



# "CLUSTERING OF GASOLINE STATIONS NEIGHBORHOODS"

## DATA SCIENCE REPORT

Author: Jaime Axt

Date: May, 2020

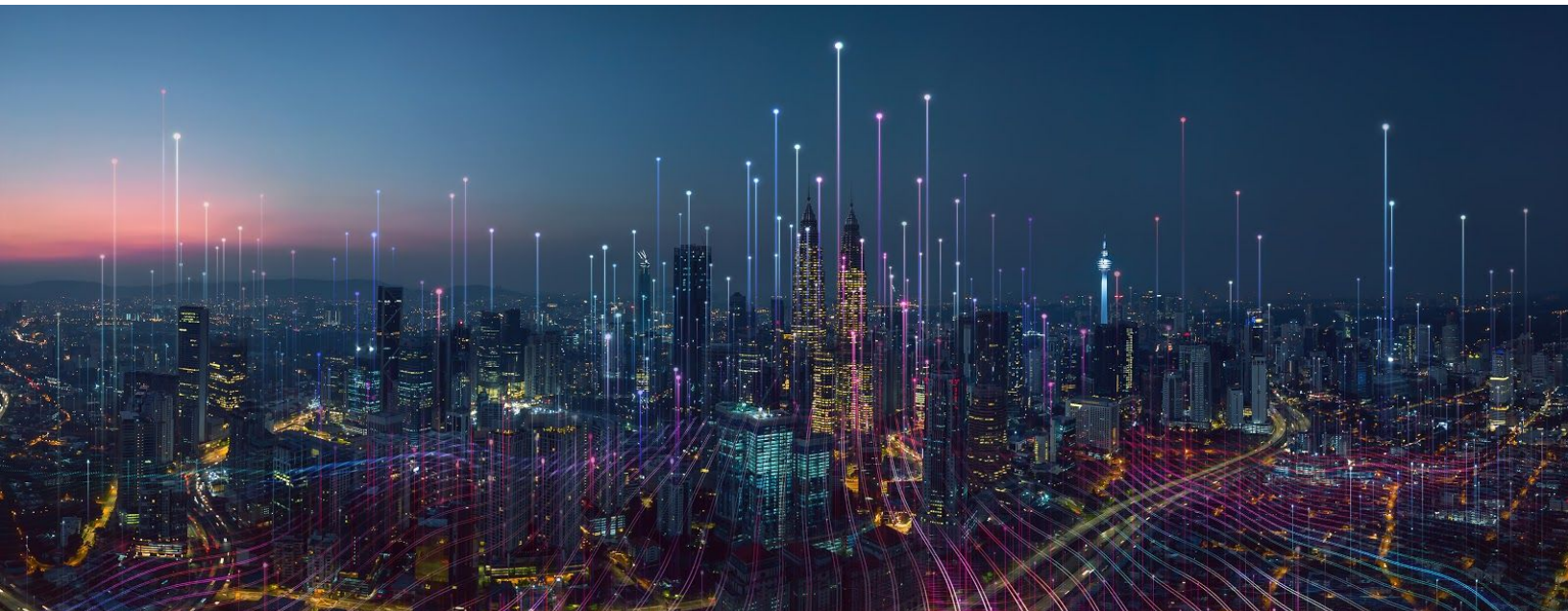


## Introduction

Every business that is successful selling their products in physical stores, at some moment, face this question: where can I open my next store? What neighborhood do I have to choose? Many times the stakeholders make decisions based on their experience or their intuition, and that is not a problem when the resources are unlimited, but the resources are always limited, so choosing a bad location could produce a financial disaster.

In this project, lets will analyze a group of gasoline stations in order identifying clusters of neighborhoods, so we could be able to identify what kind of neighborhoods represent the majors revenues to the company. Then, when the company makes a decision to open a new store, they will know what kind of neighborhood they have to search and choose.





## Data

The data we will use is the latitude and longitude of each gasoline station of an important company in Chile, obtained by API from the National Energy Commission. The data of neighborhoods like venues will be obtained from Foursquare. In addition, we will add information about population, household quantity, quantity of vehicles and quality of life index of each borough.

Data	Source	URL
Location of Gasoline Stations	National Energy Commission	<a href="https://api.cne.cl">https://api.cne.cl</a>
Venues of Neighborhoods	Foursquare	<a href="https://api.foursquare.com">https://api.foursquare.com</a>
Population and Household Quantity	National Institute of Statistics	<a href="http://www.censo2017.cl/descargue-aqui-resultados-de-comunas/">http://www.censo2017.cl/descargue-aqui-resultados-de-comunas/</a>
Quantity of Vehicles	Alberto Hurtado University	<a href="http://www.sectra.gob.cl/biblioteca/detalle1.asp">http://www.sectra.gob.cl/biblioteca/detalle1.asp</a>
Quality of Life Index	Chilean Chamber of Construction	<a href="https://www.cchc.cl/centrodeinformacion/archivos_detalle/icvu-2019-resumen-ejecutivo">https://www.cchc.cl/centrodeinformacion/archivos_detalle/icvu-2019-resumen-ejecutivo</a>



## Methodology

In first place, for this project I used a Jupyter Notebook on IBM Watson Studio.

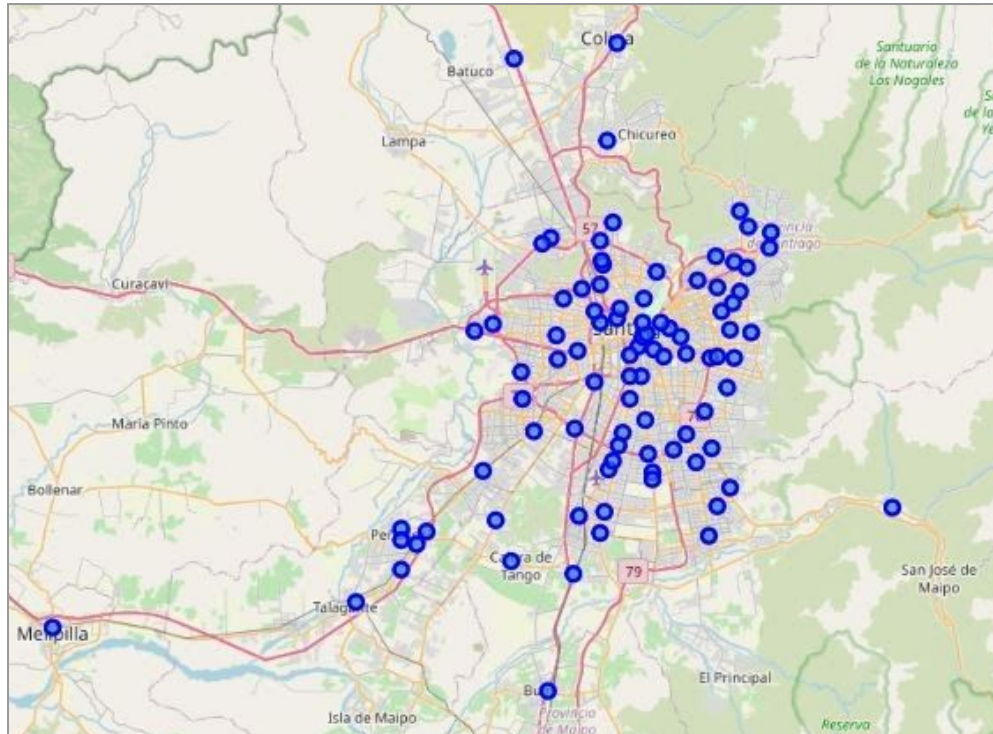
Because of the project I needed information about the geolocation, latitude and longitude, of each gasoline station in the region, I searched in Google and I found an API provided by the National Energy Commission. In a few steps I could get a token to execute some calls to the API and retrieve the information.

Then, I doubted if the information retrieved from Foursquare about venues was enough to do an accurate clusterization of the neighborhoods, so I decided to add information like population, household quantity, vehicles quantity and the quality of life index of each borough to improve the model.

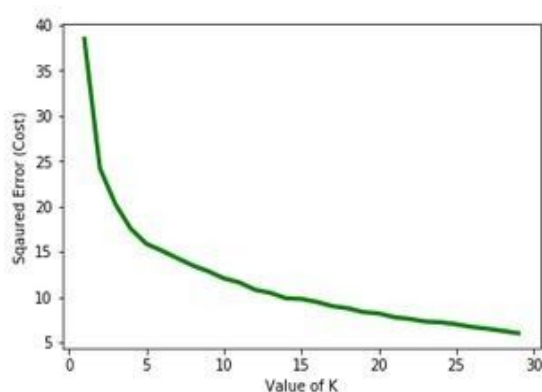
The data was wrangled using tools like Google Spreadsheet. Once the information was ready I uploaded the CSV file to IBM Watson Studio. There were two boroughs with missing data: San José de Maipo without information about quantity of vehicles, and quality of life index and Calera de Tango without information about quality of life index. To solve this, I used an average of quality of life index of all boroughs, and for quantity of vehicles I used an average of

vehicles per capita of all boroughs, multiplied by the population of San José de Maipo. Finally, the data of population, household quantity, vehicles quantity and the quality of life index were normalized using min-max normalization.

Using Folium we can see the gasoline stations before the clusterization:



In order to perform the clusterization, I used an unsupervised algorithm of machine learning named K-Means of Sklearn library, initially using  $K=5$ . To determine if  $K$  it was the right number I did an iteration increasing the value of  $K$  to obtain an elbow curve and decide which value of  $K$  to use. I decide to use  $K=6$ .

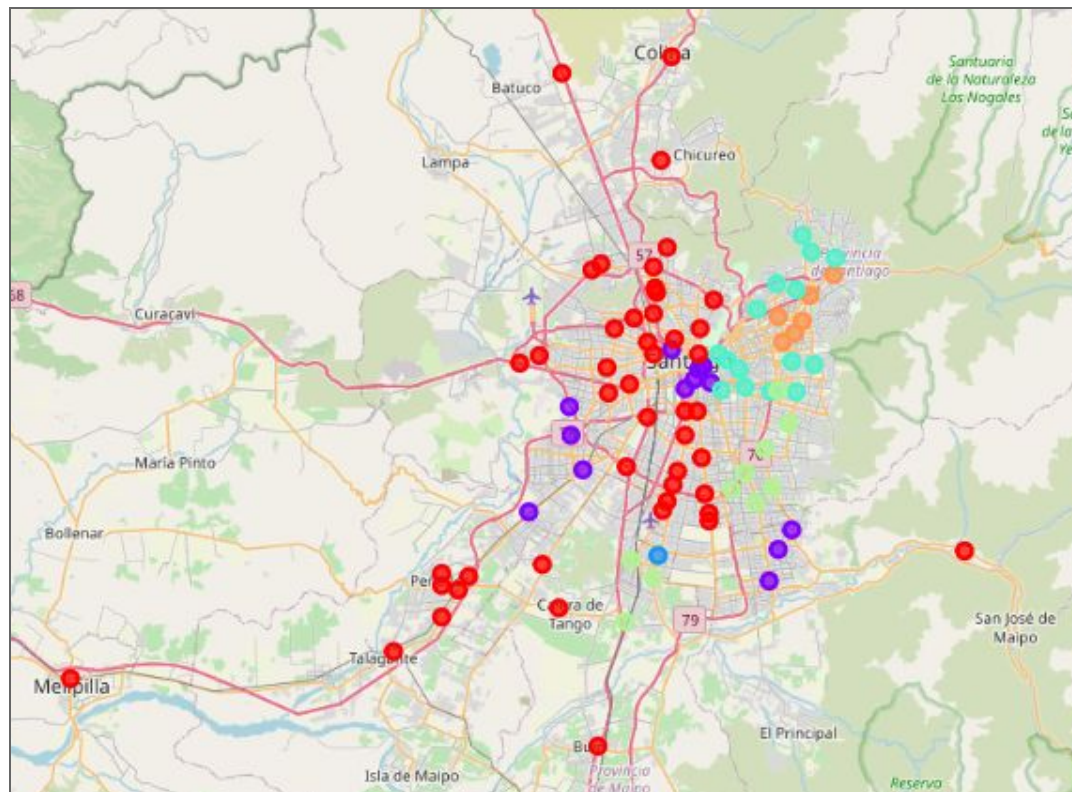






## Results

After the clusterization process we can see the gasoline stations and the clusters were belong:



We can identify 6 clusters of gasoline stations in the metropolitan region, one of them with just one gasoline station. The size of each cluster is:

Cluster Label	Number of Gasoline Stations
0	47
1	13
2	1
3	15
4	10
5	6

We can see that in the east zone, a zone known by its purchasing power, a better quality of life index and high quantity of vehicles per capita there are the clusters 3, 4 and 5. In the west zone, a zone known by a lower purchasing power, high population, and worst quality of life index, we can note there are clusters 0 and 1.

In the north zone we can identify just the cluster 0. In the south zone, a zone that has a low quantity of vehicles per capita, there are more clusters like 0, 1, 2 and 4.

Within the Américo Vespucio beltway, there are clusters 0, 1 and 3. In the center, a zone known for its high population, we can see just the cluster 1.

Considering data like population, household quantity, quantity of vehicles and quality of life index we can say that clusters have this characteristics:

Cluster Label	Population	Household quantity	Quantity of vehicles	Quality of life index
0	Low	Low	Low	Medium
1	High	High	Medium	Medium
2	Medium	Medium	Low	Medium
3	Low	Low	Medium	High
4	Medium	Medium	Medium	Medium
5	Medium	Medium	High	High

In the other hand, the most common venues of each cluster are:

Cluster Label	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Food	Flea Market	Flower Shop
1	Pharmacy	Sushi Restaurant	Bakery
2	Asian Restaurant	Yoga Studio	Donut Shop
3	Sushi Restaurant	Café	Gym
4	Food	Bakery	Pharmacy
5	Sandwich Place	Pizza Place	Plaza



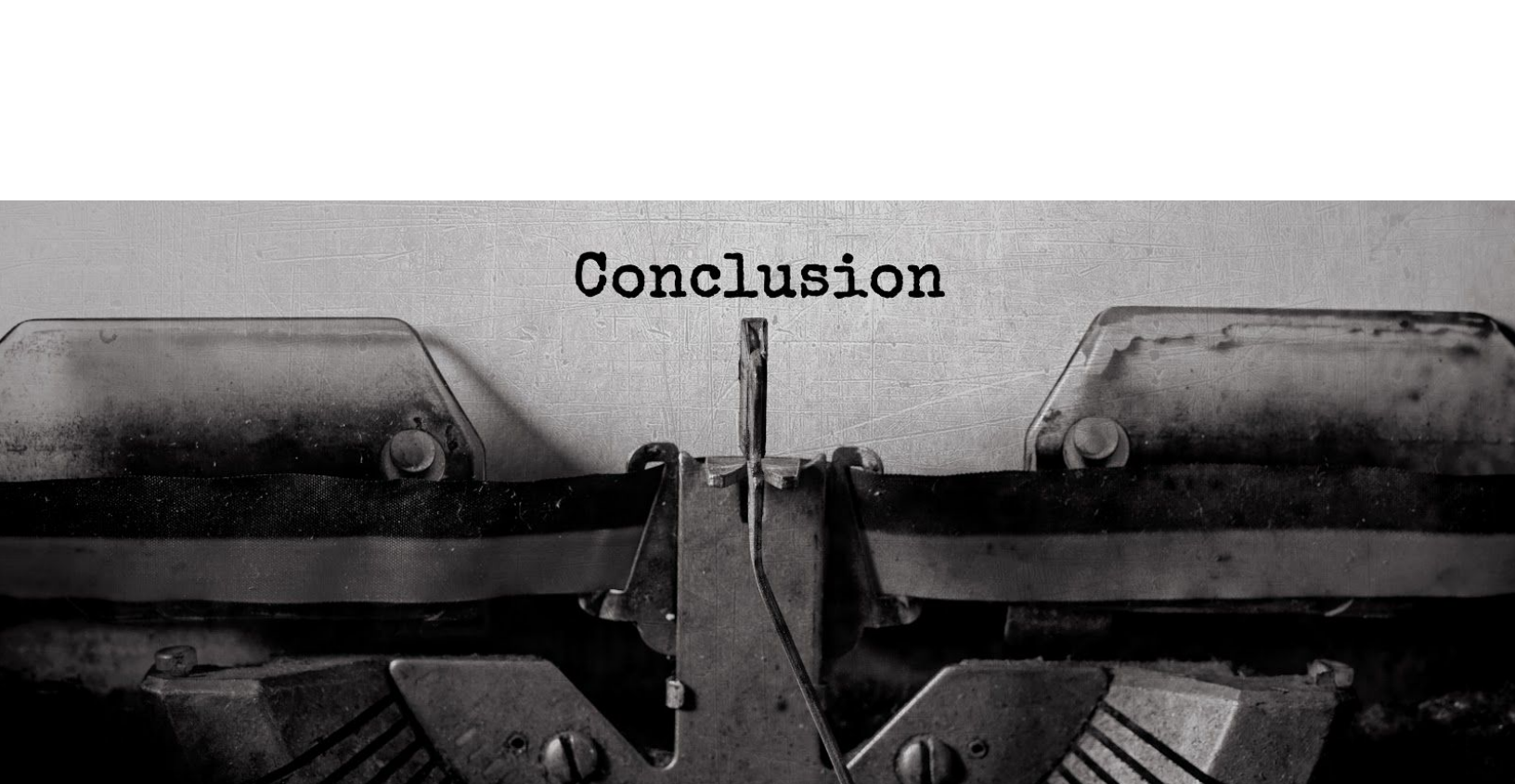


## Discussion

The principal goal of this project was to identify clusters of gasoline stations based on the characteristics of neighborhoods where they are located, in order to make recommendations to the customer who wants to open a new gasoline station in the metropolitan region.

This model will help companies to evaluate all the candidate places, considering the characteristics of the neighborhoods where the gasoline stations with best performance are located. For example, let's imagine that the gasoline station with the highest revenue is in cluster 5. With this information, we can advise the customer to open a gasoline station in the neighborhoods of cluster 5 or in another neighborhood with similar characteristics, in this case a neighborhood with medium population, medium quantity of households, high quantity of vehicles per capita, high index of quality of life and near to Sandwich Place, Pizza Places and Plazas.

I think this model can improve its performance by adding more information about population age, education, laboral situation by block of neighborhood, public transportation information, traffic zones, and so on.



# Conclusion

## Conclusion

Considering that the model performs a good clusterization result based on the data selected for the analysis, I think the model can be improved a lot using more detailed information like population by age, education, occupation by neighborhood blocks, and maybe using another service to identify places like bus stations, subway stations, medical centers, colleges, and so on. It would be interesting to add information about vehicles traffic and pedestrian traffic too.