Felipe Jimenez J.

geographical analysis restaurant chain introduction

May 2020

Introduction

In this project we will be doing an exploratory analysis to find the best districts in Costa Rica to place ten locals of a new chain restaurant in the country to do this we are basing our criteria on a market principle of placing similar business together, this is really important since our chain will be new in our country so we'll like to be near where people are going to eat so we can pop up as a new alternative for them.

We already know food courts are a good option, but recently the looking for a culinary experience, beyond of what you find in a food court, has started to grow and with it some kind of food markets with different restaurant options, this trend has been developed really organically which means the communities don't plan to have this kind of markets near them but, some restaurants and coffees started migrating to those places to offer a more unique experience to their customers.

Since these markets are not like projects scheduled to develop per se we cannot find them strictly on the internet like food markets so we will leverage on data to find those places where these markets have started to grow and develop.

Data

To accomplish these there are some pieces of information well need to find and cast:

- Data base with Costa Rica's city

- Clean data frame with the cities of interest (urban and semi-urban) and its respective location

- Data frame with nearby venues for each city

- Preprocessed data frame for the KMeans clustering method

- labeled cities with the given tag from the previous clustering method

- Explanatory summary of the clusters of interest

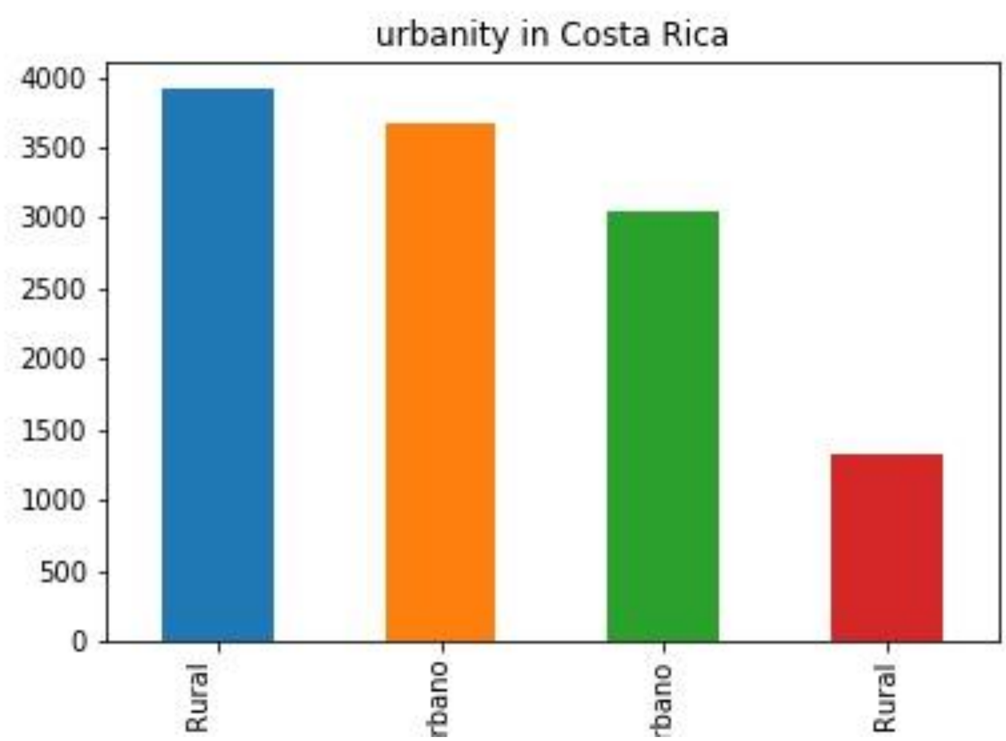
Data Source

The Data regarding costa Rican territorial division was pulled from the INEC's (

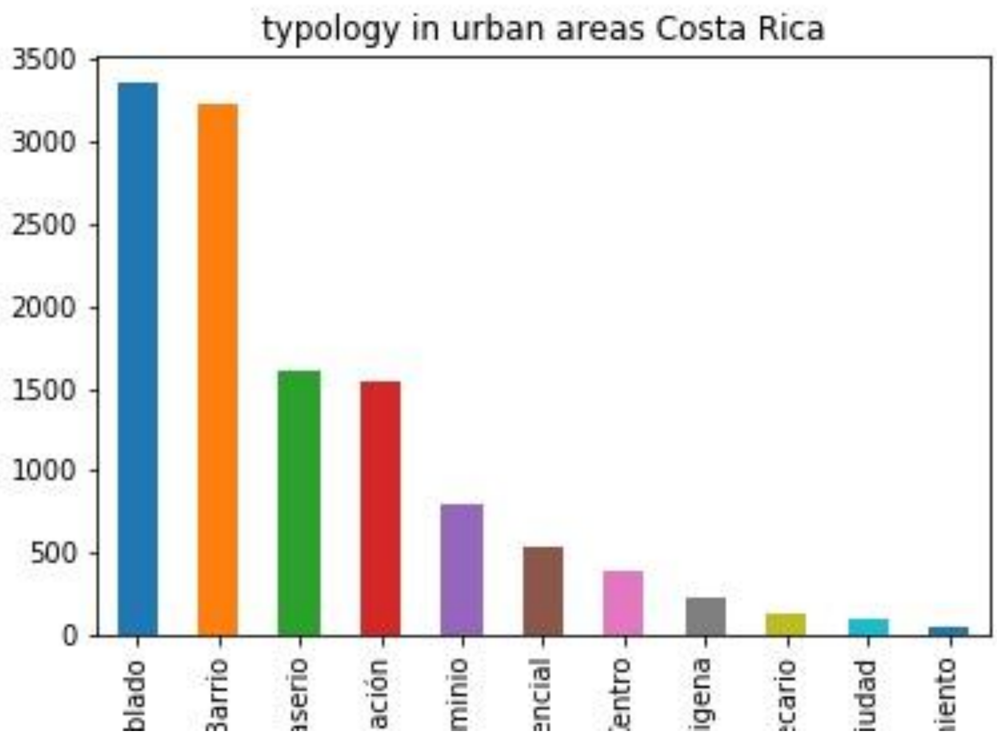
National Institute of Statistics and Censuses of Costa Rica) website.

Data Cleaning

Since on this data base we had available all cities and small towns in Costa Rica, we filtered jut to have the most populated cities (urban and semi-urban areas). These brought down the total of options around 34%, we were able to drop the blue and red columns from the next figure.



Once we removed those, we focused the study on the cities most populated. As we can see as well on the next graph, we reduction was significant we only took in consideration, Barrio, Centro, ciudad, which represents around 31% of the already reduced total. We ended up with 3321 observations.

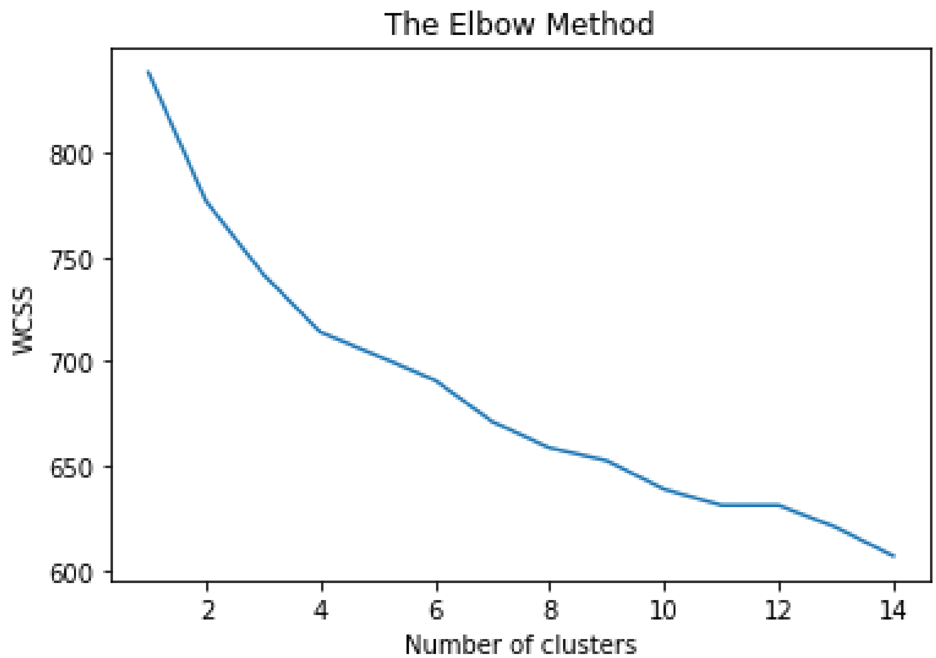


Methodology

Once we had our data cleaned, we leveraged on the Four square API and the function *getNearbyVenues* to find near venues to each location on the data frame.

With the category of every nearby venue we created a dummy data frame to regroup by the location and calculate the mean of every venue present. We did this to have valid numeric values to feed a clustering method.

We decided to use the KMeans method to label our date for further exploration. Using the Elbow method, we found optimal options to define the quantity of cluster for our model. As e can see on the figure 4, 8 and 11 are good options, at the end we decided to use 11.



Results

Once the cluster tags were set by the model, we regroup the models by tag and showed the top 4 venues for each cluster.

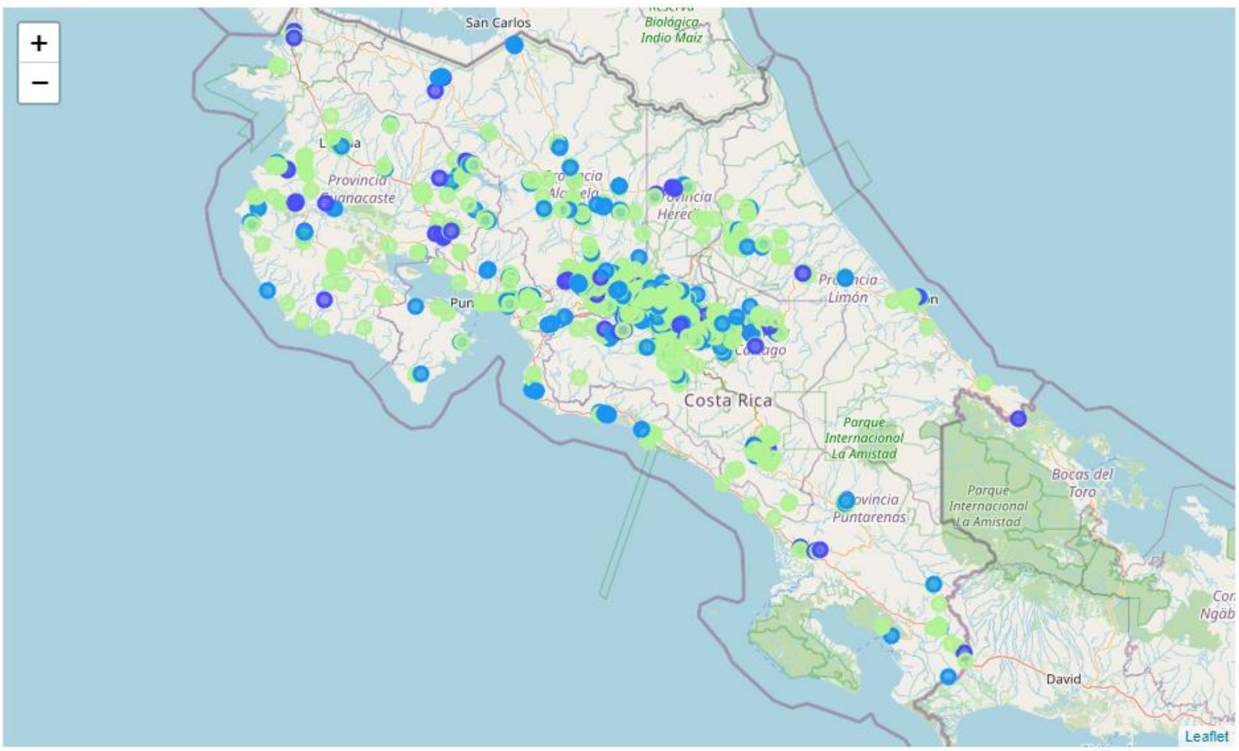
On the next table you can see the tops venues for each label.

	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
0	0	Grocery Store	Bar	Restaurant	Soccer Field
1	1	Bar	Soccer Field	Park	Convenience Store
2	2	Restaurant	Bar	Campground	Brewery
3	3	Restaurant	Bar	Pizza Place	Soccer Field
4	4	Convenience Store	Auto Garage	Burger Joint	Plaza
5	5	Bus Station	Park	Bakery	Pizza Place
6	6	Soccer Field	Grocery Store	Restaurant	Bar
7	7	Pizza Place	Fast Food Restaurant	Bakery	Bar
8	8	Mountain	Burger Joint	Steakhouse	Trail
9	9	Bar	Soccer Field	Steakhouse	Grocery Store
10	10	Campground	River	Supermarket	BBQ Joint

Te criteria used to select the clusters that better fit our solution for our initial idea was, the cluster which have al least 3 most common venues related to restaurants on this top 4.

Which are 2,3,7.

Then we projected those clusters on a costa Rica's map and do a further analysis based on the location of the clusters.



Discussion

As we predicted, the location tends to group around the GAM and the cost mostly on the known beaches and touristic regions.

We can rely on the information given by this study because the venues pulled from the Four Square API, has a minimum presence on social media, since we want to stablish a chain restaurant we want to be near those establishments.

Conclusions

We can conclude that this study did not gave us the exact location where to place our locals, it narrowed down the options in a good and measurable way.

With the information requested to Four Square, we can even pinpoint our direct competition. To be closer to it.

WE can take advantage of the technological asses that we have to shorten the time of a project.