# **Data Science Capstone**

## **Venues in each neighborhood in Oaxaca City, Oaxaca, Mexico.**

**Introduction**

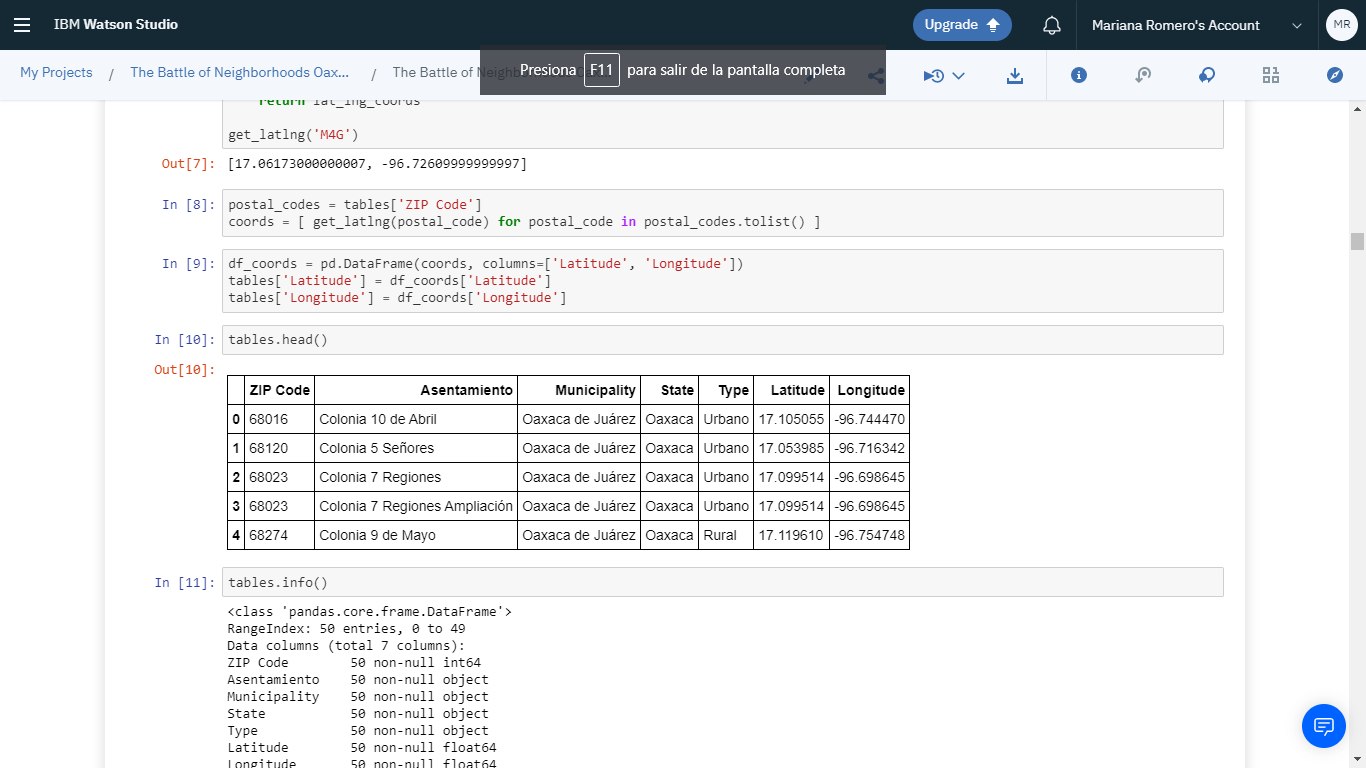
I am trying to recommend to an investor to run a business in Oaxaca, by opening a venue. Therefore, I must see what the most common venues around each neighborhood are.

**Data**

Luckily, I found a Data Base with all the postal codes in Oaxaca, included the neighborhood name (<https://codigo-postal.co/en-us/mexico/oaxaca/oaxaca-de-juarez/>).

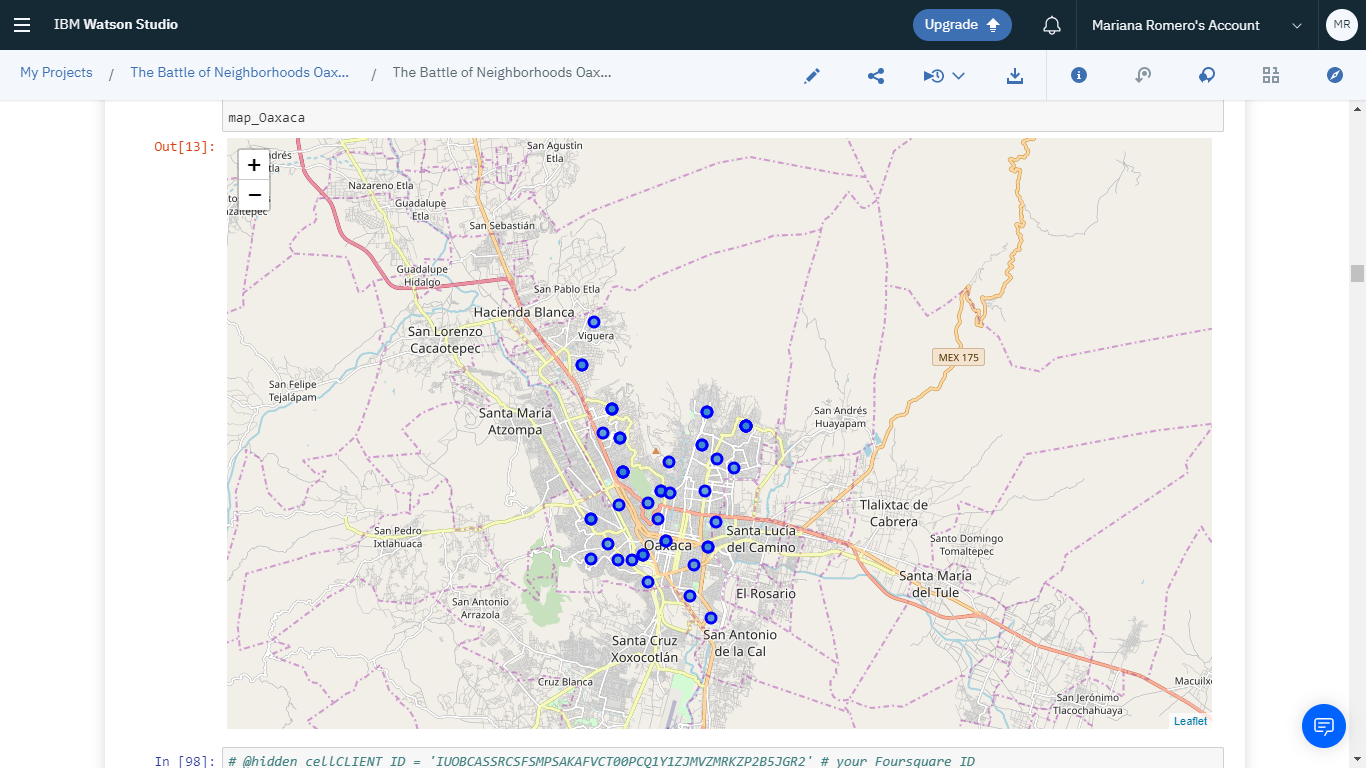
I used “pd.read\_html” to store the data in a pandas DataFrame, with 50 entries.

Then, I used “geocodert” to get the Latitude and Longitude of each neighborhood based on their postal code, in order to get the venues around each one using the data from Foursquare API.

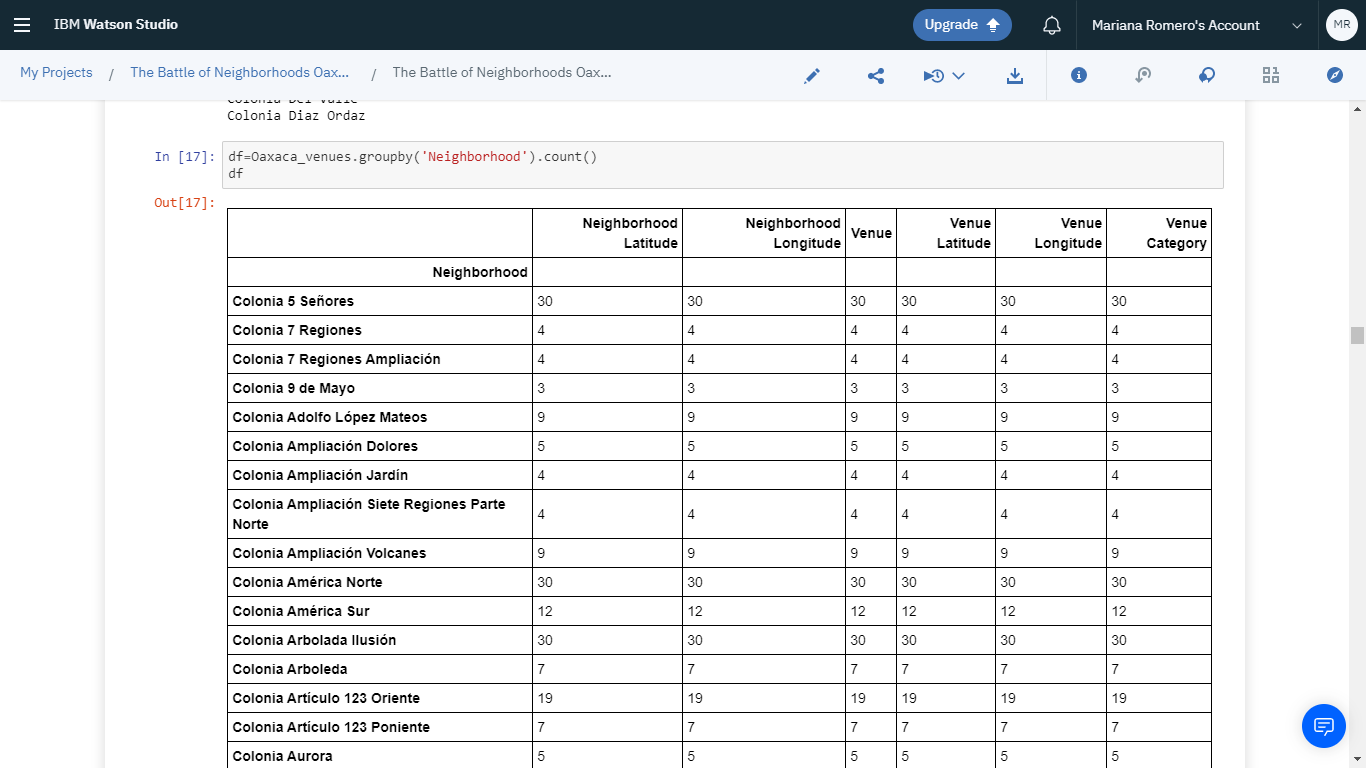


**Methodology**

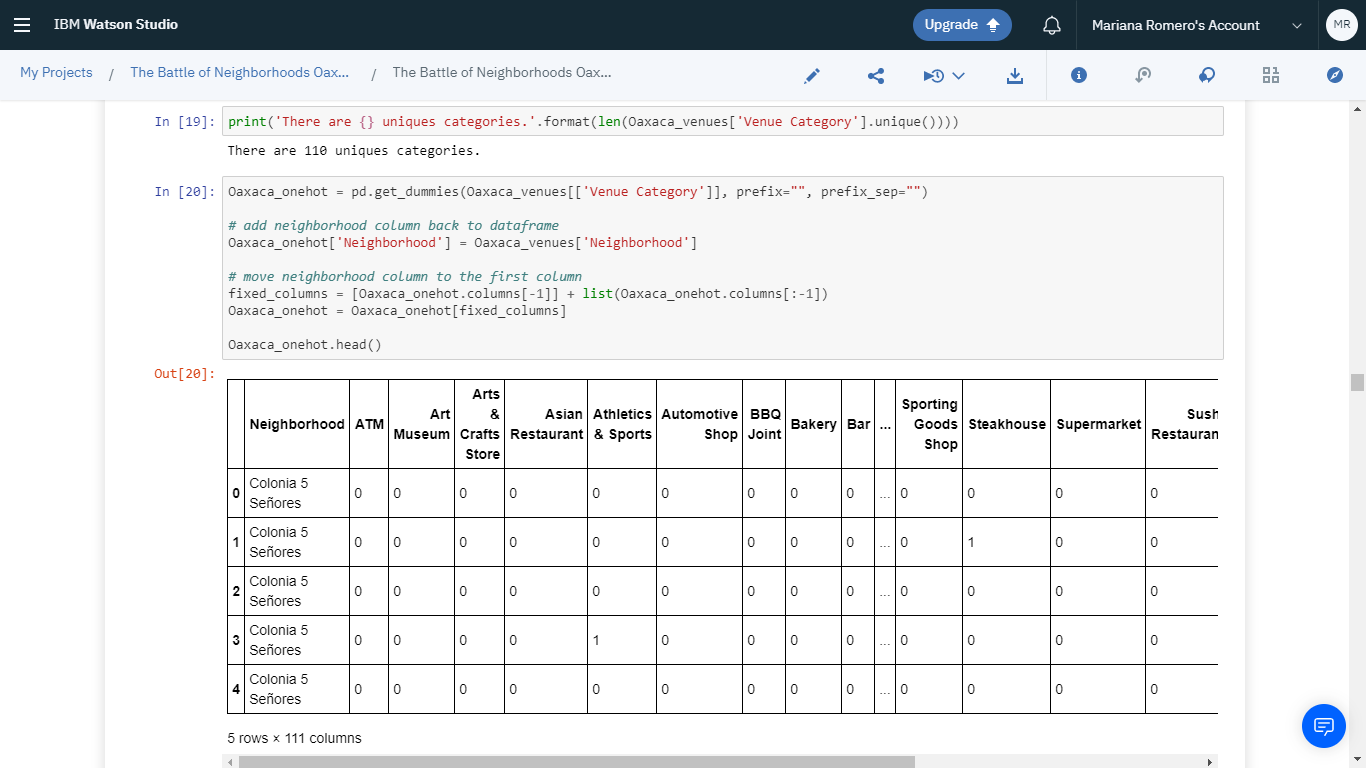
First, I used Folium package to plot my data into a real-world map and take a first look to them.

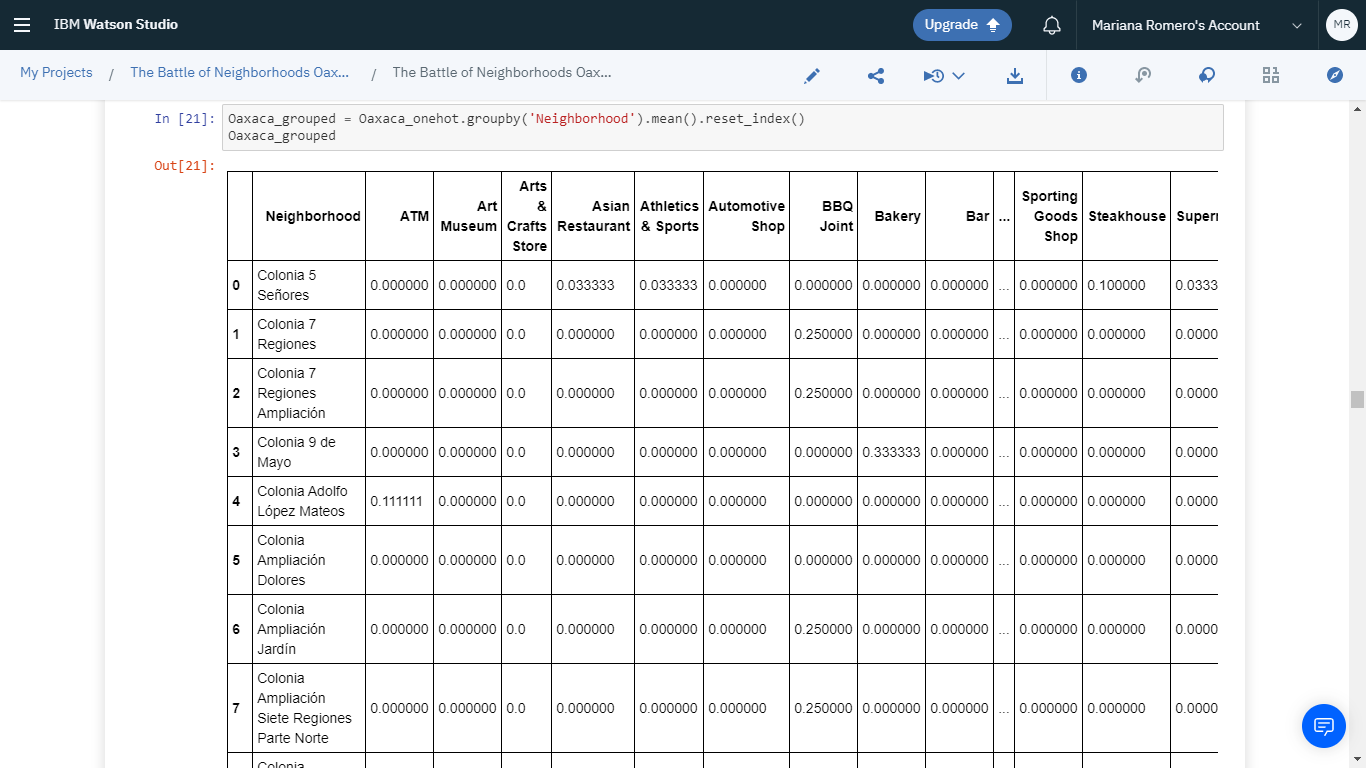


Afterwards, with a function I got the venues of each neighborhood in a 500 meters radius around each one. I realized, that there are no venues registered in foursquare around 5 neighborhoods. So, in excluded those neighborhoods from my analysis. Also, there are 110 unique venues in all our analysis.

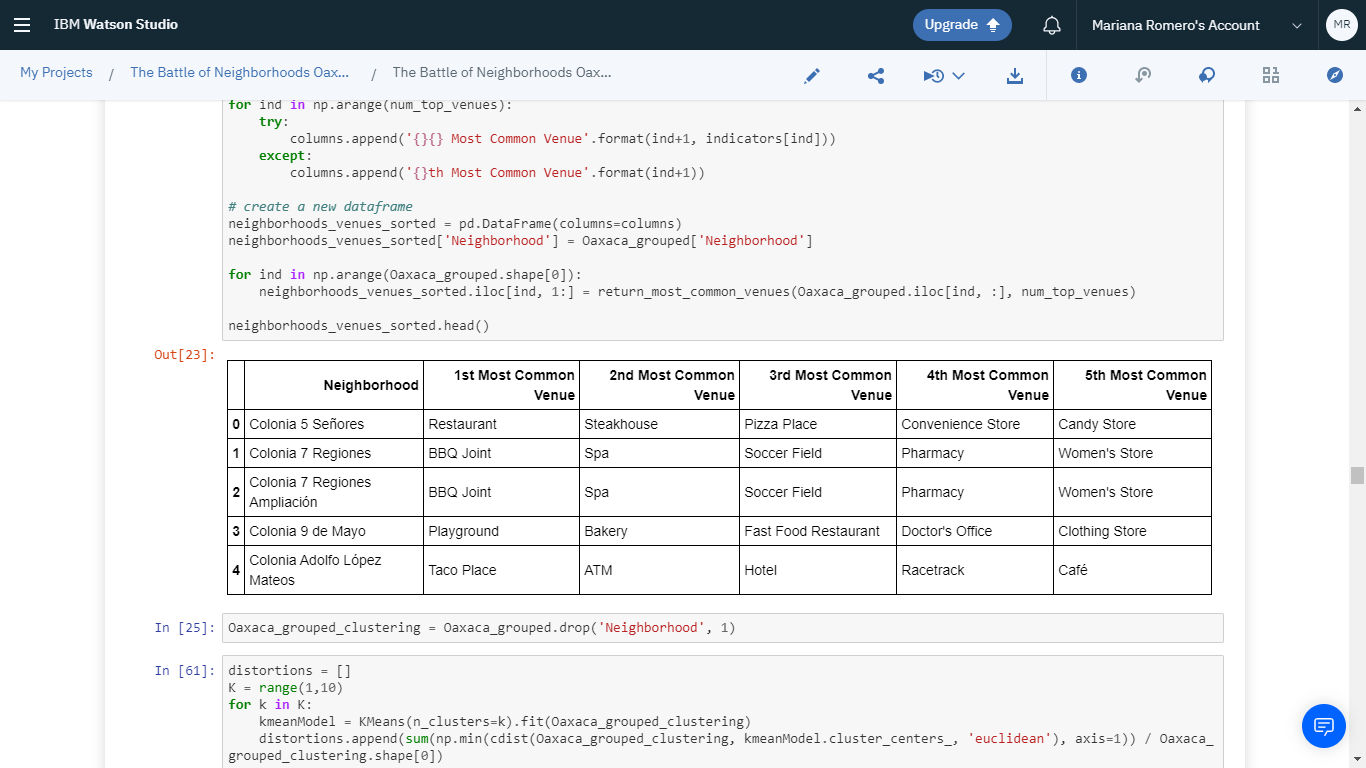


Then I use the panda function dummies to determine what venues of each category are in each neighborhood. And then calculated the mean of each neighborhood grouped by Venue Category, to calculate what percentage of venue category represents in the neighborhood.



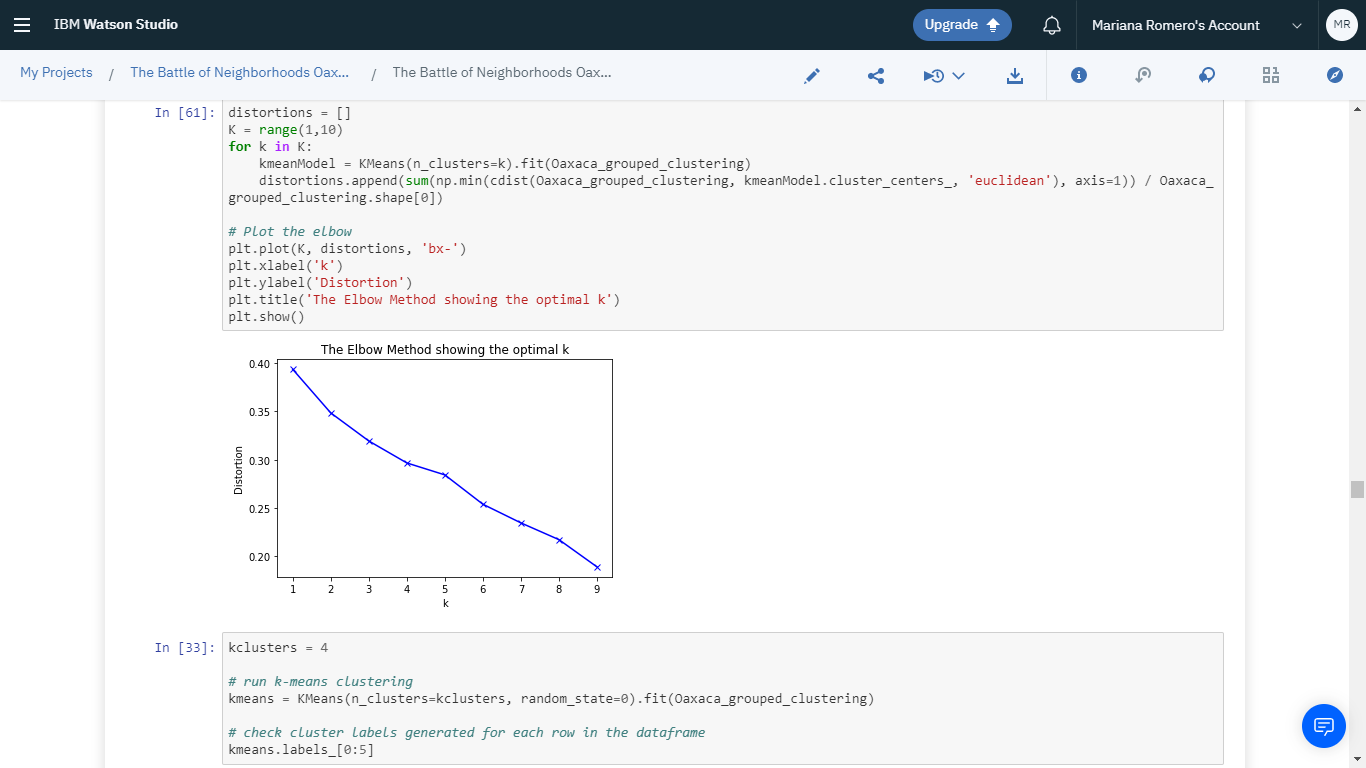


Here we can sort our dataframe by the most common venues. The result:



Here, it’s time to develop our Machine Learning algorithm. Based on our data, a K means clustering algorithm would be the best method.

To start, I would recommend computing the best K in order to minimize our error. I used the Elbow Method to show the optimal K.

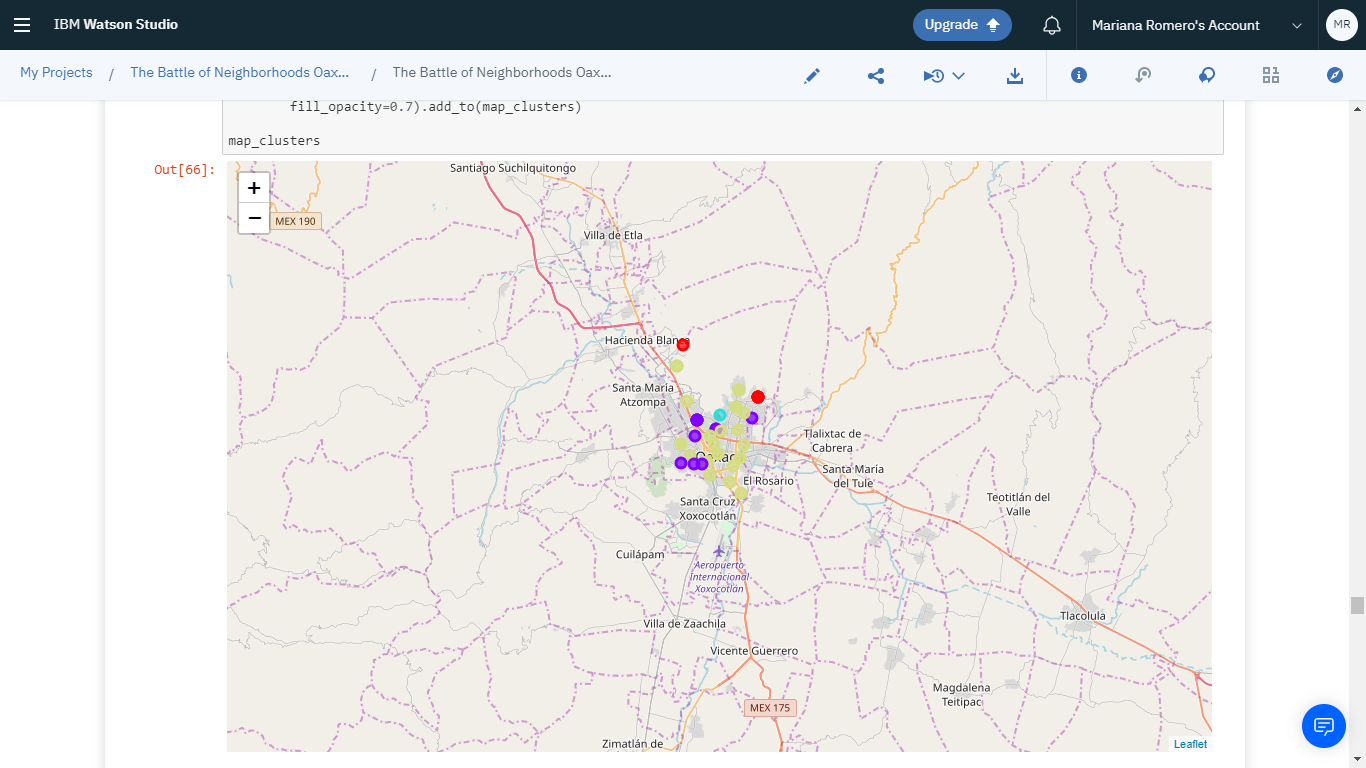


Based on this analysis, I used K=4.



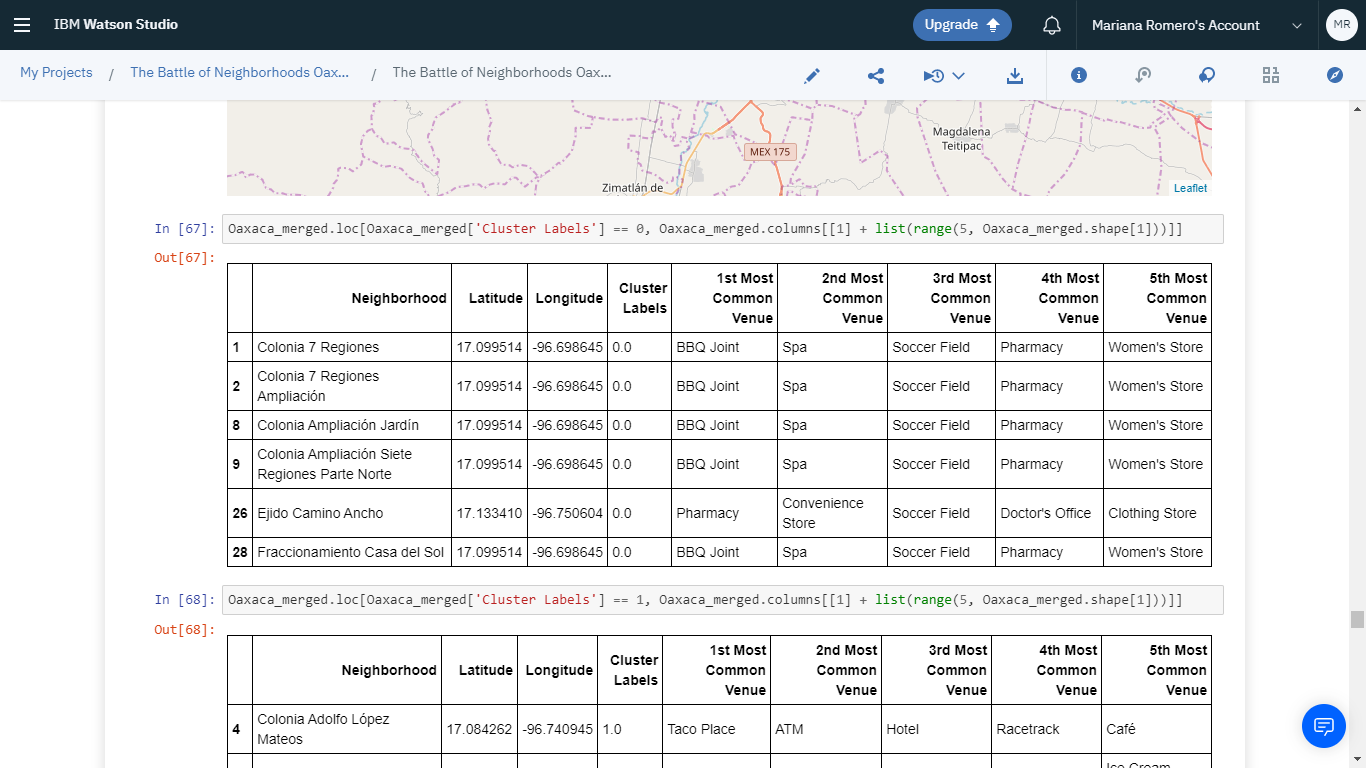
**Results**

To get a visual perspective, finally I plotted our dataframe in a folium map. The result:

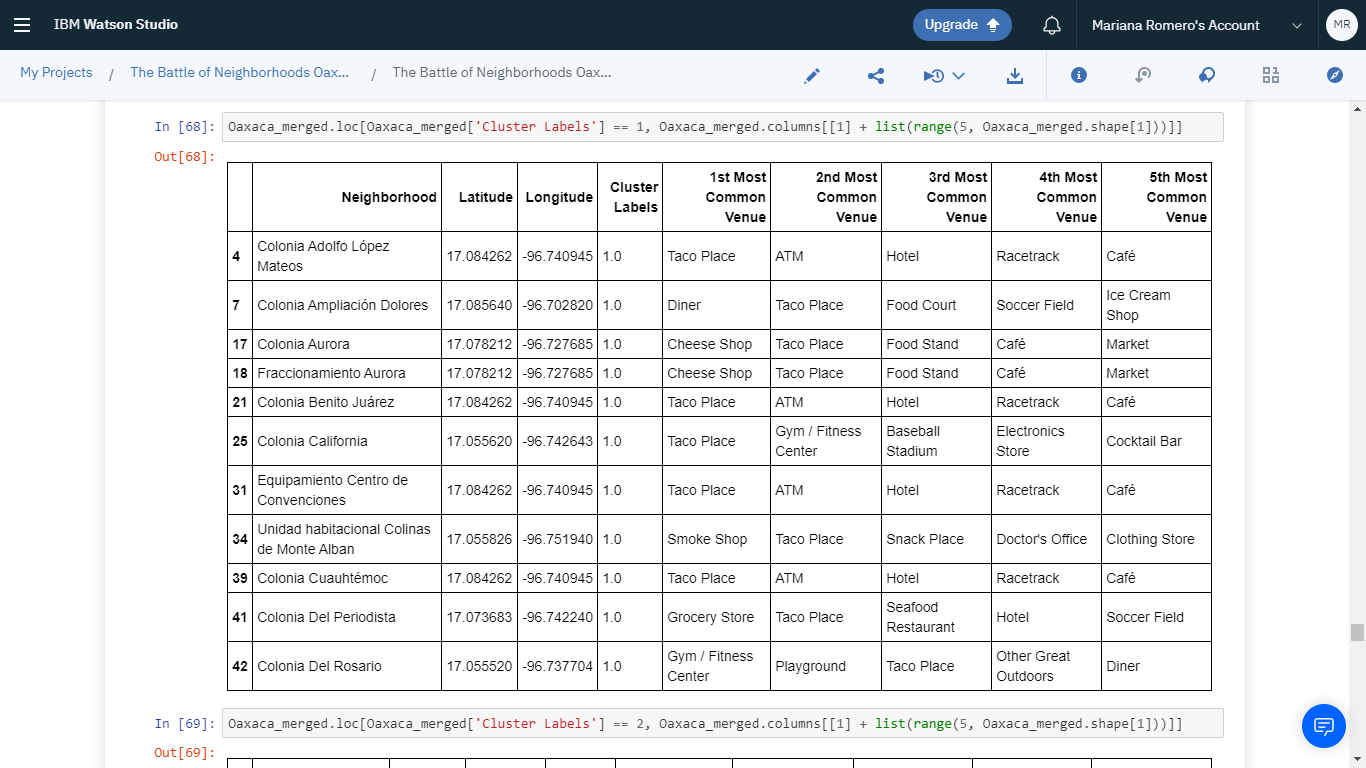


Finally, here are the most common venues, separated by clusters.

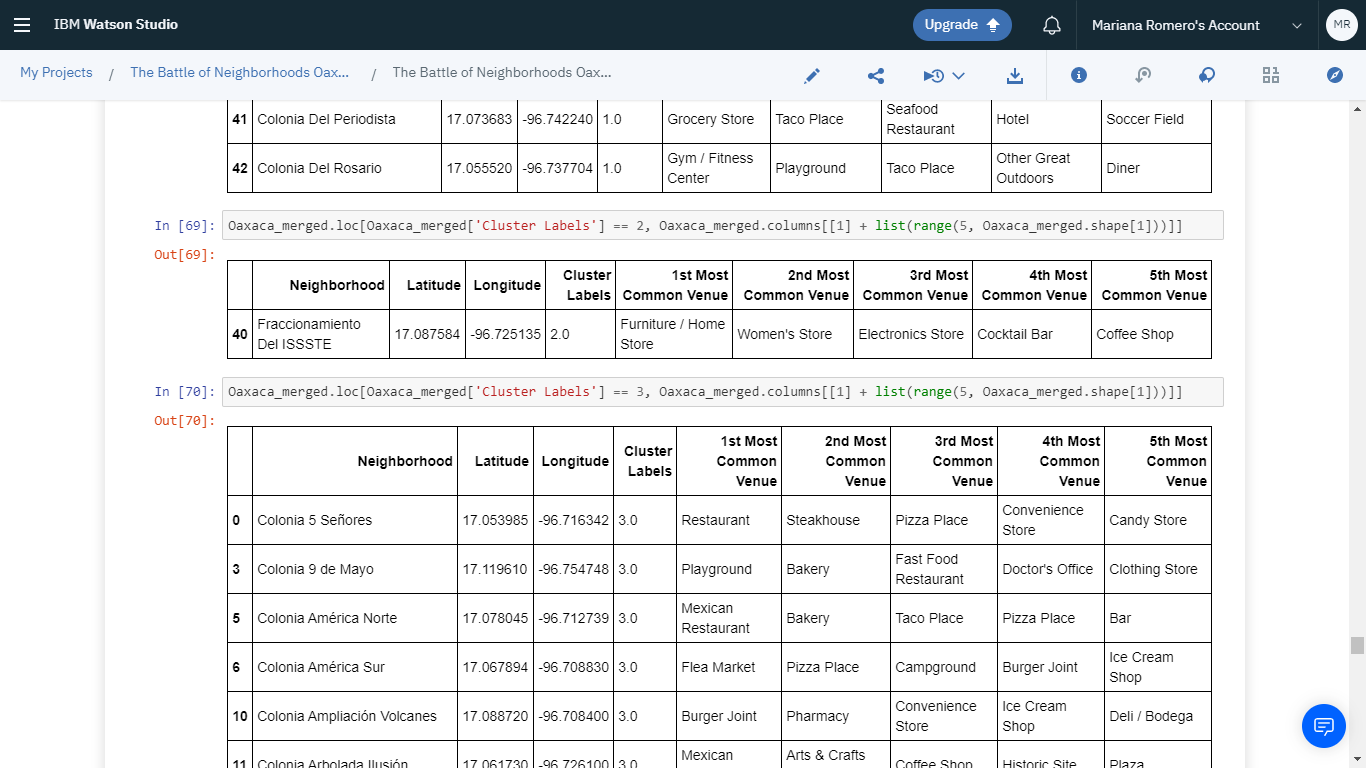
First Cluster:



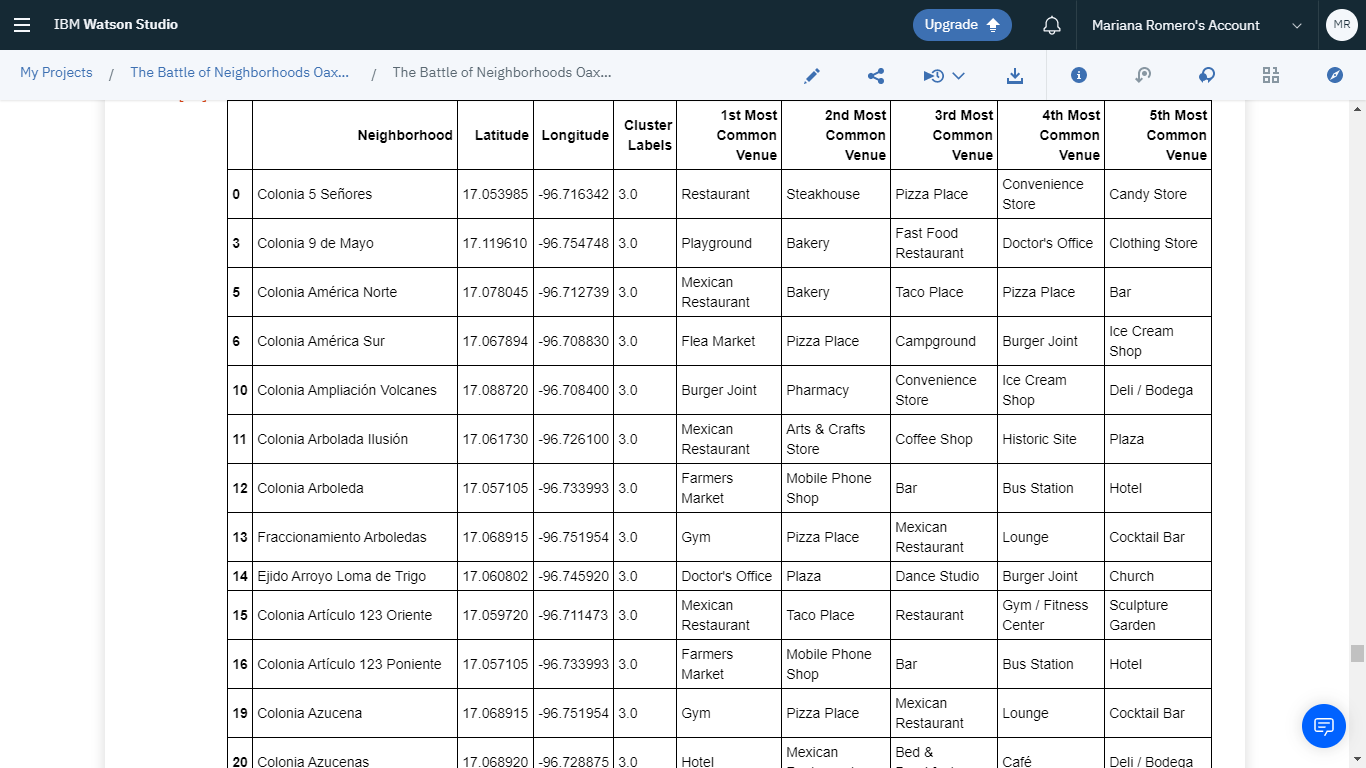
Second Cluster:

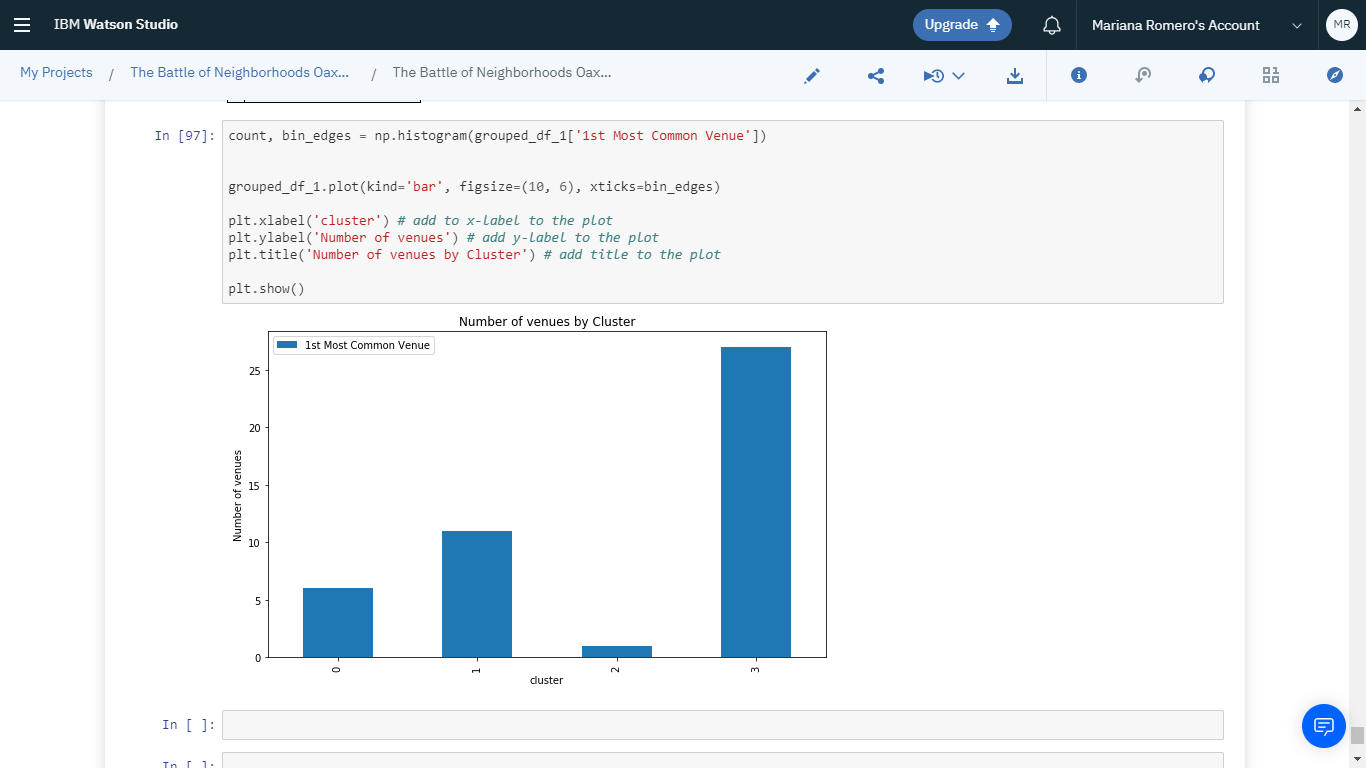


Third Cluster:



Forth Cluster:





**Discussion**

In Oaxaca there are 50 postal codes. And after I got the venues data from Foursquare API, here were only 45 neighborhoods. I believe this is very few data to get a good accuracy.

Actually, the Elbow method it helped us to tackle this error.

**Conclusion**

With that being said, I would recommend to my client, the investor, to open a new business based on the previous tables, that we got from our Cluster analysis. It depends on where the investor wants to open a business.