Clustering the new Aconcagua Region

Juan Pablo Arancibia

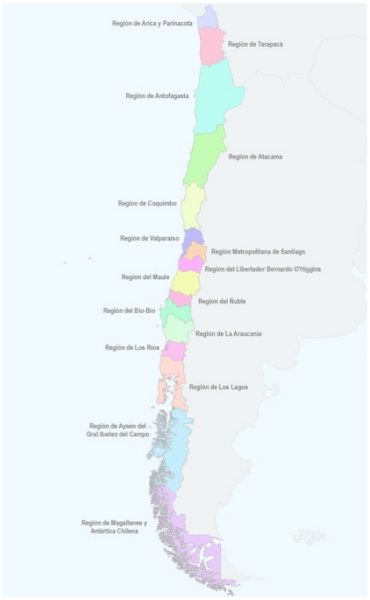
June 18, 2020

Introduction

1.1 Background

Chile is actually divided in 17 regions. All these regions have different names depending of important zones in those locations. Usually has the name of its Capitals.

1. Tarapacá
2. Antofagasta
3. Atacama
4. Coquimbo
5. Valparaíso
6. Libertador General Bernando O'Higgins
7. Maule
8. Bío-Bío
9. Araucanía
10. Lagos
11. Aysén
12. Magallanes y Antártica Chilena
13. Ñuble
14. Region Metropolitana
15. Los Ríos
16. Arica y Parinacota
17. Ñuble



In this analysis we'll focus in the Valparaíso Region. There has been an idea to group of persons to separate the region in 2 different regions. In other words, a new region. This would be an option to change the reality of many people of some cities. In all these years Valparaíso (Capital of the region) and others coastal cities has been the most benefited in Public Works, Universities, the better hospitals, etc. It is not fair, because the “back cities” (as the people call them) contribute with the most important part in the regional GDP. This is because these cities have mines and agriculture.

In the last report of the *'Subsecretaría de Desarrollo Regional y Administrativo'*, 3 territorial alternatives were disclosed: "Metropolitano", "Rural", "Histórica-Identitaria". I will take the "Historica-Identitaria" alternative since is the best option according the report.

1.2 Business Problem

In this scenario is indispensable make an analysis using machine learning tools for a visualization of different cases that we can identify. This project could assist to the "Corporación Region Aconcagua" to make effective decisions in another plane. I have mention and highlight that this project focuses in most part in a touristic view (later explained). The business problem that we are currently posing is: how could we identify if the Aconcagua region can be a new region and why? To solve this problem, we will cluster the cities of the region and make different combinations in order to recommend all venues that should has the communes for new complete regions.

2. Data

2.1 Source

The data of the communes in Valparaiso Region were scraped from Wikipedia page (https://es.wikipedia.org/wiki/Regi%C3%B3n_de_Valpara%C3%ADso).

2.2 Data cleaning

The table with the data was manipulated for clean all columns because it was very noise and dirty. For other hand the Valparaíso Region also has within its cities to the "Isla de Pascua" and “Isla de Juan Fernández”. These islands were ignored for the project, because they weren’t relevant in this analysis.
The latitudes and longitudes of each city were obtained from **geopy** with **Nominatim** library. Two communes had false coordinates so I have to change them, Llay- Llay (32.841742, -70.953908) and Rinconada (-32.833329, -70.688845). The table with this data was exported in a .csv file (V_region.csv). Link in the GitHub repository.
I worked with this data, but later, when I got the venues in each commune, I realized that the communes of Panquehue and Santo Domingo had less than 3 venues. So, I remove them from the DataFrame.

2.3 Foursquare API

I used the coordinates to get the location of all venues from the Foursquare API and merged with the principal DataFrame with a limit of 100 venues for each commune. In Chile the Foursquare app is not very popular because the most parts of venues are restaurants, hotels and tourist places. That is why this analysis was tilt, for the most part, in a tourist way.
After data cleaning, there were 760 venues and 8 features for start the analysis. The DataFrame was made up with the following columns:

- Provincia
- Comuna
- Comuna Latitude
- Comuna Longitude
- Venue
- Venue Latitude
- Venue Longitude
- Venue Category

3. Methodology and Exploratory Data Analysis

3.1 Number of venues.

Each venue was represented by one row in the dataframe and to make a comparison between the differences in quantities, I build another dataframe named “region_grouped”. In this dataframe I could show the total of venues for each commune.
Previously explained, I mentioned that the coastal cities are the most beneficiated with the best places, but also, Valparaíso is the city with the largest number of venues and the reason is simple, Valparaíso has the most important port in the country, also has the largest population and is the capital of the region. Then follows Viña del Mar that is a connected city with Valparaíso. After, the number of venues is decreasing when the radius is expanded from the capital. The next cities are Quilpué, Los Andes, Villa Almena. These communes are the next in the ranking since they can be considered as “big cities” (<39, >60). After follows Concón, El Quisco, El Tabo, Papudo, Quillota, San Antonio, San Felipe and Algarrobo (<20, >40). And then, all communes less than 20 venues, like Zapallar, Santa María, Rinconada, and so on.

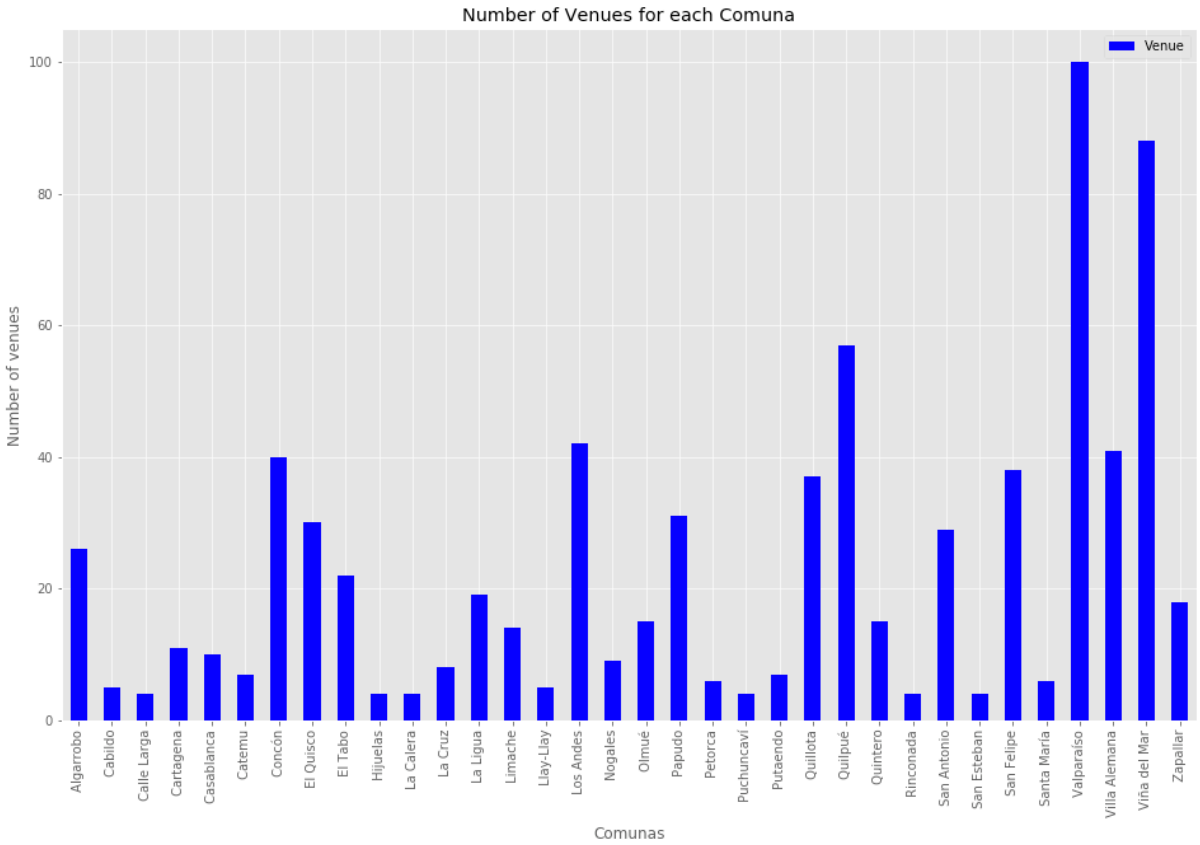


Figure 1. Rank between communes and number of venues in the Valparaíso Region.

3.2 Sorted the most commons venues.

Until now, we know the biggest cities, smallest cities and its number of venues. But also, with this data we could know the most common category venues in the communes in a ranking of 4. Our dataframe was manipulated to obtain the one-hot encoding and grouped with the mean of each commune. Now, we were able to sort the results and create a new dataframe with the most common categories in each city. This was an important part to analyze the behavior of each city and take a look at the differences between them.

3.3 Cluster the complete region.

The algorithm to use was k-means because fitted properly in this scenario. I needed a method to allowing me group and determinate which kind of cities the region might have and what would happen with this type of cities when the region would be separated. So, for to obtain the clusters I used the previously one-hot encoding dataframe.
I used the elbow method to graphic the best number of clusters to use (in a range of 1 to 10).

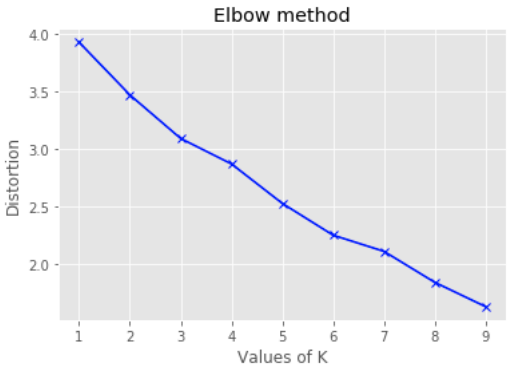


Figure 2. Elbow method for all region.

This elbow method wasn't a clear slope. So, I implemented a Silhouette function alternative to find the number of clusters. I used the mean Silhouette Coefficient of all samples to determinate the number to choose. The number was 2, with the higher average (~ 0.26) in a range to 1 of 10.
In this clustering I observed that within cluster 0 were mostly communes of the Aconcagua Region. These communes are rural places and peripheric to other cities inside of cluster 1. Also, I realize that the communes were in a range of 4 to 7 venues. So, I can say that this cluster 0 was made up for the communes with a more rural life in comparison to the communes in the cluster 1.
I named these cluster as:

- Cluster 0: Multiple rural places.
- Cluster 1: Multiple city places.

At this point it is observed that already are huge differences between the communes of the Aconcagua region and the Valparaíso region.

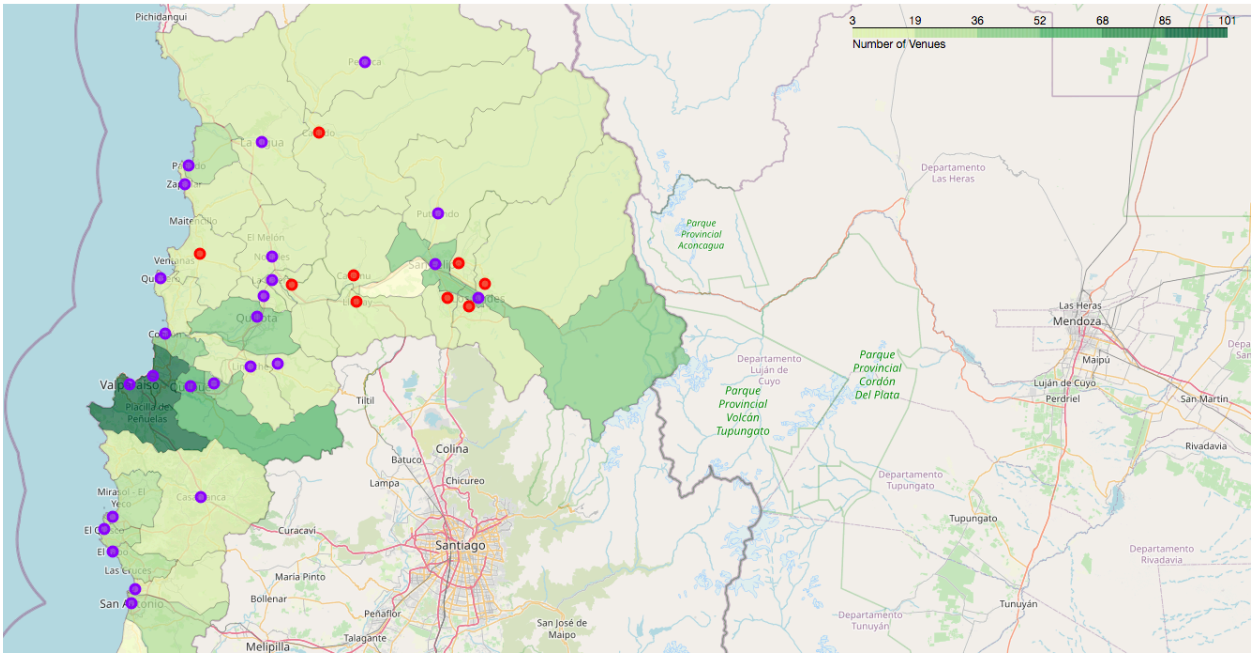


Figure 3. Valparaíso Region with 2 clusters in a choropleth map. Red: Multiple rural venues. Purple: Multiple city venues.

3.4 Separate the region

Now in this step I start to implement the separation of the region that is the purpose of this project. The Aconcagua region had **196** venues and for the other side, the Valparaíso region with **564** venues. The difference is very high, this affirms the lean to the Valparaiso communes.
Then I reimplement the one-hot encoding to obtain the most common venues in the regions.

3.5 Aconcagua Region

The Aconcagua region will be conformed for the next communes:

- Calle Larga
- Los Andes
- Rinconada
- San Esteban
- Cabildo
- La Ligua
- Papudo
- Petorca
- Zapallar
- Catemu
- Llay-Llay
- Putaendo
- San Felipe
- Santa María

Again, with the K-means algorithm I needed obtain the number of clusters to separate the communes inside the Aconcagua region, so I implemented the elbow method.

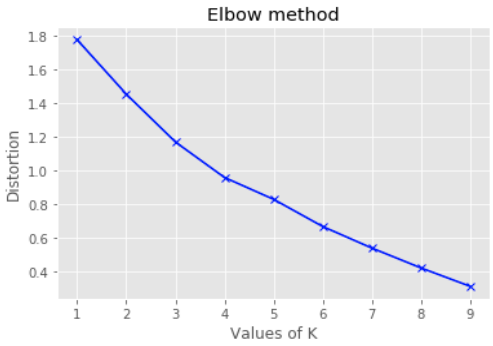


Figure 4. Aconcagua elbow method.

But this elbow method is worse than the first, since there is no clear difference between each commune, so I had to obtain the mean of Silhouette Coefficient to choose the correct number of clusters that was 4 with a score of ~0.15.

The cluster 0 was 3 communes with restaurants and a particular bus station. Coincidently these communes are next to each other, so, I can determine that the communes have a relationship between them. These communes are the entrance to the V region in the north side. Also, they are popular for its rural life and they are next to other communes with more population.

Within **cluster 1** had the largest group of the region, that also had a clear relation with the more touristic communes. These cities can call them "big cities" in the Aconcagua Region, since they concentrate the largest population. The relationship between the communes were the variety in food venues.

In the **cluster 2** was the Rinconada commune that concentrates different categories and its characteristic is that this commune has a casino, this particular venue is not explicitly in the common venues but have a close relation.

And finally, the **cluster 3** had coincidently two communes that they are the only two "doors communes" to enter and exit to Santiago (Metropolitan Region). These share food & drink and plaza venues in a quick look.

So, the name of clusters was:

- North rural places.
- Multiple tourist and city venues.
- Tourist and restaurant places.
- South rural places.

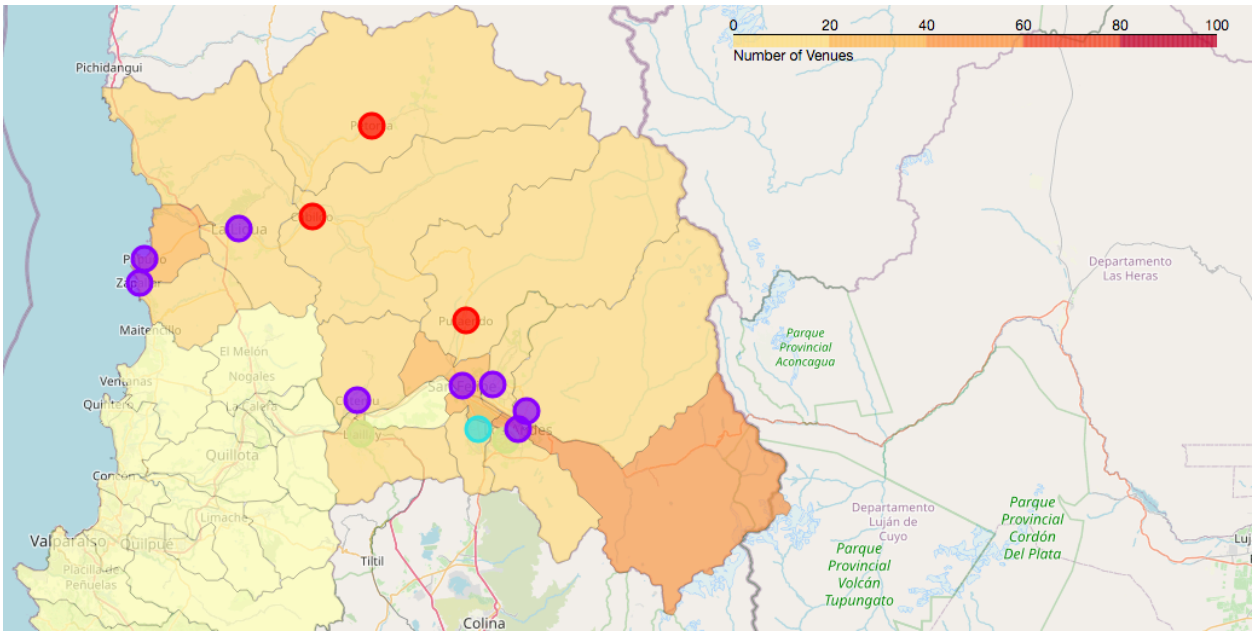


Figure 5. Aconcagua Region in a choropleth map with its clusters. Red: Multiple Rural Venues, Purple: Multiple tourist and city venues, Light Blue: Casino and rural venues, Yellow: Multiple travel venues.

3.6 New Valparaíso region

Within new Valparaíso Region will be the following communes:

- Hijuelas
- La Calera
- La Cruz
- Nogales
- Quillota
- Algarrobo
- Cartagena
- El Quisco
- El Tabo
- San Antonio
- Casablanca
- Concón
- Puchuncaví
- Quintero
- Valparaíso
- Viña del Mar
- Limache
- Olmué
- Quilpué
- Villa Alemana

I applied the elbow method to find the correct K, but again the graphic had a little slope. If I was to guide me for this elbow, I would choose 5 clusters but I wasn't very sure, so, I implemented the mean of the Silhouette Coefficient again.
The best score in the Silhouette Coefficient was to 2 clusters with an accuracy of ~0.33.

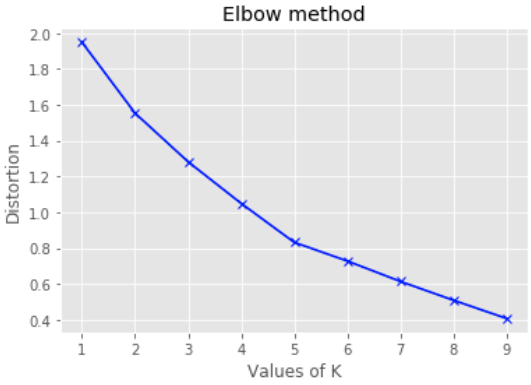


Figure 6. Elbow method in new Valparaíso region.

The **Cluster 0** with 17 communes represented different categories, a mix group of almost all communes in the region with coastal and inter-communes' cities. They were a range of 4 to 100 venues. So, this cluster represent a multiple city and tourist places.
And the **Cluster 1** had only two communes: Hijuelas and Calera. These communes are next to each other and they are next to the route 5 (the road that cross all country). They share the minimum quantity in each commune with 4 places. So, I can say that they represent a small population with a rural movement.
The name of these clusters:

- Multiple tourist and city places
- Rural places

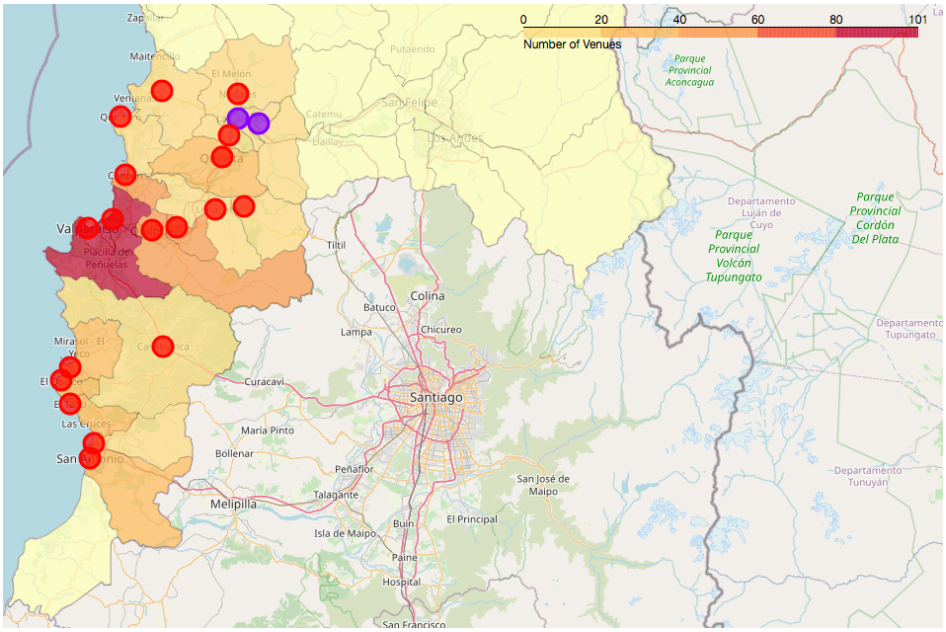


Figure 7. New Valparaíso Region choropleth map with two clusters. Red: Multiple tourist and city places. Purple: Rural and tourist places.

Results and Discussion

In discussion, we get the information with Foursquare API, that is not very popular. We have venues but in a tourist way, so this project is not very complete to compare institutional venues, colleges venues and so on. But we could focus in a general way and represent the final propose.
After that we separated the region, we find different characteristics to compare and analyze. In more details, we compare the new regions to understand witch category venues is in each location. In a simple view we have similar places in these communes, but the difference is in the amount. This happens because the centralism is in the capital (Valparaíso), and it decreasing in an east direction. Although we find communes with big cities places, the amount is very lowest in comparison with the west cities.
In summary the K-means model separate the communes mostly depending of its number of venues. Although in the Aconcagua case there be more characteristics, the same criteria was used.
In some cases there are communes with a low number of venues in the cluster with big numbers of venues. This happened because some particular category would be connected with another commune in this group.
This tell us that the Aconcagua regions **can be a different region** because is not very populated and have more space to integrate new opportunities and support a new generation to this valley.
The Aconcagua Region in comparison to the new Valparaíso Region had communes with low density of venues and therefore we can infer that this communes have a small urban development.

Conclusion

Chile is a country where the centralism works in a fear way. Just in the Metropolitan Region 6 millions of people approximately lives and in Valparaíso Region (2nd place in the ranking of population) has around 2 million. Almost the 50% of the population in **all country** and this allows inequality and lack of opportunities. This is the principal reason for the stakeholders start with the idea to separate the region in these two new regions.
In another hand, if we had the Aconcagua Region, the population will have a significant increase due to the new jobs and opportunities and this could be a problem for the life of these zones because in a most part is nature. So, if we'll carefully and respectful would build a new region with control and results of its work.