ANALYSIS OF DESIRABLE NEIGHBORHOODS IN LATIN-AMERICAN CITIES FOR EXPANSION

ANDRES REINOSO

IBM Data Science Applied Capstone Project - April 2020

1. Introduction:

A renowed Colombian food chain firm is analyzing a business expansion, this includes opening 4 new restaurants in the latinamerican capitals before 2020 ends. Covid19 effects on the food business may offer good opportunities for new participants entering the market. The venue selection for the new restaurants plays a major role in order to maximize profits and mitigate possible deployment risks, for this purpose the first stage of selection will be locating neighborhoods with the closests conditions to those restaurants already in operation locally.

2. Problem:

Perform a pre-selection of the potential neighborhoods to expand the restaurants of the firm, in four latin american cities: Rio de Janeiro, Lima, Montevideo and Buenos Aires. Pre-selection execution time cannot exceed 2 weeks.

3. Interest:

Start operations in new countries taking as an advantage the current opportunities created by Covid19 into the food chain market.

4. Data acquisition and cleaning:

In regards to the data collection process, all information was downloaded from the wikipedia website: https://es.wikipedia.org/wiki/Anexo:Barrios_de_Bogot%C3%A1, here you can find a sample for the Bogota neighborhoods, same process was executed to all 4 latin american capitals involved in this analysis. Neighborhood lists were very complete, and mostly with good quality, however in the case of Bogota, the amount of neighborhoods was overwhelming in comparison with the shortest lists for all the other capitals. Bogota lists required much more processing time. Data was saved into separated dataframes for each city, webscraping was used to get the data into the notebook, specifically BS tool.

Next step corresponds to geographical coordinates, those were found by using GEOPY several iterations were required, also limitations in terms of amount of calls were found and solved during the process.

```
# Transforming the tuple into a df
df_coordinates=pd.DataFrame(list(tuple_results),columns=["Neighborhood","Latitude","Longitude"])
df_coordinates.head(5) # Checking the values for the first 3 items
```

	Neighborhood	Latitude	Longitude
0	Canaima	4.7033	-74.092216
1	La Floresta de La Sabana	4.80953	-74.022607
2	Torca	4.73073	-74.031314
3	Altos de Serrezuela	Not available	NaN
4	Balcones de Vista Hermosa	Not available	NaN

Figure 1. Dataframe example.

A significant number of neighborhoods coordinates were missing, especially in the case of Bogota. Which were removed. Individual dataframes with the coordinates were generated at this step. Also, maps were generated in order to check if the neighborhood's coordinates were populated correctly.

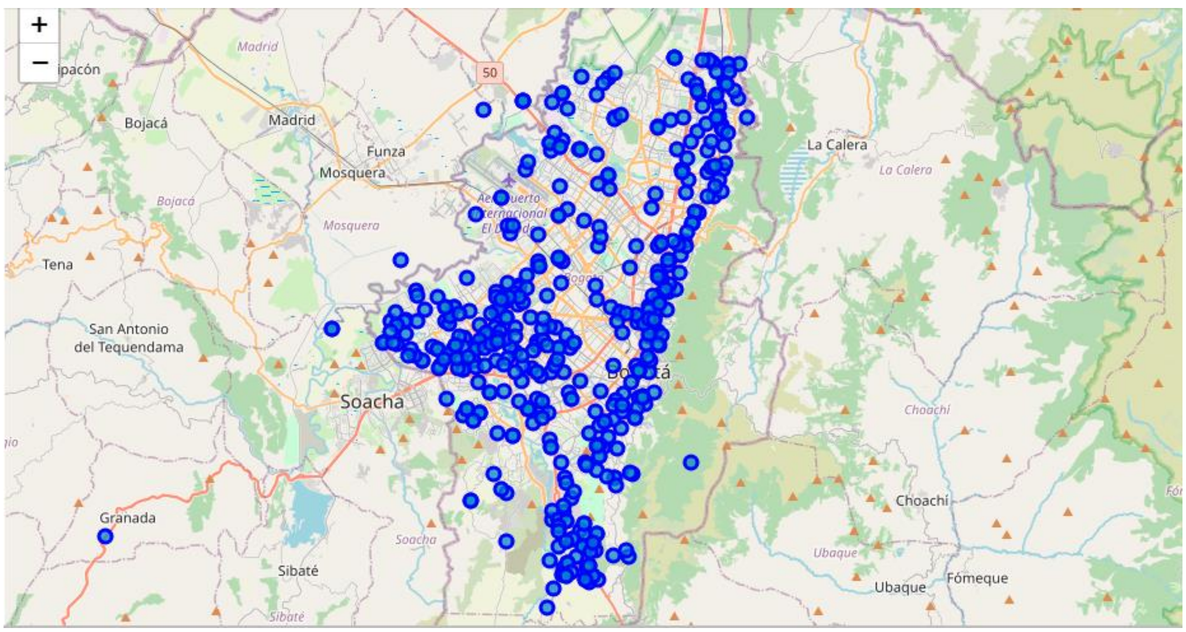

Figure 2. Finding the neighborhoods.

Finally, for the case of Bogota, additional filtering was done in order to leave only those neighborhoods were does the firm of study has restaurants in operation, reducing further our initial dataset.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Canaima	4.703298	-74.092216	Mac charlie	4.703853	-74.092535	Latin American Restaurant
1	Canaima	4.703298	-74.092216	arepas el carriel	4.701671	-74.089948	Arepa Restaurant
2	Canaima	4.703298	-74.092216	Ci Divisiones	4.703741	-74.093791	Furniture / Home Store
3	Canaima	4.703298	-74.092216	Tazz Factory Comidas Rapidas	4.704175	-74.092340	Burger Joint
4	Canaima	4.703298	-74.092216	GYM POWER ZONE	4.701110	-74.092116	Gym

```
Venues_results.to_csv('BOGVEN.csv',index=False)
```

Now we can select the neighborhoods in Colombia where does the firm has operative restaurants in order to compare and cluster specifically those ones.

```
Venues_results_Restaurant=Venues_results[Venues_results['Venue']=='Crepes & Waffles']
```

Figure 3. Foursquare search to allocate current restaurants.

As all datasets for each city was generated separately, an additional merging step was needed to start working on the data:

```
# Now let's merge the different dataframes into a single total one.
df_total = df_rio.append([df_baires,df_lima,df_montev,df_bogota])
df_total.tail() # checking if the dataset is complete and working!
```

	Neighborhood	Latitude	Longitude
43	El Motorista	4.636931	-74.098484
44	Bombay	4.657891	-74.060228
45	Las Torres	4.619145	-74.084762
46	Casa Loma	4.631093	-74.070698
47	Perpetuo Socorro	4.612864	-74.065908

Figure 4. Merging datasets.

Once ready the full list of neighborhoods, the query to determine the most common venues for every neighborhood into our dataset was done, then dataframe was sorted and selected the 10 most common places in order to organize our clusters based on these criteria.

	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	Abayubá	Gym / Fitness Center	Plaza	Food & Drink Shop	Bus Stop	Mobile Phone Shop	Factory	Basketball Court	Bakery	Pizza Place	Stadium
1	Abolição	BBQ Joint	Deli / Bodega	Burger Joint	Bus Station	Portuguese Restaurant	Gymnastics Gym	Food Truck	Frame Store	Steakhouse	Snack Place
2	Acari	Market	Ice Cream Shop	Churrascaria	Soccer Field	Pizza Place	Film Studio	Event Space	Exhibit	Fabric Shop	Factory
3	Agronomía	Bus Stop	BBQ Joint	Farmers Market	Plaza	Garden Center	Burger Joint	Trail	Athletics & Sports	Tunnel	Fish & Chips Shop
4	Aguada	Bakery	Nightclub	Train Station	Food & Drink Shop	Fast Food Restaurant	Convenience Store	Pizza Place	Ice Cream Shop	Rental Car Location	Event Space

Figure 5. Foursquare results.

K-means algorithm was applied to our latest generated dataset, this method was used due it's simple implementation and good results. Based on this we obtained 5 clusters with all the features provided by foursquare.

Results were checked, finding all neighborhoods were does our company of study has restaurants open, were grouped on the 2nd cluster, so all data from this cluster was filtered.

Finally, a list with all potential neighborhoods with the same characteristics was generated.

	Neighborhood
City	
Bogota	48
Buenos Aires	29
Lima	33
Montevideo	24
Rio de Janeiro	31

Figure 6. Qty of potential neighborhoods by City.

The complete dataframe still contained the information for the neighborhoods in Bogota, however that data can be easily removed if needed.

5. Conclusion:

Main deliverable was generated (list with desirable neighborhoods for expansion) was therefore the first step in the process for selecting potential venues for expansion in Latin American cities was completed, however, it's also recommended to check for the specific neighborhoods that were pre-selected in order to find information about but not limited to: rent/sqr meter, rental contracts on the area and other relevant factors that were not taken into account, in order to find the most suitable venues for the possible expansion.