

# **Choosing automatically the best addresses to start any business in Santiago de Cali, Colombia.**

Sebastián García Acosta

June 14, 2019.

## **1. Introduction.**

### **1.1 Background.**

Cali, one of the principal cities in Colombia, and the second cheapest cities in the world to live in (According to Forbes <http://bit.ly/Forbes-cali>) is also known as 'the capital of Salsa', one of the main reasons that justifies its fame. Therefore, it isn't strange the fact of having quite lots of venues throughout its streets, mainly related to dancing bars, nightclubs, discos, etc. These characteristics makes the city a promising set of places to those local and foreign stakeholders seeking for new investments and diversity within their portfolio. Cali, being a city with low housing prices and prosperity in public for multiple venues, converts itself in an attractive city for multiple types of investors.

### **1.2 Problem**

With all these characteristics, something comes in mind: it could be a good idea to put an establishment in this city. Nevertheless, before starting getting the hands dirty we have to study the location. And... what if we want to know which is the best location in this city to start any category of venue/establishment nearby other types of venues that we want to have around? i.e., what is the best way to have a proper environment according to any particular business surrounded by matching venues and lower competence?

### **1.3 Interest.**

Mainly local and foreign investors interested in start business in Cali, might be, therefore, strongly interested in knowing what are the key types and specific addresses that they should consider; saving time of market research and other time-consuming tasks.

## **2. Data acquisition and cleaning.**

### **2.1 Data sources**

In order to get the information about the different *boroughs* (it may sound weird, but we'll talk about this later) and its neighborhoods, i checked the [Cali government planation department website](#). In order to get the latitudes and longitudes for every borough and

neighborhood, i used the Geopy Python library. And finally, in order to get the corresponding venues for each neighborhood, i used the Foursquare API.

## 2.2 Data cleaning

Main data that was extracted from [Cali government planation department website](#) was presented in a excel spreadsheet format, which looked like this:

	A	B	C	D	E	F
3						31/12/2016
4			Código único de identificación por barrio			
5						
6	Código	Estrato	Barrio			Acuerdo
7	único	moda		No		dd-mm-año
8						
9			COMUNA 1			
10	0101	2	Terrón Colorado	049		28-08-1964
11	0102	1	Vista Hermosa	083		06-07-1966
12	0196	1	Sector Patio Bonito			
13	0199	1	Aguacatal			
14						
15			COMUNA 2			
16	0201	6	Santa Rita	049		28-08-1964
17	0202	6	Santa Teresita	049		28-08-1964
18	0203	6	Arboledas	049		28-08-1964
19	0204	6	Normandía	049		28-08-1964
20	0205	5	Juanambú	049		28-08-1964
21	0206	5	Centenario	049		28-08-1964
22	0207	4	Granada	049		28-08-1964
23	0208	5	Versalles	049		28-08-1964

Table 1. Spreadsheet downloaded from [Cali government planation department website](#)

This format was very incompatible for analysis and modelling purposes; hence, at the moment of import it to a data frame in pandas, I keep only two columns of the spreadsheet: *Estrato* (hereinafter, social stratum) and *Barrio* (hereinafter, neighborhood). Since the neighborhood column contained also the *borough* (hereinafter, borough) to which every neighborhood belongs, I separated them in different columns, ended up with two data frames: one categorized by neighborhoods, and other categorized by boroughs.

Now, our data set has size of: (328, 3)

	Stratum	Neighborhood	COMUNA
0	2	Terrón Colorado	1
1	1	Vista Hermosa	1
2	1	Sector Patio Bonito	1
3	1	Aguacatal	1
4	6	Santa Rita	2
5	6	Santa Teresita	2
6	6	Arboledas	2
7	6	Normandía	2
8	5	Juanambú	2
9	5	Centenario	2

Table 2 **Data frame organized by neighborhoods** after data cleaning (screenshot from notebook).

This merged dataframe has a size of: (22, 3)

	COMUNA	Median Stratum	Neighborhoods
0	1	1.0	Terrón Colorado, Vista Hermosa, Sector Patio B...
1	2	5.0	Santa Rita, Santa Teresita, Arboledas, Normand...
2	3	3.0	El Nacional, El Peñón, San Antonio, San Cayeta...
3	4	2.0	Jorge Isaacs, Santander, Porvenir, Las Delicia...
4	5	3.0	El Sena, Los Andes, Los Guayacanes, Chiminango...
5	6	2.0	San Luís, Jorge Eliecer Gaitán, Paso del Comer...
6	7	3.0	Alfonso López P. 1a. Etapa, Alfonso López P. 2...
7	8	3.0	Primitivo Crespo, Simón Bolívar, Saavedra Gali...
8	9	3.0	Alameda, Bretaña, Junín, Guayaquil, Aranjuez, ...
9	10	3.0	El Dorado, El Guabal, La Libertad, Santa Elena...

Table 3 **Data frame merged by boroughs** after data cleaning (screenshot from notebook).

Furthermore, having both data frames, I started the acquisition of location data for each of their categorized divisions, i.e., the latitude and longitude of each neighborhood, as well as each borough; fetching them with [Geopy](#) open source library. Finally, ended up with both datasets with this appearance:

This merged dataframe has a size of: (22, 5)

	COMUNA	Median Stratum	Neighborhoods	Latitude	Longitude
0	1	1.0	Terrón Colorado, Vista Hermosa, Sector Patio B...	3.45179	-76.5325
1	2	5.0	Santa Rita, Santa Teresita, Arboledas, Normand...	3.45179	-76.5325
2	3	3.0	El Nacional, El Peñón, San Antonio, San Cayeta...	3.45179	-76.5325
3	4	2.0	Jorge Isaacs, Santander, Porvenir, Las Delicia...	3.45179	-76.5325
4	5	3.0	El Sena, Los Andes, Los Guayacanes, Chiminango...	3.45179	-76.5325
5	6	2.0	San Luís, Jorge Eliecer Gaitán, Paso del Comer...	3.45179	-76.5325
6	7	3.0	Alfonso López P. 1a. Etapa, Alfonso López P. 2...	3.45179	-76.5325
7	8	3.0	Primitivo Crespo, Simón Bolívar, Saavedra Gali...	3.45179	-76.5325
8	9	3.0	Alameda, Bretaña, Junín, Guayaquil, Aranjuez, ...	3.45179	-76.5325
9	10	3.0	El Dorado, El Guabal, La Libertad, Santa Elena...	3.45179	-76.5325

Table 4. **Data frame organized by boroughs containing location data.** (Screenshot from notebook)

Finally, all 319 neighborhoods location data could be fetched.

	Stratum	Neighborhood	COMUNA	Latitude	Longitude
0	2	Terrón Colorado	1	3.4525	-76.5632
1	1	Vista Hermosa	1	3.4549	-76.5777
2	1	Sector Patio Bonito	1	3.46215	-76.5872
3	1	Aguacatal	1	3.45625	-76.571
4	6	Santa Rita	2	3.45147	-76.5536
5	6	Santa Teresita	2	3.45096	-76.5523
6	6	Arboledas	2	3.4492	-76.5488
7	6	Normandía	2	3.45377	-76.5443
8	5	Juanambú	2	3.45517	-76.5381
9	5	Centenario	2	3.45323	-76.5365

Table 4. Data frame organized by neighborhoods containing location data.  
(Screenshot from notebook)

At first were 32 neighborhoods whose locations could not be fetched; hence, additional cleaning techniques (which I will not explain in detail, due to its tediousness) had to be carried out in order to make their names recognizable to the Geopy library. Notice then, that the above data frame organized by neighborhoods with location data, unlike its previous version, contains 9 rows less (the previous data frame had 328, this have 319 rows); this reduction was presented owing to the fact that only 9 neighborhoods location could not be fetched by the library.

Now, having the location for each of the 319 neighborhoods, I proceeded to fetching their nearby venues within a kilometer of radius around them, and a limit of 200 venues for each neighborhood. 1413 venues were fetched, distributed in 213 different categories.

Now, we have 1415 uniques venues in our dataset.  
Now, there are 213 unique venues categories.

	Neighborhood	Venue	Venue Latitude	Venue Longitude	Venue Category	Stratum	COMUNA
0	Terrón Colorado	Hacienda Del Bosque	3.447992	-76.560816	South American Restaurant	2.0	1
1	Terrón Colorado	Zoológico de Cali	3.447884	-76.556862	Zoo	2.0	1
2	Terrón Colorado	Peces - Zoológico de Cali	3.448077	-76.557541	Zoo Exhibit	2.0	1
3	Terrón Colorado	Club Emcali	3.449028	-76.564105	Water Park	2.0	1
4	Vista Hermosa	angic's wings km 18	3.456583	-76.581650	BBQ Joint	1.0	1
5	Sector Patio Bonito	Industria DJARI	3.463099	-76.587590	Clothing Store	1.0	1
6	Aguacatal	Jardín Botánico de Cali	3.450302	-76.566403	Garden	1.0	1
7	Santa Rita	Cafe Valparaiso	3.453613	-76.549647	Deli / Bodega	6.0	2
8	Santa Rita	Bodytech Platino - Sede Oeste	3.451899	-76.547763	Gym / Fitness Center	6.0	2
9	Santa Rita	Super A	3.453384	-76.548186	Market	6.0	2

Table 5. Venues data frame (screenshot from Notebook).