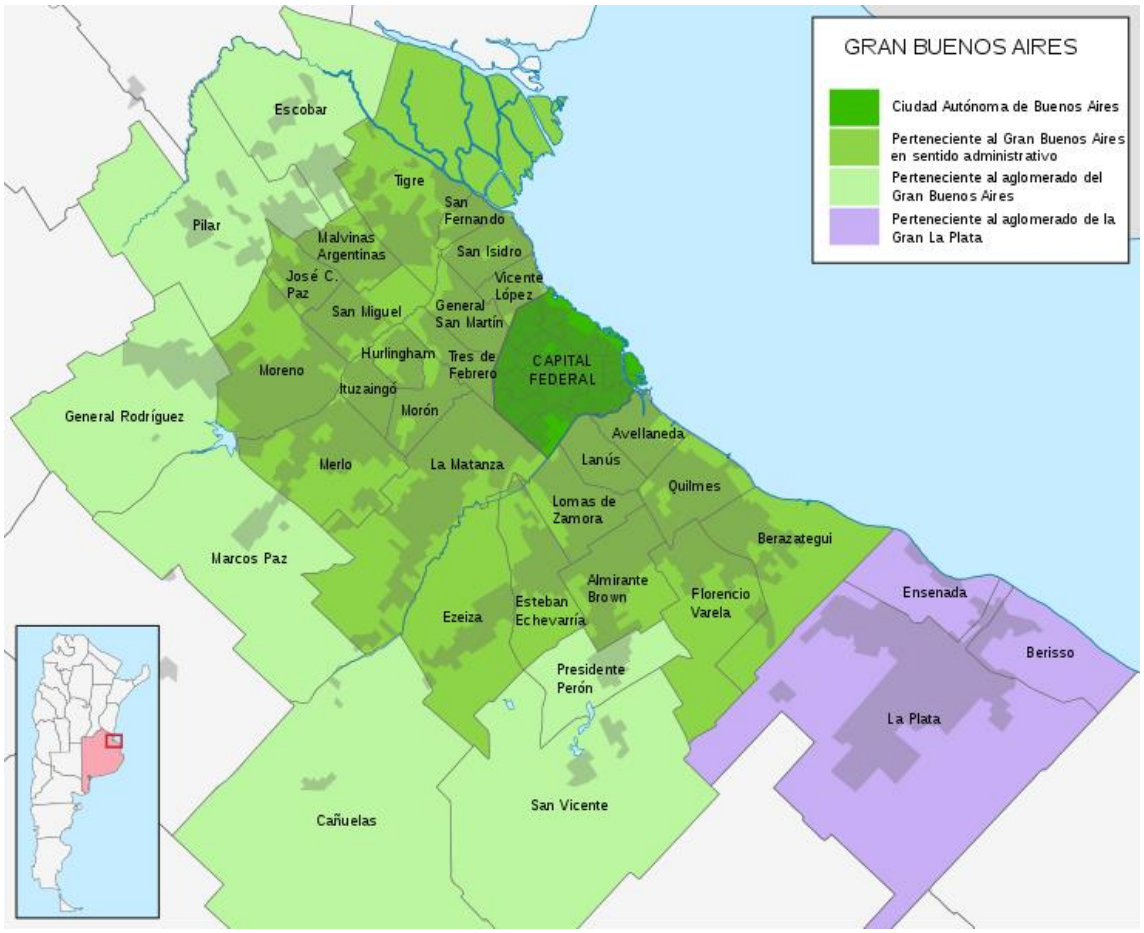
According to the Ministry of Interior of Argentina, the average amount of long-term immigrants to the country for the last ten years is 230k, and around 80% of them are moving to the Buenos Aires Metropolitan Area (AMBA - Buenos Aires City plus surrounding suburbs)[1][2].



	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Immigrants	206.243	172.368	253.406	286.155	274.081	197.572	259.505	215.240	212.845	221.892



[1] http://www.migraciones.gov.ar/pdf/estadisticas/radicaciones_resueltas_2018.pdf
[2] https://www.indec.gob.ar/dbindec/folleto_gba.pdf



Problem

Moving to another city usually is a major change in people's life and a decision that should not be taken lightly. Signing a rental contract is a commitment for a long period of time and making a mistake could be costly to fix.

Characteristics and venues of the chosen neighborhood could affect habits and daily activities. In addition, every mayor city has zones safer than others, which might be unknown prior to moving to the new city. Finally, rental cost is one of the biggest monthly expenses that cannot be easily modified and therefore it can have a big impact in personal monthly budget.

Project objective

The aim of this project is to elaborate and analyze an algorithm capable of selecting the neighborhoods in a destination city that are most similar to the surroundings of an origin address according to venues in the area and population density, pick the 10 safest and predicting the rent value of an apartment or house.

A particular case of a person moving from Barcelona, Spain, to Buenos Aires, Argentina will be used as an example, but the technics and methods could be extrapolated to any different case.

Data

Data sources

Several sources are used for this project, as different datasets are analyzed together for both cities.

Neighborhood names and centroid coordinates

Gather geolocation data:

- Borough
- Neighborhood
- Population density
- Coordinates

Barcelona

For this city I will extract boroguh, neighborhood and pop density from the townhall website through Selenium scraping, and then add coordinates with geocoder

url_bcn_pop = 'https://www.bcn.cat/estadistica/castella/dades/anuari/cap01/C0101050.htm'

Argentina

First, get province, borough, neighborhoods and its coordinates

url_arg_coor = 'https://infra.datos.gob.ar/catalog/modernizacion/dataset/7/distribution/7.5/download/localidades.json'

Population density Best granularity available is by borough, but it can be uased as an estimate for each neighborhood inside each borough

url_arg_pop = 'https://sig.indec.gob.ar/censo2010/?_ga=2.139240499.142017136.1591731561-616979075.1591488192'

Data Merging

Data Filtering

As the objective of this case study, we will make two subsets with relevant information:

Original address

Buenos Aires Metropolitan Area

https://es.wikipedia.org/wiki/Gran_Buenos_Aires

Venues in the area

Foursquare API

Population density

Crime rate

url_arg_crime = 'https://estadisticascriminales.minseg.gob.ar/datos/snic-departamentos.csv'

Prices and characteristics of current rentals.

Data cleansing

Barcelona

As seen above, the Barcelona dataset table acquired through Selenium has several row at the top and bottom that must be erased. The first column is repeated 3 times, and it comprises borough and neighborhood in the same column considering borough as a grouping row. These rows should be erased and the borough separated into a new column.

Buenos Aires

The las row contains NaN data and must be erased First column comprises borough and neighborhood as a comma separeted value and should be separated into 2 different columns.

Feature selection

The first step should be to characterize the origin neighborhood and the different ones in the destination city. This will be done by categorizing the most popular venues through Foursquare City Guide APIs and the population density of each city based on official census data (Argentina will be scraped and Spain is available in JSON format).

> https://sig.indec.gob.ar/censo2010/?_ga=2.139240499.142017136.1591731561-616979075.1591488192
>
> http://www.amb.cat/es/web/area-metropolitana/dades-obertes/cataleg/detall/-/dataset/densidad-de-poblacion/1060737/11692

Once the similarity between neighborhoods has been established, the top 5 safest ones will be chosen for the next step. This will be based on official data from the Security Department of Argentina (data available in a CSV file)

> https://estadisticascriminales.minseg.gob.ar/#

Finally, housing rent price will be predicted based on housing requirements and current prices of available places. This information will be scraped from the main Argentinean house rental online platform.

> https://www.zonaprop.com.ar/

Methodology

Results

Discussion

Conclusion