Exploring Caracas Metropolitan District venues information available in FourSquare

Rafael Blanco
IBM Data Science Coursera

1. Introduction

Caracas is the capital city of Venezuela, Caracas Metropolitan District (CMD) is composed by The Capital District (Caracas) and part of Miranda's District. Each Venezuelan district is formed by Municipalities and each municipality is composed of one or more Parishes. Moreover, The CMD is composed of five (5) municipalities and thirty two (32) parishes.

There are several applications that recollect geolocation data corresponding to cities' places of interest, like geolocation and information about administrative buildings, transport stations, residences, venues, etc. One of the most important sources of this kind of data are the interactions people have with internet connected applications while they are performing their daily activities. For example, users who post photos and reviews of restaurants they visit, or those that use a map service to find a way to reach a place, or even, those who order food from a nearby restaurant from their homes.

There are popular web applications that require a database that contains information about popular places in the community, with their coordinates (geolocated data) and descriptions about the activities performed there.

Most popular delivery food services such as UberEats are not available in Caracas Metropolitan District. Taking into account the applications requirements of geolocated data, the aim of this study is to collect data available on FourSquare portail about venues in the CMD to identify the regions to implement a prototype of a new delivery food service application.

2. Data acquisition and cleaning

The data acquisition task started with some research about the geographical and administrative distribution of the CMD. The data was collected by importing and parsing tables from the Wikipedia website; some of the entries where the information was not available on Wikipedia were collected using other internet sources and added manually into the dataset. This task results in a built dataset where each row corresponds to a single parish, the municipality it belongs to, its geographical coordinates (latitude and longitude) and the area it occupies.

	Municipality	Parish	Latitude	Longitude	Area(km2)
0	Libertador	Santa Rosalía	10.483550	-66.914420	6.1
1	Libertador	El Valle	10.467206	-66.907329	31.1
2	Libertador	Coche	10.451853	-66.925286	13.0
3	Libertador	Caricuao	10.433333	-66.983333	24.8

Figure 1: DataFrame built with the information from each Parish in CMD

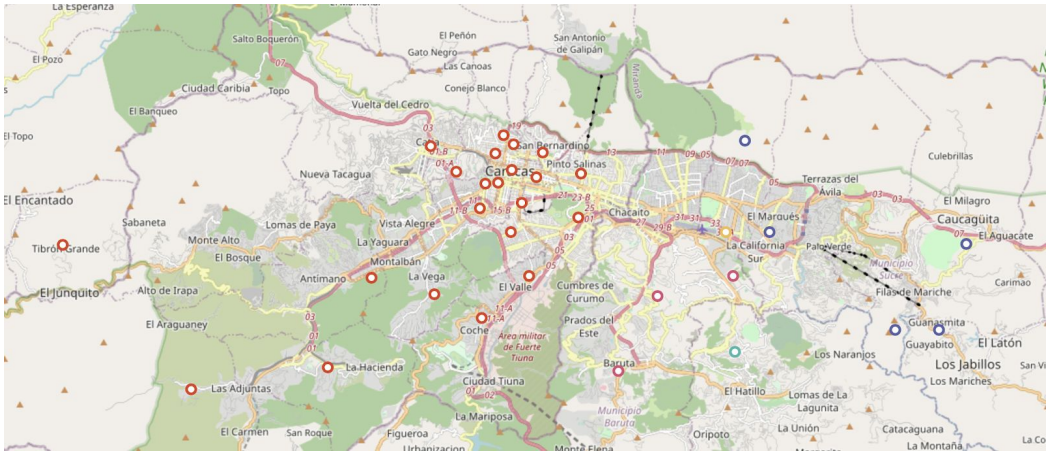

Figure 2: Geographical distribution for CMD's Parish. Border colors represent the Municipality it belongs to.

The next step was to explore the venues around the coordinates of each parish, the FourSquare explore API was used to retrieve this information. We needed to provide a radius to search. We decided to vary the radius search of each call depending on each parish and plot this information on a map to have insights about the area we were exploring.

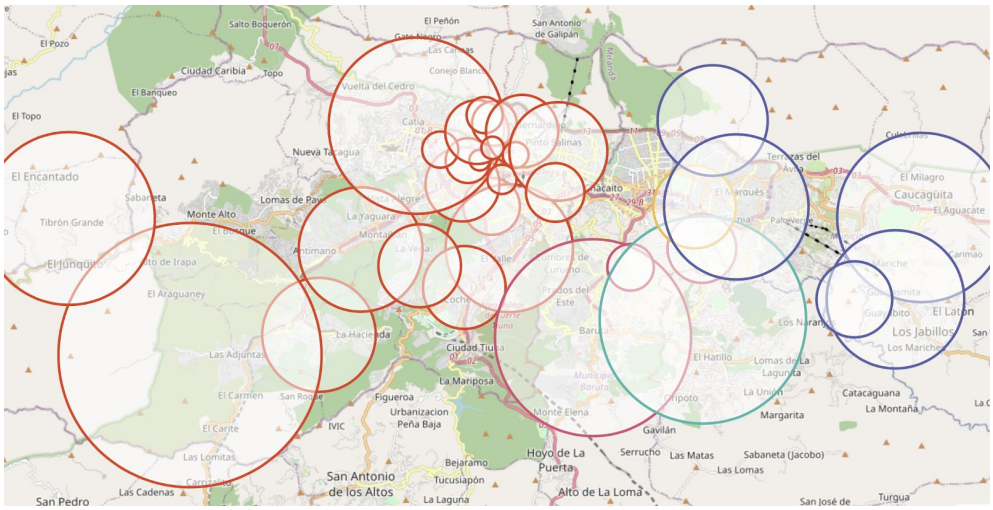

Figure 3: Radii search first iteration

We realized that there was a significant overlap produced by the formula we used to calculate the radius; after several tries, we found a formula to calculate the radius avoiding the majority of the overlaps.

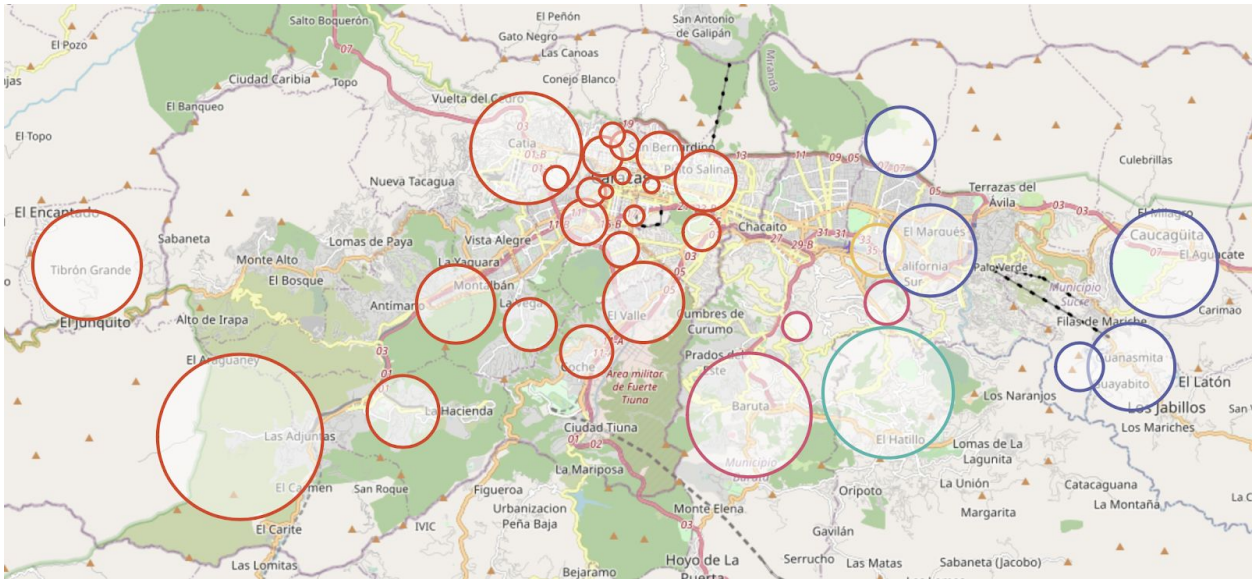

Figure 4: Radii search for exploring venues

Finally, we collected and parsed information about venues around our points of interest to build a dataset, with special interest in the number of venues inside each searching radius and the categories these venues belong to. We obtained as result a dataset with the 5 most common category venues per parish.

	Parish	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	23 de Enero	Plaza	Dog Run	Food	Flower Shop	Fast Food Restaurant
1	Altagracia	Historic Site	Convenience Store	Gymnastics Gym	Food Court	Cosmetics Shop
2	Antímano	Burger Joint	Bakery	Fast Food Restaurant	Shopping Mall	Latin American Restaurant
3	Caricuao	Pharmacy	Bakery	Plaza	Sculpture Garden	Skate Park
4	Catedral	Coffee Shop	Historic Site	Plaza	Theater	Café

Figure 5: DataFrame built with the venues category information from each Parish in CMD

3. Methodology

Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.

To begin with the data explorative process, we decided to quickly plot some information to start getting insights from the data we recollected. First, we use the unsupervised learning algorithm *k-means* to try to make clusters of parishes based on the most popular categories of venue in that region. Parishes with less than 10 venues were excluded from this experiment. We ran the algorithm to get 5 clusters (k = 5), obtaining the distribution shown in Fig. 6.

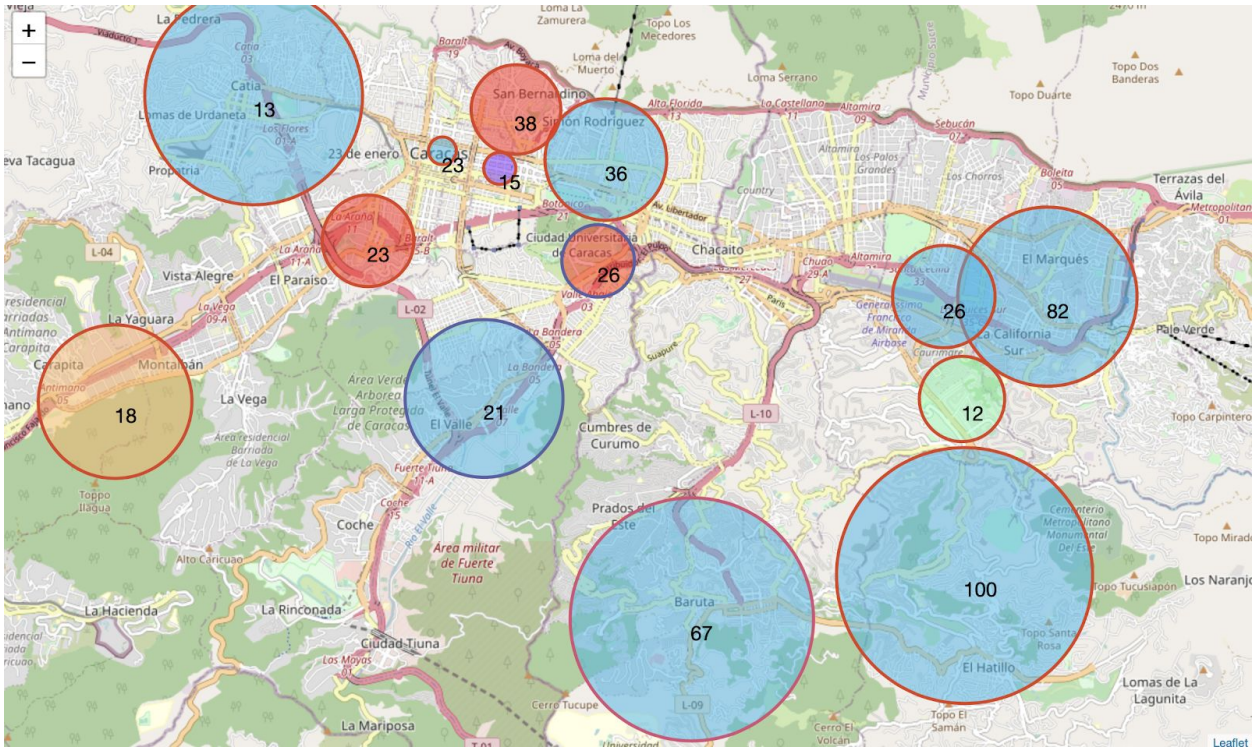

Figure 6: Clusters produced by k-means algorithm based on venues categories (k = 5).

Each color fill represents a cluster, as you can see in Fig.6, most of the parishes are colored in blue. Table 1 shows our interpretation of each cluster.

Number of parishes	Number of venues	Color	Categories with greater presence
1	15	Purple	Spanish restaurant
8	368	Blue	Bakery / Shopping Mall / Restaurants
3	87	Orange	Pizza place
1	18	Yellow	Restaurants
1	12	Green	Pharmacy, Gym

Table 1: Clusters description

4. Results

After excluding the parishes with less than 10 venues registered, we kept 15 out of 32 parishes, representing 47 % of the total parishes in CMD. It means that in more than half of the parishes in the CMD the information about venues is null.

Despite the lack of information about venues in CMD, We have found that there is a region on the south-west region of the map (see Fig 6.) with the highest venues presence in the FourSquare database. The clusters we constructed show that most of the venues are bakeries, shopping malls and some restaurants.

5. Discussion and Conclusions

We have recollected and explored geolocated data about venues in the CMD with the objective of getting insights about preferences of people who live and work there, and to find regions of interest to deploy a new web application prototype based on venues inside a city.

The data recollected to make this study is not enough to provide conclusions and convince stakeholders to invest their money. This work is an initial step of the entire process that we think should iterate starting from the data recollection phase, looking for more availables sources. However, if we had to make a decision at this point, there is evidence that our prototype should be implemented on the south-west region of the CDM because there is more available information and, apparently, people and businesses placed there have a notion about the importance of the online presence.

An interesting research direction would be to work only with the data from south-west parishes, trying to get insights about why there is more information available in this region and precise the interests of people who live there.