Capstone Project - Final Report

By Carlos Durazno

1. Introduction

1.1 Problem Description

Following the second suggestion, I will use the data from Foursquare to analyze the different types of businesses in the city of Cuenca, Ecuador. This is a small city, essentially a tourist destination due to its denomination as Cultural Heritage of Humanity, in addition to its beautiful Spanish colonial architecture. Cuenca is visited by a large number of people from different parts of the world. Of course, at this time all tourist activities are suspended by the coronavirus health crisis. But after that, I am very sure that there will be new kinds of opportunities available for this beautiful city to attract new visitors from the world. Therefore, the objective will be to find the different types of restaurants and food places by categories and show their ranking, and the areas where they are most located. With this information could guide to local investors to find new areas and places in the city to establish new business opportunities.

2. Data Section

2.1 Description of the data

The city of Cuenca is divided into neighborhoods, currently it has 15 neighborhoods, all with different land uses and established restaurants throughout the city. I will use the geographical location of the latitude and longitude of the neighborhoods to send inquiries to the Foursquare service to explore all the places in each neighborhood. After that, I will use the Foursquare ratings to determine wich categories are the best qualifyed and where are located in each neighborhood.

This will be great information to find out the preferred place in different neighborhoods and will serve restaurant owners to find possible new places to open branches.

2.2 Data source

The data to be used will be the list of neighborhoods "parroquias" in spanish, publicly available on the local government site of the "Municipalidad de Cuenca" located at http://ide.cuenca.gob.ec/geoportal-web/viewer.jsf in the section "Parroquias Urbanas". Using the Google Maps geocoding page https://developers.google.com/maps/documentation/geocoding/intro?hl=es-419. I will make a csv file that contains the name of the 15 neighborhoods, as well as the latitude and longitude of each one.

Table 1. List of "Parroquias" and geo location in Cuenca city.

	name	lat	lng
0	Bellavista	-2.881293	-79.005168
1	San Blas	-2.900793	-78.999998
2	Sucre	-2.901130	-79.015505
....			

With this information, I will send inquiries to the Foursquare service, using a function to explore each neighborhood in a 1000 meters radius, as we practice in the corresponding laboratories.

```
limit = 100

radius = 1000
```

url=https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&ll={},{}&v={}&radius={}&limit={}'.format(CLIENT_ID, CLIENT_SECRET, -2.881293, -79.005168, VERSION, radius, limit)

I will create a data frame to filter by categories and use all the places that include the words: restaurant, food, coffee shop, fast food, or something similar related to the food business. Once the information is filtered with related places, it will be necessary to present it on a map.

Continuing with the data analisys, once I get the venues related to food bussiness, will use again the Foursquare information to get the Rating value for each of the food venues. This will be accomplish with a new function to send inquiries to Foursqueare using the venue id and get the rating. Then it will be merged to the original food venues dataframe.

```
venue_id = '50784860e4b06b24d94043af'
```

```
url = 'https://api.foursquare.com/v2/venues/{}?client_id={}&client_secret={}&v={}'.format(
venue_id, CLIENT_ID, CLIENT_SECRET, VERSION)
```

Once that the rating is in the dataframe I will clustering the venues using the mean of the rating, with this information our investments will get a big picture about the best ranking places and the location of it on Cuenca, Ecuador. This cluster will be presented on a Map grouped by the different levels of the cluster with their respective rating.

3. Methodology

To analyze the venues around Cuenca city, it's was necessary to get the list of "parroquias" of the city. To obtain this I use the information available at the local government site list of "Parroquias Urbanas" wich mean neighborhoods.

The problem with this information is that there is only a list of the "parroquias" names and I have to use the google service to get the geo location coordinates for each "parroquia" name.

With this information I setup a list of names and coordinates in a csv file to use with the Foursquare service to get the venues around this.

Image 1. "Parroquias" Data Frame

```
cuenca_neighborhood.head()
```

	name	lat	lng
0	Bellavista	-2.881293	-79.005168
1	San Sebastián	-2.890083	-79.026364
2	El Batán	-2.896284	-79.033429
3	Yanuncay	-2.915857	-79.027777
4	Sucre	-2.901130	-79.015505

The next step in this project will be detect all the venues located in each "parroquia" of the Cuenca city, using a radius area of 1000 meters around the center of the geo location for each "parroquia".

For this purpose we use the Foursquare information and it will be required using the latitude and longitude, as well as the parameters of radius and number of venues limit as 100.

Once the venues are located, the second step will be sort all the venues related to food business. This will be done using a string filter for the words like: Restaurant, Joint, BBQ, Food, Place, Bar, Steakhouse, Pub and Bistro.

```
cuenca_food=cuenca_venues[cuenca_venues['Category'].str.contains('Rest
aurant|Joint|BBQ|Food|Place|Bar|Steakhouse|Pub|Bistro|Gastro')].reset_
index(drop=True)
```

The next step will be to obtain the rating of each food venue using the ID number previously obtained, so for accomplish this objective, will be necessary use a function that make a new query to Foursquare information.
The rating will be merged to a new data frame, in this new data frame will be applied a process to data cleaning and exploratory analysis to complete missing data from some venues that have not yet been rated using the mean of the others related places.

Image 2. Venues with id and rating

	Venue	Id	Rating
0	Picantes De Leo, Bellavista	50784860e4b06b24d94043af	7.4
1	Viko	505f554fe4b018ec7b9fd22b	n/a
2	El Rincon De Lucas	4e6eb68eb9933190ed2dd97c	8.2
3	Unagi Teppanyaki & sushibar	4f1b7702e4b08382322fb6c2	7.1
4	BQ Sport	4df304f6d4c01ff6b2eb0650	6.3

With this information I will create 5 clusters using the kmean function to group the places by the related rating. After that I can show the top rating venues, and his location in a folium map.

3.1 Data Clean

With the objective to obtain refine data, it was necessary to remove some duplicates venues, wich were added to the first data frame. This venues was duplicated because they are possible located in the border of the vicinity for the “parroquia” and could be located in two different queries.
After apply the drop duplicate function, we obtain a clear data that can be processed according to our needs.

After that, when I got the rating of the places, I found some places that had not been rated yet, to get a solution in this regard an average is applied using all the values that already get the rating value.

3.2 Data Analysis

Once the data frame is structured with the venue ids and respective rating, it was applied data clustering using the kmean function to the rating value. The kmean is a model of machine learning

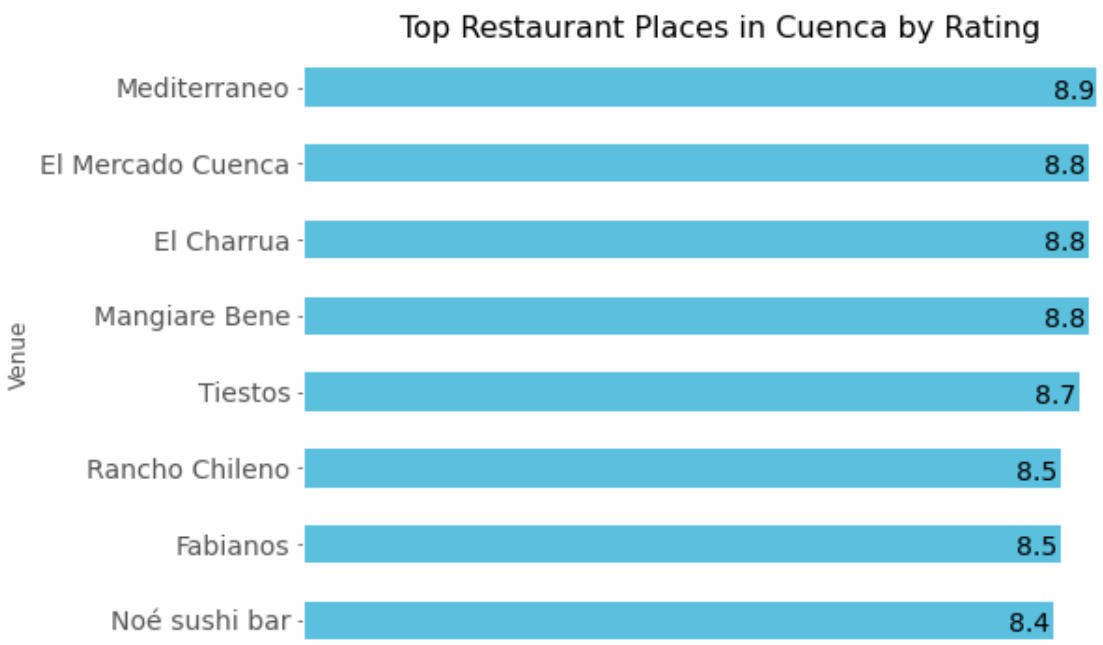
This help us to separate the top venues of the food business, enabling to generate a map for the top of venues in Cuenca city.

The next step will be to select the cluster that group the highest values, and with this elements show it on a folium map as well as a table ordered by rating column.

Whit this information is possible to create a new kind of graph that show the founded information in a different way.

First we created a graphic bar to show the best places in the city of Cuenca, so with this information we can see that the best place in the city of Cuenca is the Mediterranean restaurant, which has a rating of 8.9 according to the Foursquare rating.

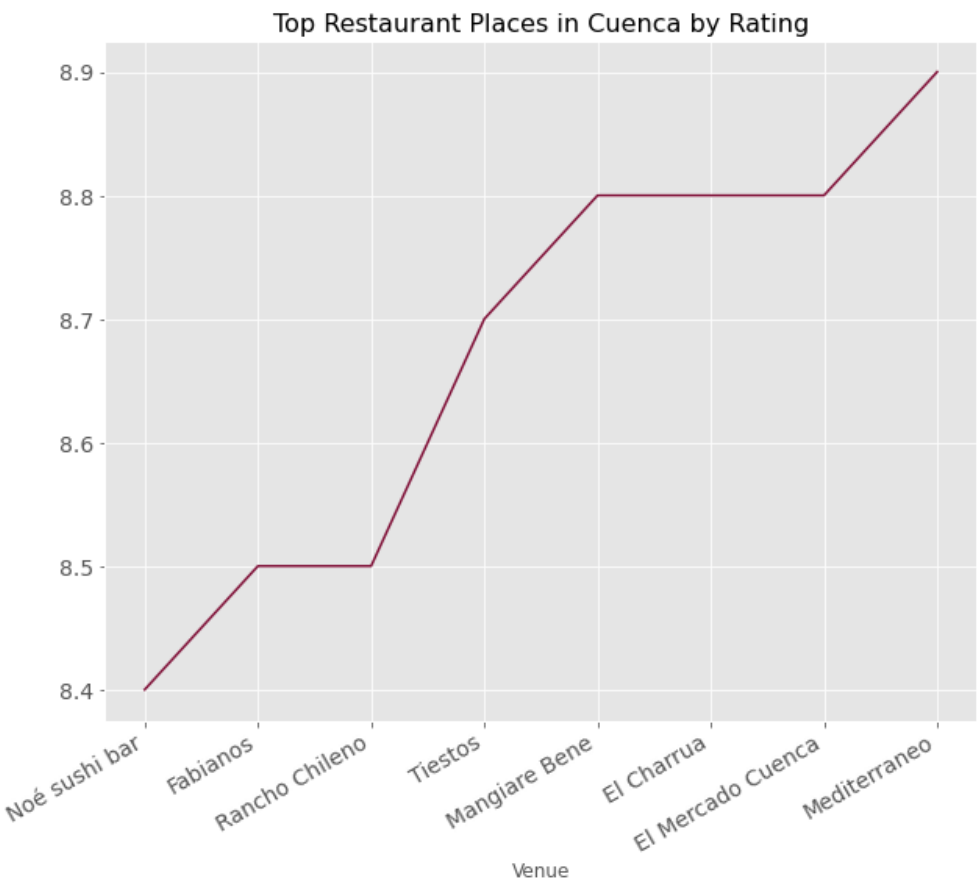
Graph 1. Top Restaurant Places in Cuenca by Rating



3.3 Results obtained

Then using the same data frame, a new graph is created, this time to show the scale of the best places of the Cuenca city, this kind of information would be important to know the level of acceptance of the venues, and what kind of places are at the moment in the city.

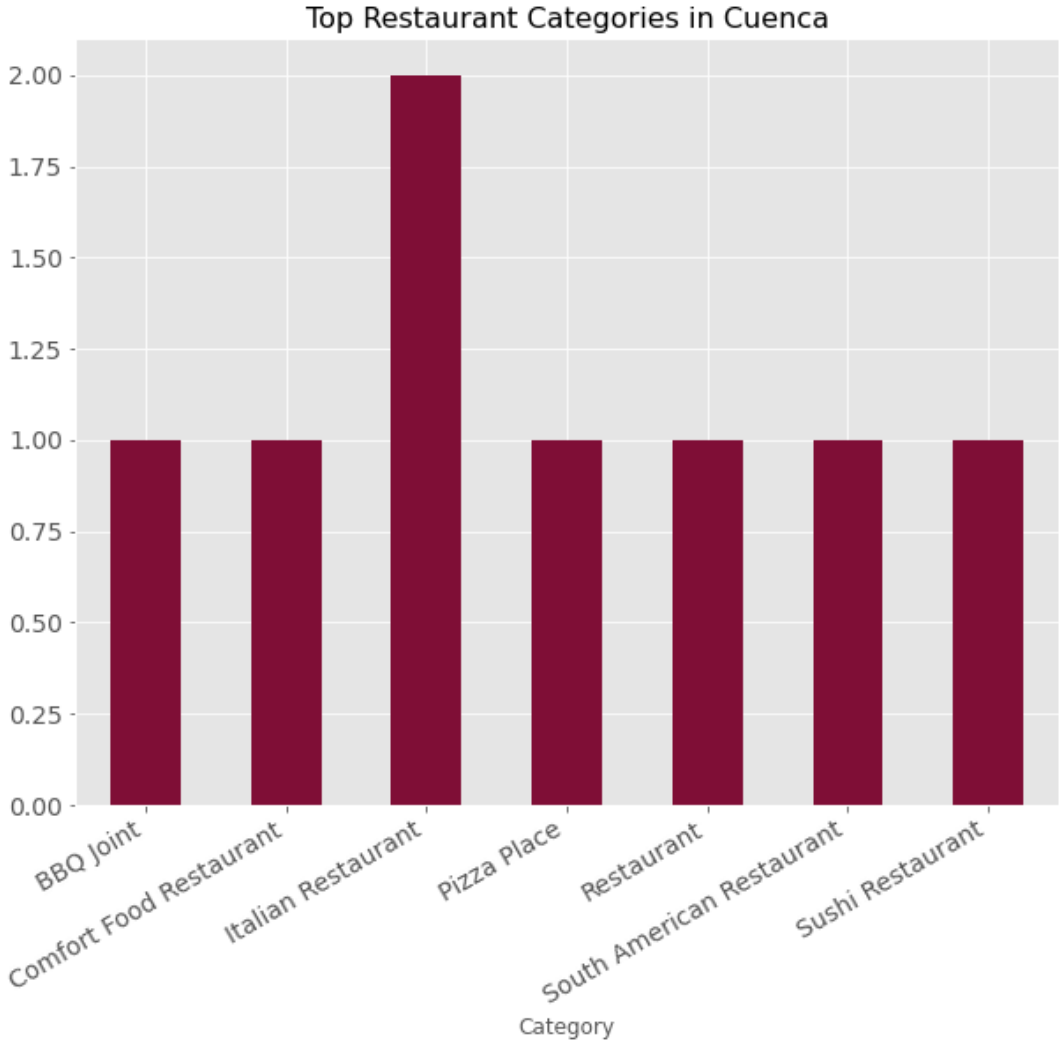
Graph 2. Top Restaurant Places Scale



As we can see the best rating goes from 8.4 to 8.9, so this is important information for an investor to know the restaurant environment and what type of place will be more accepted in the city.

Another aspect to analyze for this type of business is the categories of restaurants that we find in the highest classification. Then for this purpose a new data frame was created using the categories of the highest rated places in the city and of course this will be displayed using a bar graph.

Graph 3. Top Restaurant Categories



With this information, we can see that the Italian restaurant has two places in the ranking of the highest rated restaurants in Cuenca, so we now have an idea of what type of restaurant is the most accepted and the best in the city.

With this information, a potential investor can lean his decision on what type of business he invests.

4. Discussion

Based on the results, I can indicate that in the restaurant business in Cuenca, there is a great opportunity for investment and growth, since as can be seen in the data frameworks

obtained, there is a great diversity of food businesses that range from hot dogs joints to gourmet specialist restaurants, and all of them in a fairly acceptable rating range for a much larger startup or business establishment.

In the same way, this opens the doors for other necessary analyzes, such as the supply chains of different products and producers that can be part of the food business.

It is necessary to spend time analyzing other business opportunities based on the previous findings to structure a complete business opportunity.

5. Conclusion

The food business presents a wide business opportunity in the city of Cuenca, investors could find potential land for different types of food businesses, from small fast food places to gourmet restaurants.

The analysis of information through social networks is a fundamental source of valuable information, which provides us with knowledge about the preferences and predilections of users.

Knowing how to use this information will be essential for the adoption of new business and customer service techniques. Data science, data mining as well as machine learning are the tools that will guide the steps for the growth of businesses of all kinds. The establishment of departments and specialized areas in the management, obtaining and transformation of information will be fundamental for business decision making.

This work is a proof of concept of the need to analyze the information collected through the Foursquare social network, to face decision-making in the business world, as final work of the Capstone Applied Data Science course.