Coursera Capstone Project - Comparing business types in San José, Costa Rica

Maikael Antonio Mora Briceño

January 2021

1 Introduction

1.1 Background

San José is the capital city of Costa Rica and it's largest city, but it only a small part of the greater metropolitan area that fills the country's Central Valley. Around half of Costa Rica's population lives in the main metropolitan area but recent expansion has not been equal in it. Historically, the eastern side of the valley was the most populated, including the old capital city of Cartago, while the western side has been growing substantially in recent years thanks to upgraded infrastructure, available land and closeness to a host of new services opened in the area.

This context means that the populations and land ownership is very different between east and west of the city; even if in the past the eastern side hosted most of the high income population of the country, the western side has recently attracted new professionals and internationals firms and now hosts the country's rich. These historical and economic differences mean that entrepreneurs and investors must study consumption patterns carefully if they are to succeed.

1.2 Problem

Even if consumer studies could shed light onto important clues, these are expensive and require additional efforts in logistics and human capital. By carefully browsing the regions where investors seek to enter for concentrations certain types of businesses, they can shape their projects accordingly. In the case of investors building a new shopping center for example, they might like to attract existing businesses into the new center which requires this knowledge or if they seek to compete they need to know who their main competitors are.

1.3 Interest

The results could interest other investors as well but other stakeholders, like urban planners, publicity companies and consumers might be interested in knowing what each region offers them and what the main holes in their goods and services portfolio are.

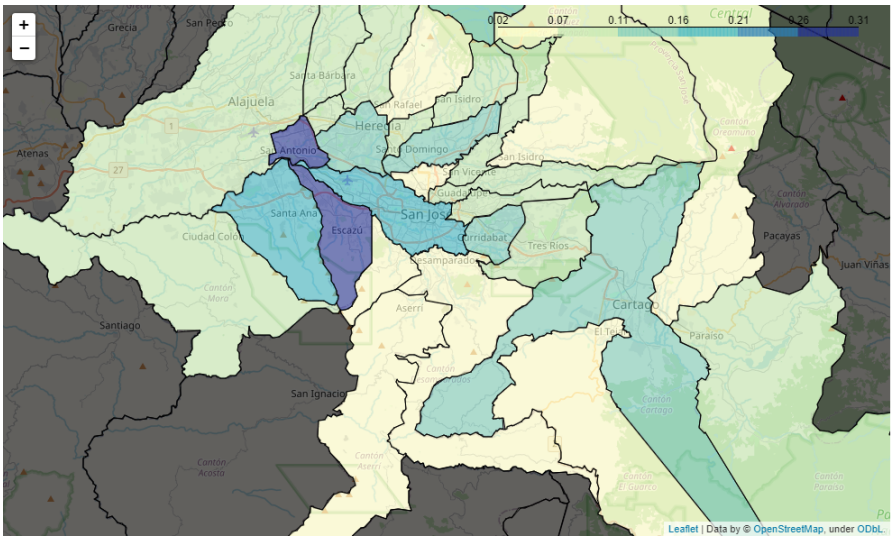

Figure 1: Map of included areas in Costa Rica by per capita income level.

As shown in the map generated using Folium on Python, there are several

areas in the west that concentrate a higher level of per capita income than in the east of the map, shown by the darker shades of blue.

2 Data

2.1 Sources

The JSON file with the coordinates for the region can be found at this site along with the information on income and development for each region. The inclusion of the regions in the investigation was done manually using the government's definition of the area.

2.2 Data management

The main difficulty in managing data for Costa Rica is the inherent lack of it. A general look at the amount of businesses found at each region show large differences in the concentration of these. The main way to handle this issue is to group them in larger areas which is why I limited myself to east and west as this allows me to concentrate the results of various districts into only two regions.

Other than that, the rest of the management did not require a high level of cleaning. The map points concentrated in the most populated areas and this was beneficial as it gives us a more realistic view of important businesses economically. A problem that could have arisen would have been coordinates for the center of the clusters that were in the center of the geographic regions which in some cases if far from the main urban areas or in the middle of a mountain. That would have required additional work to correct.

The rest of the data was successfully sourced from the Foursquare API and was correctly matched to each region. A descriptive analysis is made further into the document to shed additional light on what the regions look like from a business sense.

Overall, the data totalled 32 unique business types, 28 districts in total of which 13 were allocated to the West Region and 15 to the East Region.

3 Methodology

Concretely, the first step was simply organizing the geographical data. This involved sourcing the coordinates and JSON file in order to build a map and to cluster the data correctly.

After this, we ran queries into Foursquare to attach business types to each region, we limited ourselves to three in order to prevent any problems due to lack of data. In order to make sure we had regions with enough information, we grouped the districts into two regions depending on their longitude relative to the center of the capital.

Once we had the businesses sorted into their regions, it was time to create a ranking. This was done by giving three points to each type of business for every time it ranked first in a district of said region. Two points were given to second and one to first. Once that was done, we had information regarding business types and their relative importance in their regions.

4 Results

After completing the code and organizing the data, we can begin to do some basic analyses of the results. For example, we have 32 business types in total that range from certain restaurants to pools, gyms and shopping plazas. To begin, I first present the locations from where the businesses will be centered from.

As I previously stated, the locations are concentrated towards the middle of the area to be studied which is important as this directly relates to the areas with highest population. There is little value in including businesses towards

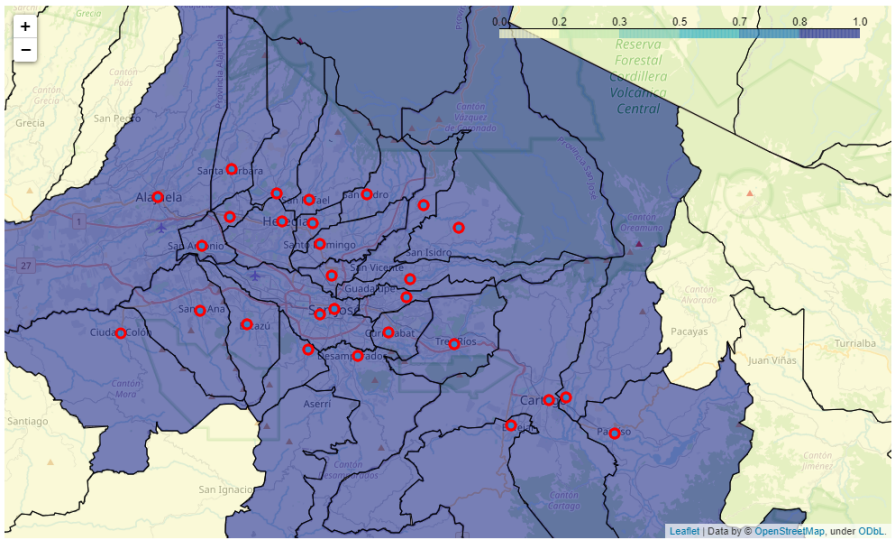

Figure 2: Starting points for the search of businesses on Foursquare.

the edge of the map as these would be of little importance to stakeholders.

The search for the top three businesses in each of those areas and subsequent grouping into east and west yielded the following results:

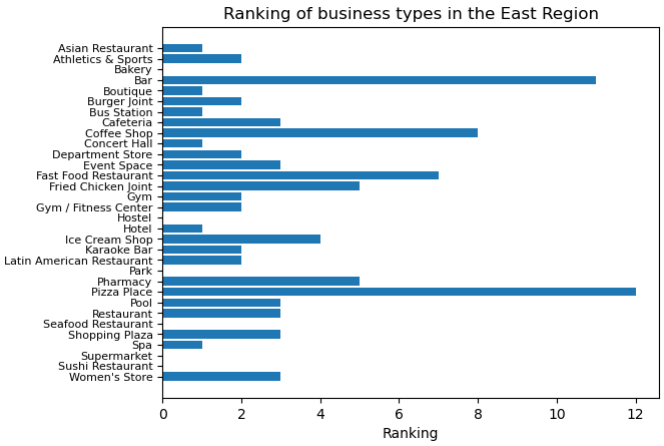

Figure 3: Ranking for businesses in the east region

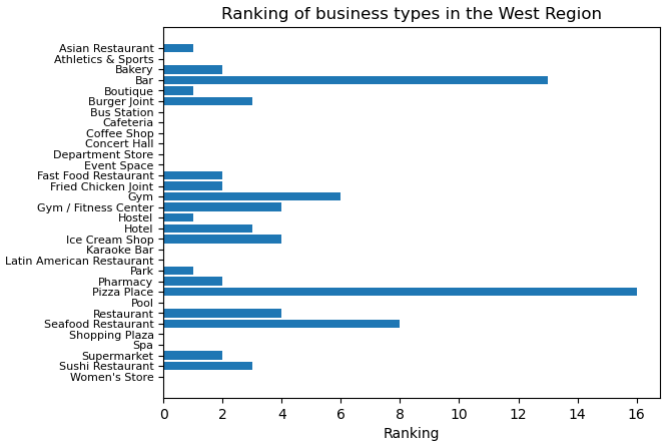

Figure 4: Ranking for businesses in the west region

The ranking represents the results of a points based system where every business type received 3 points for every time it ranked first in a district in that region. The types also received 2 for ranking second, 1 for ranking third and none if they did not appear. If a type appears with no points it means that it did not rank in any district of that region.

There are several characteristics that are interesting to analyze in this situation. First off, the eastern region shows a lower concentration of types than the western region. There are fewer non ranking types and the points were more evenly allocated between types compared to the western region. There are cases that don't follow this such as gyms and seafood restaurants in the western region, that buck the trend and have a higher ranking in the west than east.

There are some similarities that must be pointed out as well. For example, pizza restaurants are clearly a favorite in both regions as are bars but as mentioned before, show a higher concentration in the west. In general though, the fact that differences are found means that there is value in what we did, specific conclusions referring to stakeholders are included towards the end.

For additional visualization, this is how the regions look on a map:

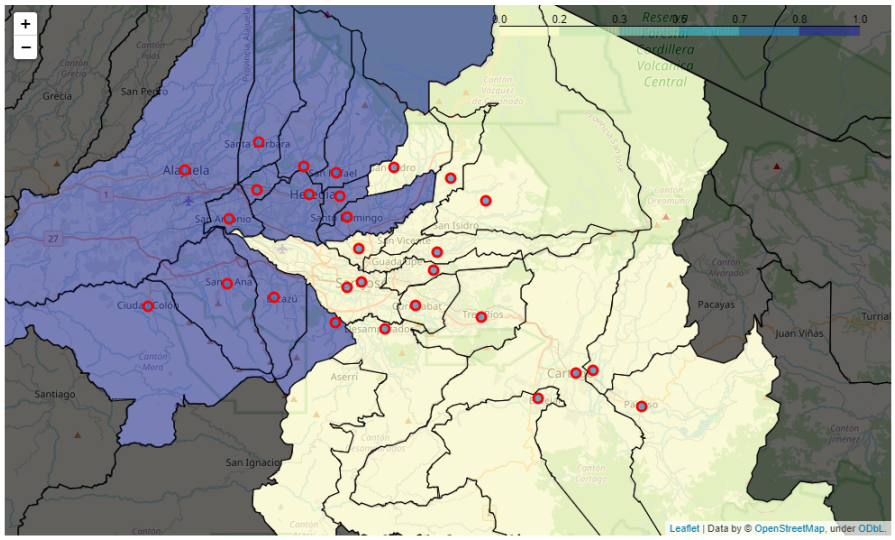

Figure 5: East and West Region (Blue is West).

Just by looking at the points and regions, there is a strong argument for generating additional regions but as mentioned before, some points show less data concentrations than others.

5 Observations

There are several benefits of the path I followed in order to get to the final objective. On the one hand, the code can be easily repeated with an additional number of types and the information exists in a district by district level, so other macro regions could be created to fulfill other needs by additional stakeholders.

In this specific case, where I created a fake situation where investors wanted to build two shopping centers in opposite sides of town, there is great value in the results. Firstly, for the western region, there are types that show no importance like department stores, karaoke bars and concert halls for example. This does not mean that they should be excluded, but their inclusion should

also bring additional support to draw consumers (publicity for example) and they could be considered riskier inclusions in any project. On the other hand, seafood restaurants and gyms have shown a heavier importance so spaces for these types of businesses should be considered.

In the case of the eastern region, most types show some kind of importance which could mean that investor's will find it difficult to go wrong as consumers have shown some type of attraction. Specific types that are important in this region include cafeterias and coffee shops, fast food restaurants and pharmacies. This again means that investors should look to include these types of businesses into their plans.

What is true for both regions is that the inclusion of pizza restaurants and bars is key to success as they are by far the highest ranked in both regions. What is also key to understand is that types that are highly ranked are also a sign of competition for entrants into those specific types, so the conclusions can be analyzed in two ways. On the one hand, opening a bar or pizza restaurants in a new shopping center has the potential to attract huge amounts of clients, but these will also face heavy competition from existing businesses that exist.

That last point also means that unranked types could be opportunities to enter without heavy competition, so profits and success could be found if mixed with good publicity and marketing. In the end, the decisions will most likely depend on the type of investors that are looking at the data, the availability of resources and how much they are expecting and willing to compete with the established businesses. Another strategy could be to negotiate with existing businesses to move into the shopping center.

6 Conclusion

The investigation used location data to group districts in Costa Rica and the businesses that characterize each one to create a picture of the commercial sector

in the main urban region of the country. I included a small background and some information on income to show some of the underlying differences between the regions being treated.

Results show that there is indeed a difference between both regions, where the west shows a huge concentration of very specific types (seafood restaurants and gyms for example) while the east shows a larger inclusion of types. Both regions show a huge importance of bars and pizza restaurants in the types of establishments but the concentration is higher in the west.

The differences evidenced show that this information has value and that supposing that both regions were the same could have diminished the return on any investment made by stakeholders. After going through the investigation, additional ideas pop up. For example, different weighs could be given to the importance of several districts depending on population and income level to pin point an ideal point for building the shopping center and not only the businesses to target.

The benefit of how the data was collected and managed is that it can easily be varied to include more types and/or other types of data depending on the needs of the stakeholders (consumers, government or investors). Additional regions can be built by grouping districts by other characteristics and weighing distances, income and consumer preferences. Additionally, smaller divisions of land can be used if necessary, unfortunately income data does not exist for that level of geographical division.