Looking for best place to open a new bakery in Guadalajara, based on users ranking

Ayrton Mondragon

April 18, 2019

1.- Introduction

1.1 Background

Guadalajara has been growing rapidly, and a lot of good place have been being opened. All those are very popular and people loves to go to spend some time enjoying of a good dinner, a good coffee or a good desert. Bakeries are very popular and and the interest of people for this kind of places has been growing as well. People like to enjoy a piece of cake and talk with friends, take a hot or cold coffee and it's very popular also that a lot of people goes to the bakery to buy a cake for special dates, like birthdays, anniversaries and so on. There are some bakeries located in Guadalajara, that are very good and very popular.

1.2 Problem

My wife wants to open a bakery in Guadalajara, México. She wants to know where is the best place to open it based on all existing bakeries already in Guadalajara and the reviews of the users, she wants to know if it's better to open it close to the current ones, or if it's better to go to a location were there are just a few ones.

2. Data acquisition and cleaning

2.1 Data Source

I'll be working with Guadalajara neighborhoods, so I'll web scrapping this site: https://codigo-postal.co/en-us/mexico/jalisco/guadalajara/ to retrieve all Guadalajara zip codes and neighborhood names. I need to get the latitude and longitude for each of the neighborhoods and since the geocoder API is not retrieving me any information for the zip codes listed from the web page above, I will create my own file from the info on this site: http://codigo-postal.es.mapawi.com/mexico/7/tlajomulco-de-zuniga/3/308/ciudad-de-guadalajara/45659/31459/.
After I get all this info together, I will make a call to the foursquare API to retrieve all the food venues close to each neighborhood, so I can make the analysis about where is the base place to open a new bakery.

2.2 Data Cleaning

The table scrapped from web had some columns that had the same value for all neighborhoods, so I dropped those columns. Once I dropped the columns, I joined all the neighborhoods with same zip code on the same row. After that I create a new data frame with the latitude and longitude and the zip codes and merged both data frames so I can display a map with all the neighborhoods on it. I had to rename all the neighborhoods in order to remove the accent so I avoided any issue displaying the map. For this

project I'll retrieve the information for all venues from folium api, but I'll do a data analysis with just bakeries and desert venues. To do this what I did was, once I retrieve all venues information I filtered the information on the venues dataframe with the specific categories that I wanted so I could create a new dataframe with just that type of venues.

3. Exploratory Data Analysis

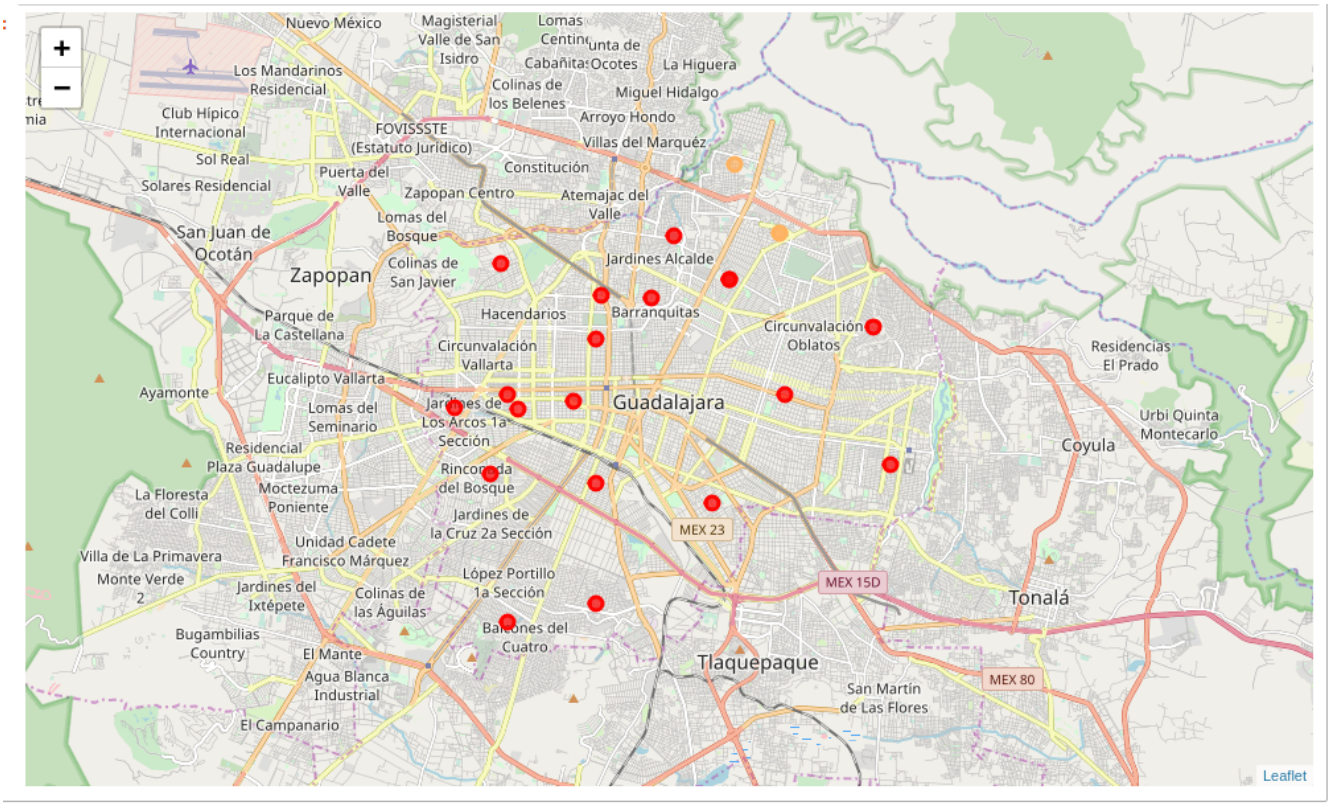
3.1 Insights from Data

Once I get all the venues from foursquare api I'm counting all different venue categories I got for each neighborhood, and I got 147 unique categories in total. Then I'm getting the hot venues on each neighborhood so I can get which category is the most visited.
Once I get all the most visited categories, I'm grouping by frequency on each neighborhood so I can have a better picture of which venue is the most rated on each neighborhood.
Now that I got all the frequencies, I'm getting the top 5 for each neighborhood. With this information I can see that almost in all neighborhoods the most common venue is a food place, so this is a good indicator, at first look I could say open a bakery in any of these Neighborhoods could be a good idea, also we can see on the new data frame that the second most common venue in almost all neighborhoods is a food place, so lets continue with our analysis.

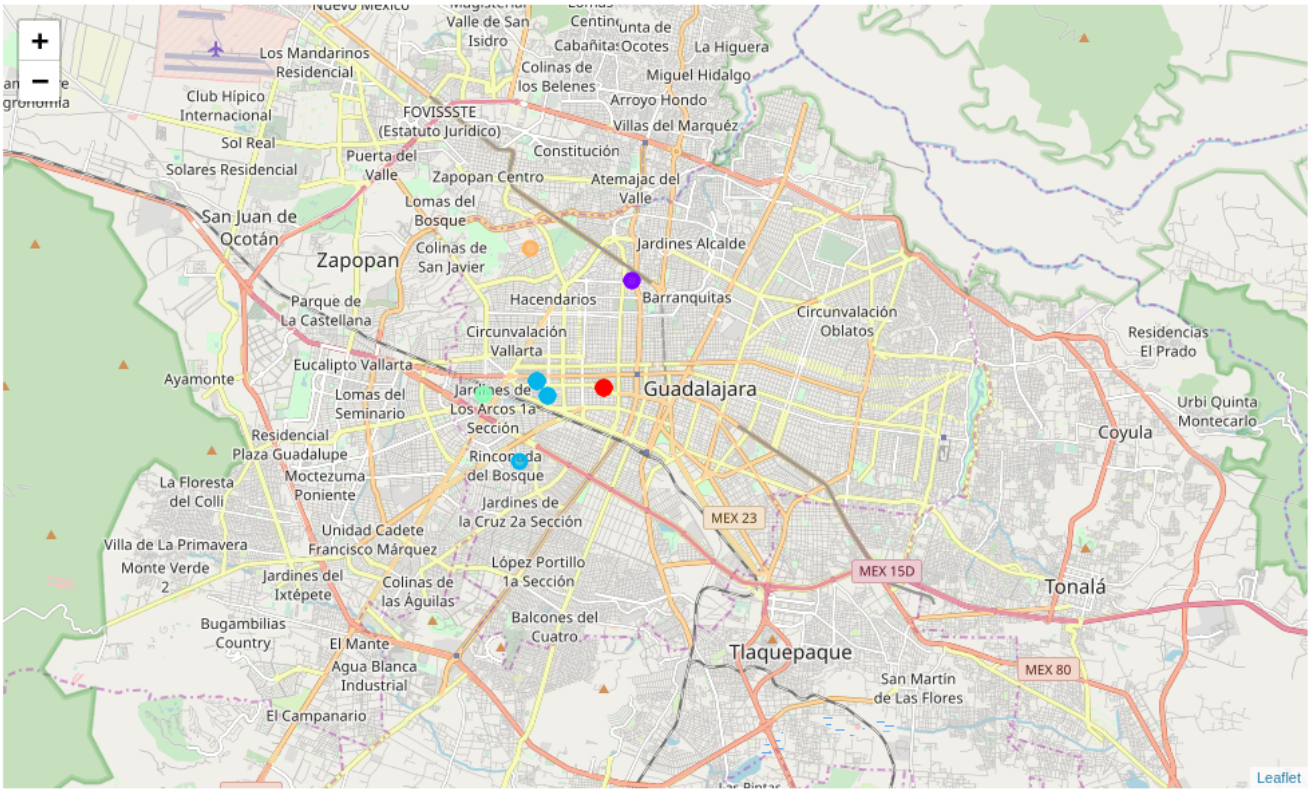
4. Predictive Model

4.1 KMeans

For this project I will use KMeans clustering algorithm. Kmeans is a clustering algorithm used to group sample data based on characteristics of features from each data point on it. The algorithm it iterates and calculate the best center for each centroid, the centroid is defined by the user or can calculated randomly. The selection of this centroid is critical for the algorithm because it depends of how well those are select the accuracy of the model.
In this case I'm using kmeans clustering algorithm in two scenarios, the first one is clustering the neighborhoods based venue category frequency on each neighborhood. For this scenario I'm setting 5 clusters and displaying each cluster on the Guadalajara map.
On the second scenario I'm filtering the bakeries, coffee shops, and deserts stores only , then converting each category to a numerical value, like 1 for coffee shops for example.
With all this information I'm creating a new clustering model and displaying this information on a map.
The idea of create those clusters is to see which neighborhood is the best place to open a new bakery, we can check which group has most records on it so we can decide if is better to open it there or by the other hand, realize where are less food venues and analyze if it's better place to open a new bakeries.
After created the clusters with all venue data, we got this:



After creating the clusters with the second scenario data, we got this:



5. Conclusion

For the first scenario I have created 5 clusters, but looking at the map above, seems that the algorithm has grouped almost all the neighborhoods on the same cluster, so in terms of taking a decision about where is the best place to open a new bakery, we don’t have a clear trend there. This grouped data is not giving us any insight to take any decision so we can conclude that including all the venue data on each neighborhood doesn’t work for us.

By the other hand using bakery and coffee shops information, I have created 5 clusters, on the data set there are 3 different categories for venues, so creating more clusters than categories, seems for me that we can get a better picture of how venues and neighborhoods are well distributed. As we can see on the second map above, there are clearly 5 clusters, where one has more elements on it, cluster 2. Analyzing this information we can conclude that most of the bakeries and coffee shops in Guadalajara are located between Colonia Barrera and Colonia Arcos Vallarta. We can conclude that open a Bakery around those locations could be a good idea because a lot of people use to hang out on the places near to those neighborhoods. So we can have more people walking around and we can gain customers rapidly.