How can I open a similar venue in my local economy like the most popular from the Downtown?

By Gustavo Israel Montenegro Vargas

Introduction

When I go to the Downtown (I´m living in Mexico by now), always I take a snack in some place. The Downtown always is crowded, maybe you need to wait for a table even in the small places. While I´m waiting I don´t avoid to think how much people arrives each minute. Like everybody I have my favorite venues, for instance, Gino´s pizza is one of them. The most delicious pizza that I´ve never ate before. Gino´s is a special venue, I had one of the better dates there. There are many branches of Gino's around the city and they always are crowded. I think that one branch of Gino's near of my house could be a good idea. But wait, there are not many pizza places near my home. Why? I don´t know, maybe it could be an opportunity or maybe it could be the worst business idea. What kind of venues does the people near my house like? Currently there are many popular brandings near my house like Smart Fit or Starbucks, but they are not crowded like in the Downtown. I need to think more about this. Why do they are not crowded? Maybe the neighborhoods around there are dangerous. It could be possible. But I want to be less pessimistic, maybe they don´t like those kinds of venues. So, while I wait, I like to think in the possibility to open a popular venue. Maybe like the pet shop on the Roma neighborhood or maybe like FCE Bookstore on Lazaro Cardenas avenue.

In the city there isn´t enough public information of the local economy of each Borough. How can I make a decision if I want to start a business? In this notebook I try to get useful data to get a fast opinion about this situation. I know that the information isn´t enough because the nearest businesses don´t have a public information about their sales or check-ins. I hope this kind of quickly analysis help to determine someone who want to open a busines like me. I think that in the Downtown there are a great variety of venues and one of them could be a great idea here. I will less blind if I can get relevant information from the users or the almost inexistent venues stats in my local economy.

Data description

'T' and 'C' Data Frames

I had to create the Data Frames with the location of each neighborhood. I took the Data with postal code from a map's web site. [1]. I used the 'Nominatim' library to get the coordinates. The address in the geocoder was not exactly, I needed to use the postal Code to get the coordinates. Then I saved the Data Frames in a CSV files. So, the Data Frames have

- Neighborhood
- Postal Code
- Latitude
- Longitude

The 'C' Data Frame has 41 rows.

The 'T' Data Frame has 17 rows.

These Data Frames don´t have NAN values. The 'T' Data Frame has the data from Tlahuac borough. The 'C' Data Frame has the Data from Cuauhtemoc borough. The Downtown is in the last one. My local economy is Tlahuac.

There is a big difference between the boroughs. I will explain it in the 'Discuss' section.

'CN' and 'TN' Data Frames

These Data Frames has the most popular neighborhoods of each borough. The CN Data Frame has extra neighborhoods than the Downtown because you can hear someone to speak about If you are local or not, It doesn´t matter. All the fancy venues in the Cuauhtemoc borough are there. The situation with the CN borough is different. How I live there, I could see how the neighborhoods have improved. The new brands have a small mall over the main avenue. These Data Frames have the same features that the C and T Data Frames.

[1] https://micodigopostal.org/ciudad-de-mexico/cuauhtemoc/

https://micodigopostal.org/ciudad-de-mexico/tlahuac/

The 'CN' Data Frame has 15 rows

	Neighborhood	Postal C	Latitude	Longitude
0	Buenavista	6350	19.446167	-99.152696
1	Centro (Área 1)	6000	19.432700	-99.137600
2	Centro (Área 2)	6010	19.439600	-99.136600
3	Centro (Área 3)	6020	19.438400	-99.128500
4	Centro (Área 4)	6040	19.430900	-99.149900
5	Centro (Área 5)	6050	19.433700	-99.144300
6	Centro (Área 6)	6060	19.431700	-99.129500
7	Centro (Área 7)	6070	19.428900	-99.145200
8	Centro (Área 8)	6080	19.426000	-99.138800
9	Condesa	6140	19.414864	-99.176429
10	Hipódromo	6100	19.409371	-99.171213
11	Hipódromo C.*	6170	19.409266	-99.179582
12	Roma Norte	6700	19.418323	-99.162565
13	Roma Sur	6760	19.405833	-99.163304
14	Santa María R*	6400	19.448417	-99.157975

* Hipódromo Condesa, Santa María la Ribera.

The 'TN' Data Frame has 10 rows

	Neighborhood	Postal C	Latitude	Longitude
0	Agrícola Met*.	13280	19.290832	-99.052493
1	La Nopalera	13220	19.298380	-99.052562
2	La Turba	13250	19.320556	-99.151701
3	Los Olivos	13210	19.320556	-99.151701
4	Santa Ana	13060	19.267541	-99.008578
5	Santa Ana C.*	13300	19.301105	-99.036331
6	Santa Ana N.*	13300	19.301105	-99.036331
7	Santa Ana P.*	13300	19.301105	-99.036331
8	Santa Ana Sur	13360	19.284663	-99.030658
9	Zapotitla	13310	19.306475	-99.041657

* Agrícola Metropolitana, Santa Ana Centro, Santa Ana Norte, Santa Ana Poniente.

'CNV_ byRating' Data Frame

I got the data from Foursquare here. I used Foursquare to get the ten most popular venues of each neighborhood from CN Data Frame with a radius of 500 meters. I used the 'explore' endpoint to make a regular call. This kind of call gave me the recommended venues and the IDs. The I used the 'details' endpoint to get the 'rating' and 'main categorie' features of each venue. It really was a challenge for me. But now I know how to get the Foursquare information from a JSON file. This Data Frame is sorted by rating. The radius was very low because the neighborhoods are not to big. I could be sure that I didn´t have dup licate rows. Here is the head of the Data Frame. The Data Frame Features are ID, Name, Rating and Main Categorie.

	ID	Name	Rating	Main Categorie
0	4baec3d5f964a52018d63be3	Nevería Roxy	9.4	Ice Cream Shop
1	4b684e6ef964a520d0702be3	Museo de Arte Popular	9.4	Art Museum
2	4b819a4cf964a520fcb130e3	Fondo de Cultura Económica Rosario Castellanos	9.3	Bookstore
3	50ca02bf245f2d4aa8c2aeef	El Cardenal	9.3	Mexican Restaurant
4	51ba6ba9498e9bcd16eee9ec	Cachito Mío Quiches & Tartas	9.2	Coffee Shop

'TN_VenuesUC' Data Frame

I got the data from Foursquare here. I used Foursquare to get the ten most popular venues of each neighborhood from TN Data Frame with a radius of 4000 meters. Many of these neighborhoods don´t have near venues on Foursquare. I had to give a higher value of radius to get the data. This evaluation had 50 venues of each neighborhood, but there were many duplicate values. So, I had to drop the duplicate values. Then I got the rating, likes, dislikes features with the 'details' endpoint. Of course, the ID too.

I wanted to use the stats field from the JSON file, but many of these neighborhoods don´t have the stats field. Only the 'Tip Counts' field. Also, I added the 'Tip Counts'. The Data Frame has the Main Categorie feature too. Here is the head. The Data Frame is not sorted.

	ID	Name	Main Categorie	Latitude	Longitude	dislikes	Tip_Counts	Rating	likes
0	4cae778aeb65b1f760595ccd	Los Taquitos	Taco Place	19.295809	-99.05543	False	37	8.8	102
1	4eaadaf4b803cf1ffb47a0d8	Carnitas La Flor de Cotija	Mexican Restaurant	19.294897	-99.05770	False	13	8.5	32
2	50521cfbe4b02dee62ca3635	Carnitas Estilo Michoacán El Paisa	Taco Place	19.287660	-99.05095	False	10	8.2	11
3	4f20c6d4e4b08ed6b4ca47a2	Don Cuco	Taco Place	19.291856	-99.05517	False	24	8.1	68
4	4f6fe82ee4b07e41f09878a5	Sushisimo	Sushi Restaurant	19.297009	-99.05433	False	16	8.0	35

You can see the code in my GitHub repository

https://github.com/GustavoMontenegroVargas/Coursera_Capstone/blob/master/Final%20assignment%20.ipynb

Methodology

By the moment I want to go back to my first point, when I´m waiting at the restaurant. That´s good, I found some data about the recommended venues in the Cuauhtemoc Borough of each neighborhood. The same for the Tlahuac borough. But, in fact, it seems like something isn't very good. The venues from the Tlahuac borough have few data than I expected. The amount of likes and the Tip counts are not enough. So, maybe the rating isn´t the best idea like a parameter because I don´t have the number of check-ins. I can think in another parameter: the number of venues, I mean, the most common kind of venue in the region. In this way I can choose a place with a good rating from the Cuauhtémoc borough in order to open it in Tlahuac Borough. I need to know where exist an opportunity area.

With the 'TN_VenuesUC' Data Frame

- Group a new Data Frame by 'main categorie' and apply the sum() function. (You need to create a new Data Frame with the main categorie feature and a column with one's)
- Group a new Data Frame by ''main categorie' and apply the mean() function. (You need to create a new Data Frame with the main categorie and rating features)
- Merge the Data Frames on the 'main categorie' column.
- You can sort the result Data Frame by rating or by the column with one's.

Here is the result (head())

By rating

Main Categorie	Rating	Number of Venues
Pet Store	9.200000	1
Sports Club	9.000000	1
Art Museum	8.900000	1
College Gym	8.800000	1
Market	8.800000	1

By number of venues

Main Categorie	Rating	Number of Venues
Mexican Restaurant	7.957143	28
Taco Place	7.961905	21
Seafood Restaurant	7.655556	9
Pizza Place	7.800000	6
Ice Cream Shop	8.233333	6

Let's choose the Mexican Restaurant to make a quick analysis.

The 'Mexican Restaurant' categorie has the highest number of venues but its rating is not very good. There is a Mexican restaurant in the most popular venues (rating higher than 9 in the 'CNV_byRating' Data Frame) from the Cuauhtémoc borough.

50ca02bf245f2d4aa8c2aeef	El Cardenal	9.3	Mexican Restaurant

- Where are the better Mexican restaurants in the Tlahuac borough?
- Can I make a rating prediction with a Multiple linear regression model with only two features (likes and tips)?

I will explain these questions in the Discussion section.

You can see the code in my GitHub repository
https://github.com/GustavoMontenegroVargas/Coursera_Capstone/blob/master/Final%20assignment%20.ipynb

- Now, I must create a Data Frame with the Mexican restaurants.
- Then, I need to sort the Data Frame by rating.

I called it MR. The sorted Data Frame with the Mexican restaurants data.

Here are the details of the numerical features of the MR Data Frame.

	Latitude	Longitude	Tip_Counts	Rating	likes
count	28.000000	28.000000	28.000000	28.000000	28.000000
mean	19.296607	-99.070497	59.214286	7.957143	193.928571
std	0.018850	0.063055	151.556213	0.626825	520.827613
min	19.259260	-99.167681	0.000000	6.500000	5.000000
25%	19.288848	-99.144074	3.000000	7.500000	13.000000
50%	19.296633	-99.049559	8.500000	7.950000	21.500000
75%	19.311836	-99.014284	22.250000	8.500000	56.000000
max	19.338141	-99.004887	774.000000	9.100000	2670.000000

I don´t have NAN values in the Data Frame. Actually, I´m really interested in the rating with this Data Frame.

- Let´s create a map with the rating distribution.

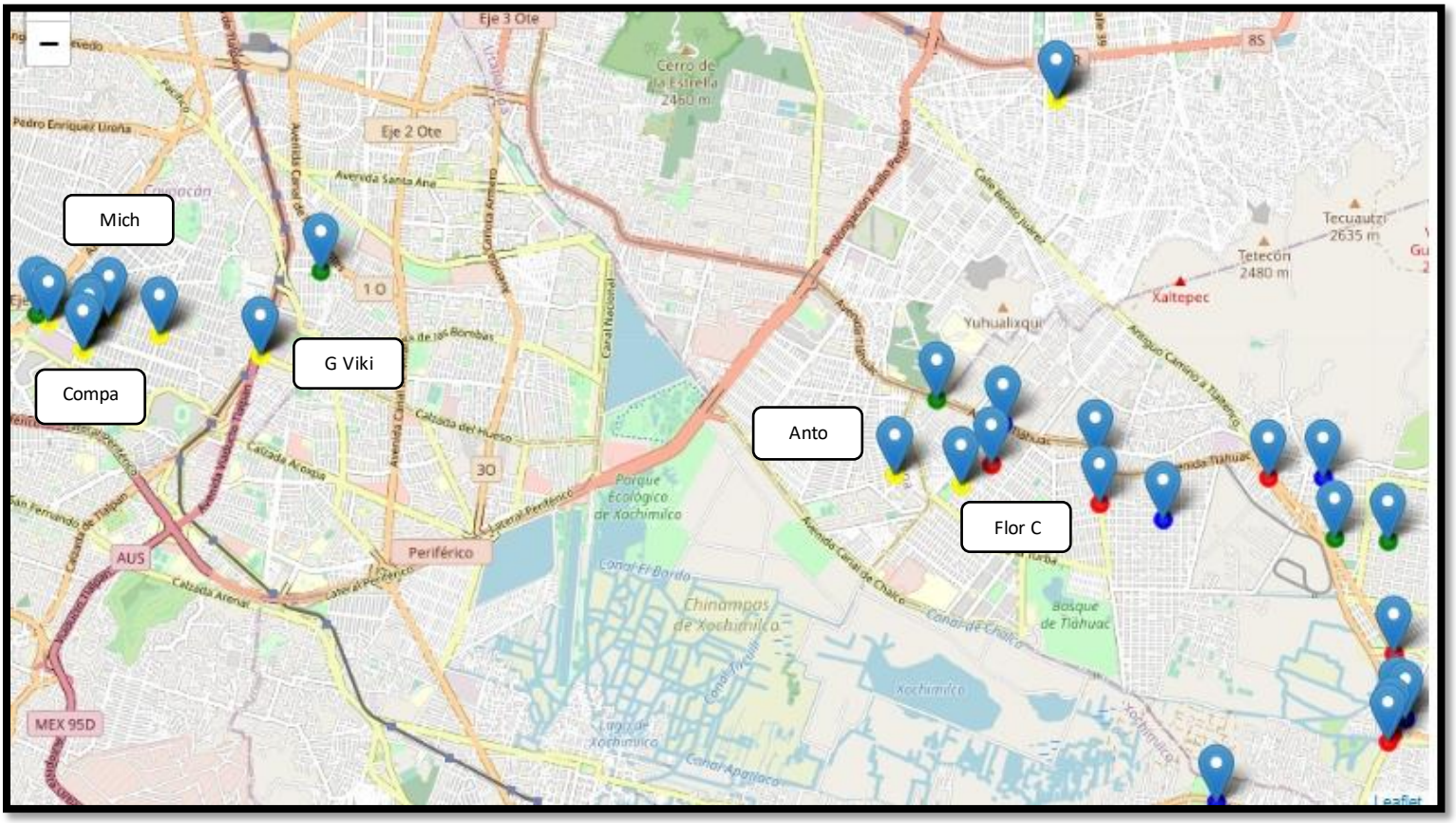
Before to show the map, I need to explain the map. I added a color marks and 'name' labels. The color marks are between the percentage. The red marks are in the lower rating and the yellow marks are in the higher rating.

Now, I can expect venues very far from Tlahuac Borough. If you remember, I said that the radius in the Tlahuac borough calls was higher than the Cuauhtémoc borough. I will discuss this later. Here are almost all the venues.

0	Michoacaníssimo	9.1
1	Gorditas "La Vikina"	8.9
2	Taqueria pepe	8.6
3	Pozole Rey Papatzin	8.6
4	Barbacoa El Compadre	8.6
5	Pozoleria La Fuente	8.6
6	Antojitos Don Rafa	8.5
7	Taquería La Leyenda	8.5
8	Carnitas La Flor de Cotija	8.5

9	Michoacanísimo Birria de Chivo	8.3
10	El Tacuche Arabe	8.2
11	El Rincón de Peribán	8.2
12	Las Cazuelas Don Jose	8.0
13	Pozoleria "Flor de Maíz"	8.0
14	Huaraches Pablo	7.9
15	Caballo Azteca	7.9
16	La Oaxaqueña	7.7
17	Tortas San Lorenzo	7.7

18	Mercado De Tlahuac	7.6
19	Carnitas Y Barbacoa "Hermanos R."	7.6
20	El Típico	7.5
21	Mercado de Tlaltenco	7.5
22	Taqueria Carnitas Chela	7.4
23	las tres lolas	7.4
24	Los Sopes de Chelita y sus muchachas	7.4
25	Restaurante Bar La Playa	7.3
26	El puente	6.8



Finally, I tried to make a rating predictor model with Multiple linear regression and the numerical features. It didn´t works . The value of accuracy is very low: 0.24. Maybe I could fix it with another kind of machine learning algorithm, but it will be later with more data like price, check-ins and so on.

You can see the code in my GitHub repository
https://github.com/GustavoMontenegroVargas/Coursera_Capstone/blob/master/Final%20assignment%20.ipynb

Discussion about Mexican Restaurants

Now, after the methodology I can explain the results. First, there are a lot of differences between the Tlahuac borough and the Cuauhtemoc borough. In Mexico City the boroughs that are nearest from the Downtown have a better economy. So, it avoids to open better venues. I can see this on the map.

There are two regions. The left region has the most popular venues in the Mexican Restaurant categorie. The right region has the worst. Only two good places exist in the right side. This situation is very interesting. In fact, it could be a good opportunity. I can get more information about this kind of venue. The people maybe could be tired of the same kind of food. There is the most common kind of venue too. We need to investigate more about the Cardenal restaurant. About its way to work. I could review the tips in order to know what the people wants. I know now that the people who rate these places prefer the Mexican venues without the Tlahuac borough. The venues in the left side in fact are very far from the Tlahuac Borough.

There is not enough amount of likes to make a good predictor model. This feature is a very important factor that the venues in Tlahuac are not using. It seems Foursquare isn´t very popular near the Tlahuac Borough.

It seems if you open a similar restaurant like the Cardenal restaurant in the Tlahuac borough and you try to connect your venue with the main venue browsers you can earn the attention of the persons who are not happy with restaurants near the borough.

Conclusions about report

The notebook and the report are very useful. You can know how to start If you want to open a venue near the Tlahuac borough.

I know, there are many others features that you need to consider, but I didn´t know that the people prefer the Mexican Restaurants far from Tlahuac borough. That´s an important thing.

Many venues are not registered in Foursquare. Maybe with other site with API I can get more detail information in certain region. Also, many venues in Tlahuac borough don´t use the social media ways to get better promotion.

There are many categories in the Data Frame that only have a one venue. That venue could have a higher rating. We need to check the tips. The check-ins are very important. Maybe the rating could be high, but if you don´t have enough registers, that information could be not very relevant.

A good idea of business is to offer social media services in the Tlahuac borough.

Tlahuac doesn´t have variety. It´s enough with see the Data Frames. The most common kind venue is food.

If a categorie has one venue and that venue has a high rating, you should investigate more about it even there are not enough information about the check-ins.

Gustavo Israel Montenegro Vargas

01/08/2020