Ángel Guardián Expansion
Finding the Best Location for a New Base

Daniel Zachrisson

July 8, 2020

1. Introduction

1.1 Background

Ángel Guardián is an organization in Guatemala City that provides a service that consists in connecting people who are under the influence of alcohol or drugs with a volunteer driver (off duty volunteer firefighter). The volunteer driver takes the person under the influence on their own vehicle to their home. The purpose of this service is to reduce the amount of people out in the streets driving under the influence of alcohol or drugs.

1.2 Problem

The purpose of the project is to analyze location data of Guatemala City to find the most strategic location for Ángel Guardián to expand throughout the city so that waiting times for clients are reduced.

1.3 Interest

The organization Ángel Guardián will have interest in this project if they are looking to expand, or any other organization that wants to start a similar service may use the insights found on this analysis for their benefit.

2. Data Acquisition and Data Processing

2.1 Data Sources

The postal codes for each 'zona' was obtained from https://worldpostalcode.com/guatemala/ciudad-de-guatemala. The dataset was combined with the longitude and latitude for approximately the center of each 'zona' obtained from https://www.gps-coordinates.net. The location information was obtained by using the Foursquare API by using the 'zonas' number to retrieve the venue information for each zona. The radius was set to be 800 meters, with a limit of 100 venues per call.

2.2 Cleaning the Data

The data from both sources were combined into a single data frame using the Pandas library. The Foursquare API was used to retrieve information such as the names of the venues close to each center of each 'zona' and to obtain the latitude and longitude of such venues. The information collected from the Foursquare API was saved into a separate data frame. The venues were grouped by their respective 'zona' and later counted to have an idea of how many venues are in each 'zona'.

A new data frame was created using the 'get_dummies()' function on the 'Venues Category' column to get the categorical values of 'category' into a numerical value that could be used for the analysis later. The dummy values of 0s and 1s was grouped again by 'zona'. The mean was calculated by each 'zona' to find the most popular venue.

The data was placed into a new data frame where the first column has the 'Zona' and the following columns are the ranking of popularity of each kind of venue, starting with the most popular and to the least popular.

	Zona	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	Zona 1	Café	Bar	Restaurant	Bakery	Coffee Shop	Pharmacy	Latin American Restaurant	Performing Arts Venue	Park	College Auditorium
1	Zona 10	Café	Hotel	Italian Restaurant	Steakhouse	Breakfast Spot	Bar	Seafood Restaurant	American Restaurant	Restaurant	Shopping Mall
2	Zona 11	Bus Station	Arts & Crafts Store	Sandwich Place	Athletics & Sports	Taco Place	Soccer Stadium	Dive Bar	Donut Shop	Diner	Department Store
3	Zona 12	Bar	Café	Bike Rental / Bike Share	Seafood Restaurant	Restaurant	Athletics & Sports	Yoga Studio	Factory	Flea Market	Fast Food Restaurant
4	Zona 13	Airport Lounge	Airport	Bar	Café	Fast Food Restaurant	Chinese Restaurant	Rental Car Location	Breakfast Spot	Fried Chicken Joint	Bed & Breakfast

Figure 1. Clean Data

3. Methodology

A GitHub public repository. The repository contains the Excel files that contain the coordinates and the postal codes for each 'zona' in Guatemala City. The documents were read into two different data frames and then merged into a single one.

	Zona	Postal_Code	Latitude	Longitude
0	Zona 1	1001	14.643991	-90.517660
1	Zona 2	1002	14.657443	-90.514399
2	Zona 3	1003	14.630372	-90.528475
3	Zona 4	1004	14.618579	-90.517832
4	Zona 5	1005	14.636268	-90.493370

Figure 2. Postal Codes with respective coordinates

The was processed using a combination of the Pandas, Numpy and the SKLearn libraries. The Pandas library was used for the manipulation of the data frames. Pandas has many functions that make dropping and modifying full columns or rows fast and efficient. The Numpy library was mostly used to create some ranges to be used in loops.

The Foursquare API was utilized to obtain the name and coordinates of the venues in a radius of 800 meters of each "Zona's" coordinates. The information was merged into a new data frame, in which it was sorted and manipulated to transform some categorical data into numerical data.

Zona	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Zona 1	14.643991	-90.51766	Cafe Imeri	14.642682	-90.516324	Bakery
Zona 1	14.643991	-90.51766	Panaderia Berna	14.639136	-90.515013	Bakery
Zona 1	14.643991	-90.51766	Pasaje Tatuana	14.645485	-90.513617	Arcade
Zona 1	14.643991	-90.51766	El Portalito	14.640646	-90.513624	Bar
Zona 1	14.643991	-90.51766	Dunkin' Donuts - Pasaje Rubio	14.640926	-90.514065	Breakfast Spot

Figure 3. Venues with their Respective Coordinates

The KMeans package was imported from the SKLearn library. It was used to create the different clusters in which each 'zona' would later fall into based on similarities in venues from the other 'zonas'.

Zona	Postal_Code	Latitude	Longitude	Cluster Labels
Zona 1	1001	14.643991	-90.517660	2.0
Zona 2	1002	14.657443	-90.514399	2.0
Zona 3	1003	14.630372	-90.528475	1.0
Zona 4	1004	14.618579	-90.517832	2.0
Zona 5	1005	14.636268	-90.493370	4.0

Figure 4. Merged Data Frames with Cluster Labels

The ‘cm’ package from the Matplotlib library was used in the elaboration of the map in conjunction with the ‘colors’ package from Matplotlib, the Nominatim package from Geopy library and the Folium library. The Nominatim package was used to determine the coordinates of Guatemala City that would later be used in Folium to place Guatemala City in the center of the map. The data from the KMeans clusters was used to create circular marks in the map generated with the Folium library. The marks are for each ‘zona’ and the colour they are represent to which cluster they belong. Marks of the same colour belong to the same cluster. The clusters are determined by using the means of common venues in each ‘zona’ and later comparing it with other ‘zonas’ that have similar venues.

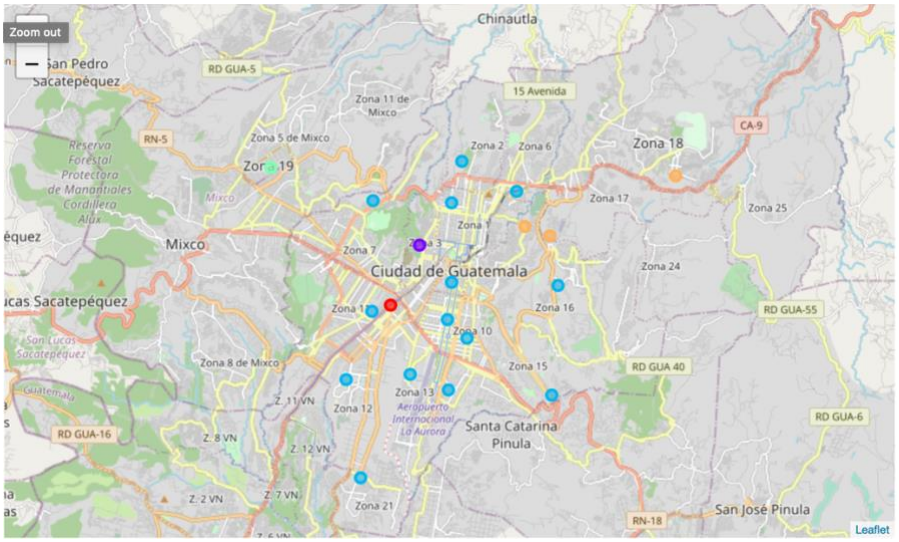

Figure 5. Guatemala City Map with Clustering

The data for the average amount of bars and sport bars was passed onto a new data frame, that using the Pyplot package from the Matplotlib library was plotted into a horizontal bar chart to make it easier to compare the ‘zonas’ with a higher percentage of bars.

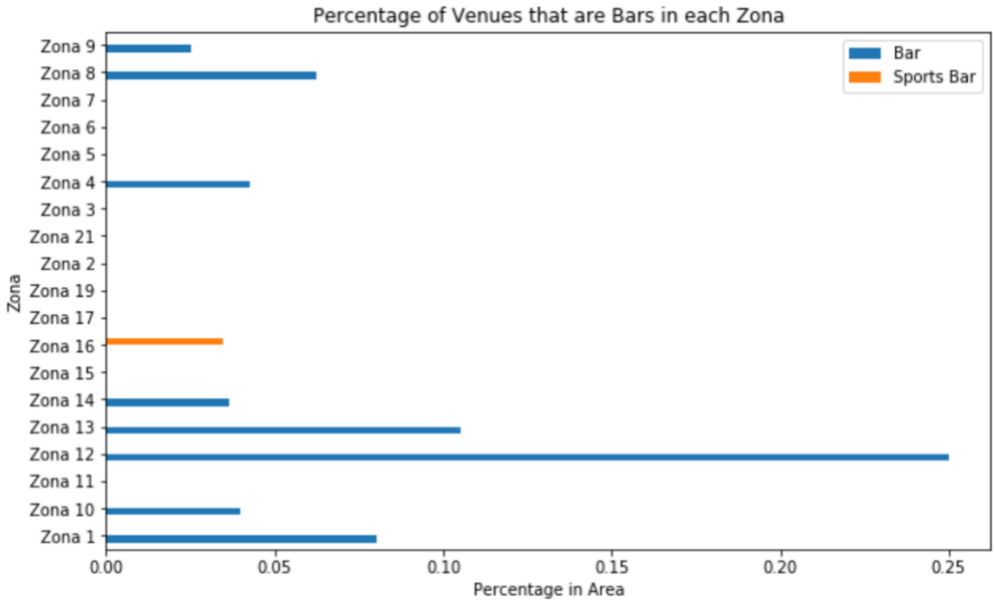

Figure 4. Bars and Sports Bar per 'Zona'

4.0 Results and Discussion

The clusters provided little insight on which 'zonas' had a greater number of bars and nightlife in general. The most common venues in most of the 'zonas' were pretty similar, reason why one of the clusters is very large. Guatemala has many repeated venues such as fast food places, convenience stores and pharmacies in every 'zona' so for bars to pop out, we need a large number of them. It's also important to mention that Guatemala has many venues that may not have made it into the Foursquare database yet. It's also important to know that the percentages were rounded to three decimal places, reason why many of the 'zonas' show as if they have no bars whatsoever.

Since clustering didn't work as expected, the bar chart was generated. In the bar chart, it can be seen that the 'zona' with the largest number of bars is 'Zona 12'.

In future studies, the data should be combined with other APIs to have a more exact representation of the venues around Guatemala City. It is also relevant to explore the impact of changing the radius on each call to the Foursquare API, as some of the 'zonas' have weird shapes and might fall in the radius of a different 'zona' if the radius is not properly set. It's also relevant to mention that a better way to determine the center of each 'zona' would be needed to get more accurate results as well.

5.0 Conclusion

Ángel Guardián would benefit of setting base around ‘Zona 12’ as is the area with the largest number of bars. They can reduce the waiting time for their client once they want to use the service by being closer to where the bars are. If they want to go big and set base in many new places, the other two best locations would be around ‘Zona 13’ and ‘Zona 1’ as they have large number of bars as well.