Capstone Project: The Battle of the Neighborhoods (Week 2)

Applied Data Science Capstone by IBM/Coursera

Author: Maximiliano Rivas

Content

1. Introduction: Business Problem

The current project has as objective to discover the best possible location in Santiago de Chile for **starting a new bar entrepreneurship focused on selling premium craft beers to customers.**

Santiago de Chile has around 7MM people living in an area of 15.400 km2 partitioned in 52 different boroughs. **These boroughs have a population, number of houses, area and a large array of different amenities given.**

Our mission is to **leverage different data science tools that helps to determine which boroughs are the most promising and venues inside these ones to locate the stakeholders' bar.**

The criteria used for the analysis are:

- **Boroughs with high density of people** would be prefer in order to assure a flow of customers into the bar.
- **Boroughs with low number of people per house** will be prefer due to singles and young couples used to visit more bars than families with kids.
- We will prefer **locations with many restaurants, cinemas, theaters, discos and bars** in order to gain exposure for the targeted customers but **avoiding proximity to bars delivering our same services.**

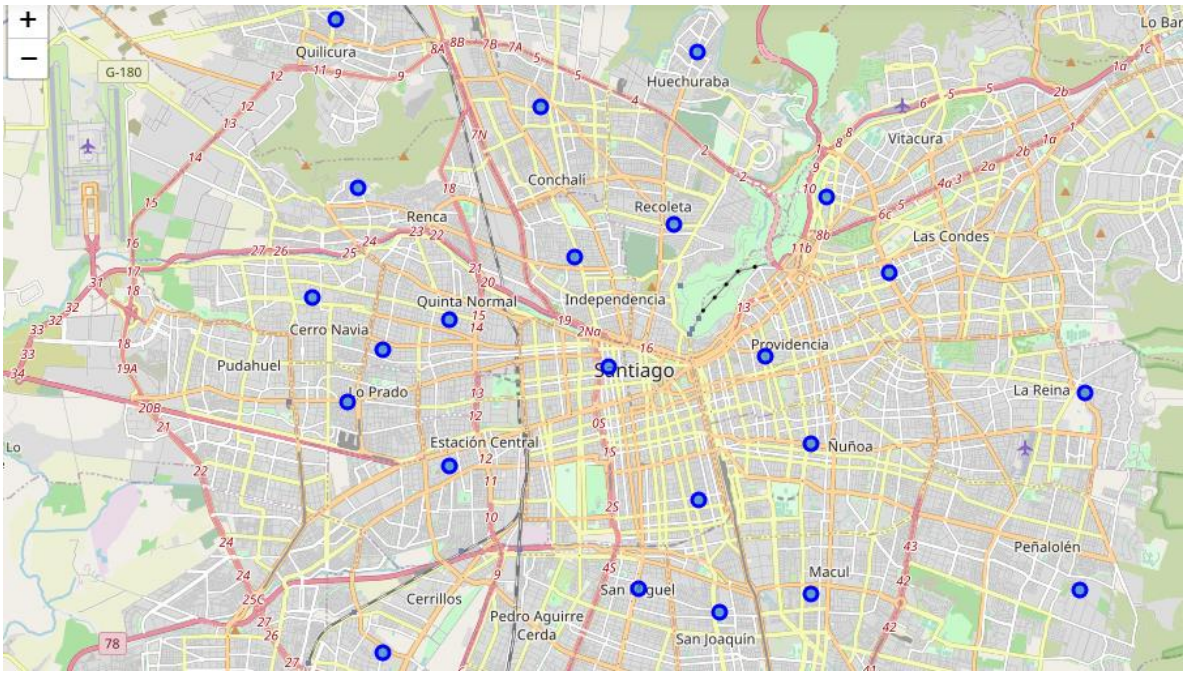
2. Data

a. Data about the boroughs and their location

We must filter for the boroughs in Santiago, convert the format for the coordinates and drop those columns that are useless for our propose. It is important to notice that each borough has a unique identifier call "CUT" which is useful to cross data later.

	CUT	Borough	Latitude	Longitude
294	13101	Santiago	-33.437222	-70.657222
295	13102	Cerrillos	-33.500000	-70.716667
296	13103	Cerro Navia	-33.422000	-70.735000
297	13104	Conchalí	-33.380000	-70.675000
298	13105	El Bosque	-33.567000	-70.675000

Here we can use folium to map all boroughs. We can notice that points are distributed in a circular area.


Quilicura
Huechuraba
Vitacura
Conchalí
Recoleta
Renca
Las Condes
Quinta Normal
Independencia
Cerro Navia
Providencia
Pudahuel
Santiago
Lo Prado
La Reina
Estación Central
Ñuñoa
Macul
Peñalolén
Cerrillos
Pedro Aguirre Cerda
San Joaquín
G-180
78

b. Data about population, number of houses and area of boroughs

In this case the data was extracted from a csv file posted by “Instituto Nacional de Estadisticas” (public institution focused on creating indices and data in Chile).

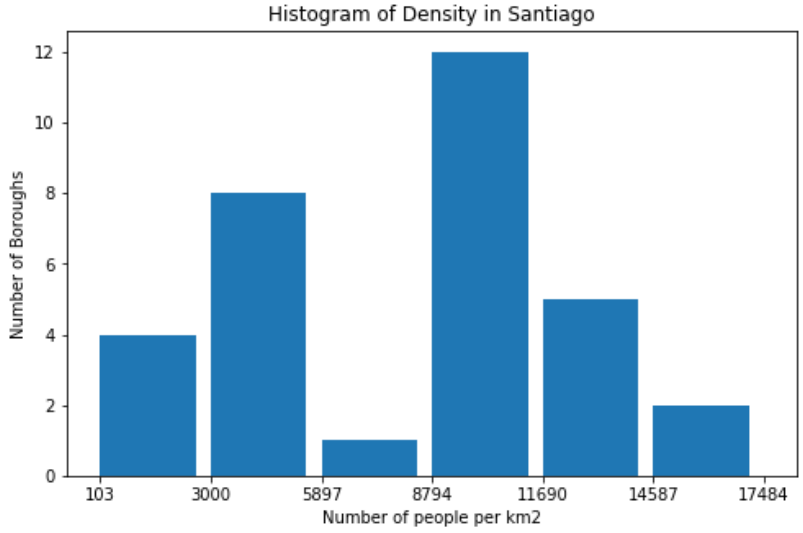
	CUT	Population	Houses	Area (km2)	Density
274	13102	80832	24547	16.779650	4817.263672
275	13131	82900	23855	6.277112	13206.710938
276	13132	85384	31777	28.417034	3004.676758
278	13109	90119	31480	9.979139	9030.739258
283	13113	92787	29801	23.438091	3958.812012

Here we can cross this dataframe with the previous one thanks to the ‘CUT’ code, generating a unique table with information of 32 boroughs.

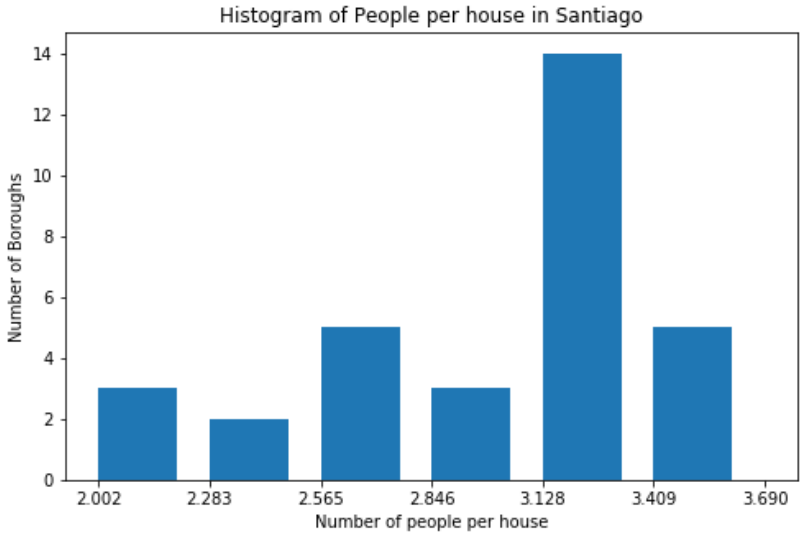
	CUT	Borough	Latitude	Longitude	Population	Houses	Area (km2)	Density	People/House
0	13101	Santiago	-33.437222	-70.657222	404495	193628	23.135237	17483.935547	2.089032
1	13102	Cerrillos	-33.500000	-70.716667	80832	24547	16.779650	4817.263672	3.292948
2	13103	Cerro Navia	-33.422000	-70.735000	132622	38020	11.097359	11950.771484	3.488217
3	13104	Conchalí	-33.380000	-70.675000	126955	37759	11.109763	11427.335938	3.362245
4	13105	El Bosque	-33.567000	-70.675000	162505	47941	14.324402	11344.626953	3.389687

We are able to extract 3 insights

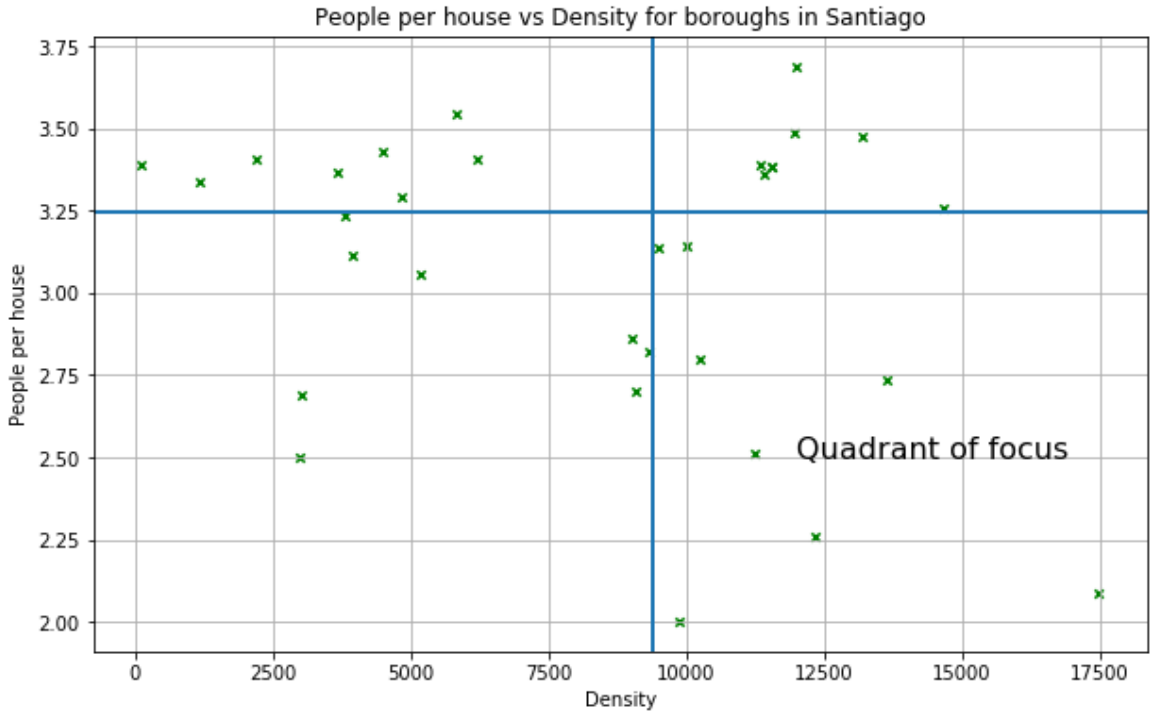
Around half of the boroughs (15) have low density of people (<8.8k people/km2) and the other half high density (>8.8k people/km2). We must focus on this half.



Around one third of the boroughs (10) have a low number of people per houses (<2.85 people/house). We must focus on this third.



Finally, with the 2 previous pieces of information we have a quadrant of interest to put focus on



c. Data about the venues and their location for each borough

Thanks to the Foursquare API we can extract the location and category for every Venue located in each borough of Santiago. To extract the info, we use the proper credentials and a range of 2 kilometers around each borough and a limit of 100 venues per borough.

	Borough	Borough Latitude	Borough Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Santiago	-33.437222	-70.657222	Plaza de Bolsillo - Santiago Centro	-33.436778	-70.655481	Plaza
1	Santiago	-33.437222	-70.657222	Starbucks	-33.437938	-70.657007	Coffee Shop
2	Santiago	-33.437222	-70.657222	Amanda's	-33.439206	-70.658247	Arepa Restaurant
3	Santiago	-33.437222	-70.657222	Museo Chileno de Arte Precolombino	-33.438776	-70.652363	Museum
4	Santiago	-33.437222	-70.657222	GYROS BISTRÓ	-33.439793	-70.655792	Sandwich Place
...	...	...	...	...	...	...	...
1883	Vitacura	-33.400000	-70.600000	Eric Kayser	-33.416495	-70.597481	Deli / Bodega
1884	Vitacura	-33.400000	-70.600000	Check INN	-33.388067	-70.615203	Restaurant
1885	Vitacura	-33.400000	-70.600000	Brooks Night	-33.389916	-70.615861	Athletics & Sports
1886	Vitacura	-33.400000	-70.600000	Ultra Music Festival - Chile 2015	-33.391467	-70.617569	Concert Hall
1887	Vitacura	-33.400000	-70.600000	Santiago Papperchase	-33.386364	-70.610017	Racetrack

1888 rows × 7 columns

3. Methodology and Analysis

The methodology we are going to follow has 3 stages:

a. Clustering the boroughs with k-means

Here we want to identify the type of boroughs we are dealing with so we can discard those useless for the purpose of this project

b. Slicing the clusters of boroughs according to density and people per house indexes

Here we want to select the subset of boroughs that meet the demographic requirements stated at the introduction, locations with high density of people and few people per house (singles, couples without children, etc.)

c. Select neighborhoods within these boroughs with a density of "complementary" amenities

Finally, from this small portion of borough we would identify those that contain neighborhoods packed with complementary amenities for starting a craft beer bar (i.e. restaurants, cinemas, theaters, other types of bars, etc.)

a. Clustering the boroughs with k-means

First, we need to create a dummy variable for each venue category in our table and calculate and average for every Borough in order to get a weighted collection of categories per Borough.

	Borough	Airport	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	...	Video Game Store	Video Store	Volleyball Court	V
0	Cerrillos	0.017857	0.017857	0.0	0.0	0.0	0.000000	0.017857	0.0	0.0	...	0.0	0.0	0.0	
1	Cerro Navia	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	...	0.0	0.0	0.0	
2	Conchalí	0.000000	0.019231	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	...	0.0	0.0	0.0	
3	El Bosque	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.038462	0.0	0.0	...	0.0	0.0	0.0	
4	Estación Central	0.000000	0.000000	0.0	0.0	0.0	0.016949	0.000000	0.0	0.0	...	0.0	0.0	0.0	

5 rows × 240 columns

Then we can sort this weighted collection to get the top 10 categories for each borough.

	Borough	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	Cerrillos	Fast Food Restaurant	Ice Cream Shop	Sandwich Place	Pharmacy	Burger Joint	Movie Theater	Food & Drink Shop	Park	Café	Department Store
1	Cerro Navia	Bus Station	Park	Japanese Restaurant	Pharmacy	Burger Joint	Fried Chicken Joint	Flea Market	Mountain	Grocery Store	Chinese Restaurant
2	Conchalí	Furniture / Home Store	Department Store	Ice Cream Shop	Pharmacy	Restaurant	Farmers Market	Gym	Bakery	Sushi Restaurant	Donut Shop
3	El Bosque	Flea Market	Pharmacy	Pizza Place	Burger Joint	Supermarket	Gastropub	Sushi Restaurant	Mobile Phone Shop	Chinese Restaurant	Bar
4	Estación Central	Pharmacy	Bakery	Fast Food Restaurant	Snack Place	Hot Dog Joint	Chinese Restaurant	Museum	Asian Restaurant	Bar	Sushi Restaurant

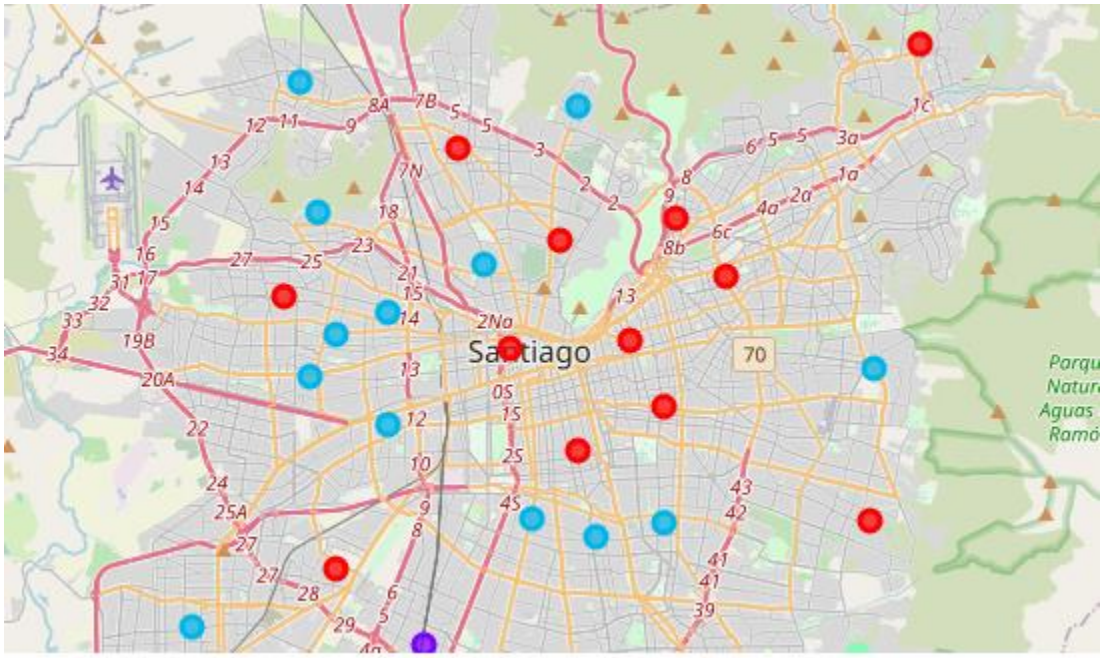
Finally, we applied the machine learning clustering technique called K means for the boroughs and with a k parameter of 5. It is done with the objective to discriminate which boroughs we should choose without doing calculations borough by borough.

	Cluster Labels	Borough	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	0	Cerrillos	Fast Food Restaurant	Ice Cream Shop	Sandwich Place	Pharmacy	Burger Joint	Movie Theater	Food & Drink Shop	Park	Café	Department Store
1	0	Cerro Navia	Bus Station	Park	Japanese Restaurant	Pharmacy	Burger Joint	Fried Chicken Joint	Flea Market	Mountain	Grocery Store	Chinese Restaurant
2	0	Conchalí	Furniture / Home Store	Department Store	Ice Cream Shop	Pharmacy	Restaurant	Farmers Market	Gym	Bakery	Sushi Restaurant	Donut Shop
3	0	El Bosque	Flea Market	Pharmacy	Pizza Place	Burger Joint	Supermarket	Gastropub	Sushi Restaurant	Mobile Phone Shop	Chinese Restaurant	Bar
4	2	Estación Central	Pharmacy	Bakery	Fast Food Restaurant	Snack Place	Hot Dog Joint	Chinese Restaurant	Museum	Asian Restaurant	Bar	Sushi Restaurant

It is possible to notice that the k-means model creates the cluster 0 and 2 with a group of boroughs while the clusters 1, 3 and 4 only have 1 or 2 boroughs. We will consider these last clusters as outliers for our analysis.

	CUT	Borough	Latitude	Longitude	Population	Houses	Area (km2)	Density	People/House	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
Cluster Labels															
0	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14
3	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

And it possible to create a map to visualize the results.



b. Slicing the clusters of boroughs according to density and people per house indexes

Here we are going to work with the cluster 0 and 2. Using the medians of 'Density' and 'People per house' we can slice these clusters in order to discard those boroughs with less potential up front.

The cluster 0 satisfy the fact of having 'complementary amenities' for our beer bar (i.e. restaurants, parks, hotels, nightclubs, etc.)

Applying the median slicing we get a subset of 4 boroughs, with a 'Density' mean of 12.427 and 'People per house' mean of 2.37.

```
12427.0280761725 2.3735781606150073
```

	index	CUT	Borough	Latitude	Longitude	Population	Houses	Area (km2)	Density	People/House	...	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Mc Comm Ven
0	0	13101	Santiago	-33.437222	-70.657222	404495	193628	23.135237	17483.935547	2.089032	...	Coffee Shop	Art Museum	Tea Room	Booksto
1	19	13120	Ñuñoa	-33.454000	-70.604000	208237	92248	16.856802	12353.292969	2.257361	...	Coffee Shop	Peruvian Restaurant	Pizza Place	Ca
2	22	13123	Providencia	-33.435000	-70.616000	142079	70965	14.394146	9870.609375	2.002100	...	French Restaurant	Park	Pizza Place	Ho
3	26	13127	Recoleta	-33.406000	-70.640000	157851	50178	15.784667	10000.274414	3.145821	...	Park	Nightclub	Pharmacy	Bake

4 rows × 21 columns

The cluster 2 satisfy the fact of having 'complementary amenities' too (i.e. restaurants, parks, food trucks, etc.)

Applying the median slicing in the cluster 2 we get a subset of 4 boroughs as well, but a lower 'Density' mean of 11.152 and a higher 'People per house' mean of 2.79

	index	Borough	Houses	Area (km2)	Density	People/House	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 Mc Comm Ven
0	5	Estación Central	52486	14.353261	10244.431641	2.801528	2	Pharmacy	Bakery	Fast Food Restaurant	Snack Place	Hot Dog Joint	Chinese Restaurant	Muse
1	7	Independencia	36666	7.355460	13633.545898	2.734986	2	Restaurant	Sandwich Place	Park	Peruvian Restaurant	Chinese Restaurant	Farmers Market	C
2	28	San Joaquín	30096	9.942262	9504.075195	3.139686	2	Food Truck	Restaurant	Theater	Farmers Market	Sandwich Place	Chinese Restaurant	BBQ Jc
3	29	San Miguel	42947	9.613154	11229.821289	2.513656	2	Sushi Restaurant	Pizza Place	Restaurant	Plaza	Peruvian Restaurant	Park	La Americ Restaur

So, we decided to continue working with the cluster 0 filtered and its 4 boroughs

c. Select neighborhoods within these boroughs with a density of "complementary" amenities

Here we are going to look up for the neighborhoods with more potential for locating a premium craft beer bar within each borough, those neighborhoods packed with bars.

Here we extract all the bars within the **Santiago Borough**

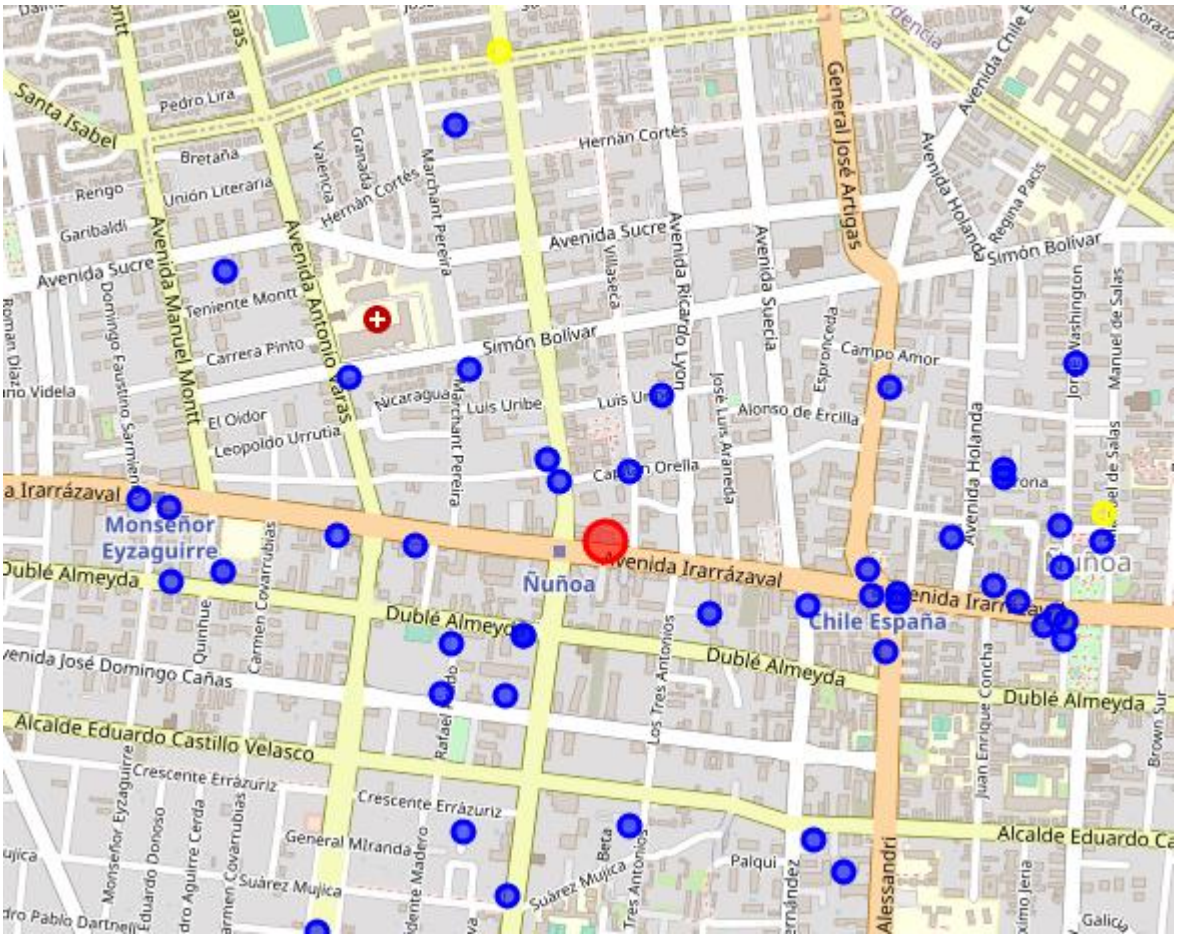
	name	categories	lat	lng	distance	cc	id
0	Bar Nacional 1	South American Restaurant	-33.439843	-70.653029	486	CL	4b686332f964a52065752be3
1	Bar Nacional 2	Bar	-33.439575	-70.652301	526	CL	4b587fc9f964a520b55a28e3
2	Bar Nacional 3	Restaurant	-33.442458	-70.649687	910	CL	4b685244f964a5208d712be3
3	Bar La Unión Chica	Bar	-33.443312	-70.651074	886	CL	4dc00a7c43a147cd6eceeadd
4	bar devic	Beer Garden	-33.441018	-70.665086	843	CL	505e7aa3e4b0a70768115621

And map them, coloring in yellow those bars specialized in selling beers. There are plenty of bars in the nearby but only 2 of them specialized in beers so this is a potential neighborhood

Here we extract all the bars within the Ñuñoa Borough

	name	categories	address	lat	lng	labeledLatLngs	distance	cc	city	state	country	formattedAddress	crossStree
0	Bar C-ñal	Concert Hall	Irarrazabal 2191	-33.454075	-70.608077	[{'label': 'display', 'lat': -33.454075, 'lng'...	378	CL	Santiago	Santiago Metropolita	Chile	[Irarrazabal 2191, Santiago, Santiago Metropol...	NaN
1	Bar Sin Nombre	Bar	Av. Irarrázaval 3420	-33.455060	-70.595159	[{'label': 'display', 'lat': -33.45506039, 'ln...	829	CL	Santiago de Chile	Metropolitana de Santiago de Chile	Chile	[Av. Irarrázaval 3420 (Holanda), Santiago de C...	Holanda
2	Mephisto Bar	Bar	José Pedro Alessandri 103	-33.455956	-70.597991	[{'label': 'display', 'lat': -33.4559564584904...	599	CL	Ñuñoa	Metropolitana de Santiago de Chile	Chile	[José Pedro Alessandri 103, Ñuñoa, Metropolita...	NaN
3	Bar The Clinic	Bar	Jorge Washington 58	-33.454466	-70.594196	[{'label': 'display', 'lat': -33.454466, 'lng'...	912	CL	Ñuñoa	Metropolitana de Santiago de Chile	Chile	[Jorge Washington 58 (Plaza Ñuñoa), 7790827 Ñu...	Plaza Ñuñoa
4	Mizu Sushi Bar & Delivery	Japanese Restaurant	NaN	-33.455685	-70.605759	[{'label': 'display', 'lat': -33.4556847549477...	248	CL	Santiago de Chile	Metropolitana de Santiago de Chile	Chile	[Santiago de Chile, Metropolitana de Santiago ...	NaN

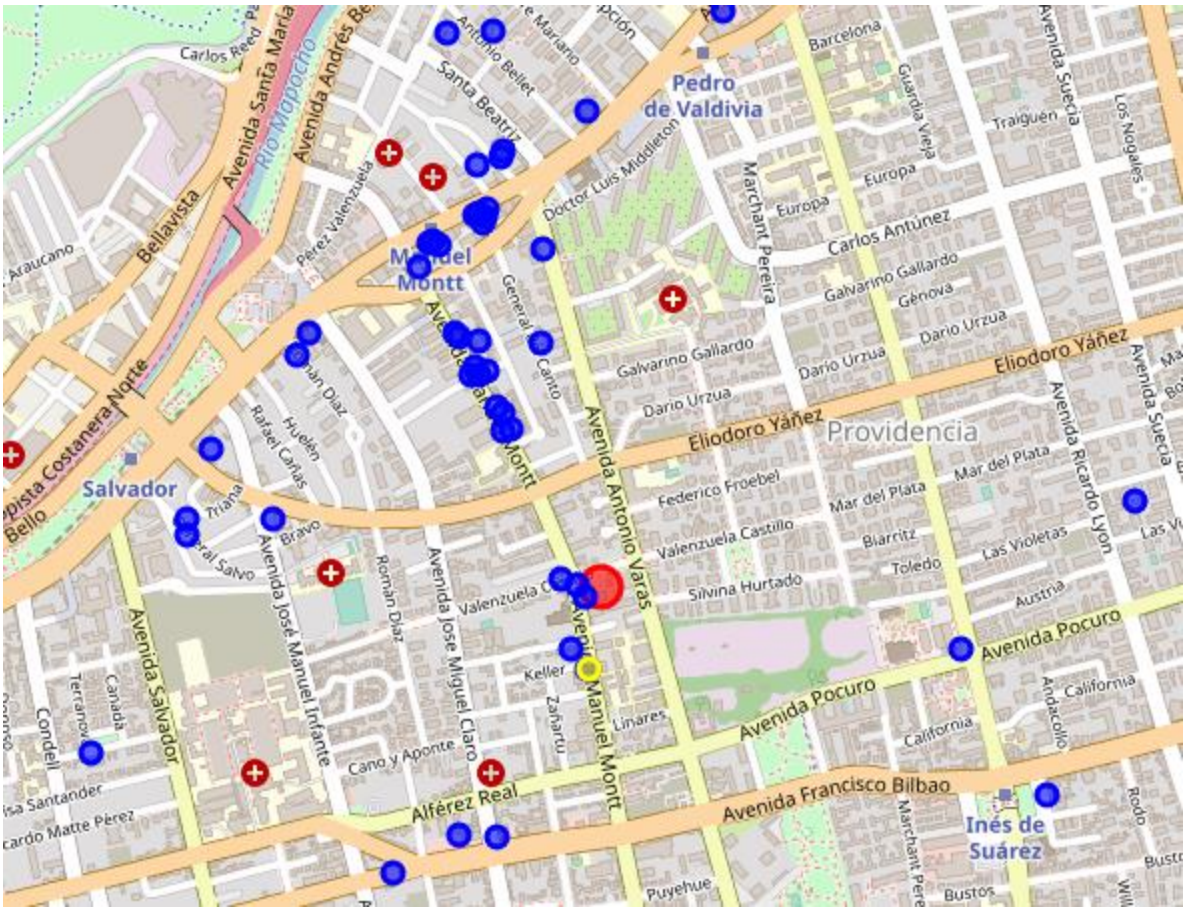
And map them, coloring in yellow those bars specialized in selling beers. Again, there are plenty of bars in the nearby but only 2 of them specialized in beers, which are far apart each other, so this is a potential neighborhood



Here we extract all the bars within the Providencia Borough

	name	categories	lat	lng	distance	cc	neighborhood	id
0	Patxaran Bar	Bar	-33.436440	-70.616253	162	CL	NaN	4e5286aac65ba127fb330cf0
1	Domani Pizzeria Bar	Pizza Place	-33.431115	-70.618464	489	CL	Providencia	5896af397ff1e43860cd3ede
2	Bar Nativo	Bar	-33.431058	-70.618685	504	CL	NaN	4ed05a866c2510ace0ec7195
3	Pacto Arte Bar	Bar	-33.431200	-70.618630	488	CL	NaN	4ceefc887b94370430513053
4	Bar Sísmico	Karaoke Bar	-33.442676	-70.614595	864	CL	NaN	4c099e78bbc676b0ad9248d5

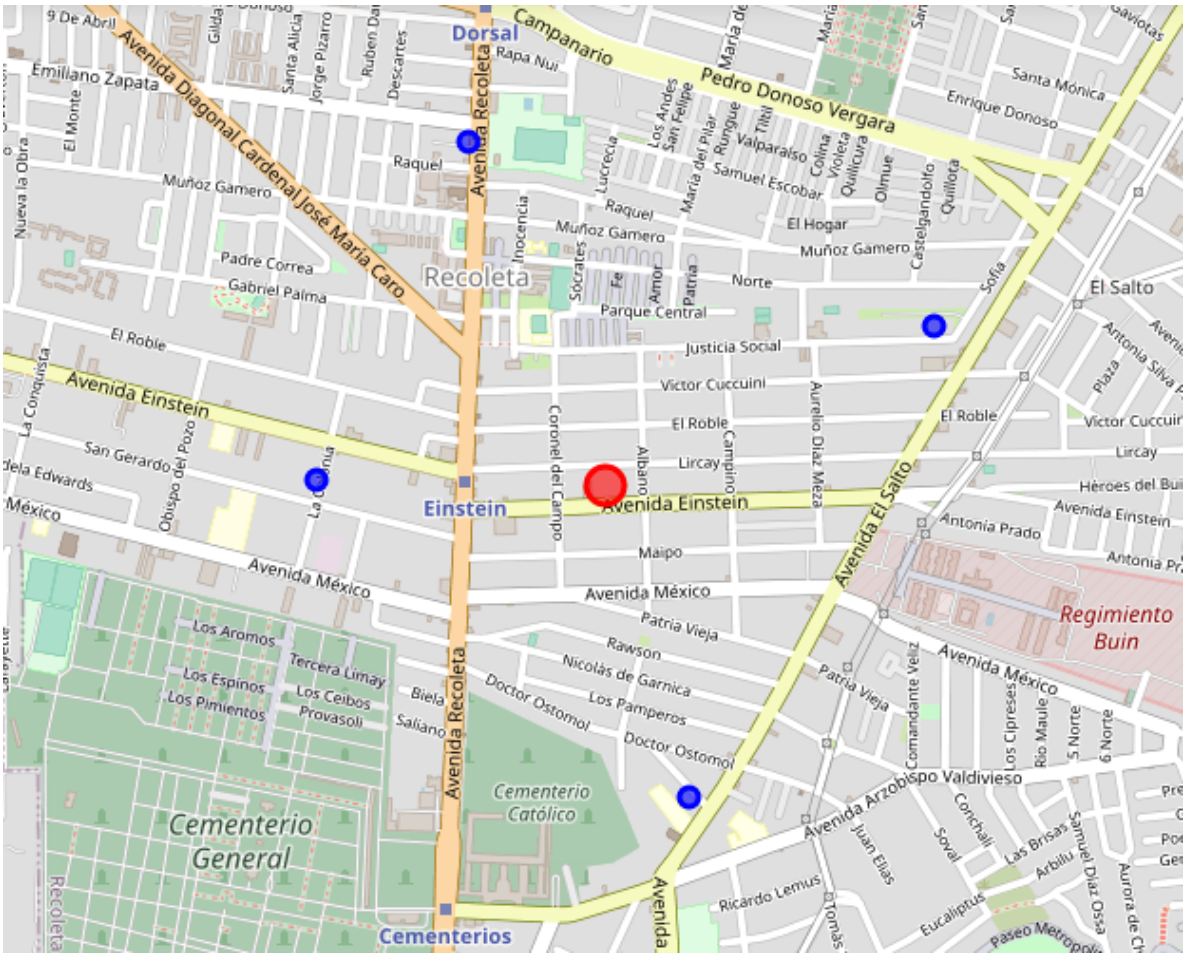
And map them, coloring in yellow those bars specialized in selling beers. Again, there are plenty of bars in the nearby but only 1 of them specialized in beers and far from the high density of bars, so this is another potential neighborhood



Here we extract all the bars within the Recoleta Borough

	name	categories	lat	lng	distance	cc	id
0	Royal Salute ~bar~	Bar	-33.403015	-70.632568	766	CL	55720d30498ebb629de95970
1	Barrio Einstein	Housing Development	-33.405891	-70.646516	605	CL	544bc3c7498eefae2e4f5b8e
2	Barraca de Fierro	Hardware Store	-33.399540	-70.643070	773	CL	4fbe38d1e4b0bb10a26f0fe6
3	Barrio los olivos	Church	-33.411895	-70.638116	679	CL	4ff061b1e4b0616aaaba94f4
4	Barrio El Salto Chico	Road	-33.415812	-70.642098	1109	CL	4fa121bee4b0f30726e0e1bd

In this case we map the bars but there are only 4 of them far apart each other, we consider it is not a neighborhood with potential for our beer bar


Dorsal
Campanario
Pedro Donoso Vergara
Avenida Recoleta
Avenida Diagonal Cardenal José María Caro
Recoleta
Avenida Einstein
Einstein
Avenida México
Avenida El Salto
El Salto
Regimiento Buin
Cementerio General
Cementerio Católico
Cementerios
Parque Central
Justicia Social
Victor Cuccuini
Lircay
Maipo
Patria Vieja
Rawson
Nicolás de Garnica
Los Pamperos
Doctor Ostomol
Avenida Arzobispo Valdivieso
Antonia Prado
Héroes del Bui
Raquel
Muñoz Gamero
Padre Correa
Gabriel Palma
El Roble
San Gerardo
Obispo del Pozo
Coronel del Campo
Aurelio Díaz Meza
Comandante Veliz
Los Cipreses
Río Maule
5 Norte
6 Norte
Los Aromos
Tercera Limay
Los Espinos
Los Ceibos
Los Pimientos
Provasoli
Biela
Saliano
Ricardo Lemus
Juan Elías
Soval
Corichali
Las Brisas
Arbilu
Samuel Díaz Ossa
Eucaliptus
Paseo Metropol
Emiliano Zapata
Rapa Nui
Santa Mónica
Enrique Donoso
Valparaíso
Samuel Escobar
El Hogar
Norte
Sofía
Antonia Silva
Plaza
Victor Cuccuini
Lircay
Héroes del Bui
Avenida Einstein
Antonia Prado
Antonia Pr
Avenida México
Patria Vieja
Avenida México
Avenida Arzobispo Valdivieso
Recoleta

4. Results and Discussion

As a result, for our project we got that there 3 boroughs with neighborhoods within them capable of sustain and assurance a new premium craft beer. These are Santiago, Ñuñoa and Providencia. It is because there are plenty of 'complementary' amenities in a range of 1km around the 'heart' of each borough which are helpful to gain exposure for target customers, a constant flow of people and a lack of competitive landscape (only 1 or 2 bars specialized in beers), which can be helpful in catch a the lovers of beer niche.

Contrasting the results with reality it is highly feasible that a new specialized beer bar could start here, these are well-known zones of restaurants, bars, clubs and nocturnal entertainment services. It is necessary though to do a further analysis in the place to locate this new bar, all the economies associated with the entrepreneurship as investment and costs, and a measure of proximity to workplace and public transport modes.

5. Conclusions

The main conclusions for our project are:

- Analyzing important business decision as location could be done easily with data science knowledge and a few open sources of information.
- It is possible to leverage machine learning techniques to clustering information without the need to build hard mathematical models or having a deep knowledge about the data.
- Foursquare API is a powerful tool to gain insights about location of amenities, it is helpful for analyzing potential places for new business as in our case.
- Data science analysis need to be done with a common and business sense in order to find shortcuts that allow us to save time and energy, as in our case with the density and people per house analysis rational.
- It is the first step for developing the beer bar idea but one of the most important and with huge impact in the revenues and growth of business on the future.

6. References

- Foursquare API.
- Wikipedia (comunas de Chile).
- INE (Población comunas de Chile).