## Driving a Taxi in Buenos Aires

### Background

Some time ago, during the Smart City Expo in Buenos Aires a “Dataton” was organized under the title "A Day in cities through your data". The event was organized by the public national open data teams, together with the City and Province of Buenos Aires.

Several datasets with urban information were made public on the event. Among them, a sample of taxi trips requested through the BA TAXI app. This is a UBER like app used by the drivers of official taxi´s in Buenos Aires.

### The Problem

My friend Juan who lives in the suburbs lost his job and is thinking to drive a taxi to earn some income until he can get another job. Fortunately for him the taxi owner allows him to choose his job schedule. He would also like to move to a neighborhood with high taxi demand to minimize driving but good gastronomy options as he is lazy to cook his own meals.  
So in this assignment I will be answering two questions:

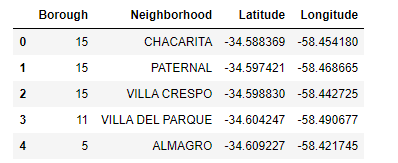
* ***What are the most profitable days and shifts (morning, evenings or nights) to drive a taxi?***
* ***What is the neighborhood with the best balance of taxi traffic and gastronomy options?***

### The Data

* I will be using for this the Buenos Aires Government open Portal that provides geo referential data of the city of Buenos Aires: [data.buenosaires.gob.ar](https://data.buenosaires.gob.ar/dataset)
* Foursquare to explore the neighborhoods: [foursquare API](https://foursquare.com/)
* Data from the BA Taxi app service, a government sponsored initiative to securely travel in Taxis: [BA Taxi](https://www.buenosaires.gob.ar/aplicacionesmoviles/ba-taxi-pasajeros)
* A dataset of +19K taxi trips in Buenos Aires requested through the BA Taxi app: [BA Taxi Trips](https://data.world/vazquez-brust/viajes-solicitados-por-ba-taxi)

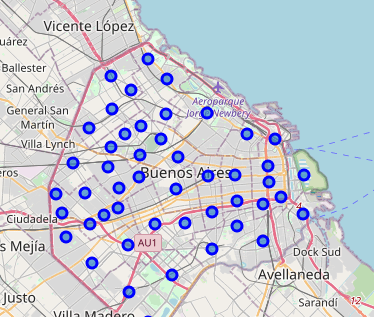
### The Buenos Aries Dataset

This is provided as a json file with the neighborhoods provided as a polygon of coordinates. After reading the file and transforming it into a dataframe and using the centroid of the polygons as the geo localization of each neighborhoods, we can produce the following tabular representation of the neighborhoods and their latitude and longitude:



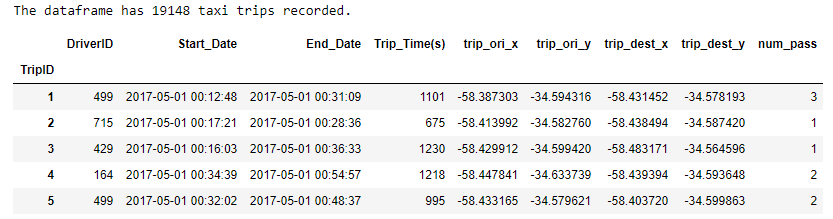
From this data we can find that there are 15 boroughs and 48 neighborhoods in the city of Buenos Aires.

Below we can see a map of the city of Buenos Aires with blue markers for each of its 48 neighborhoods:

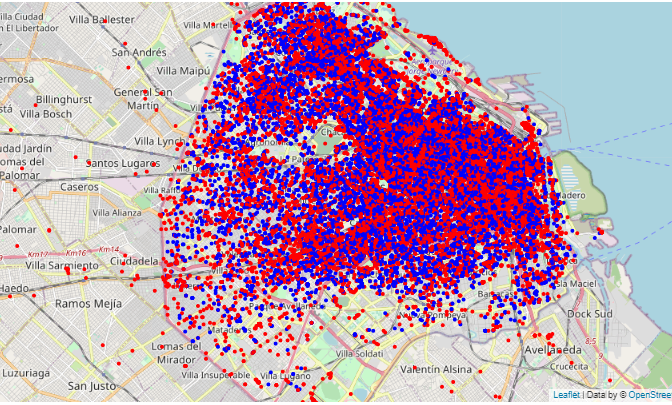


### The BA Taxi Dataset

The below table shows a tabular representation of the trips with the longitude and latitude of trip origin and destination



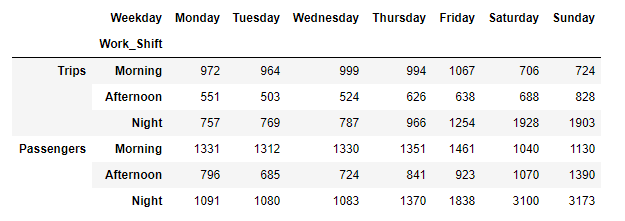
We can add markers to represent the dataset. The blue markers represent the places where a trip started while the red represents the place where the trip ended.



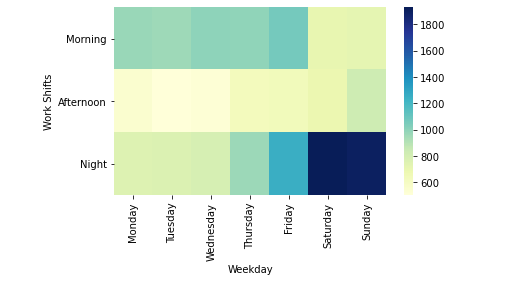
### Part 2: Exploring the Data

We will start by exploring the BA Taxi dataset. We will add the time of the day for each trip (Morning, Afternoon, Night) to segment the trips and check the frequency of trips by day period. We will also identify the day of the week for each trip to check the distribution by weekday.

Below we can see a tabular representation of the number of trips and total number of passengers carried for each work shift and weekday:



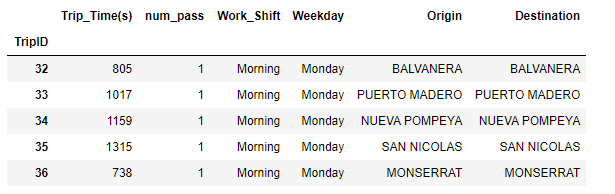
We can also see how displaying this same information as a heatmap produces an improved visualization that help us better understand the data and get conclusions. As the number of trips and number of passengers is highly correlated, we will use the number of trips for the representation:



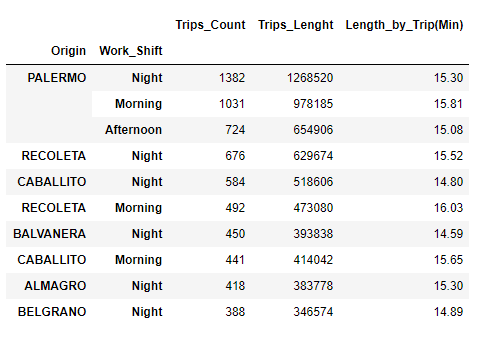
We can clearly see how the greatest demand for taxi trips happen during weekends on the night shift followed by the mornings during weekdays.

### Part 2b: Looking at taxi trips by Neighborhood

Using the Buenos Aires neighborhood dataset, we can assign the origin and destination of each trip to its corresponding neighborhood to produce the following table:



We are interested in understanding where do we have the most passenger pick-ups. To get this, we can group the dataframe by origin and sort it in descending order.

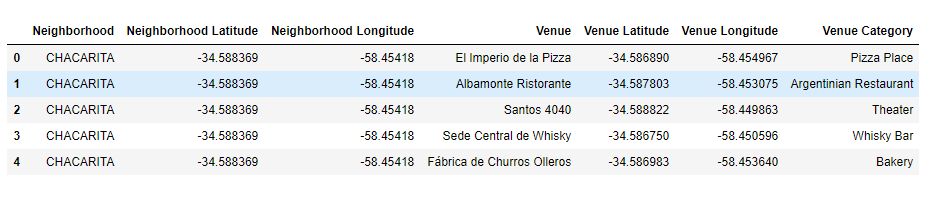


We can see how PALERMO stand out as the neighborhood with the most traffic for all day periods. RECOLETA and CABALLITO are the two other neighborhoods that appear for multiple day periods so could be options to consider. We can also note that the night shift is also the shift with most occurrence while the trip length in consistent around 15 minutes across shifts and neighborhoods

### Part 3: Neighborhood Exploration

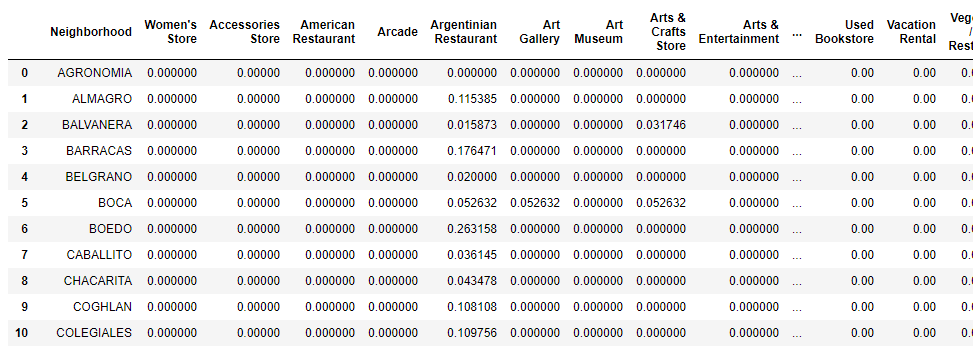
We will now explore the neighborhoods for our friend Juan using Foursquare to find neighborhoods with good variety of gastronomy to live in.

The first step is to get a list of venues for each neighborhood:

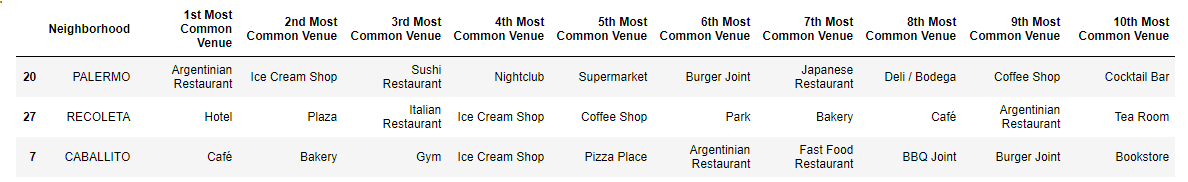


The search produced 2022 venues and 230 unique categories.

Using one hot encoding we can represent the venues as categorical variables in columns and find the frequency of each venue for each neighborhood as displayed in the following table (showing first 10 lines only):



From here we can query the 3 neighborhoods we are interested in and review the top 10 venues in terms of frequency of appearance.



The three seems to be excellent choices! They have plenty of restaurants and gastronomy options. That probably explains the high taxi traffic in the night shift.

### Part 4: Result Considerations

The Buenos Aires taxi trips combined with the Buenos Aires geo-localization dataset could be used in many ways besides the one explored in this project. Some other interesting options beyond the scope of this work are:

* Use of K-Means to identify 3 segments of trip length (short-medium-long) and identify the neighborhood with the largest number of short trips (as they are the most profitable).
* As we have the Driver ID we can identify waiting time per driver and map it to the neighborhood where the trip ends as an indirect way to identify demand surplus (neighborhoods were time between trips from a same driver are low would indicate more demand than taxis in that area)

In terms of the limitation of this analysis, while +19000 records seem a significant amount of records, this represents only 4 months of data so we should validate these results with a larger dataset and verify there is no seasonality in the results (current records span only across a winter season).

### Part 5: Conclusions

We analyzed in detail the taxi traffic in Buenos Aires, we sliced the information by the day of the week and by the different time periods in a day (working shifts).

We found that the night shift and the weekends are the busiest slots, hence the most profitable for driving a Taxi. We also explore the neighborhoods with the most taxi trips for the considered timeframe and explore the type of venues in it. We found that the 3 top neighborhoods in terms of taxi trips have plenty of gastronomy options and nightclubs, which correlates well with the idea of weekend night taxi traffic.

**For all the above, our final recommendation for Juan is to move to Palermo where he will enjoy good gastronomy options and select the night shifts and weekend days to drive the taxi to maximize his income.**