Applied Data Science Capstone Project

By Óscar Luis Rojas Fernández

"El Tramito a Granel."

Problem description

Tramito a Granel literarily mean "The little section, by measure".

In many places of Latin America, the central markets were divided in "tramos", or small sections where the products where taken in bulk and sold by measure ("a granel") as opposed to prepackaged.

This is particularly useful to local producers that cannot afford prepackage costs, so central markets usually have products that are hard to find otherwise. Also, this way of selling caters to environmentally conscious buyers that can bring their own reusable bags or jars.

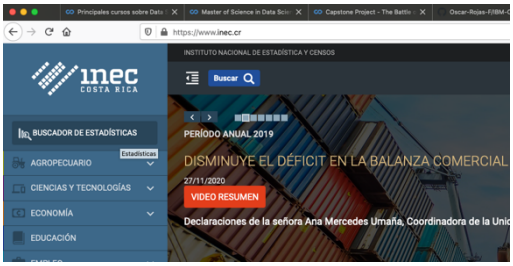
However, with the expansion of the cities and the modernization of life, the central markets are becoming out of the practical reach of most of the country.

El Tramito a Granel opened in 2018 in a district called “San Juan” of a canton called “La Unión”. It takes the experience of the central market’s “tramo” to the suburbs and it has been a great success. This year they opened a second location in a district called “Escazú” with good results as well.

For this capstone project I will play the role of an advisor to the Tramito’s owners and help them find which districts in Costa Rica are similar to the ones where they have succeeded. This will help them make informed decisions on future expansion plans.

Data sources

INEC is the national statistics institute in Costa Rica. Among many other things they provide population statistics by province, canton (sort of borough) and district.



This will be the source of the list of districts in Costa Rica as well as population density (which may be used as part of the exercise).

There is also a wikipedia page (INEC is one of its sources, by the way), so both options are valid. I will explore both and decide for the easiest.

The geolocation information will be scrapped out of Wikipedia. Each district in Costa Rica has its own page with basic information, which includes its coordinates.

With the districts location, the foursquare API will be used to gather information of each district in terms of the amount and type of venues that it has at a 2km radius.

Methodology.

Venue information will be used to make a district "footprint" that will serve as the basis for a district clustering exercise.

A visual examination will be made in order to validate the radius for the venue footprint of the district candidates prior to the clustering exercise.

The result will be k clusters of districts with similar characteristics among each cluster. The clusters that include the districts where the actual locations are now will be the ones grouping the district candidates for "El Tramito"s eventual expansion.

Results.

The first part of the work got a list of districts out of Costarrican official sources. That list was depurated and loaded into a primary dataframe that looked like the figure below.

```
df.reset_index(drop=True, inplace=True) # and the index is reset.
#Lets take a look:
df.tail(30)
```

Out[5]:

	Distrito	Cantón	Provincia	Población total	Densidad de población	Porcentaje población urbana	Relación hombres mujeres	Relación dependencia demográfica	Porcentaje de población de 65 años y más	Porcentaje de población nacida en el extranjero	P c P u c
443	Jacó	Garabito	Puntarenas	11685	83.2206	81.5319	99.0291	47.5006	3.06376	28.9688	5
444	Tárcoles	Garabito	Puntarenas	5544	31.5179	55.3391	104.274	49.4743	4.59957	13.8889	5
445	Limón	Limón	Limón	61072	1021.95	98.3249	88.6743	53.0013	6.3024	6.79362	4
446	Valle La Estrella	Limón	Limón	17908	14.5247	10.124	105.391	69.2308	4.22158	4.82466	5
447	Río Blanco	Limón	Limón	8307	62.2947	51.571	107.987	57.9278	4.28554	8.98038	5
448	Matama	Limón	Limón	7128	20.9801	14.5903	98.6069	61.3035	5.85017	7.32323	5
449	Guápiles	Pococí	Limón	36469	140.282	86.5859	92.6722	47.5761	5.30588	6.28205	5
450	Jiménez	Pococí	Limón	10501	97.4842	44.4434	97.7217	53.7257	5.35187	7.18979	5
451	Rita	Pococí	Limón	24041	51.2547	33.838	104.465	53.7443	5.93153	6.14783	5

A pivot dataframe was built on the second part of the project. It was based on the district information of the primary dataframe and the "page" method of the Wikipedia API and the "coordinates" attribute of the objects that result from the call of the API.

This part proved to be very tricky as many Wikipedia search calls produce ambiguous results that need to be resolved. I ended doing a three pass structure that combined different ways to call the search plus a manual correction at the end.

The result proved to be quite satisfactory and yielded a pivot dataframe with the list of districts and its coordinates.

```
In [11]: df_pivote
```

Out[11]:

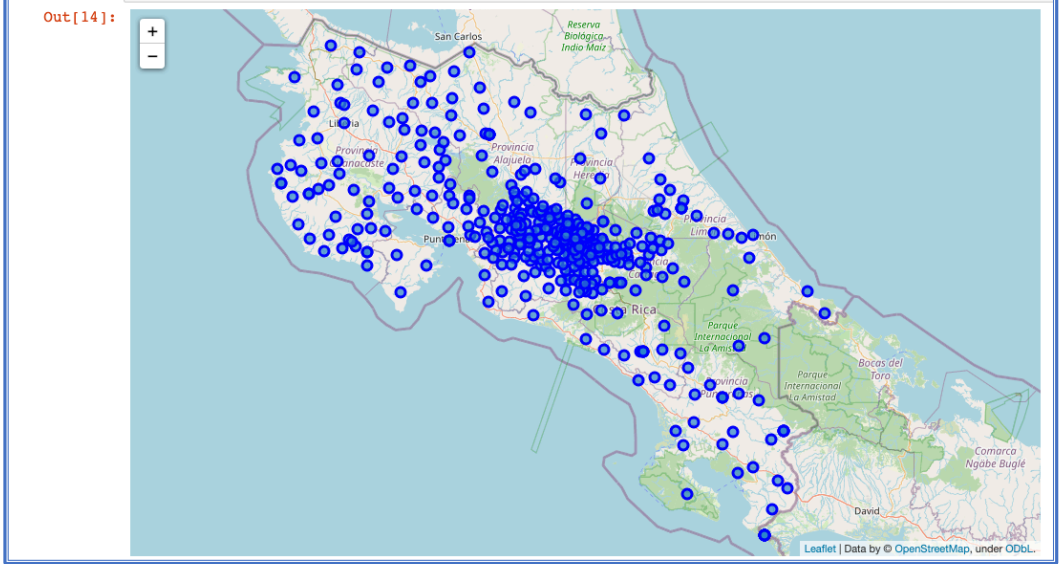
	Distrito	Cantón	Provincia	Latitude	Longitude
0	Carmen	San José	San José	9.9363	-84.07
1	Merced	San José	San José	9.939959099999999381225279648788273334503173 82...	-84.08835820000
2	Hospital	San José	San José	9.926555499999999199189915088936686515808105 46875	-84.08903549999
3	Catedral	San José	San José	9.924567299999999647752702003344893455505371 09375	-84.07134220000
4	Zapote	San José	San José	9.920340799999999589135768474079668521881103 5...	-84.05922859999
5	San Francisco de Dos Ríos	San José	San José	9.908293300000000414229361922480165958404541 01...	-84.05817140000

The third part of the project started with a visual inspection of the map of Costa Rica and the districts. It also prompted a correction of some districts that were erroneously located in other parts of the world (14 in total). That correction was feeded back into the pivot dataframe of the part two and then the code was rerun so we can get a clean list and keep the analysis going.

The map looks like this:

```
map_CR
```

Out[14]:



Once the map was depurated, another problem arose: some district names are the same for multiple locations. For example: San Pedro is a district near the center of the province of San José, but also a rural district in Alajuela is called San Pedro. To solve this, a dataframe was created called CR_districts with what I called “distrito extendido” or extended district which is a compound name of the district, the canton (kind of a borough) and the province names.

With that list in hand, a list of venues was created for each district with the help of the foursquare API using a two kilometer radius from the center of each district. (two kilometer is a reasonable distance to consider a venue to be nearby for the purpose of going to shop for artisan spices).

The results were compilated in a list of categories and a list of districts paired with the occurrence of venue categories was created. That list was normalized, and the result looked like this:

```
In [31]: # Now the table is grouped and the data is normalized
         CR_grouped = CR_onehot.groupby('Distrito_extendido').mean().reset_index()
         CR_grouped
```

Out[31]:

	Distrito_extendido	Accessories Store	Airport	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Amphith
0	Aguabuena,Coto Brus,Puntarenas	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000
1	Aguas Claras,Upala,Alajuela	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000
2	Aguas Zarcas,San Carlos,Alajuela	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000
3	Alajuela,Alajuela,Alajuela	0.000000	0.000000	0.000000	0.00	0.000000	0.010000	0.000000
4	Alajuelita,Alajuelita,San José	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000
5	Alegría,Siquirres,Limón	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000
6	Alfaro,San Ramón,Alajuela	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000

Out of that list, each district could be assigned its most popular venues, so that a sorting exercise could then be made. That list looks like this.

Out[35]:

	Distrito_extendido	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	8 C V
0	Aguabuena,Coto Brus,Puntarenas	Recreation Center	Breakfast Spot	Farm	Drugstore	Electronics Store	Entertainment Service	Event Service	E
1	Aguas Claras,Upala,Alajuela	Resort	Volcano	South American Restaurant	Yoga Studio	Falafel Restaurant	Drugstore	Electronics Store	E S
2	Aguas Zarcas,San	Restaurant	Pizza	Bar	Department	Italian	Gym	Pharmacy	F

k-means was selected as the results of an unsupervised clustering algorithm were in line with the results that are needed for the purposes of this project.

Since there are more than 400 districts, I decided to use 20 clusters. (that number was validated by further exploration up and down).

The results were as follows:

In [38]:
```
print('This is the cluster were the first location is:')
CR_merged.loc[CR_merged['Distrito_extendido'] == 'San Juan,La Unión,Cartago', CR_merged.columns[[0] + list(range(3, CR_merged.shape[1]))]]
```

```
This is the cluster were the first location is:
```

Out[38]:

	Distrito_extendido	Cantón	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 M Comm Venu
247	San Juan,La Unión,Cartago	La Unión	17.0	Sandwich Place	Supermarket	Ice Cream Shop	Gym	Movie Theater	Restaurant	Shop Mall

In [39]:
```
print('This is the cluster were the second location is:')
CR_merged.loc[CR_merged['Distrito_extendido'] == 'San Antonio,Escazú,San José', CR_merged.columns[[0] + list(range(3, CR_merged.shape[1]))]]
```

```
This is the cluster were the second location is:
```

Out[39]:

	Distrito_extendido	Cantón	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 Most Common Venue
12	San Antonio,Escazú,San José	Escazú	1.0	Hotel Bar	Mountain	Forest	Trail	Scenic Lookout	Soccer Stadium	Supermark

In [40]:
```
# Lets understand a little bit these special clusters.
Best_cluster1=CR_merged.loc[CR_merged['Cluster Labels'] == 17, CR_merged.columns[[0] + list(range(3, CR_merged.shape[1]))]]
Best_cluster1.shape
```

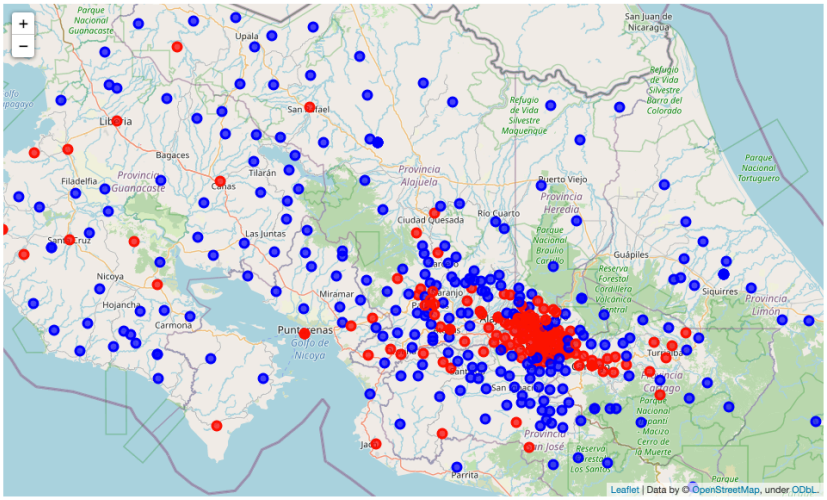
Out[40]:
```
(144, 13)
```

In [41]:
```
Best_cluster2=CR_merged.loc[CR_merged['Cluster Labels'] == 1, CR_merged.columns[[0] + list(range(3, CR_merged.shape[1]))]]
Best_cluster2.shape
```

And visually it looks like this:

```
map_clusters
```

Out[43]:



The red dots are the “qualifying” districts (Or similar districts to the ones were the current venues are located). A visual inspection of this map can bring out some strategy for the business expansion. However, this can also be expressed with a list of “cantones” like this:

In [55]:
```
# First, the two clusters are merged.
Candidate_districts=pd.concat([Best_cluster1, Best_cluster2])[['Distrito_extendido','Cantón']]
# Now, they are grouped by cantón.
Candidate_cantones=Candidate_districts.groupby('Cantón').count()
Candidate_cantones.sort_values(by='Distrito_extendido',ascending=False).head(10)
```

Out[55]:

	Distrito_extendido
Cantón	
San José	11
Alajuela	10
Cartago	6
La Unión	6
Desamparados	6
Goicoechea	6
Tibás	5
Santo Domingo	5
Palmares	5
San Rafael	5

And that is the final result: a list of 10 “cantones” sorted by how many qualifying districts are there in each “canton”.

Observations.

By far, the gathering of the data and its corresponding depuration was the most time-consuming part of this project.

If is worth noting that using official sources like the national statistics and census institute should be incorporated whenever possible as they more often than not do provide unbiased and reliable data. The downside of these data sources is that they are not as updated nor complete as “live” data sources that leverage collaborative approaches for its content creation and updating.

In this particular case, the most recent census was for 2011, so the population information had to be left out. However, next year (2021) will bring us a new census, so this same exercise could be run, and it should be very interesting.

“Live” data sources like Wikipedia or foursquare do provide very updated information and are the diametral opposite from official sources. The consideration must be made to adjust for errors in its content, as it was shown on the multiple locations that were erroneously placed.

A combination of official and live data sources proved appropriate for this project.

Conclusions.

“El tramito a granel” is a successful business that found a niche and its blooming.

Its two current locations in San Juan de La Unión and Escazú have proven to be right for this kind of business.

If an expansion was to be done, it would be appropriate to consider locations that share the same characteristics as the ones that have proven to be successful.

Unsupervised clustering algorithms can be used to group similar locations and thus determine a list of candidates.

The following is a list of “cantones” with several districts that are similar to San Juán de La Unión and Escazú:

Cantón	
San José	11
Alajuela	10
Cartago	6
La Unión	6
Desamparados	6
Goicoechea	6
Tibás	5
Santo Domingo	5
Palmares	5
San Rafael	5

With the exception of San José, (where the central market is) and La Unión (where there is already one of the location), this list can be considered a valid reference to look for future expansion locations.