The Battle of the Neighborhoods

Applied Data Science Capstone- Capstone Project

Jhonny Calvo

calvorondon@gmail.com

Proposal for Opening a Bar-Restaurant in Caracas City.

1. Introduction

A description of the problem and a discussion of the background.

1.1 Problem Background

Caracas, is Venezuela's Capital. Here is the main commercial and cultural center of the country. Most of governmental and private institutions has their main office in this city. Caracas has an excellent climate, with an average temperature of 18.6 ° C throughout the year, which makes it a destination sought by tourists [1].

In Caracas there are people of many nationalities and many ethnicities, so you can say that this is a multicultural city. The city center is very crowded by commerce, therefore, the cost of services and settlement is high. Competitiveness is also high, so the creation of a new business requires previous studies as well.

1.2 Problem Description

A Bar-Restaurant is a commercial establishment where foods, alcoholic and non-alcoholic drinks are sold. They are usually aimed at adults, especially young people. The bars usually have a specific style that together with the structural design give the main characteristics and become striking for certain types of customers.

The problem lies in finding a suitable place to establish a new Bar-Restaurant, which is preferred for people because of its authenticity and unique style. It must also be located in an accessible place, within a busy area with low competition.

1.3 Objective of the project

The client of this project wishes to establish a new bar-restaurant, aimed at people between 25 and 50 years old, who are executives and professionals who wish to have fun after leaving the office. The client also requires the bar restaurant to be located in an accessible place where you can get customers who meet the profile.

1.4 Products

A recommendation from the best area of Caracas will be delivered to the client to open a new bar restaurant, based on their requirements. We will provide lists with those places with high density of Bar-Restaurant and classified by specific characteristics (as type of bar or restaurant).

With this the client will have a clearer vision of what type of bars and restaurant already exist, where they are located and where he should locate his new restaurant bar

2. Data Description

This is a description of the data and how it will be used to solve the problem. We will use two datasets to analyze Caracas City.

2.1 Dataset 1:

With this data we can know the name of the main neighborhoods of the city and the location of all existing venues. This Dataset is taken from the Antipodas.Net website [2], which contains the principal neighborhood's coordinates of Caracas City. We use this datasets to explore the neighborhoods distribution. We have found that Caracas City contains 30 Neighborhoods.

To obtain this data, we have used BeautifulSoup, LXML and requests, which are libraries for web scraping. So, using this libraries we obtain the html code for the web page, and then we need to keep the essential information into a dataframe using pandas library.

Dataset 1 is structured as follows:

- Neighborhood: which is the neighborhood name.
- Distance: is the distance in kilometers from Caracas center to each neighborhood.
- Coordinates: contains geographical coordinates in DMS format (DMS is "Degree- Minutes-Seconds").

An example of this data is shown below:

	Neighborhood	Distance	Coordinates
1	Parroquia San Pedro	1 Km	N10°29'20.36" O66°53'20.29"
2	Parroquia El Recreo	2 Km	N10°30'18.11" O66°53'15.83"
3	Chacao	3 Km	N10°29'45.53" O66°51'12.49"
4	Parroquia Candelaria	3 Km	N10°30'21.46" O66°54'17.78"
5	Parroquia San Agustín	3 Km	N10°29'38.76" O66°54'36.94"

Figure 1. Neighborhoods coordinates from Antipodas.net .

2.2 Dataset 2:

We use Foursquare [3] to explore the principal Venues of Caracas City. So, this datasets contains all Venues for Caracas City, its coordinates, and the Neighborhoods where the Venue belongs. The Dataset needs to be filter, to keep those with restaurant and bar characteristics. With this Datasets we explore the places with high density of bar and restaurants. We have search places with the follow tags: Bar, Restaurant (chinese, spanish, american, etc..), Lounge, and many others.

The coordinates found in Dataset 1, have been utilized as input for the Foursquare API. An example of this datasets is shown below:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parroquia San Pedro	10.489889	-66.889694	Estadio Olímpico Universitario	10.490449	-66.885710	Soccer Stadium
1	Parroquia San Pedro	10.489889	-66.889694	Plaza del Rectorado	10.492148	-66.890817	College Quad
2	Parroquia San Pedro	10.489889	-66.889694	Video San Pedro	10.485993	-66.891776	Video Store
3	Parroquia San Pedro	10.489889	-66.889694	Estadio Universitario de Caracas	10.489607	-66.884255	Baseball Stadium
4	Parroquia San Pedro	10.489889	-66.889694	Plaza Cubierta	10.491076	-66.890764	College Quad

Figure 2. Foursquare results (first 5 rows).

Notice that geographical coordinates in this dataset are in DD format (Decimal Degrees).

3. Methodology

The aim in this project is to classify all Bar-Restaurant venues in Caracas city, so we can evaluate the density of clusters by neighborhoods, and provide the best places to open a new Bar-Restaurant. We can divide the project in three parts:

- **Part 1**: consist in ordering and cleaning Dataset 1, so it can be manageable and useful in all libraries required. We have use Foursquare, with Dataset 1 as input, and so we can obtain all venues in Caracas city (Dataset 2). Dataset 2 have been filtered to keep Bar Restaurant venues. Folium library is used to visualize neighborhoods and venues in a map.

- **Part 2**: Consist in the analysis of each neighborhood, so we can know the frequency of venues and the common venues per neighborhoods.

- **Part 3**: Clustering neighborhoods. We used k-means to classify neighborhoods by venues, and observe it in a map, and know how the clusters are distributed in the city.

3.1 Part 1

3.1.1 Data Ordering

As we can see in figure 1, Dataset 1 contains three columns: Neighborhoods, Distance and Coordinates. We don´t need Distance columns, so we dropped it from the Data Frame. We only need to order the coordinates columns, so we can have latitude and longitude in different columns.

	Neighborhood	Lat	Long
0	Parroquia San Pedro	N10°29'20.36"	O66°53'20.29"
1	Parroquia El Recreo	N10°30'18.11"	O66°53'15.83"
2	Chacao	N10°29'45.53"	O66°51'12.49"
3	Parroquia Candelaria	N10°30'21.46"	O66°54'17.78"
4	Parroquia San Agustín	N10°29'38.76"	O66°54'36.94"

Figure 3. Neighborhood coordinates in DMS format.

3.1.2 Conversion from DMS to DD (handle for foursquare)

Neighborhood Coordinates in Dataset 1, are in DMS format (Degree- Minutes-Seconds), and we going to use this coordinates to search the main venues for each neighborhood. But Foursquare and Folium use DD format (Decimal Degrees) in coordinates, so, we need to convert coordinates in Dataset 1 to DD format. To do this, a conversion function was defined, to transform from DMS to DD format.

	Neighborhood	Latitude	Longitude
0	Parroquia San Pedro	10.489889	-66.889694
1	Parroquia El Recreo	10.505306	-66.889806
2	Chacao	10.497306	-66.854694
3	Parroquia Candelaria	10.507111	-66.906889
4	Parroquia San Agustín	10.496000	-66.912611

Figure 4. Neighborhood coordinates in DD format.

3.1.3 Data Visualization

We used Folium library to visualize the center of the neighborhoods and its distribution in Caracas city. We have use coordinates from Dataset 1 as input.

Figure 5. Average position of Caracas neighborhoods.

3.1.4 Filtering

Using Foursquare and Dataset 1, we obtain all venues in all categories from Caracas City. Remember, in this study we only want to explore those venues with Bar-Restaurant characteristic. So, we have explored all 'Venue Category' to select and retain those that match the requirements. The selected categories was:

1. American Restaurant
2. Asian Restaurant
3. Bar
4. Chinese Restaurant
5. Cocktail Bar
6. Disco
7. French Restaurant
8. Gay Bar
9. Hotel Bar
10. Italian Restaurant
11. Japanese Restaurant
12. Latin American Restaurant
13. Lounge
14. Mediterranean Restaurant
15. Mexican Restaurant
16. Middle Eastern Restaurant
17. NightClub
18. Paella Restaurant
19. Peruvian Restaurant
20. Pizza Place
21. Restaurant
22. Salad Place
23. South American Restaurant
24. Spanish Restaurant
25. Speakeasy
26. Sports Bar
27. Steakhouse
28. Sushi Restaurant
29. Tapas Restaurant

In the follow figure, we can see how the above venues categories are repeated in the result data:

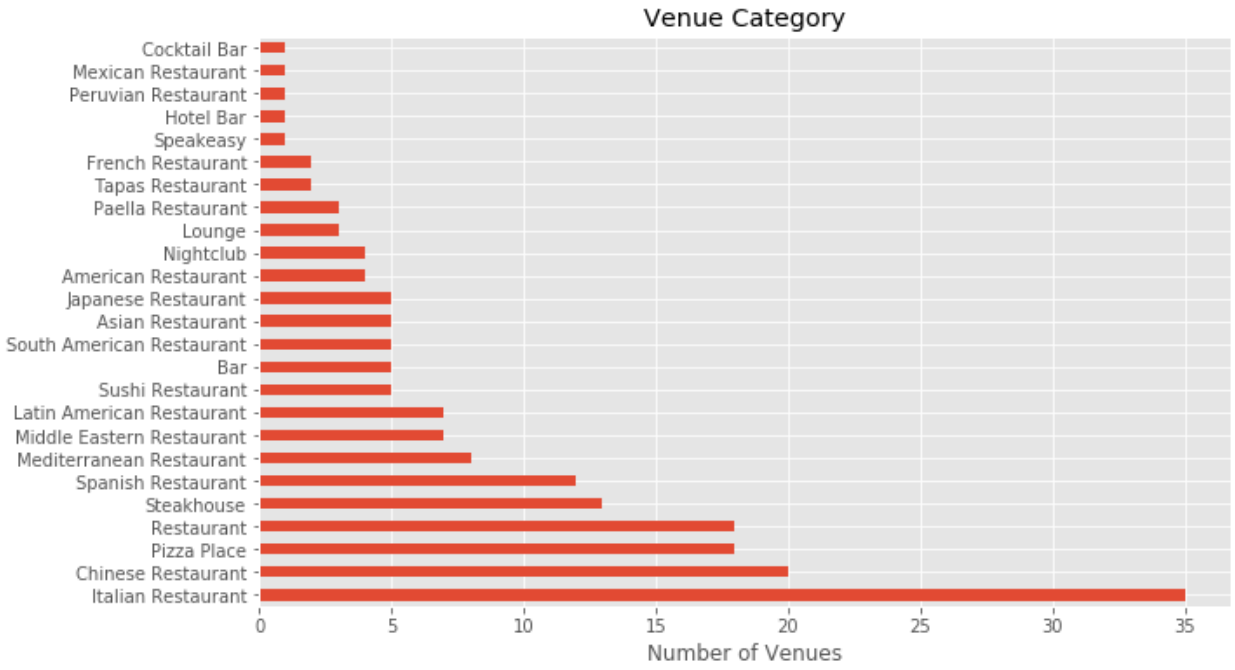


Figure 6. Venues categories for Bar-Restaurant.

Figure 7. Bar Restaurant Venues in the City.

3.2 Part 2: Analyzing Each Neighborhood

The k-means algorithm isn't directly applicable to categorical variables because Euclidean distance function isn't really meaningful for discrete variables. Using pandas.get_dummies, we have convert categorical variable into dummy/indicator variables. So, for each neighborhood, we have obtain the indicator variable for each venues category contains in that neighborhood. This process help to group row by neighborhoods and taking the mean of the frequency of occurrence of each venue category, and most common venue. It also help to run the clustering (figure 8).

```
----Baruta----
                venue  freq
0   Italian Restaurant  0.43
1         Pizza Place  0.29
2  American Restaurant  0.14
3     Tapas Restaurant  0.14
4         Cocktail Bar  0.00


----Carrizal----
                      venue  freq
0                Steakhouse   1.0
1       American Restaurant   0.0
2  Middle Eastern Restaurant   0.0
3          Sushi Restaurant   0.0
4                 Speakeasy   0.0
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue
0	Baruta	Italian Restaurant	Pizza Place
1	Carrizal	Steakhouse	Tapas Restaurant
2	Chacao	Italian Restaurant	Pizza Place
3	Las Mercedes	Italian Restaurant	Steakhouse
4	Parroquia Altagracia	South American Restaurant	Italian Restaurant
5	Parroquia Candelaria	Italian Restaurant	Spanish Restaurant
6	Parroquia Catedral	Chinese Restaurant	South American Restaurant
7	Parroquia Coche	Chinese Restaurant	Tapas Restaurant
8	Parroquia El Paraíso	Chinese Restaurant	Italian Restaurant
9	Parroquia El Recreo	Steakhouse	Spanish Restaurant
10	Parroquia El Valle	Chinese Restaurant	Tapas Restaurant

Figure 8

3.3 Part 3: Clustering neighborhoods.

Our Dataset 2, can be viewed as an unlabeled data, and we want to classify venues in different groups or clusters, in order to observe the most common Bar-Restaurants by neighborhoods and make this analysis easier. So, we have use **sklearn.cluster – Kmean Clustering** [4] methods to obtain clusters of Bar-Restaurant venues.

The input data was the Dataframe containing all Bar-Restaurant grouped by neighborhood (figure 9). As explain in Part 2, categorical variable 'Neighborhood' has been dropped from dataframe, so we can run the k-means algorithm. The k-means clustering, partitioned the venues in groups according to the selected number of clusters. The venues in each grouped have characteristics in common.

	Neighborhood	American Restaurant	Asian Restaurant	Bar	Chinese Restaurant	Cocktail Bar	French Restaurant	Hotel Bar	Italian Restaurant	Japanese Restaurant	...
0	Baruta	0.142857	0.000000	0.000000	0.000000	0.00000	0.000000	0.00000	0.428571	0.000000	...
1	Carrizal	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.00000	0.000000	0.000000	...
2	Chacao	0.071429	0.047619	0.047619	0.023810	0.02381	0.023810	0.02381	0.142857	0.047619	...
3	Las Mercedes	0.000000	0.046512	0.069767	0.023256	0.00000	0.023256	0.00000	0.186047	0.046512	...
4	Parroquia Altagracia	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.00000	0.333333	0.000000	...

Figure 9.

4. Results

Using K-means algorithm, we have obtain the venues clustering for Cararacas City. We have selected 6 as the number of cluster. We obtained that clusters labeled as '0', contain 135 venues, representing 72.6% of total clusters. Figure 10 shows the quantity of venues (or Bar-Restaurants) in each cluster label and its proportion (in pct).

	Cluster Labels
0	135
6	23
4	20
2	5
5	1
3	1
1	1

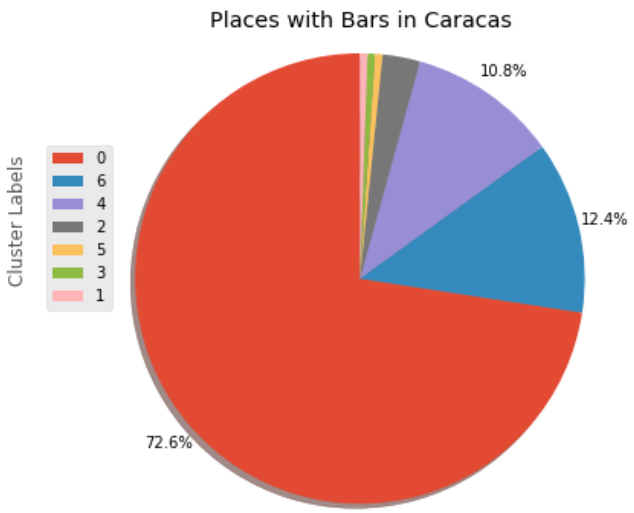


Figure 10. Cluster results from k-means algorithm.

The figure above, shows that clusters 5, 3 and 1 contains only one venues, so, we can group them in one single category. In the figure below, we can see the clusters distribution throughout the city.

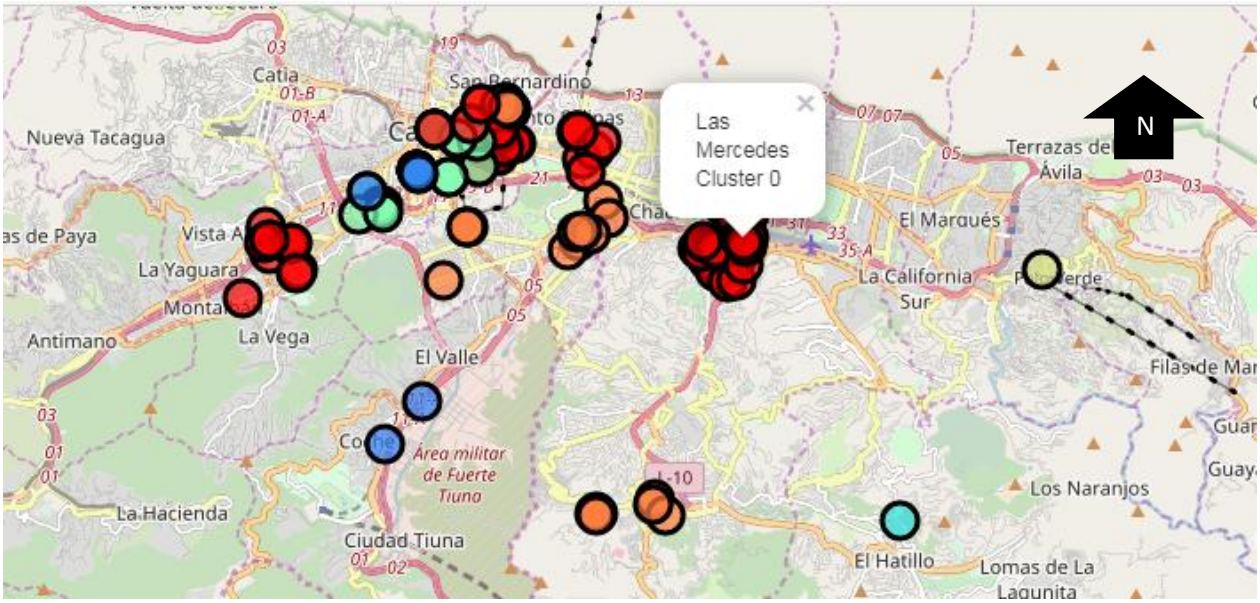


Figure 11. Clusters Distribution in the city.

For each cluster, we obtained a dataframe with the neighborhoods name and the 1st to 2nd common venues, as shown in the figure below. So, in order to find the best place to locate a restaurant, we explore the most common places by cluster, and the neighborhoods that contain them. We also wanted to know which are the least common places, in order to evaluate the competition.

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
181	Las Mercedes	0	Italian Restaurant	Steakhouse	Mediterranean Restaurant	Sushi Restaurant	Spanish Restaurant
182	Las Mercedes	0	Italian Restaurant	Steakhouse	Mediterranean Restaurant	Sushi Restaurant	Spanish Restaurant
183	Las Mercedes	0	Italian Restaurant	Steakhouse	Mediterranean Restaurant	Sushi Restaurant	Spanish Restaurant
184	Las Mercedes	0	Italian Restaurant	Steakhouse	Mediterranean Restaurant	Sushi Restaurant	Spanish Restaurant
185	Las Mercedes	0	Italian Restaurant	Steakhouse	Mediterranean Restaurant	Sushi Restaurant	Spanish Restaurant

Figure 12 Cluster '0' Neigborhoods and common venues.

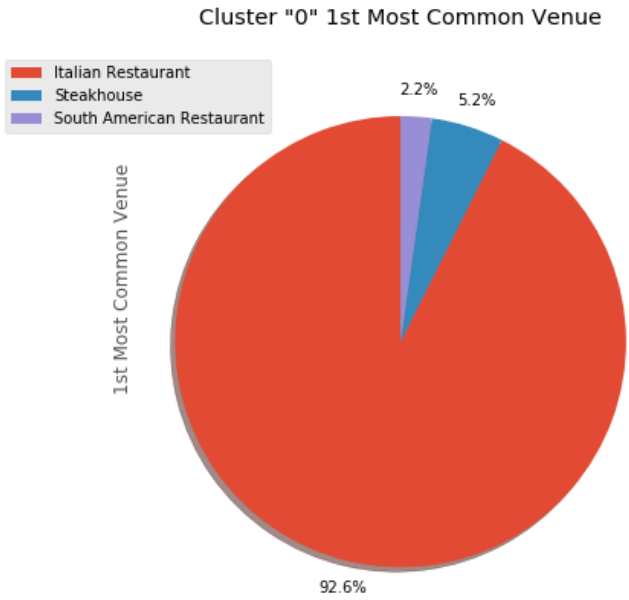
4.1 Cluster '0'.

We have found that "Italian Restaurants" are the most common Bar-Restaurant venues in the city, labeled as Cluster '0'. Has shown below, this represents the 92.6% of the most common venue. Steakhouse and South American Restaurant follows, but they only represents 5.2% and 2.2% respectively. Exploring the 2nd most common venues, we have found that Steakhouse, Pizza Place and Spanish Restaurant also represents a large group of venues labeled with these cluster. The least common venues in this cluster are Steakhouse and Tapas Restaurant. Most of these venues are located in "Las Mercedes", "Chacao", and "La Candelaria" neighborhoods.

	1st Most Common Venue
Italian Restaurant	125
Steakhouse	7
South American Restaurant	3

	2nd Most Common Venue
Steakhouse	43
Pizza Place	42
Spanish Restaurant	37
Latin American Restaurant	10
Italian Restaurant	3

	5th Most Common Venue
Mediterranean Restaurant	45
Spanish Restaurant	43
Chinese Restaurant	30
Pizza Place	10
Tapas Restaurant	7



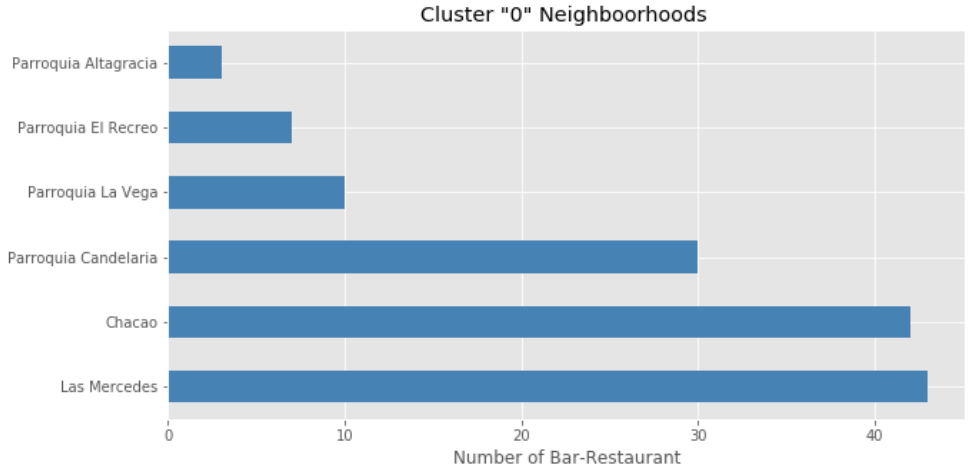


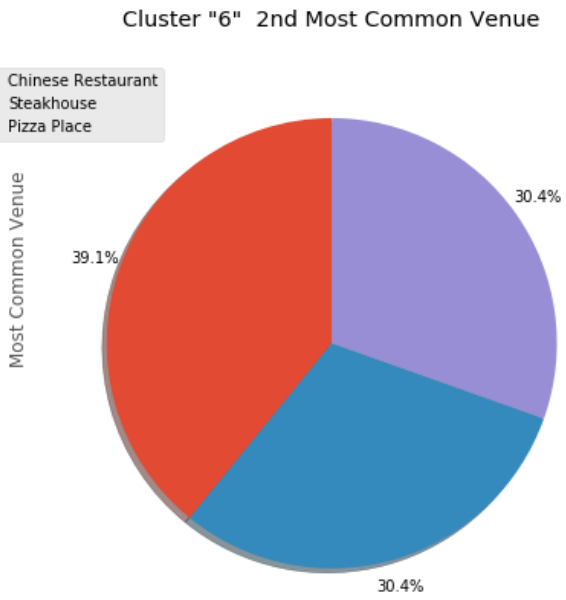
Figure 13. Cluster '0' Results.

4.2 Cluster '6'

This is the second important cluster, containing 23 venues. The Most common, again, is the Italian Restaurant together with Pizza Place. So, we have explored the second most common place in order to know the other venues we can frequently find with this label. For example, Chinese Restaurant has an important presence in the neighborhoods containing cluster '6' label. These neighborhoods San Pedro, Baruta, San Bernardino and Santa Rosalía.

Most Common Venue	
Italian Restaurant	12
Pizza Place	11

2nd Most Common Venue	
Chinese Restaurant	9
Steakhouse	7
Pizza Place	7



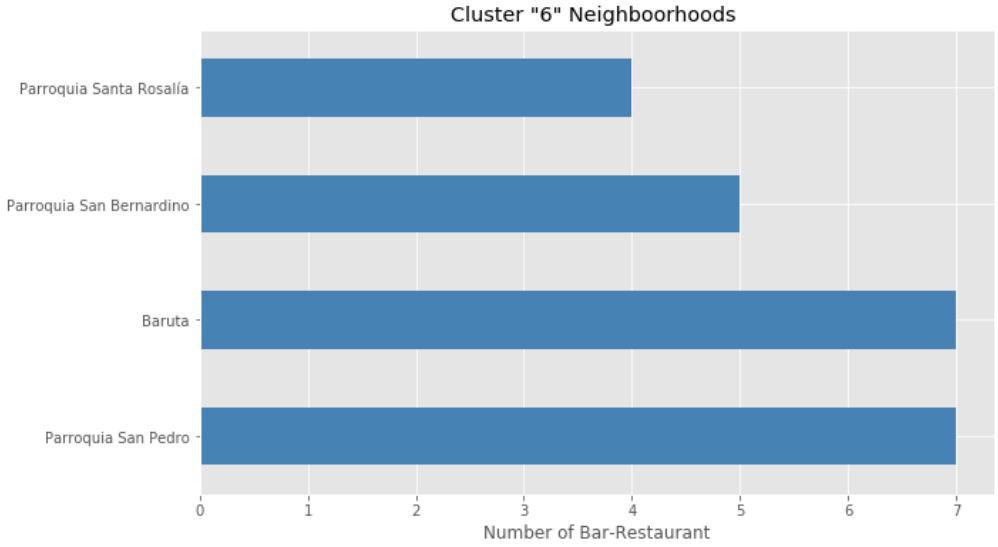
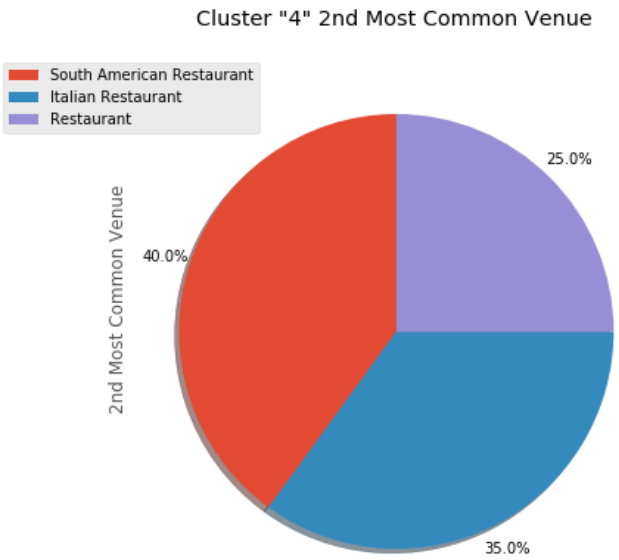


Figure 14. Cluster '6' Results.

4.4 Cluster '4'

The most common places with '4' label cluster, is the Chinese Restaurant, with a total of 20 venues in the neighborhoods containing this label. The second most common places are South American Restaurant, Italian Restaurant and 'generic' restaurants. These Bar-Restaurants are located in El Paraiso, San Agustin, Catedral and Santa Teresa neighborhoods.

	2nd Most Common Venue
South American Restaurant	8
Italian Restaurant	7
Restaurant	5



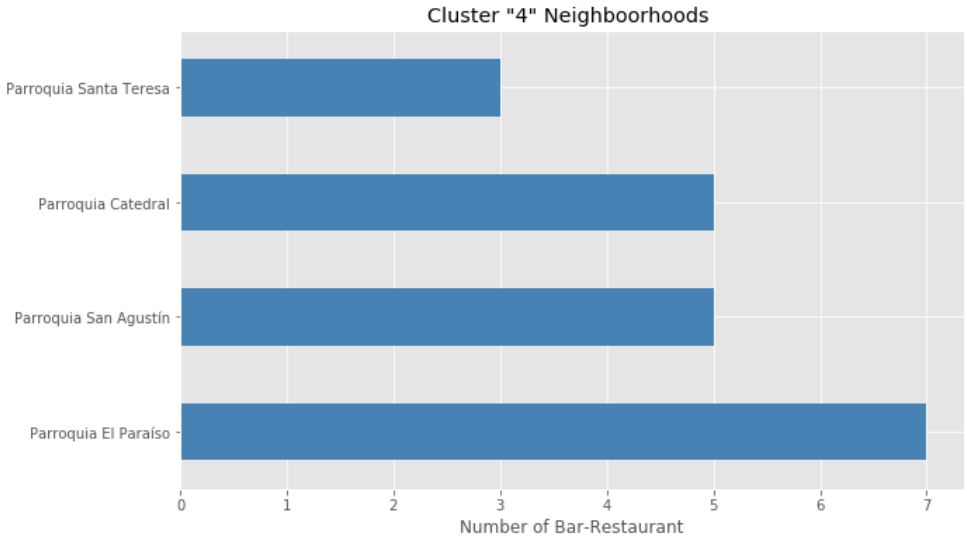


Figure 15. Cluster '4' Results

4.5 Clusters 1, 2, 3 and 5

The rest of the clusters contain very few venues. But they are interesting to explore given the little competition that these places may have. We have observed that Petare, El Hatillo and Carrizal neighborhoods have Asian restaurants and Bars as the least common places in the city.

Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Petare	5	Pizza Place	Tapas Restaurant	Mediterranean Restaurant	Asian Restaurant	Bar
Pueblo el Hatillo	3	Middle Eastern Restaurant	Tapas Restaurant	Mediterranean Restaurant	Asian Restaurant	Bar
Parroquia San Juan	2	Chinese Restaurant	Tapas Restaurant	Mediterranean Restaurant	Asian Restaurant	Bar
Parroquia San Juan	2	Chinese Restaurant	Tapas Restaurant	Mediterranean Restaurant	Asian Restaurant	Bar
Parroquia San Juan	2	Chinese Restaurant	Tapas Restaurant	Mediterranean Restaurant	Asian Restaurant	Bar
Parroquia El Valle	2	Chinese Restaurant	Tapas Restaurant	Mediterranean Restaurant	Asian Restaurant	Bar
Parroquia Coche	2	Chinese Restaurant	Tapas Restaurant	Mediterranean Restaurant	Asian Restaurant	Bar
Carrizal	1	Steakhouse	Tapas Restaurant	Mediterranean Restaurant	Asian Restaurant	Bar

Figure 16. Cluster 1, 3 and 5 results.

5. Discussion

The results shows that Italian Restaurant, Pizza Place, Steakhouse, Spanish Restaurant, Mediterranean Restaurant and Chine Restaurant are the most common venues or Bar-Restaurant with high presence in the city. These venues were labeled mostly as cluster '0', and are located in the neighborhoods that are towards the city center and near the business centers (Chacao, Las Mercedes and Candelaria neighborhoods). In contrast, we have observed that Steakhouse and Tapas Restaurant are the least common venues near the city and business center. We recommend evaluating these neighborhoods more, since they are in places near the city and the business center. These places have high competition in the restaurants grouped as the 1st most common venues, but with few restaurants grouped among the 5th most common venues.

The others clusters, had located the places around the city. And they also show that Italian Restaurant, Chinese Restaurant and Pizza Places are common. Although Tapas Restaurant, Asian Restaurant and Bars are also common. This result may be due to the few places assigned in these clusters. But they are interesting to explore given the little competition that these places may have.

It would be interesting to combine these results with demographic studies, surveys on gastronomy in the city, and with data showing the location of the main distributors of food and liquors as well. It is also possible to improve the clustering's precision, placing more geographical points in strategic places to obtain more venues from Foursquare, for example east of the city.

6. Conclusion

We carried out a study about the most common venues in the city of Caracas, and grouping them using the k-means algorithm, assuming that venues can be viewed as an unlabeled data. With the foursquare tool we were able to obtain the location of all the Bar-Restaurants in the city, which made the clustering work more effective.

We have found that most of the restaurants near the business center are Italian Restaurant, Pizza Place, Steakhouse, Spanish Restaurant, Mediterranean Restaurant and Chine Restaurants, which are mostly found in Chacao, Las Mercedes and Candelaria neighborhoods. Therefore, we recommend locating a new bar-restaurant within these areas, as long as it is of a different type from the mentioned categories, as a Tape Restaurant.

Web References

[1] https://en.wikipedia.org/wiki/Caracas

[2] https://www.antipodas.net/coordenadaspais/venezuela/caracas.php

[3] https://foursquare.com/

[4] https://scikit-learn.org/stable/modules/clustering.html#k-means