Coursera – IBM data science certification

Selecting county to strengthen healthcare system.

Adolfo Alamos Mella

Introduction

Nowadays the world is facing the COVID-19 pandemic, a disease that is spread all around the globe. The inexistence of a vaccine has pushed the local governments to take different kind of measures to stop the advance of this virus, being the most effective the social distancing and lockdowns. Despite all the measures taken, the high rates of infection have led to the collapse of health systems in different countries.

We will explore the city of Santiago, the capital of Chile. This city is composed of 32 counties, where approximately 30% of the total population of the country resides. Each one of this counties has different levels of confirmed covid-19 cases. We will use data analysis techniques to cluster the counties and find valuable insights to select which county should be priority in the efforts to reinforce the health system to face the advance of COVID-19.

There are a variety of things to considering before making a recommendation, the local government would like to benefit the most people possible and provide equal access to health care across the county, we need to study demographical data and the actual scenario of each candidate to find the best suitable location to invest.

By the end of this project, we will have identified those counties that would receive the greatest benefit by receiving resources and efforts to enhance their response to health emergencies.

Data

Health related establishments:

- The data about health establishments is retrieved from the official health department website.
- We use Google Maps API to get more information for each location.
-

Demographic data:

Data is provided by the social development department, this data is in a report called "county indicators CASEN" . This file contains a range of data per county, from income, poverty rates and access to basic services, that might be useful.

- Income: Average income per household (CLP)
- Poverty rate: Population in situation of income poverty (%)
- Access to basic services: Distribution of households by index of access to basic services (%)
-

Population and Covid-19 updated data

- Data is retrieved from the national department of statistics.

Methodology

Data exploration

First, we explore each county and its actual scenario against COVID-19 pandemic.

Using data from the national statistics department, we proceed to visualize deaths and the adjusted mortality rate per county.

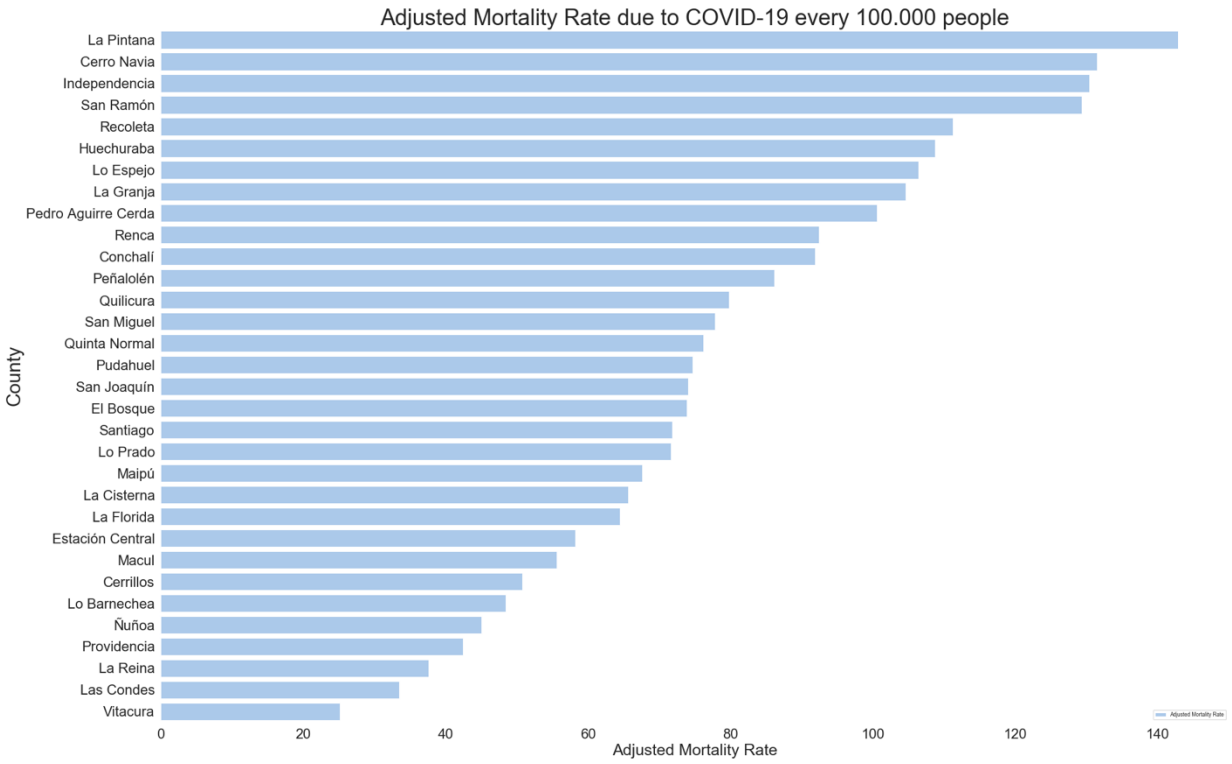
Each county has different population levels, for that reason, we need to work with the observed and adjusted mortality rate per county, which refers to the amount of deaths every 100.000 persons.

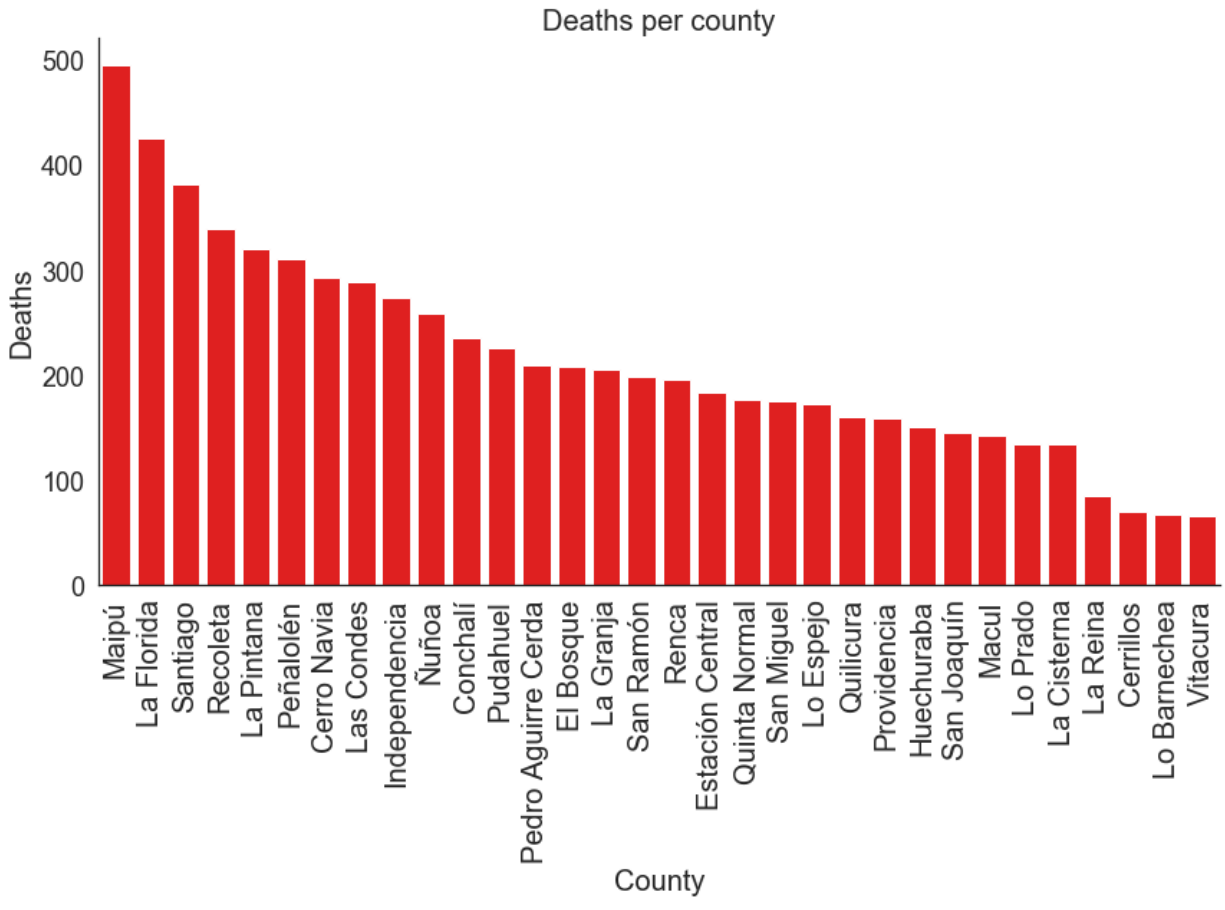
The population of each county is different in many aspects, this is a very important fact to keep in mind while studying dinamic diseases like COVID-19, where gender & age plays a main role while determining the risk of a patient. Because of this, we will calculate the adjusted mortality rate per county, that is calculated adjusting by sex & gender considering a standard population (in this case, we will work with the population of Chile from 2002).

To obtain the adjusted mortality rate we need to calculate:

- Deaths by gender and age group
- Population by gender and age group.
- [Deaths / population] by gender and age group
- Standard population (Chile 2002 census)

- Ponderation of each mortality rate observed by gender and age group proportional to the standard population.





We can see here, that the county with highest deaths by COVID-19 Maipú, is well positioned in terms of adjusted mortality rate, the main attention should be focused in La Pintana, Cerro Navia, Independencia and San Ramón.

Mapping Health related establishments in Santiago.

We classified the health related establishments into 3 levels of attention.

Levels of medical care attention:

The healthcare network is organized into 3 levels of care attention

- **Primary**: The first level is of low complexity, since it has general professionals and technicians, not specialists with equipment. It has an open type of care, that is, outpatient. It is from here that referrals are made at levels of greater complexity.

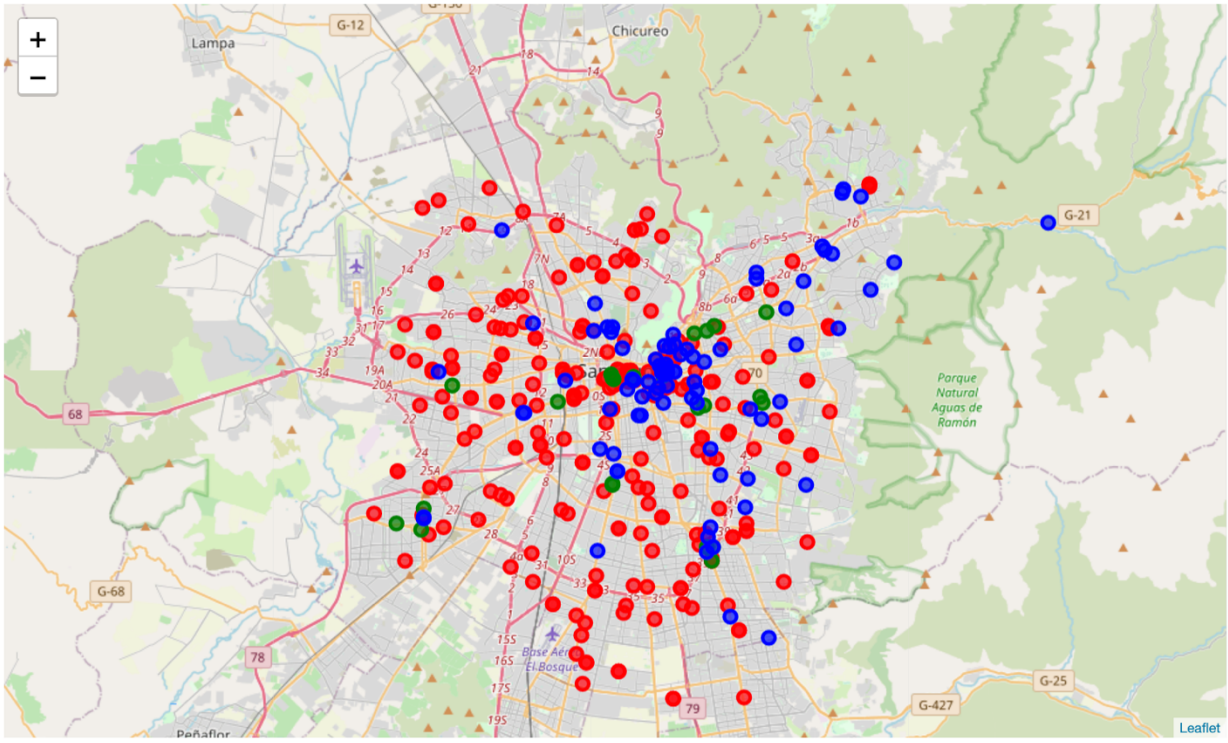
- **Secondary**: They have medium coverage and medium complexity, that is, they serve a smaller number of people, but with a higher level of technology and specialized professionals. Outpatient care is provided.

The patients who require more complex treatments are referred from primary or secondary care to the last level, which is the tertiary level.

- **Tertiary**: It has low coverage and high complexity, that is, it serves a much smaller number of people, but with high technology and specialization. They perform open and closed care.

Then, the establishments were mapped using Folium. The data was first collected and addresses, latitudes and longitudes were assigned to each establishment using Google Maps API. Each level of attention has a color assigned.

Primary level = Red, Secondary level = Greeb, Tertiary level = Blue.



In this map, the circles represent health related establishments of each level of attention.

What we can observe on the map is a regular distribution in the first level of care throughout Santiago, but as more complexity of treatment and care is needed, as in the case of COVID-19, the establishments tend to be located towards the eastern sector of the city.

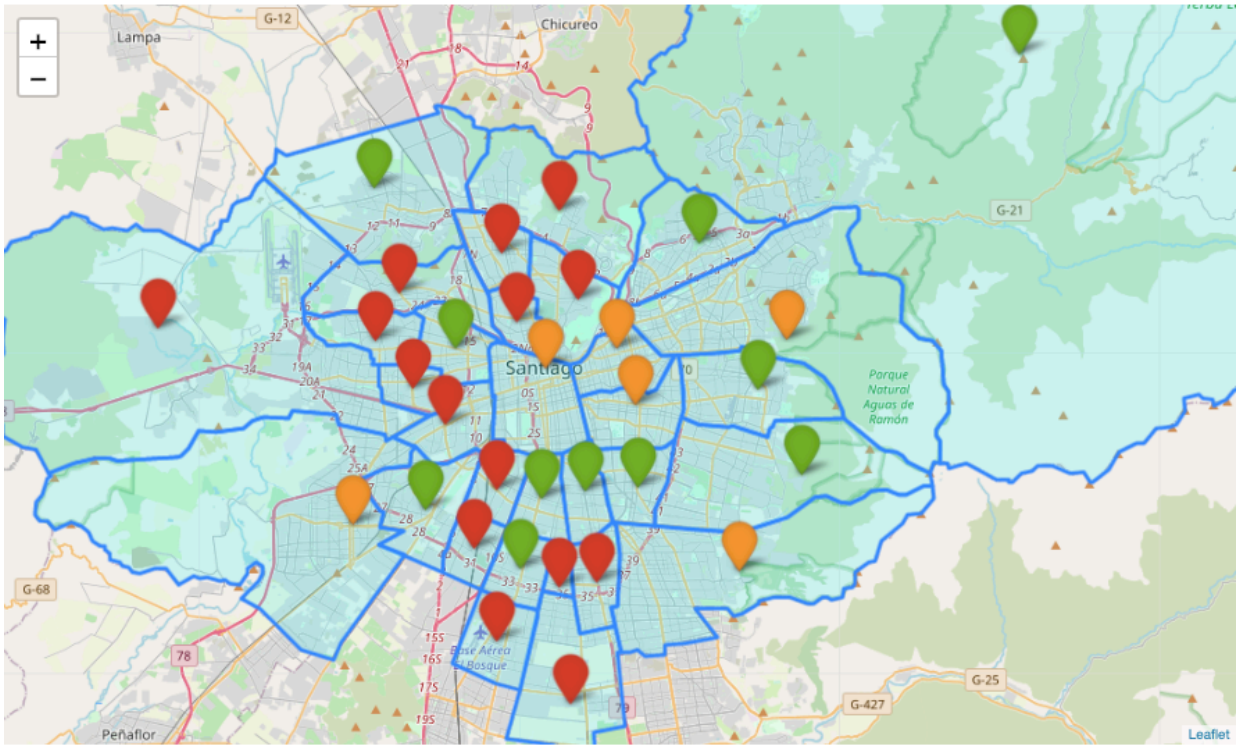
Clustering the counties

With the data collected, we used a clustering machine learning algorithm called K-means.

K-means is a unsupervised machine learning algorithm, that make inferences from datasets using input vectors without referring to a known or labelled outcome.

The objective of K-means is simple: group similar data is to group similar data points together and discover underlying patterns. Each cluster refers to a collection of data points aggregated together because of certain similarities.

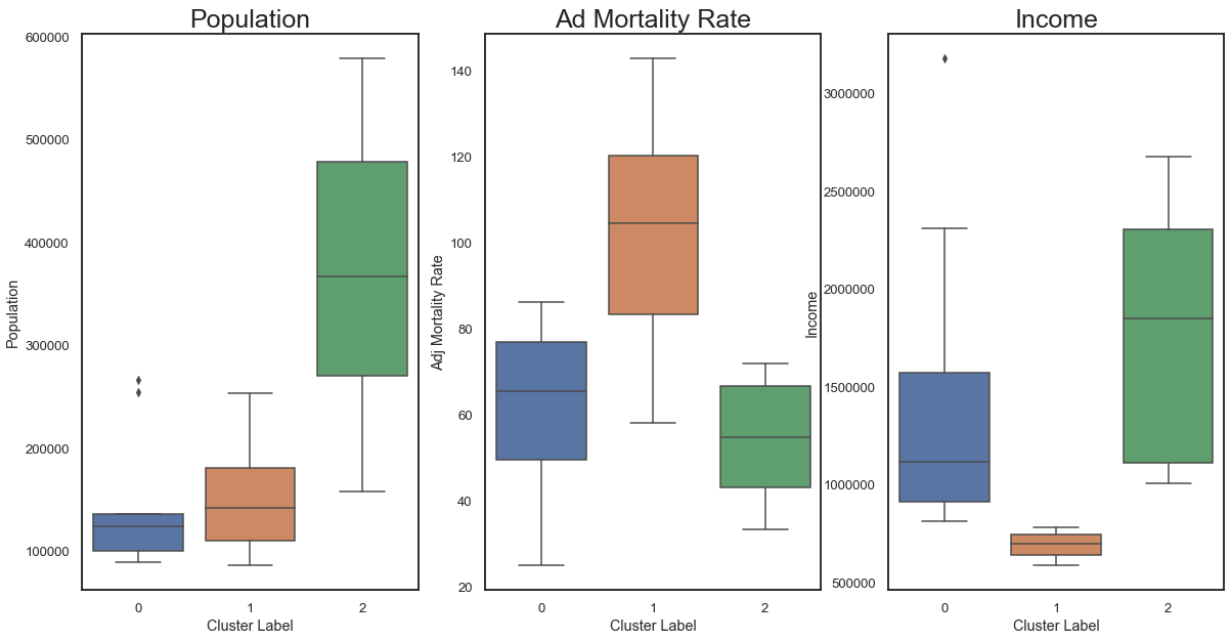
In this case, our objective is to cluster the similar counties into 3 clusters and find insights that leads us to select our main candidates to make our recommendation.

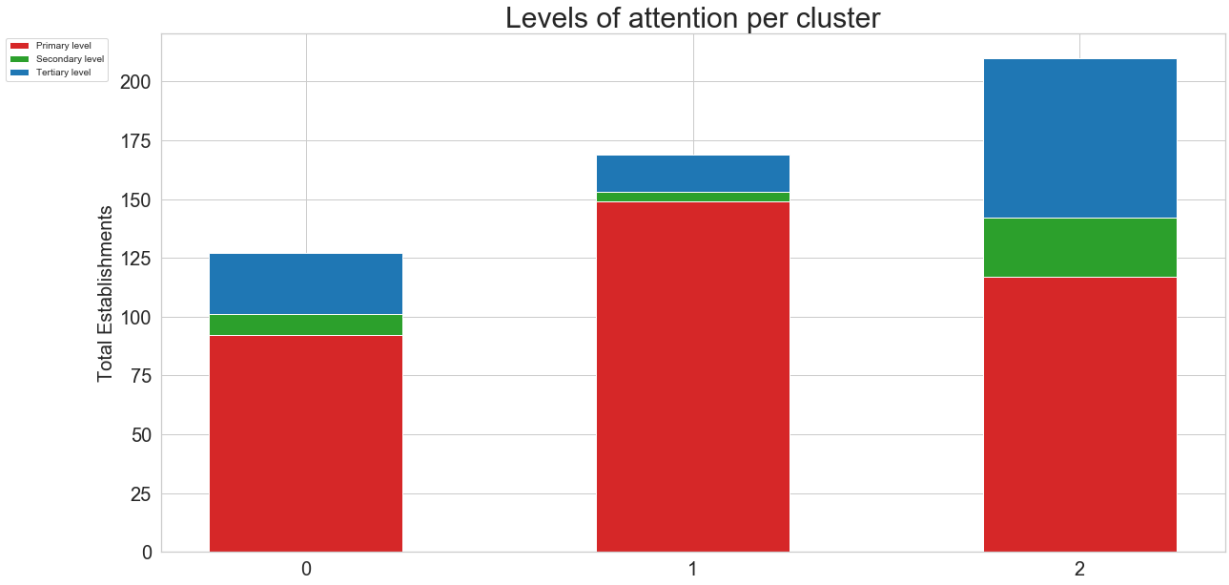


Here, we can see mapped each county, its boundaries with the assigned cluster label.

Clusters exploration

We proceeded to explore each cluster and its feature to find the best suitable candidates with a series of visualization.

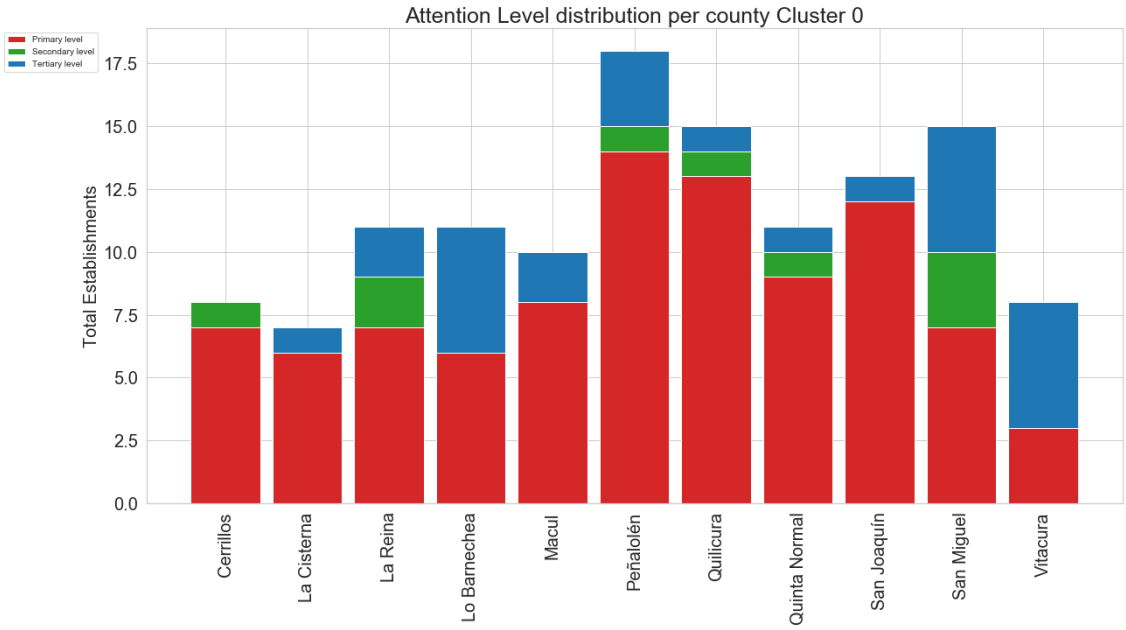




We can determine the main features of each cluster by taking closer looks to each graph, we can observe the different levels of population, adjusted mortality rates and income levels of the counties that are in each cluster.

Exploring each cluster

Cluster 0

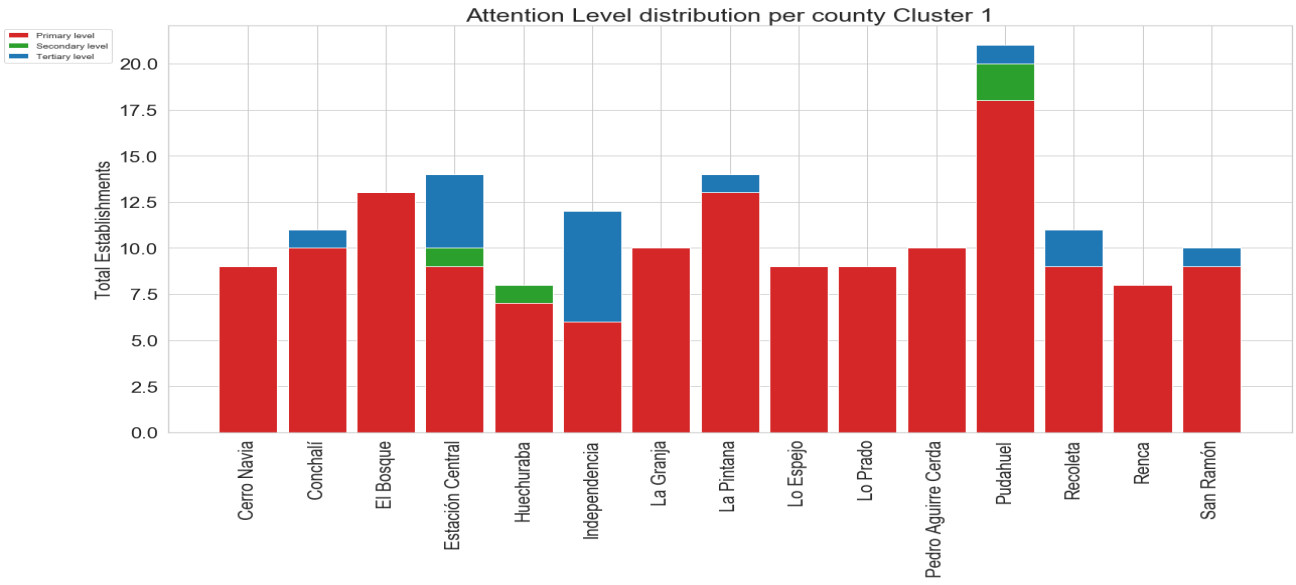


This cluster is composed by 11 counties that contains 25.7% of the total population.

As we can see, the cluster 0 is mainly made up of counties with low population and medium/high levels of income.

Almost all of the counties have tertiary level of attention, and are well balanced between the three levels and had a effective response to the pandemic, with a mean adjusted mortality rate below the mean of the total data, but in low numbers, proportional to the level of population. It does not seems that this counties should be a priority at the moment to reinforce the health system, as it low population level are seemingly very well covered.

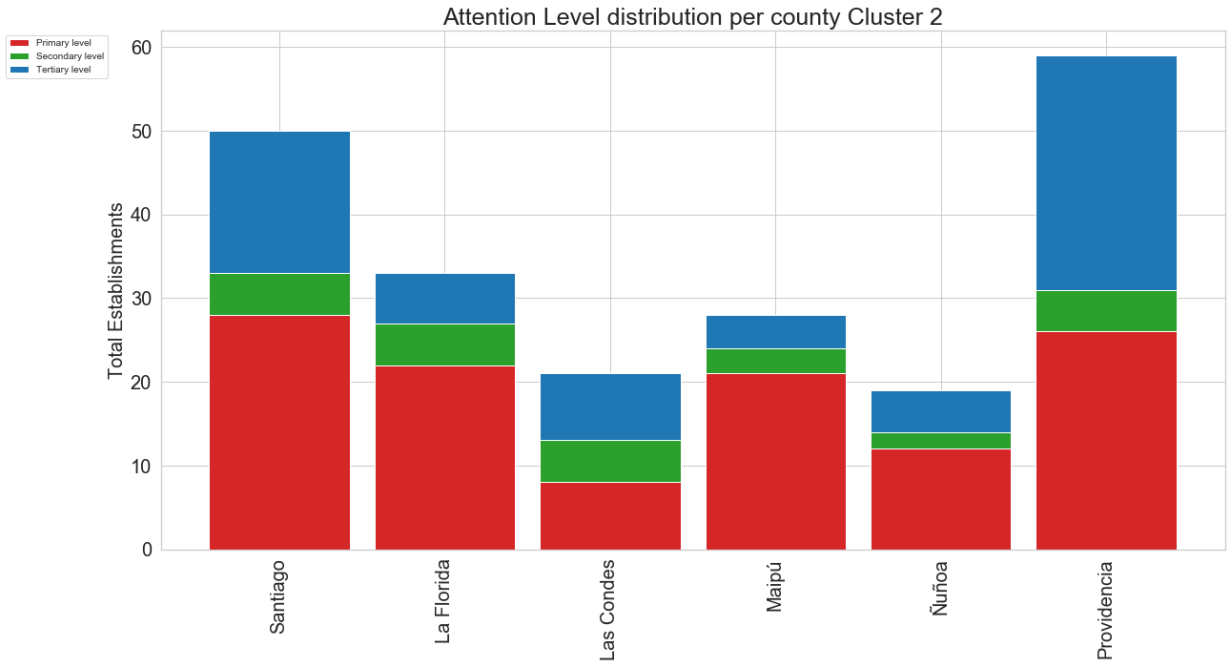
Cluster 1



We can see that this cluster is made up mainly of low-income counties located far from the eastern area, and that this is reflected in the low number of second and third level establishments (around 50% of these counties only have first-level establishments of care). This is directly related to the high mortality rate presented by this cluster (the highest by far amongst all clusters) as there is not much access to specialized attention. Very low tertiary level and almost inexistent secondary attention.

Considering the aforementioned points, we observe that 37.25% of the total population resides in this cluster, so we should consider counties in this cluster as serious candidates for our recommendation.

Cluster 2

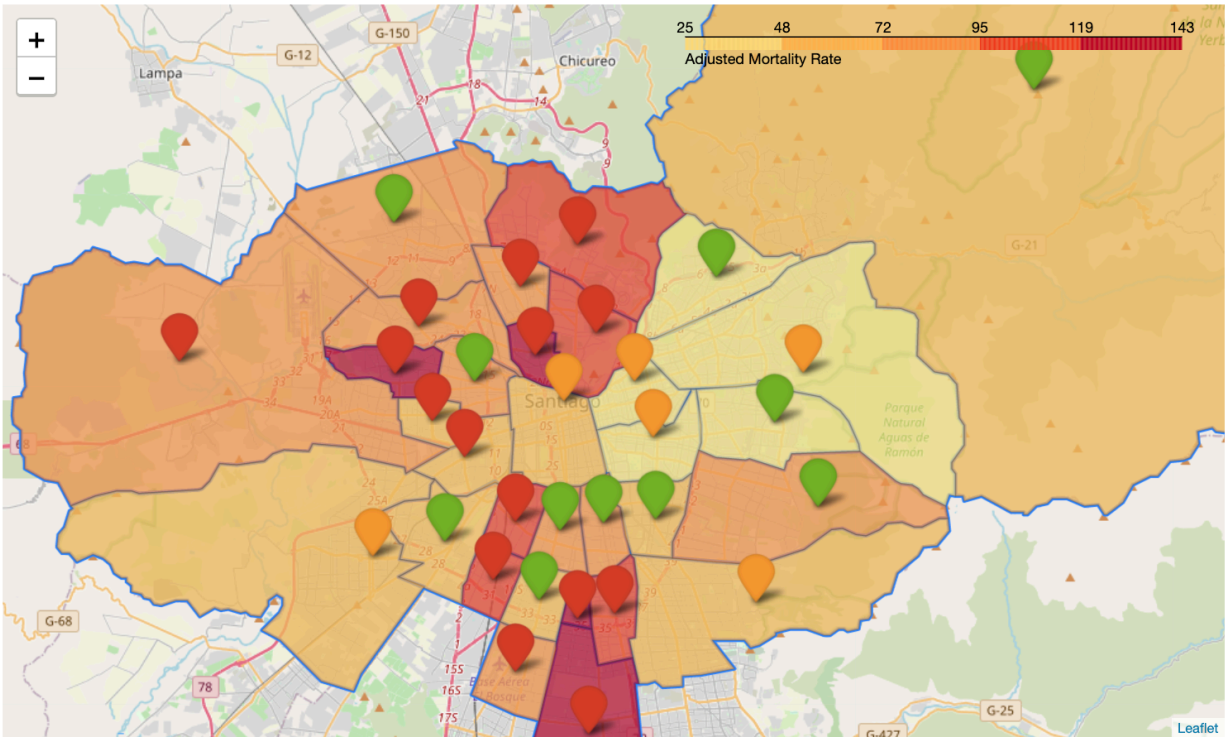


We observe that this cluster is made up of counties with highest income levels, in addition to this, all the counties are very well balanced between all the 3 levels of attention, that is well represented by the lowest adjusted mortality rate.

Nonetheless, this cluster is the one whose posses the most number of health related establishments.

Having made these observations in greater detail for each cluster, we will fully focus on cluster number 1 to find the county that would benefit the most with an improvement in its health system.

After all this exploration, using the census data and calculated adjusted mortality rate, we overlaid a density map. This allows an easy way to visualize the relationship between each cluster and the adjusted mortality rate of the area:



The magnitude of mortality rate is clearly distinguishable between each cluster, our selected cluster in red has the highest levels of mortality rates.

Results

From the top 5 candidates (top 5 highes adjusted mortality rates in cluster number 1), "La Pintana" has a considerable difference in the adjusted mortality rate against the other candidates, while "Cerro Navia", "Independencia" and "San Ramón" have very similar rates, finally, "Recoleta" has a significantly lower rate and better performance.

On the other side, "San Ramón" has a very high rate with almost half of the population.

We have to take into consideration, that La Pintana and San Ramon are adjoining counties, so a reinforcement in either one of this counties would benefit both of them, also note that both have a high level of primary attention and both have tertiary level attention.

The last two candidates, Cerro Navia and Independencia are demographically very similar, but in terms of health establishments they are very different, Independencia is balanced 50/50 between primary and tertiary level, while, on the other side, Cerro Navia only has primary level attention.

If we look closer to the map, Cerro Navia is a neighbor county with "Lo Prado" and "Renca", of the same cluster. This two counties also have only primary level attention.

In the context of a highly infectious virus as COVID-19 that is spread very fast among population, is an important fact to take into consideration that we want to avoid the population to move long distances to get healthcare in case of having sympthoms, so reinforcing Cerro Navia would benefit this three counties, addressing their healthcare attention to this closer county.

Also, in general terms, Cerro Navia, posses the 2nd highest adjusted mortality rate among all the counties in Santiago, 4th highest poverty rate and low income level.

With that in mind, our conclusion is that the best location to reinforce the health system in Santiago, is the county of Cerro Navia.

A a strengthening of health systems in Cerro Navia would benefit the most in the actual scenario of the pandemic in the city of Santiago.

This project can be expanded to any city of interest where the data is available.