Location recommendations for establishing a new Pub venue in Montevideo

A. Introduction

A.1. Introduction / Business Problem

Montevideo is the capital of Uruguay, where about one and a half million people live and a population high density of 6.253 people per square kilometer. In this project, I chose Montevideo because I grew up in this city and I'm living there at present. The city is divided into **62 neighborhoods** in total.

In the last decade, **Uruguay** has experienced an **important economic growth** which lead to a significant number of **new venues** from all categories specially near the **city's most populated neighborhoods**.

Therefore, **stakeholders** require some advice on **where to settle and launch their new businesses**. Experience from local residents is always useful, but it is interesting to also **rely on data** to take a final **decision**.

This project will be targeted to new **investors** trying to find an optimal location for a new **Pub** in **Montevideo**, Uruguay.

When we think of it by the investor, we expect from them to prefer the neighborhoods where there is a lower real estate cost and the type of business they want to install is less intense.

We are also particularly interested in **areas with plenty of venues in vicinity** as an indicator of **economic activity**, and **neighborhoods with not too high housing prices**. We would also prefer locations **as close to city center as possible**, assuming that first two conditions are met. When we consider all these problems, we can create a map and information chart where the real estate index is placed on Montevideo and each district is clustered according to the venue density.

We will use our data science powers to generate a few most promising neighborhoods based on this criteria. Advantages of each area will then be clearly expressed so that best possible final location can be chosen by investors.

A.2. Data Description

Based on the definition of the problem, factors that will influence the final decision are:

1. **number of existing Venues** from every category in each neighborhood, within a certain radius from the center of the neighborhood
2. **number of existing Pubs** within the same radius, as a subcategory of "Bars"
3. **house pricing in each potential neighborhood**
4. **distance of the neighborhood from city center** is a plus, important but not decisive

The following data sources will be needed to extract/generate the required information:

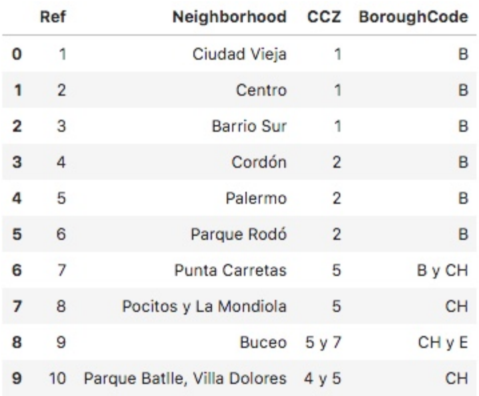
1. **"Anexo: Barrios de Montevideo" Wikipedia page**, to extract the name and other information of the 62 neighborhoods of the capital of Uruguay, via website scraping with Beautiful Soup. https://es.wikipedia.org/wiki/Anexo:Barrios_de_Montevideo
2. **Average house sale prices from each neighborhood**. I found information of the year 2017 and before from **Instituto Nacional de Estadistica** http://www.ine.gub.uy/ and set-up my own data table. For those neighborhoods in which no data was reported, the average of neighborhoods with the **same borough code** was assigned as the corresponding value for the missing data. When there was no data for 2017, I took the first previous value and applied linear extrapolation, by considering the full set of values from a neighborhood where all the information was complete. http://www.ine.gub.uy/c/document_library/get_file?uuid=dc2d978d-7027-4370-b2de-2f70c4d4ede8&groupId=10181
3. **Coordinates of the center of each neighborhood** are not included in the Wikipedia page, so I used **Google Maps geocoding API** https://cloud.google.com/maps-platform/maps/?apis=maps to get the list of (latitude, longitude) of the 62 rows. To center the Folium maps, I used Google Maps API geocoding to get the coordinates of a well known Montevideo location: bus terminal called **"Terminal Tres Cruces"**, as "Montevideo center".
4. **Number and categories of the Venues** within a certain radius from each neighborhood center, by using the **Foursquare API**. I filtered and dropped every neighborhood with less than 50 venues to keep only those neighborhoods with enough commercial activity as to consider establishing a business there. This information is grouped to find the most common venues and the **cluster the potential neighborhoods**. Finally, for those potential neighborhoods, I explored the **number of Bars** and then filtered specifically those which are **Pubs**.
5. **A .json file with the spacial coordinates of every Montevideo neighborhood** with the purpose of building a Choropleth Map with "average housing sale prices" (*HSP*) as the parameter. I found that file from a Github repository https://github.com/vierja/geojson_montevideo/blob/master/barrios.geojson and it said the original source was https://catalogodatos.gub.uy, but the file was deleted from the main source.

As a database, I used GitHub repository in my study.

B. Methodology

B.1. Website scraping with Beautiful Soup to extract the names and other available information of Montevideo's 62 neighborhoods

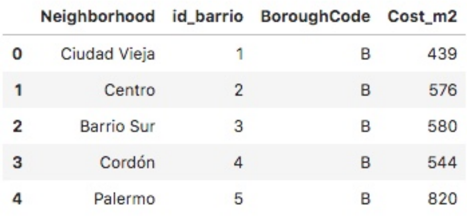
From *"Anexo: Barrios de Montevideo"* Wikipedia page, we arrive to the following dataframe with two specially important columns: 'Neighborhood' name and 'BoroughCode' of each one of the 62 neighborhoods. These are the first 10 rows:



	Ref	Neighborhood	CCZ	BoroughCode
0	1	Ciudad Vieja	1	B
1	2	Centro	1	B
2	3	Barrio Sur	1	B
3	4	Cordón	2	B
4	5	Palermo	2	B
5	6	Parque Rodó	2	B
6	7	Punta Carretas	5	B y CH
7	8	Pocitos y La Mondiola	5	CH
8	9	Buceo	5 y 7	CH y E
9	10	Parque Batlle, Villa Dolores	4 y 5	CH

B.2. Get the average housing sales price (USD/square meter) for each neighborhood, and perform *DATA CLEANSING:*

In the original data source, the .csv file showed some missing information we have to deal with. To handle these missing data, one possibility is to assign each neighborhood with missing value of [Cost_m2] the value of the mean of the housing sale prices from all the other neighborhoods sharing the same *Borough Code*. Finally, the cleaned dataframe is merged with the above dataframe and these are the resulting first 5 rows:



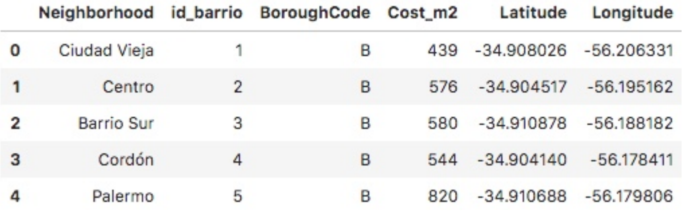
	Neighborhood	id_barrio	BoroughCode	Cost_m2
0	Ciudad Vieja	1	B	439
1	Centro	2	B	576
2	Barrio Sur	3	B	580
3	Cordón	4	B	544
4	Palermo	5	B	820

Notice the column *id_barrio*, which will be important in the later processing of the .json file to create a Choropleth map.

B.3. Get the coordinates of each Neighborhood by using *Google Maps Geocoding API*

I used *Python Folium library* to visualize maps. To center the Folium maps, I used *Google Maps API geocoding* to get the coordinates of a well known Montevideo location: bus terminal called **"Terminal Tres Cruces"**, as "Montevideo center".

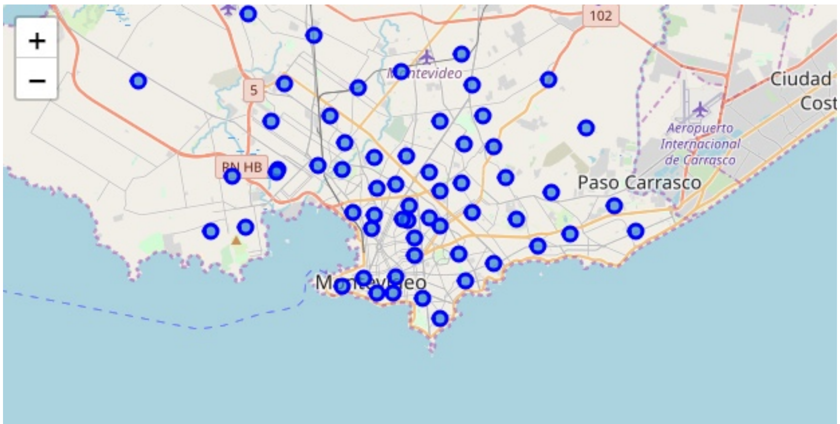
Then, we want to obtain the latitude and longitude from the corresponding address, for each neighborhood in the data table , following the same procedure as above. Finally, let's concatenate the dataframe with the resulting coordinates with the main one. These are the first 5 rows:



	Neighborhood	id_barrio	BoroughCode	Cost_m2	Latitude	Longitude
0	Ciudad Vieja	1	B	439	-34.908026	-56.206331
1	Centro	2	B	576	-34.904517	-56.195162
2	Barrio Sur	3	B	580	-34.910878	-56.188182
3	Cordón	4	B	544	-34.904140	-56.178411
4	Palermo	5	B	820	-34.910688	-56.179806

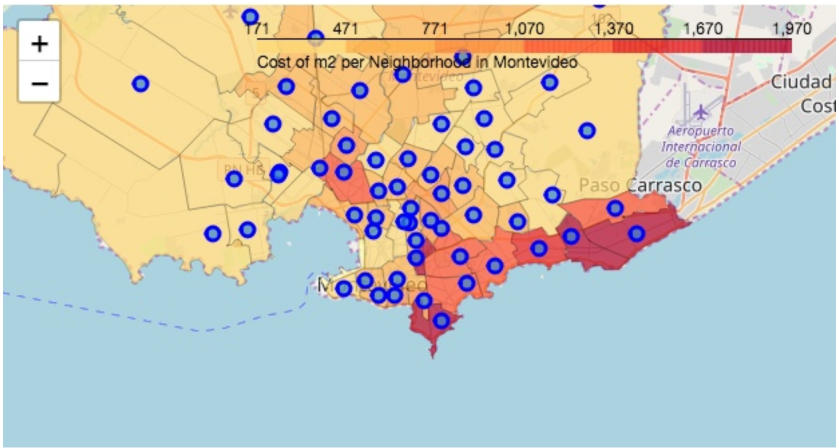
B.4. Create a map of Montevideo and show marks representing the location of each of the 62 Neighborhoods

Folium map of Montevideo with neighborhoods superimposed on top. I used latitude and longitude values of the 62 neighborhoods to get the visual as below:

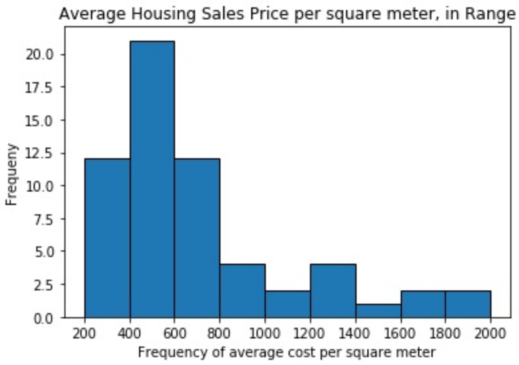


B.5. Analyze the average housing sale prices *(HSP)* per square meter in Montevideo

To analyze the HSP in Montevideo, a useful visualization method is to build a **Choropleth map of Montevideo**, with average housing sales price per square meter as the parameter. With the purpose of building the choropleth, we use a .json file with the spacial coordinates of every Montevideo neighborhood:



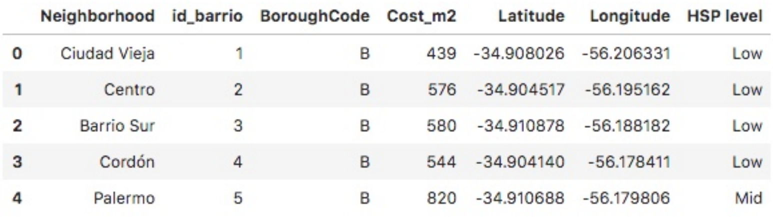
Now let's create **Histogram** showing the frequency of Average Housing Sales Price per square meter, in different ranges:



As it seems in above histogram, we can define the ranges as below:

- 200-800 AHP : “Low Level HSP”
- 800–1600 AHP : “Mid Level HSP”
- 1600–2000 AHP : “High Level HSP”

Therefore, we use a *Python* self-written *function* add a new column called '*HSP level*', based on the previous classification:

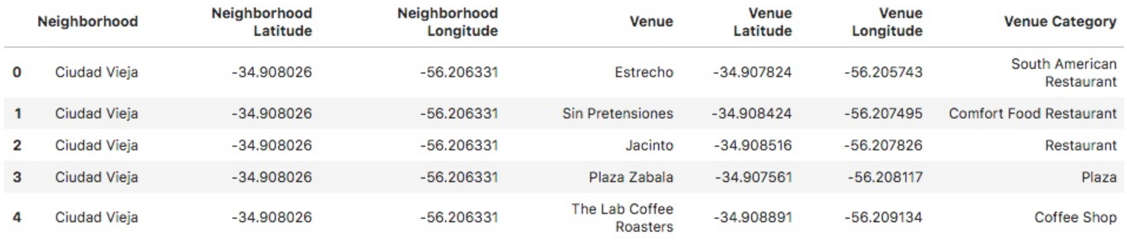


	Neighborhood	id_barrio	BoroughCode	Cost_m2	Latitude	Longitude	HSP level
0	Ciudad Vieja	1	B	439	-34.908026	-56.206331	Low
1	Centro	2	B	576	-34.904517	-56.195162	Low
2	Barrio Sur	3	B	580	-34.910878	-56.188182	Low
3	Cordón	4	B	544	-34.904140	-56.178411	Low
4	Palermo	5	B	820	-34.910688	-56.179806	Mid

B.6. Use Foursquare API to explore each neighborhood's *Venues*

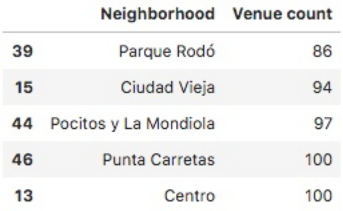
Next, we are going to use the Foursquare API to explore the 62 neighborhoods to make a pre-selection of the potential neighborhood candidates. A **limit of 100 venues** and **radius of 750 meters** from each neighborhoods center coordinates sounded reasonable.

Here is a head of the list Venues name, category, coordinates, merged with its corresponding neighborhood information:



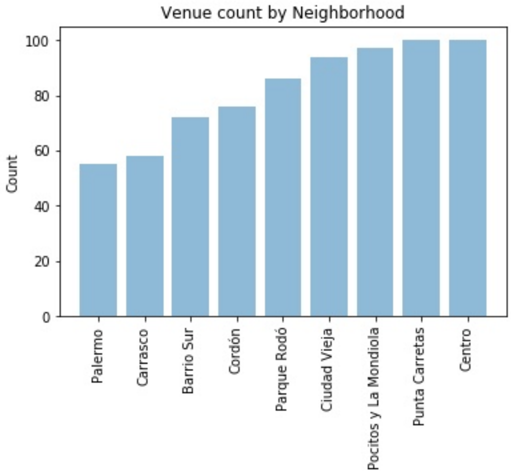
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Ciudad Vieja	-34.908026	-56.206331	Estrecho	-34.907824	-56.205743	South American Restaurant
1	Ciudad Vieja	-34.908026	-56.206331	Sin Pretensiones	-34.908424	-56.207495	Comfort Food Restaurant
2	Ciudad Vieja	-34.908026	-56.206331	Jacinto	-34.908516	-56.207826	Restaurant
3	Ciudad Vieja	-34.908026	-56.206331	Plaza Zabala	-34.907561	-56.208117	Plaza
4	Ciudad Vieja	-34.908026	-56.206331	The Lab Coffee Roasters	-34.908891	-56.209134	Coffee Shop

Then, we may calculate how many venues were returned for each neighborhood. The following are the 5 neighborhoods with more venues:



	Neighborhood	Venue count
39	Parque Rodó	86
15	Ciudad Vieja	94
44	Pocitos y La Mondiola	97
46	Punta Carretas	100
13	Centro	100

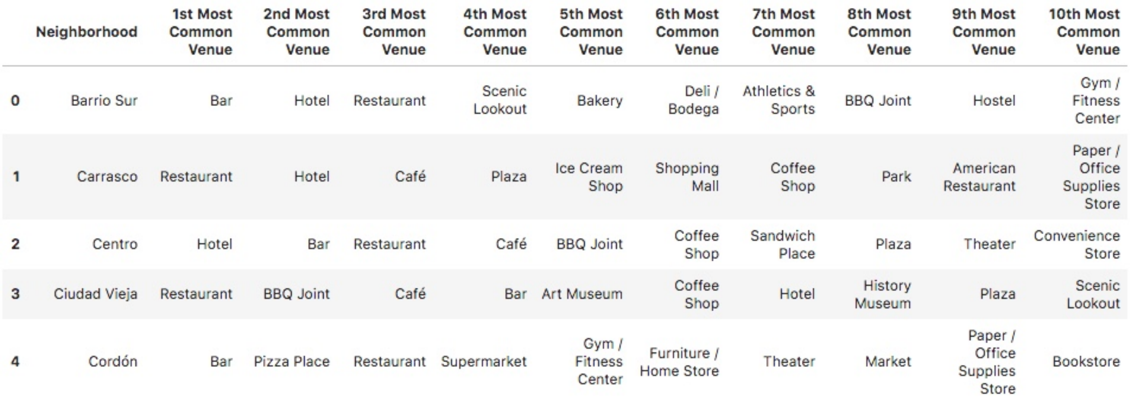
We can see that "Centro" and "Punta Carretas" reached the 100 limit of venues. On the other hand, 53 of the 62 neighborhoods were below 50 venues. **Only those neighborhoods with more than 50 venues will be pre-selected**, as an indicator of high economic activity. Therefore, these 9 will be the final neighborhood candidates in our pre-selection. Let's plot a with the results:



B.7. Analyze each candidate Neighborhood

From now, we will concentrate on these 9 neighborhood candidates. In summary, 146 unique categories were returned by Foursquare.

Next, let's group rows by neighborhood and by taking the mean of the frequency of occurrence of each category, and after that create a new dataframe and display the top 10 venues for each neighborhood. Here is a head of the resulting table:

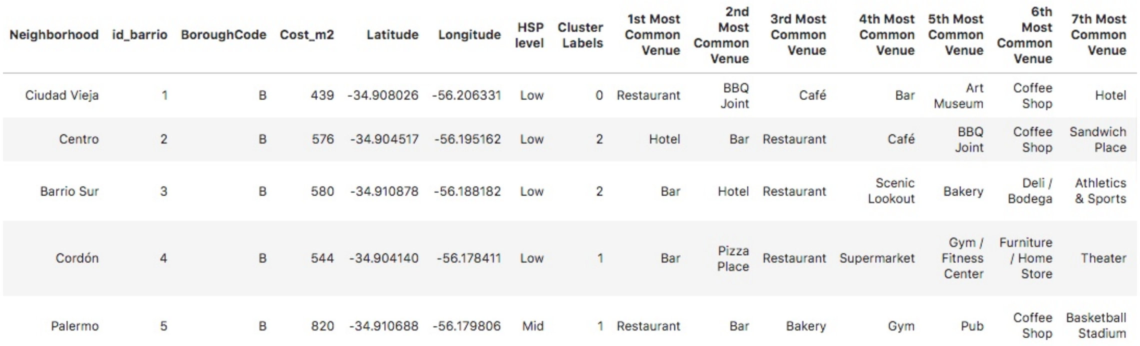


	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Barrio Sur	Bar	Hotel	Restaurant	Scenic Lookout	Bakery	Deli / Bodega	Athletics & Sports	BBQ Joint	Hostel	Gym / Fitness Center
1	Carrasco	Restaurant	Hotel	Café	Plaza	Ice Cream Shop	Shopping Mall	Coffee Shop	Park	American Restaurant	Paper / Office Supplies Store
2	Centro	Hotel	Bar	Restaurant	Café	BBQ Joint	Coffee Shop	Sandwich Place	Plaza	Theater	Convenience Store
3	Ciudad Vieja	Restaurant	BBQ Joint	Café	Bar	Art Museum	Coffee Shop	Hotel	History Museum	Plaza	Scenic Lookout
4	Cordón	Bar	Pizza Place	Restaurant	Supermarket	Gym / Fitness Center	Furniture / Home Store	Theater	Market	Paper / Office Supplies Store	Bookstore

B.8. Cluster the candidate Neighborhoods

We have some common venue categories in the neighborhoods. For this reason we used unsupervised learning K-means algorithm to cluster the neighborhoods. Therefore, we run K-means to cluster the neighborhood into 4 clusters.

Then, we create a new dataframe that includes the cluster labels as well as the top 10 venues for each neighborhood, and add to the dataframe a column with the cluster labels, and merge the main dataframe containing Montevideo data for the pre-selected neighborhoods, with the dataframe with the cluster labels and the top 10 venues for each neighborhood. This is the head of the result (last 3 columns are not shown):



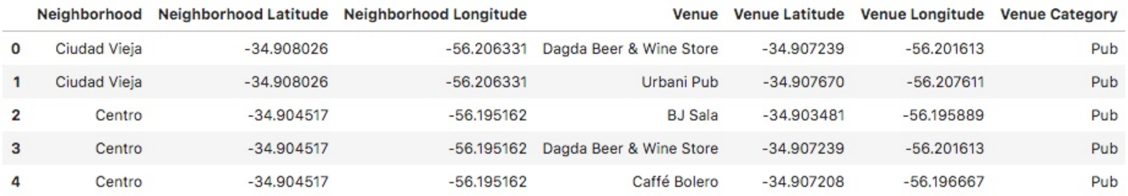
Neighborhood	id_barrio	BoroughCode	Cost_m2	Latitude	Longitude	HSP level	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
Ciudad Vieja	1	B	439	-34.908026	-56.206331	Low	0	Restaurant	BBQ Joint	Café	Bar	Art Museum	Coffee Shop	Hotel
Centro	2	B	576	-34.904517	-56.195162	Low	2	Hotel	Bar	Restaurant	Café	BBQ Joint	Coffee Shop	Sandwich Place
Barrio Sur	3	B	580	-34.910878	-56.188182	Low	2	Bar	Hotel	Restaurant	Scenic Lookout	Bakery	Deli / Bodega	Athletics & Sports
Cordón	4	B	544	-34.904140	-56.178411	Low	1	Bar	Pizza Place	Restaurant	Supermarket	Gym / Fitness Center	Furniture / Home Store	Theater
Palermo	5	B	820	-34.910688	-56.179806	Mid	1	Restaurant	Bar	Bakery	Gym	Pub	Coffee Shop	Basketball Stadium

It is interesting to also show the description of the top 3 venue categories for each neighborhood on the map. Therefore, we group each neighborhood by the top 3 venue categories count and combine these informations in a column named '*Join*'.

B.9. Use Foursquare API to explore the neighborhood candidates "*Pubs*"

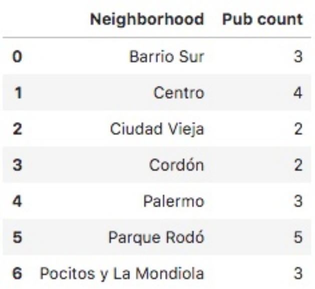
Let's get the number of **Pubs** that are in Montevideo candidate neighborhoods within a radius of 750 meters.

In the request, we get **Pubs** but also some secondary categories such as **Bars**, etc, because some categories share the same **Foursquare venue code**. Therefore, we filter the other categories and keep exclusively **Pubs**. Only 22 different pubs where returned by Foursquare for those 9 pre-selected neighborhoods:



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Ciudad Vieja	-34.908026	-56.206331	Dagda Beer & Wine Store	-34.907239	-56.201613	Pub
1	Ciudad Vieja	-34.908026	-56.206331	Urbani Pub	-34.907670	-56.207611	Pub
2	Centro	-34.904517	-56.195162	BJ Sala	-34.903481	-56.195889	Pub
3	Centro	-34.904517	-56.195162	Dagda Beer & Wine Store	-34.907239	-56.201613	Pub
4	Centro	-34.904517	-56.195162	Caffé Bolero	-34.907208	-56.196667	Pub

Let's count the number of **Pubs** in each neighborhood candidate:

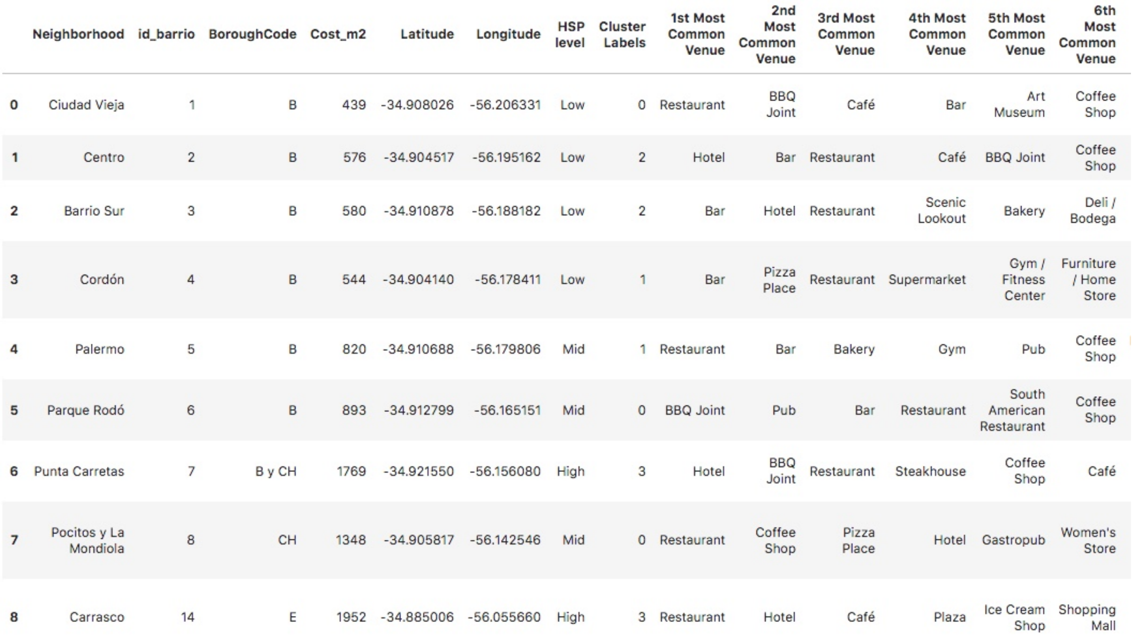


	Neighborhood	Pub count
0	Barrio Sur	3
1	Centro	4
2	Ciudad Vieja	2
3	Cordón	2
4	Palermo	3
5	Parque Rodó	5
6	Pocitos y La Mondiola	3

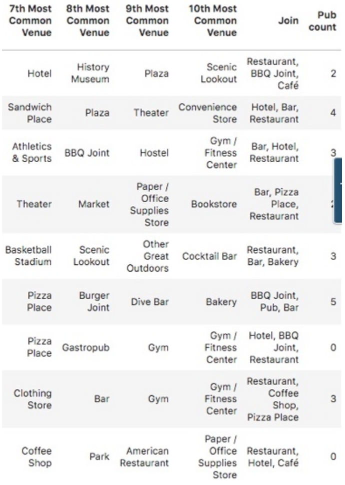
Notice that there are only 7 rows, because "Carrasco" and "*Punta Carretas*" showed not a single **Pub**. We will assign then the value 0 for the column '*Pub count'*.

C. Results

Let’s aggregate the columns 'Join' and 'Pub count' into our main **master table**, the dataframe created in B.8. It also includes the columns '*Cluster labels*' and '*HSP Levels*'. These are the results (shown in two pictures for reasons of space):

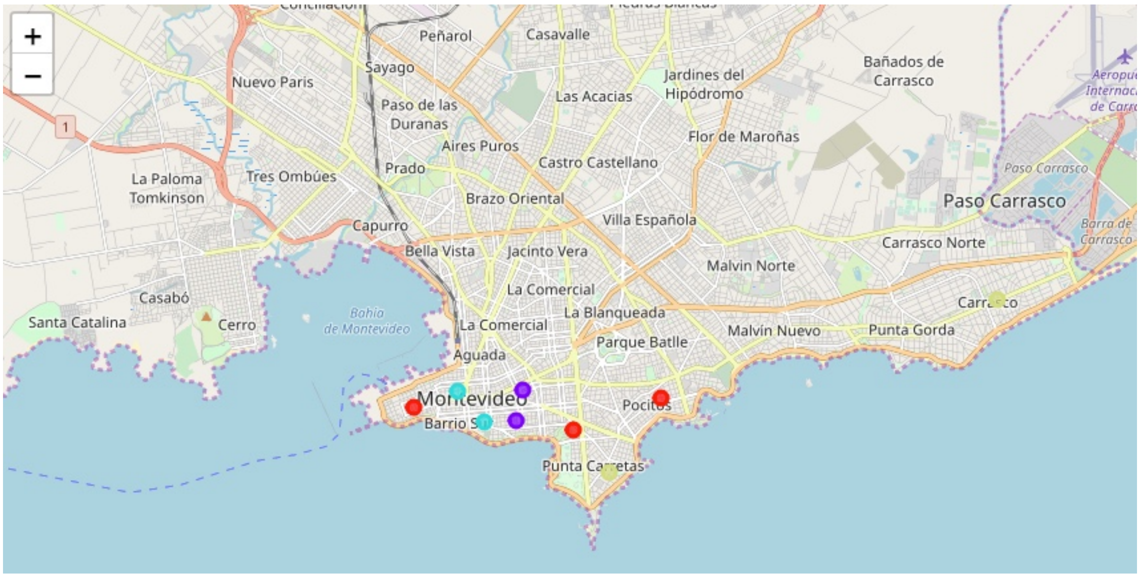


	Neighborhood	id_barrio	BoroughCode	Cost_m2	Latitude	Longitude	HSP level	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
0	Ciudad Vieja	1	B	439	-34.908026	-56.206331	Low	0	Restaurant	BBQ Joint	Café	Bar	Art Museum	Coffee Shop
1	Centro	2	B	576	-34.904517	-56.195162	Low	2	Hotel	Bar	Restaurant	Café	BBQ Joint	Coffee Shop
2	Barrio Sur	3	B	580	-34.910878	-56.188182	Low	2	Bar	Hotel	Restaurant	Scenic Lookout	Bakery	Deli / Bodega
3	Cordón	4	B	544	-34.904140	-56.178411	Low	1	Bar	Pizza Place	Restaurant	Supermarket	Gym / Fitness Center	Furniture / Home Store
4	Palermo	5	B	820	-34.910688	-56.179806	Mid	1	Restaurant	Bar	Bakery	Gym	Pub	Coffee Shop
5	Parque Rodó	6	B	893	-34.912799	-56.165151	Mid	0	BBQ Joint	Pub	Bar	Restaurant	South American Restaurant	Coffee Shop
6	Punta Carretas	7	B y CH	1769	-34.921550	-56.156080	High	3	Hotel	BBQ Joint	Restaurant	Steakhouse	Coffee Shop	Café
7	Pocitos y La Mondiola	8	CH	1348	-34.905817	-56.142546	Mid	0	Restaurant	Coffee Shop	Pizza Place	Hotel	Gastropub	Women's Store
8	Carrasco	14	E	1952	-34.885006	-56.055660	High	3	Restaurant	Hotel	Café	Plaza	Ice Cream Shop	Shopping Mall



7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Join	Pub count
Hotel	History Museum	Plaza	Scenic Lookout	Restaurant, BBQ Joint, Café	2
Sandwich Place	Plaza	Theater	Convenience Store	Hotel, Bar, Restaurant	4
Athletics & Sports	BBQ Joint	Hostel	Gym / Fitness Center	Bar, Hotel, Restaurant	3
Theater	Market	Paper / Office Supplies Store	Bookstore	Bar, Pizza Place, Restaurant	2
Basketball Stadium	Scenic Lookout	Other Great Outdoors	Cocktail Bar	Restaurant, Bar, Bakery	3
Pizza Place	Burger Joint	Dive Bar	Bakery	BBQ Joint, Pub, Bar	5
Pizza Place	Gastropub	Gym	Gym / Fitness Center	Hotel, BBQ Joint, Restaurant	0
Clothing Store	Bar	Gym	Gym / Fitness Center	Restaurant, Coffee Shop, Pizza Place	3
Coffee Shop	Park	American Restaurant	Paper / Office Supplies Store	Restaurant, Hotel, Café	0

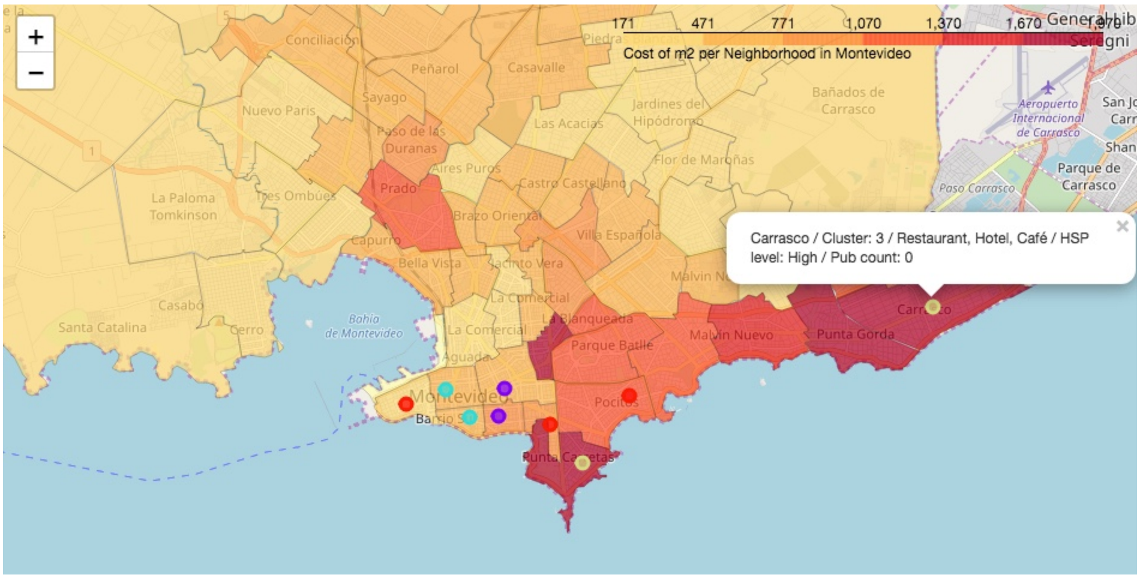
First, let's visualize the resulting **clusters** of the pre-selected neighborhoods of Montevideo in a **Folium** map:



In *Data Description* section, one of the targets was also visualize the **Average Housing Sale Prices** per square meter for each neighborhood with **Choropleth** style map, as we did in B.5.

Let's now add to that map the resulting **Clusters**, and descriptive **marks** containing the following information:

- Neighborhood name,
- Cluster number label,
- Top 3 venue categories,
- Average Housing Sale Price (*HSP*) level,
- Number of *Pubs* within 750 meters from the neighborhood's center coordinates



D. Discussion

Montevideo is far away the biggest city in Uruguay with a high population density specially near the coast. We performed data analysis through the available information by adding the **coordinates** of the neighborhood's centers and their average housing sale price of year 2017.

Foursquare venue data exploration showed as expected that neighborhoods with the highest number of venues where located there, along the shore of *Rio de la Plata* river. "*Buceo*", "*Malvin*" and "*Punta Gorda*" are exceptions because they are more of *residential* kind of neighborhoods. Only 9 of the 62 neighborhoods where pre-selected, based on having more than 50 venues, as a sign of enough economic activity as to consider establishing a *nightclub* business there.

Now starting from the pre-selected candidates, I used the ***K-means*** unsupervised learning algorithm as part of this clustering study separating into 4 clusters, based on the categories of the venues near the neighborhood's centers. We can see from the different cluster groups, that "*Barrio Sur*" and "*Centro*" have a high percentage of *Bars*, *Hotels* and *Restaurants*, and "*Cordon*" and "*Palermo*" also have a high rate of *Bars* and *Restaurants*, but with a more distributed sort of venue categories. "*Parque Rodo*" shows an important amount of *Bars* and *Pubs*, but in the rest venue categories this neighborhood is very similar to "*Pocitos*" and "*Ciudad Vieja*" in terms of non-night-time activities such as *Coffee Shops*, so they were clustered together. On the other hand, "*Punta Carretas*" and "*Carrasco*" both show no evidence of *Pubs* in their top 10 venue categories.

Taking this into consideration, we then explored specifically the number of Pubs in each candidate, and confirmed that not a single Pub venue was found within 750 meters from the neighborhood's center. Therefore, this represents an interesting opportunity of establishing the business in that zones.

But would we choose "*Punta Carretas*" or "*Carrasco*"? We could use the '*HSP level*' column, but we find out both have the same level: **High** average sale price per square foot. Therefore, as explained in *Data Description*, the last factor to take into account is the distance of both neighborhoods from city center: "*Punta Carretas*" is far more proximate to ***Tres Cruces bus terminal*** and to the other candidates than "*Carrasco*". So a good *preliminar* advice may be to establish the business in "*Punta Carretas"*. **But this two options go against the restriction of having a not-so-high average sale price**, so perhaps "**Pocitos**" with its low rate of **Pubs** and '*Mid*' *HSP* would be a better choice.

I ended the study by visualizing the data a on a **Folium** Montevideo map.

E. Conclusion

Even though we approximated to a solution that is fair enough, we need some more elements to take a more certain decision, because the **number of Pubs within neighborhoods centers** and the **HSP level** seem to result opposing elements, therefore we would need to accept a **compromise solution**.

Nevertheless, *Data Science* analysis provides very important descriptive and quantitative tools that lead to take better decisions and arrive to more profitable outcomes. Not only investors but other stakeholders and common citizens can benefit from the available information.

Thanks for your attention!!

Best regards,

Daniel Gallino

daniel.gallino@gmail.com