# Capstone Project - The Battle of Neighborhoods

## Suitable new store locations in Buenos Aires for a Sportwear Retailer

This notebook contains:

- A description of the problem and a discussion of the background - Week 1
- A description of the data and how it will be used to solve the problem - Week 1
- Methodology and Exploratory Data Analysis - Week 2
- Inferences and Discussion - Week 2

The separate report goes into more description of the methodology. This notebook gives only brief outlines of the methodology of each step, but includes all the code, dataframes and visualizations.

## 1 Introduction and Discussion of the Business Objective and Problem

### Locations for new Sportwear stores branch in Buenos Aires, Argentina

#### The Task at Hand

A foreign investor is willing to open a new store selling sportswear in Buenos Aires. It has stores in the main cities of the world and has no greater knowledge of the mentioned city. For this reason, contact our study that has extensive knowledge in the field of data science. The result of our advice will be to indicate which are the neighborhoods of Buenos Aires with the greatest potential to open this new branch. This will be an important part of your decision-making process, the other will be the qualitative

analysis of the neighborhoods once these data and reports are reviewed and studied. The neighborhoods to be considered will be those in which there are public and private spaces for physical activities and with fewer stores selling sportswear. Foursquare data will be very useful for making decisions based on data on the best of those areas. This database is considered to have updated data.

**Criteria**

Based on our qualitative data, we suggest that the best locations to open a new branch are places near recreational areas such as sports clubs, gyms, public squares in general sports training places. The analysis and recommendations will focus on the neighborhoods with the highest amount of the class of establishments mentioned. Limiting the number of neighborhoods will allow us to carry out further research, such as finding specific sites for the installation of the new branch.

**Why Data?**

Taking advantage of the ability of data science, investors can save countless hours of analysis for each neighborhood, consulting many real estate agents with their respective costs and perhaps making a wrong decision. Data will provide better answers and better solutions to their task at hand.

**Outcomes**

The objective is to identify the best neighborhoods to open a new branch as part of the company's plan. The results will be translated to the administration in a simple way that will transmit the data-based analysis to the best locations to open the store.

# 2 The Data Science Workflow

## Data Requirements

The main neighborhoods in Buenos Aires are divided into "Comunas" (administrative areas). The data regarding the neighborhoods needs to be researched and a suitable useable source identified. If it is found but is not in a useable form, data wrangling and cleaning will have to be performed. The cleansed data will then be used alongside Foursquare data, which is readily available. Foursquare location data will be leveraged to explore or compare neighborhoods, identifying the high traffic areas where consumers go for Soccer Stadium, Gym, Athletics & Sports, Fitness Center and Outdoors & Recreation are most interested in opening new store.

**The Data Science Workflow for Part 1 & 2 includes the following:**

Outline the initial data that is required: Neighborhoods data for Buenos Aires including names, location data if available, and any other details required.

Obtain the Data:

- Research and find suitable sources for the neighborhoods data for Buenos Aires.

- Access and explore the data to determine if it can be manipulated for our purposes.

Initial Data Wrangling and Cleaning:

- Clean the data and convert to a useable form as a dataframe.

**The Data Science Workflow for parts 3 & 4 includes:**

Data Analysis and Location Data:

- Foursquare location data will be leveraged to explore or compare neighborhoods around Buenos Aires.
- Data manipulation and analysis to derive subsets of the initial data.
- Identifying the high traffic areas using data visualization and statistical analysis.

Visualization:

- Analysis and plotting visualizations.
- Data visualization using various mapping libraries.

Discussion and Conclusions:

- Recommendations and results based on the data analysis.
- Discussion of any limitations and how the results can be used, and any conclusions that can be drawn.

# Data Research and Preparation

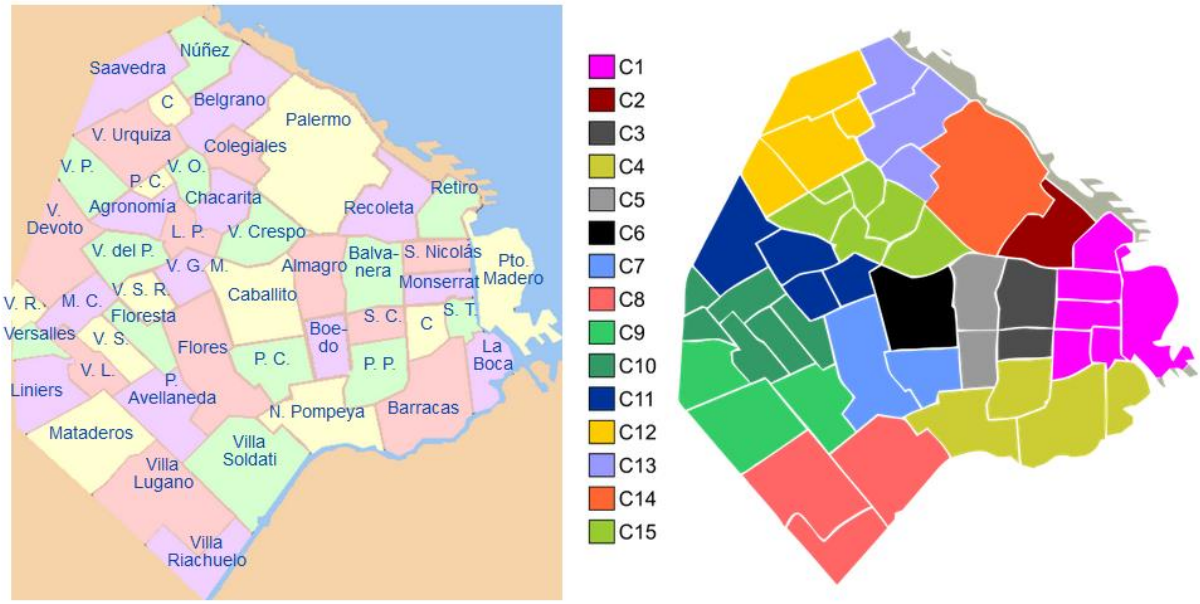
## Import the Buenos Aires districts data

**File: Listado_barrios.csv (administrative "communes")**

### Districts

Barrios of Buenos Aires and Communes of Buenos Aires

The city is divided into barrios (neighborhoods) for administrative purposes, a division originally based on Catholic parroquias (parishes). Buenos Aires only consists of 48 official barrios. There are several subdivisions of these districts, some with a long history and others that are the product of a real estate invention. A notable example is Palermo – the city's largest district – which has been subdivided into various barrios, including Palermo Soho, Palermo Hollywood, Las Cañitas and Palermo viejo, among others. A newer scheme has divided the city into 15 comunas (communes)




Initially looking to get this data by scraping the relevant Wikipedia page https://es.wikipedia.org/wiki/Barrios_de_la_ciudad_de_Buenos_Aires , fortunately, after much research, this data is available on the web and can be manipulated and cleansed to provide a meaningful dataset to use.

Data for the neighborhoods is necessary to select the most suitable of area for new store.

## Week 1:

### Discussion of the Business Objective and Problem / The Data Workflow

We now have located and imported the relevant data for the neighborhoods of Buenos Aires and have constructed a dataframe. Our business objective, strategy and methods to achieve our goal have been laid out, and a data workflow established. Next up, we will leverage Foursquare location data to identify the high traffic areas where consumers go for Soccer Stadium, Gym, Athletics & Sports, Fitness Center and Outdoors & Recreation are most interested in opening new store.

**The Battle of Neighborhoods continues in the next section.**

# Week 2 - Data Analysis

## 3. Methodology and Exploratory Data Analysis

The Data Science Workflow for parts 3 & 4 includes:

- Data Analysis and Location Data:
- Foursquare location data will be leveraged to explore or compare districts around Paris.
- Data manipulation and analysis to derive subsets of the initial data.
- Identifying the high traffic areas using data visualization and statistical analysis.

Visualization:

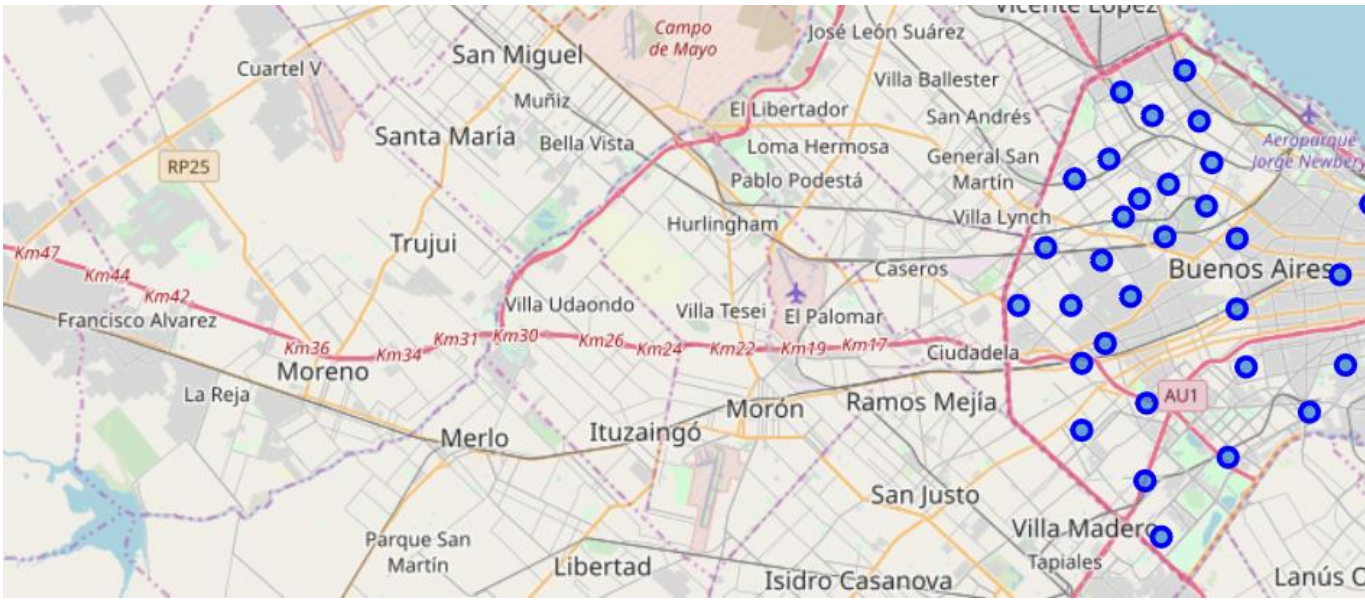
- Analysis and plotting visualizations.
- Data visualization using various mapping libraries.

Discussion and Conclusions:

- Recommendations and results based on the data analysis.
- Discussion of any limitations and how the results can be used, and any conclusions that can be drawn.

### Use the geopy library to get the latitude and longitude values of Buenos Aires

### Create a map of Buenos Aires, Argentina with districts superimposed




**Use the Foursquare API to explore the neighborhoods of Buenos Aires.**

# Exploratory data analysis

**Explore the first neighborhood in our dataframe to become familiar with the data.**

In [52]:
```
# Explore the first Neighborhood in our dataframe.
# Get the Neighborhood's name.

barrios.loc[0, 'Barrio']
```

Out[52]: 'Agronomia'

In [19]:
```
# Get the neighborhood's latitude and longitude valu
barrio_latitud = barrios.loc[0, 'Latitud'] # neighbo
barrio_longitud = barrios.loc[0, 'Longitud'] # neigh
barrio_name = barrios.loc[0, 'Barrio'] # neighborhoo
print('Latitude and longitude values of {} are {}, {
```

```
Latitude and longitude values of Agronomia are -3
```

**Get the top 100 venues that are in the neighborhood *Agronomia* within a radius of 500 meters**

In [53]:
```
# Now, let's get the top 100 venues that are in Agron
# First, let's create the GET request URL. Name your

LIMIT = 100 # limit of number of venues returned by F
radius = 500 # define radius
url = 'https://api.foursquare.com/v2/venues/explore?&
    CLIENT_ID,
    CLIENT_SECRET,
    VERSION,
    barrio_latitud,
    barrio_longitud,
    radius,
    LIMIT)
url # display URL
```

Out[56]:

| | name | categories | lat | lng |
|---|---|---|---|---|
| 0 | Feria del Productor al Consumidor | Farmers Market | -34.593981 | -58.483098 |
| 1 | Club Arquitectura | Athletics & Sports | -34.589630 | -58.484929 |
| 2 | Vivero Agronomía | Garden Center | -34.591700 | -58.488838 |
| 3 | Social Parrilla | BBQ Joint | -34.588955 | -58.484677 |
| 4 | Corredor Aeróbico de Agronomía | Trail | -34.592877 | -58.483940 |
| 5 | Parada Línea 80 | Bus Stop | -34.592044 | -58.489767 |
| 6 | Túnel Gustavo Cerati | Tunnel | -34.592892 | -58.490347 |

## Create a nearby venues function for all the neighborhoods in Buenos Aires

## Create a new dataframe called for the venues of Buenos Aires called *Buenos_aires-venues*

In [26]:
```
# Now, run the above function on each neighborhood and create a new dataframe called buenos_aires_venues.
buenos_aires_venues = getNearbyVenues(names=barrios['Barrio'],
                                      latitudes=barrios['Latitud'],
                                      longitudes=barrios['Longitud']
                                     )
```

```
Agronomia
Balvanera
Belgrano
Caballito
Chacarita
Coghlan
Colegiales
Constitucion
La Paternal
Mataderos
Monte Castro
Nueva Pompeya
Nunez
Parque Avellaneda
Parque Chacabuco
Parque Chas
Parque Patricios
Puerto Madero
Recoleta
Saavedra
```

Out[27]:

| | Barrio | Barrio Latitud | Barrio Longitud | |
|---|---|---|---|---|
| **0** | Agronomia | -34.591516 | -58.485385 | Feri |
| **1** | Agronomia | -34.591516 | -58.485385 | |
| **2** | Agronomia | -34.591516 | -58.485385 | |
| **3** | Agronomia | -34.591516 | -58.485385 | |
| **4** | Agronomia | -34.591516 | -58.485385 | Co |
| **5** | Agronomia | -34.591516 | -58.485385 | |
| **6** | Agronomia | -34.591516 | -58.485385 | |
| **7** | Balvanera | -34.609215 | -58.403140 | |
| **8** | Balvanera | -34.609215 | -58.403140 | |

**Check how many venues were returned for each neighborhood**

| Barrio | Barrio Latitud | Barrio Longitud | Venue | Venue Latitud | Venue Longitud |
|---|---|---|---|---|---|
| **Agronomia** | 7 | 7 | 7 | 7 | 7 |
| **Balvanera** | 14 | 14 | 14 | 14 | 14 |
| **Belgrano** | 42 | 42 | 42 | 42 | 42 |
| **Caballito** | 42 | 42 | 42 | 42 | 42 |
| **Chacarita** | 22 | 22 | 22 | 22 | 22 |
| **Coghlan** | 15 | 15 | 15 | 15 | 15 |
| **Colegiales** | 31 | 31 | 31 | 31 | 31 |
| **Constitucion** | 8 | 8 | 8 | 8 | 8 |
| **La Paternal** | 4 | 4 | 4 | 4 | 4 |
| **Mataderos** | 3 | 3 | 3 | 3 | 3 |

**Analyze each of the Neighborhoods**

| | Barrio | American Restaurant | Antique Shop | Arcade | Argentinian Restaurant | Art Gallery | Art Museum | Arts & Crafts Store | Ente |
|---|---|---|---|---|---|---|---|---|---|
| **0** | Agronomia | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| **1** | Agronomia | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| **2** | Agronomia | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| **3** | Agronomia | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| **4** | Agronomia | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| **5** | Agronomia | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| **6** | Agronomia | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| **7** | Balvanera | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| **8** | Balvanera | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| **9** | Balvanera | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |

**Group rows by neighborhood and take the mean of the frequency of occurrence of each category**

| | Barrio | American Restaurant | Antique Shop | Arcade | Argentinian Restaurant | Art Gallery | Art Museum | Arts Cra Sto |
|---|---|---|---|---|---|---|---|---|
| 0 | Agronomia | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.0000 |
| 1 | Balvanera | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.0714 |
| 2 | Belgrano | 0.000000 | 0.000000 | 0.000000 | 0.047619 | 0.000000 | 0.023810 | 0.0000 |
| 3 | Caballito | 0.000000 | 0.000000 | 0.000000 | 0.071429 | 0.000000 | 0.000000 | 0.0000 |
| 4 | Chacarita | 0.000000 | 0.000000 | 0.000000 | 0.090909 | 0.000000 | 0.000000 | 0.0000 |
| 5 | Coghlan | 0.000000 | 0.000000 | 0.000000 | 0.066667 | 0.000000 | 0.000000 | 0.0000 |
| 6 | Colegiales | 0.000000 | 0.000000 | 0.000000 | 0.129032 | 0.000000 | 0.000000 | 0.0000 |
| 7 | Constitucion | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.0000 |
| 8 | La Paternal | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.0000 |
| 9 | Mataderos | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.0000 |
| 10 | Monte Castro | 0.000000 | 0.000000 | 0.000000 | 0.166667 | 0.000000 | 0.000000 | 0.0000 |
| 11 | Nueva Pompeya | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.0000 |
| 12 | Nunez | 0.023810 | 0.000000 | 0.000000 | 0.071429 | 0.000000 | 0.000000 | 0.0000 |
| 13 | Parque Avellaneda | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.0000 |
| 14 | Parque Chacabuco | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.0000 |

**Print each neighborhood with it's top 10 most common venues**

```
----Agronomia----
                venue  freq
0      Farmers Market  0.14
1  Athletics & Sports  0.14
2            Bus Stop  0.14
3               Trail  0.14
4              Tunnel  0.14


----Balvanera----
                  venue  freq
0  Fast Food Restaurant  0.21
1                 Café  0.14
2            Restaurant  0.07
3     Electronics Store  0.07
4           Pizza Place  0.07


----Belgrano----
                          venue  freq
0                   Coffee Shop  0.07
1  Vegetarian / Vegan Restaurant  0.05
2                   Pizza Place  0.05
3                      Tea Room  0.05
4                          Café  0.05
```

# The top 10 venue categories for each neighborhood

This is a very useful results table that can provide at a glance information for all of the districts. Even once any conclusions are drawn further into the data workflow, we can refer back to this table for meaningful insights about the top categories of businesses in all the neighborhoods. Even without actual counts and numbers, it makes a great reference table for the client.

| | Barrio | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Cor V |
|---|---|---|---|---|---|
| 0 | Agronomia | Athletics & Sports | Trail | Garden Center | Farmers M |
| 1 | Balvanera | Fast Food Restaurant | Café | Hotel | BBQ |
| 2 | Belgrano | Coffee Shop | Bookstore | Argentinian Restaurant | Ice Cream |
| 3 | Caballito | Café | Bakery | Pizza Place | |
| 4 | Chacarita | Pizza Place | Argentinian Restaurant | Bus Stop | B |
| 5 | Coghlan | Café | Bakery | Pizza Place | Resta |
| 6 | Colegiales | Pizza Place | Argentinian Restaurant | Café | Br |

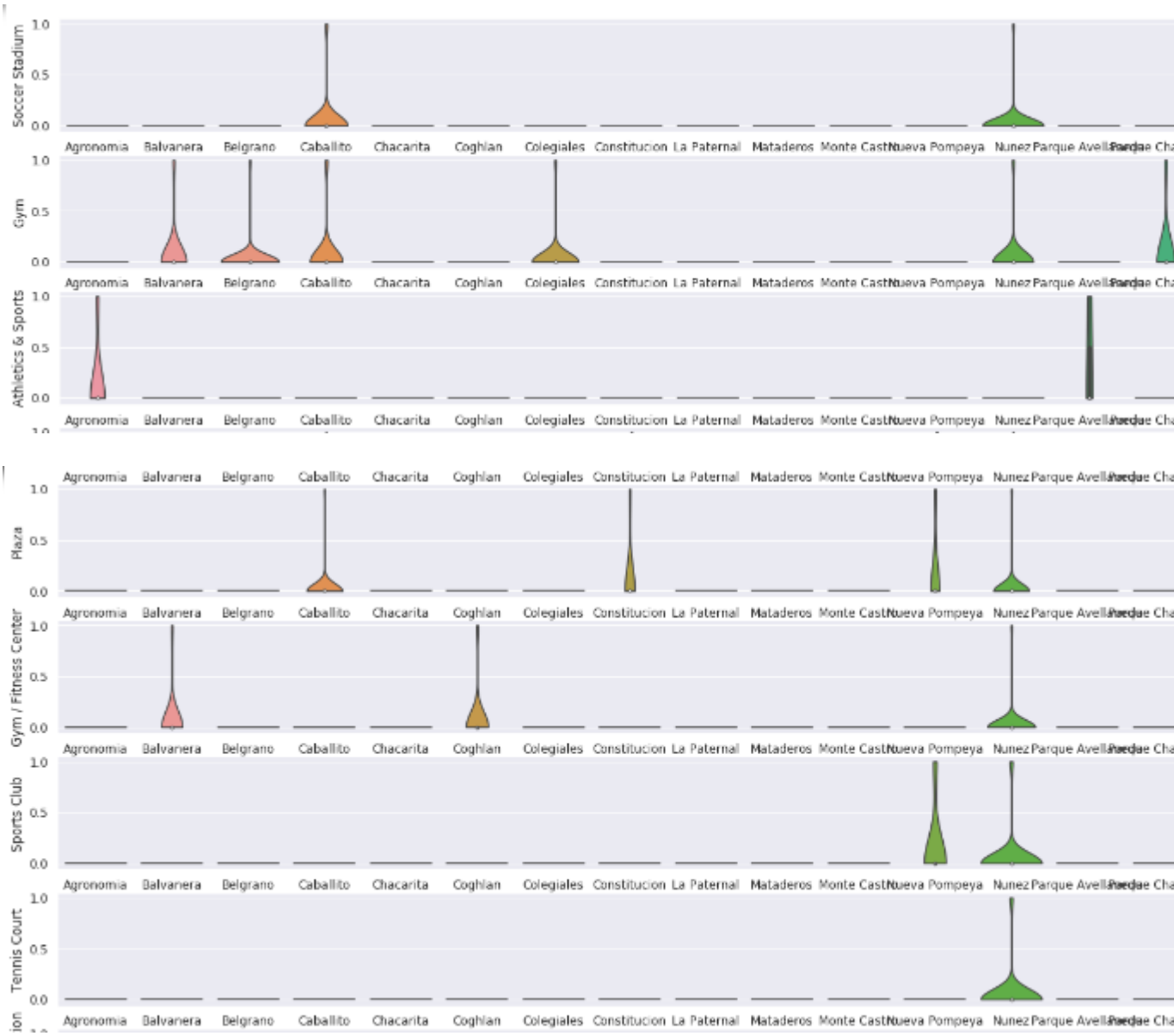
**The criteria of types of business agreed with the client:**

- Soccer Stadium
- Athletics & Sports
- Fitness Center
- Outdoors & Recreation
- Gym
- Plaza
- Sports Clubs
- Tennis Court
- Yoga Studio
- Soccer Field

Let's look at their frequency of occurance for all the neighborhoods, isolating the categorical venues

These are the venue types that the client wants to have an abundant density of in the ideal store locations. I've used a violin plot from the seaborn library - it is a great way to visualise frequency distribution datasets, they display a density estimation of the underlying distribution.

## Frequency distribution for the top 10 venue categories for each neighborhood


Soccer Stadium
Gym
Athletics & Sports
Plaza
Gym / Fitness Center
Sports Club
Tennis Court
1.0
0.5
0.0
Agronomia
Balvanera
Belgrano
Caballito
Chacarita
Coghlan
Colegiales
Constitucion
La Paternal
Mataderos
Monte Castro
Nueva Pompeya
Nunez
Parque Avellaneda


## Neighborhoods

So, as we can see in the analysis, there are 6 candidate neighborhoods, Nunez stands out, according to the criteria agreed with our client with great frequency.

They are the following:

Neighborhoods

Nunez, is the main candidate

- Villa Real
- Villa Devoto
- Caballito
- Villa Lugano
- Villa Crespo

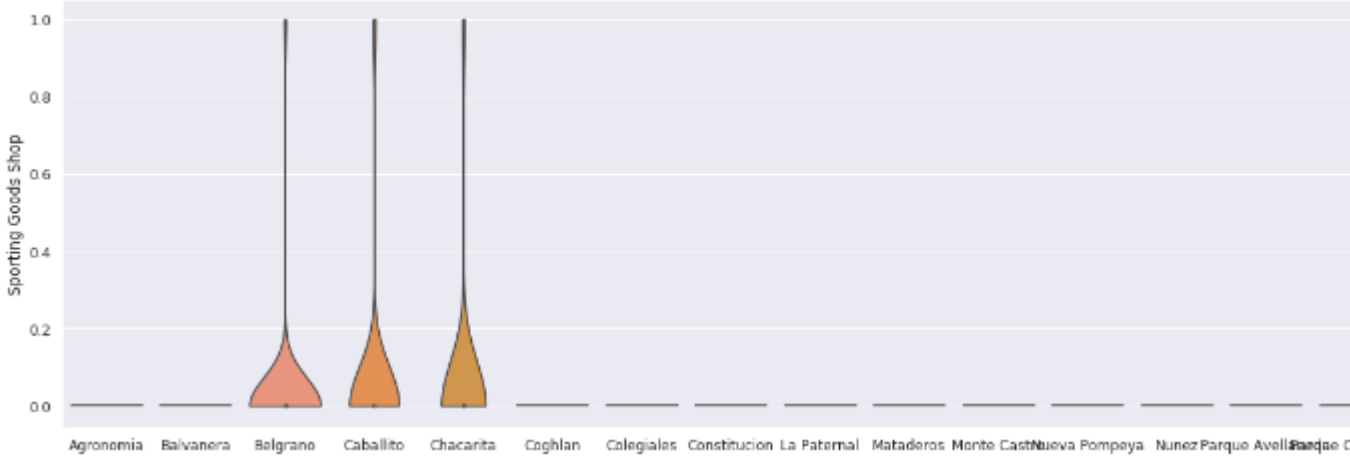
Let's take this further with some exploration and inferential analysis

We have the 6 neighborhoods that include all the criteria of place category. But if we include the following categories of place:

- Clothing Store
- Sporting Goods Shop

We could make some inferences based on data and knowledge of the domain of marketing and industry, to focus the list. That is, we should analyze whether there are other stores selling sportswear.

## Frequency of Clothing and Sporting Goods Shops for each neighborhood





## 4. Inferences and discussion

### Chosen neighborhoods - Results

The inferential analysis that uses the data, as well as knowledge of the domain of retail and marketing, allow the list to focus on only 3 neighborhoods of the 6 selected.

The reasoning is that they have fulfilled most of the requirements presented by our client:

- Soccer stadium
- Athletics & Sports
- Fitness center
- Outdoors & Recreation
- Gym
- Square
- Sports clubs

- Tennis court
- Yoga Studio
- Soccer field

Then, the 3 final candidate neighborhoods to open the new store where most of the requirements are met They are:

- Nunez
- Villa Devoto
- Villa Real

Nunez is the one that meets most of the needs posed by our client.

# Where are our chosen neighborhoods? Let's visualize them on a map of Buenos Aires.

Out[59]:

| | Comuna | Barrio | Latitud | Longitud |
|---|---|---|---|---|
| **0** | Comuna 13 | Nunez | -34.545348 | -58.462149 |
| **1** | Comuna 11 | Villa Devoto | -34.600994 | -58.515516 |
| **2** | Comuna 10 | Villa Real | -34.618943 | -58.525877 |

## Observations

The three selected neighborhoods are residential with ample spaces for sports and outdoor activities. From this visualization, it is clear that, on a practical level, without data on which to base decisions, the number of neighborhoods to be analyzed is very large, and investigating and then visiting them all would be a daunting and slow task.

We have significantly reduced the search area to only 3 that should adapt to our client's business.

### Inferences

We have made inferences from the data when making location recommendations, but that is exactly the point. There is no right or wrong answer or conclusion for the task at hand. The job of data analysis here is to run a course for the selection of new store locations

- to meet the criteria initially established by our client places where sports practices abound.
- Reduce the search to only a few of the main areas that best fit the criteria.

### Conclusions

There are many ways in which this analysis could have been done based on different methodologies and perhaps different data sources. The method used is a direct way to reduce the options, complying with the initial directives of our client. The analysis and the results is not conclusive, it is a starting point that will guide the next part of the process to find the location of specific stores. The next part will involve knowledge of the domain of the industry, and perhaps, of the city itself. But data analysis and the resulting recommendations have greatly reduced the best data-based options and what we can infer from them.

Without taking advantage of the data to make specific decisions, the process could have been extended and resulted in the opening of a new store in An incorrect area. The data has helped provide a better strategy and a way forward, these data-based decisions will lead to a better solution in the end.