Capstone Project: Hospitals in Lima districts

Name: Alexis Martínez

1. Introduction:

The coronavirus COVID-19 pandemic is the defining global health crisis of our time and one of the greatest challenges we have ever faced. Since its beginnings in Asia, it has spread to most countries in the world, making the presidents order a long quarantine to keep it at bay. Unfortunately, the economy of every nation has plummeted while the positive cases and deaths keep increasing. There doesn't seem to be a cure in the near future, leaving us with the only option of coexist with the virus.

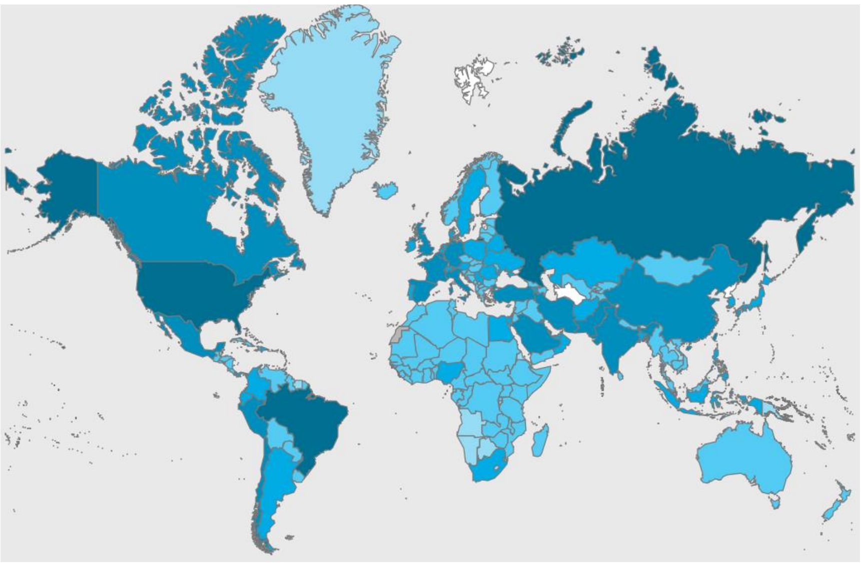


Figure 1: Choropleth map of the world according to Covid-19 cases

Peru is one of the most affected nations by COVID-19, being in the top 10 countries with most cases as of June 2, 2020. The Peruvian president Martín Vizcarra declared a national state of emergency on March 15, when the country had just 71 confirmed cases of COVID-19. The order closed Peru's borders and banned Peruvians from leaving the house except to access essential goods or perform essential work. But public health experts say living and working conditions in the country of 33 million, where a fifth of people live on only around $100 a month, has made it near impossible for many Peruvians to comply with quarantine measures.

1.1. Business Problem:

The government of Peru has organized a group of data scientists to analyze the COVID-19 positive cases in the city of Lima and select a district to open a new hospital. Lima is the capital city of Peru, and it holds 60% of the total positive cases in the country. Analyzing the data, which districts would the team recommend opening a new hospital?

2. Data

To solve this problem, the team has set up to find the following data:

- List of districts in Lima, Peru. This encompass the whole project.
- Latitude and Longitude of the districts in Lima, Peru. We need this information to plot the map and get the hospitals nearby.
- COVID-19 positive cases in Lima per district.
- Hospitals in Lima per district.

2.1. Source of Data

To gather the data needed, we are going to use the following sources:

List of Districts in Lima:

The Wikipedia Page https://en.wikipedia.org/wiki/List_of_districts_of_Lima has the data we need in a table format. The following table is a dictionary of each column:

Column Name	Definition
Districts	Name of the District
UBIGEO	Unique code only used in Peru
Area	Area in Km of the district
Population	Population of the district (2005)
Population Density	Population density of the district (2005)
Created	Date of Creation
Postal Code	Postal Code of the District
Location	Map of Lima showing the District’s location

To get the data into Python, we’ll use web scraping techniques. The columns we’ll need are Districts, UBIGEO and Postal Code.

Latitude and Longitude of the districts in Lima:

This information can be retrieved with Python package Geocoder. This package can be very unreliable, and Google Maps API isn’t really an option since 2018. Because of that, we have a CSV file with the correct latitude and longitude of each district, in case the Python package doesn’t work.

COVID-19 positive cases in Lima:

The government of Peru has the information open to everyone in their webpage https://www.datosabiertos.gob.pe/dataset/casos-positivos-por-covid-19-ministerio-de-salud-minsa. The dataset they provide is updated every day, and has the positive cases per department, province and district of Peru. The following table is a dictionary of each column in the dataset.

Column Name	Definition
UUID	Unique Identifier
DEPARTAMENTO	Department in Peru
PROVINCIA	Province in Peru
DISTRITO	District in Peru
METODODX	Type of test applied
EDAD	Age
SEXO	Gender
FECHA_RESULTADO	Date of the Positive Case (DD/MM/YYYY)

In this case, we are only going to use the columns ‘DISTRITO’.

Hospitals in Lima per district

To get this data, we are going to use Foursquare API. Using this API, we can make RESTful API calls to retrieve data about hospitals in different districts of Lima. To use it, we need a developer account active. Take into account that the account has a limited amount of normal and premium calls per day. In the following link, we can find the venue categories they have and the ID to call them https://developer.foursquare.com/docs/build-with-foursquare/categories.

3. Methodology

After retrieving the data, we need to start manipulating it to resolve the problem.

3.1. Data Preparation

For starters, we’ll use the data from Wikipedia to get a base dataframe with the district names. The steps are:

- Keep columns "Districts", "UBIGEO" and "Postal Code". All 3 columns can be useful in the future.
- Change column names
- Change the only district "Santa María del Mar District" to the correct name.
- Check for any data type inconsistencies.

Next, we’ll merge our base dataframe with the coordinates of districts in Lima. The steps are:

- Merge the district dataframe with the coordinates dataframe.
- Check for any errors.

Now we have a complete dataframe with the coordinates for each district in Lima. The next phase is to add the data for COVID-19 positive cases per district. The steps are:

- Subset the dataframe for where the ‘Department’ is Lima and ‘Region’ is Lima. This will give us the 43 districts of Lima.
- Drop the columns 'DEPARTAMENTO' and 'PROVINCIA'.
- Count the districts to get the number of COVID-19 positive cases per District.
- Rename the columns for consistency purposes.
- Sort the values to easily concatenate the data with our base dataframe.
- Concatenate to the base dataframe.

- Drop the column "Districts"

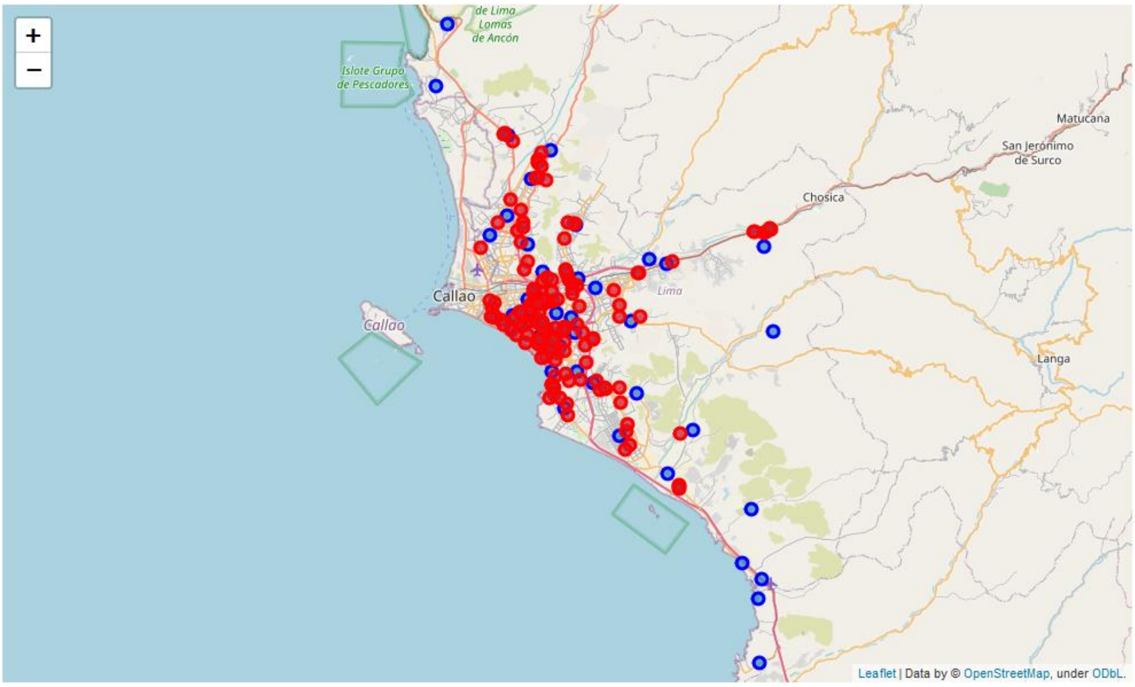
We have an almost completed dataframe. The next thing on our list missing is the hospitals data. To get it, we are going to use the Foursquare API. To do this, we set up a function ‘getNearbyHospitals’ to quickly get all Hospitals near the desired location. The steps are:

- Call the function and save the data in a dataframe.
- Check for duplicate hospitals in the dataframe.
- Check for wrong names in the dataframe.
- Check that the hospitals belong to the respective district.
- Check that the hospitals retrieved are actually hospitals.
- Check that the hospitals have specializations in regard to COVID-19 (pneumology)
- Finally, rename the columns

This part of the data preparation took a lot of time. An explanation is given in the **“Discussion”** section.

3.2. Data Analysis:

After all our work preparing the data, we can finally start analyzing. For now, lets start by plotting our district and hospital data in map:

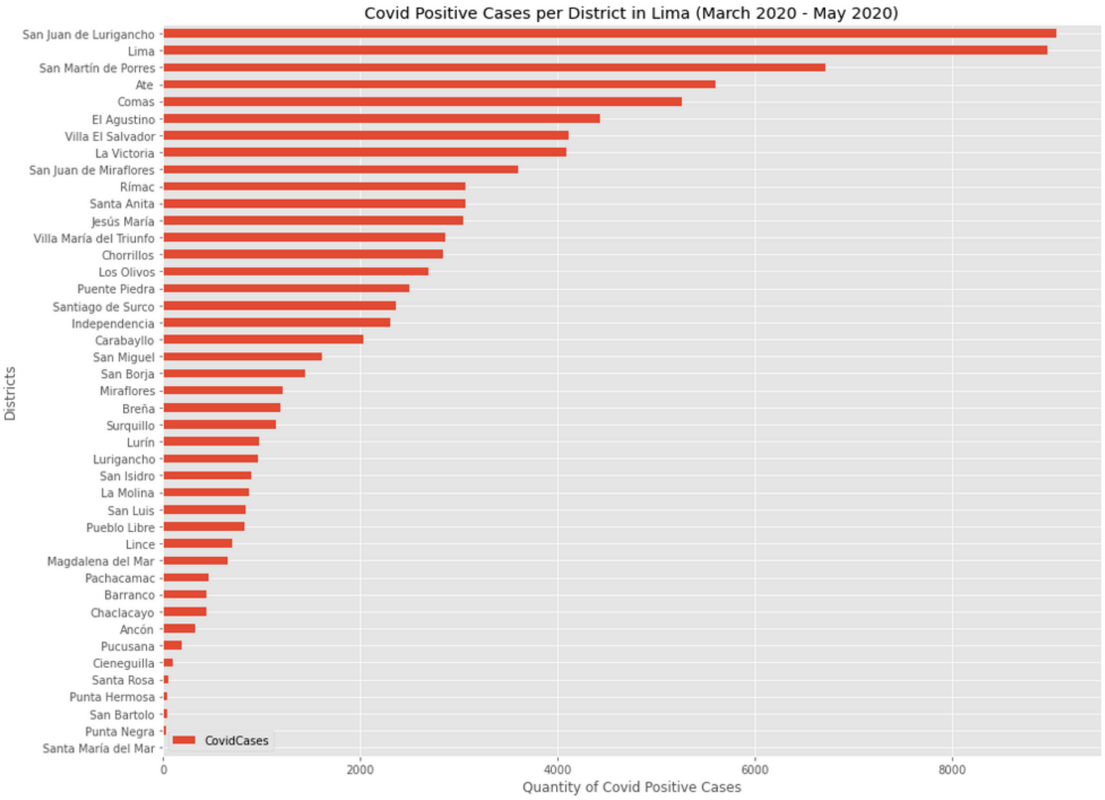


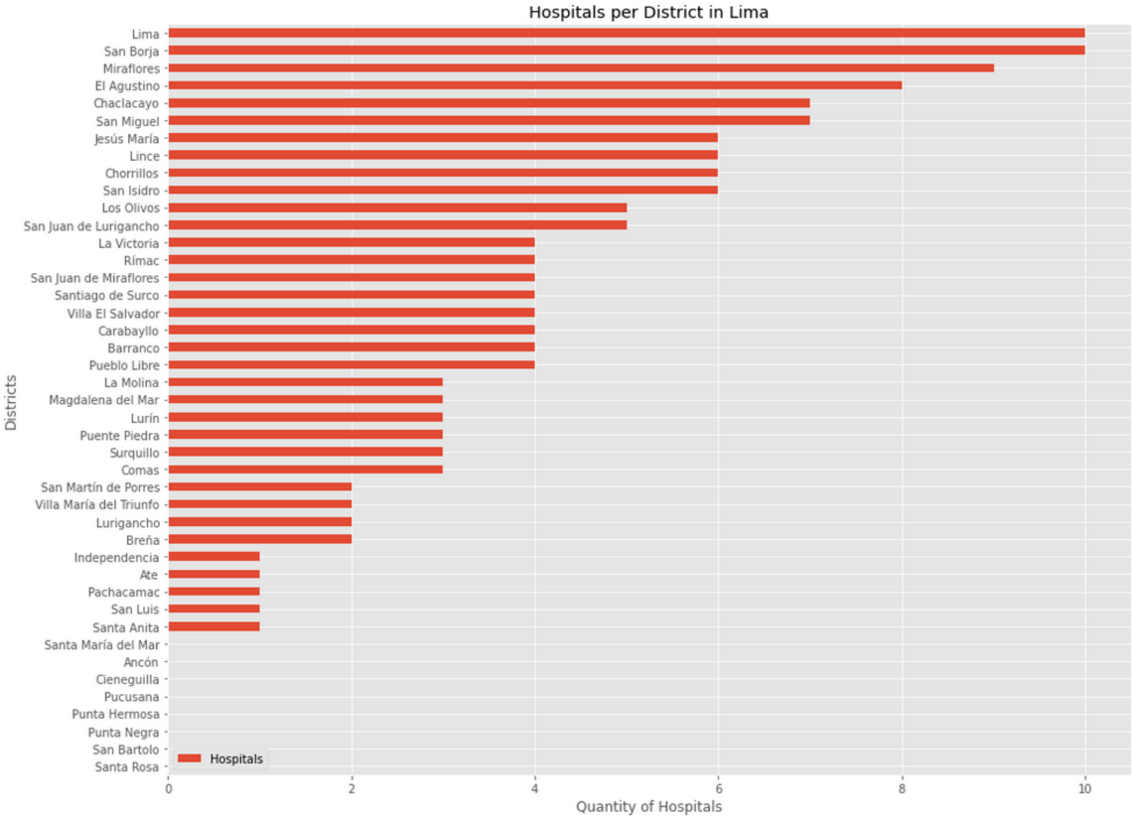
For this map, the blue dots are the districts and the red dots are the hospitals. We can gather some conclusions from this map:

- The agglomeration of hospitals seems to be in the center of Lima.
- As you progressively leave from the center, there’s less hospitals.
- Districts in the outskirts of the city have little to no hospitals.

Gaining insight from the map, we can move on to evaluate the quantity of hospitals and COVID-19 cases per district. We'll use horizontal bar charts for that. To create them, we first need to modify our dataframe slightly:

- Group the dataframe by 'District' and 'CovidCases' to count the hospitals.
- Add the districts with no hospitals at the bottom.





Interesting, we can gather some information from this graphics:

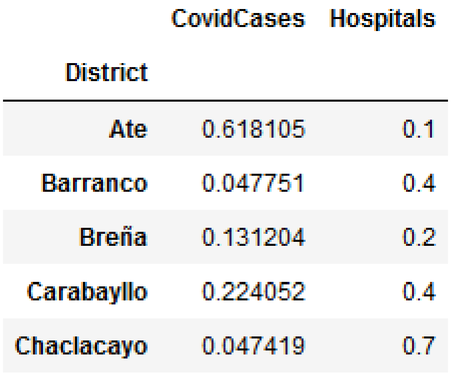
- San Juan de Lurigancho and Lima have the most amount of COVID-19 cases.
- Santa Maria del Mar, Punta Negra, San Bartolo, Punta Hermosa and Santa Rosa have the least amount of COVID-19 cases.
- Lima, San Borja and Miraflores have the most hospitals than any other district
- Santa Maria del Mar, Punta Negra, San Bartolo, Punta Hermosa, Santa Rosa, Ancon, Cieneguilla, and Pucusana have no hospitals.

- Lima has both the most COVID-19 cases and a large number of hospitals.
- Santa Maria del Mar, Punta Negra, San Bartolo, Punta Hermosa and Santa Rosa have no hospitals. But, at the same time, they have little to no cases of COVID-19.

3.3. Machine Learning

For our machine learning part, we are going to do a basic Content-Based machine learning. This will help us recommend the most needed district. The steps are:

- Import from the package sklearn, MinMaxScaler.
- Normalize the values of the dataframe.
- Turn it into a dataframe.



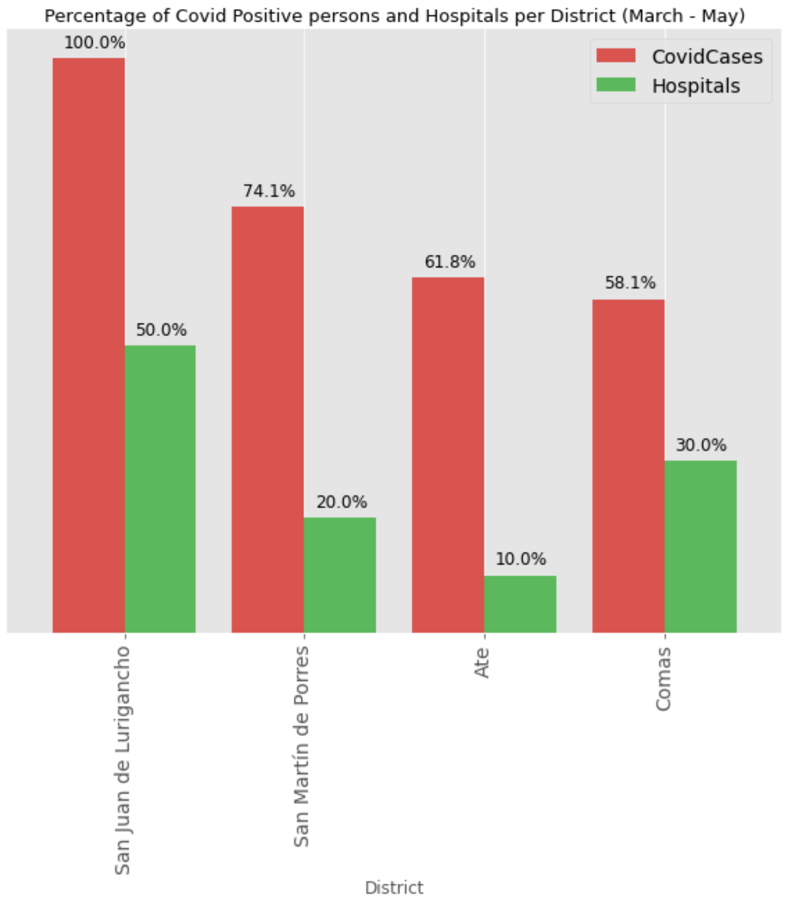
	CovidCases	Hospitals
District		
Ate	0.618105	0.1
Barranco	0.047751	0.4
Breña	0.131204	0.2
Carabayllo	0.224052	0.4
Chaclacayo	0.047419	0.7

Now that we have our dataframe normalized, we can head to the results and answer the question to our problem.

4. Results

In the result section, we are going to use the normalized dataframe to plot a revealing graphic. The steps are:

- Sort the values of our normalized dataframe.
- Set up a condition. In this case we want the most amount of cases with the least amount of Hospitals per district.
- Create the bar plot.



Now that we have our plot, we can answer the question.

Q: Analyzing the data, which districts would the team recommend opening a new hospital?

A: The district that need new hospitals amidst this pandemic are **San Juan de Lurigancho, San Martin de Porres, Ate and Comas.**

The reason we can give for this decision is:

- These four districts are among the top 5 most positive COVID-19 cases
- Lima has the same amount of positive cases as San Juan de Lurigancho, but also has the most amount of hospitals.
- The difference between COVID-19 cases and hospitals within these districts is about 30%to 50%
- If Lima is the example, these 4 districts need more hospitals to help defeat COVID-19.

5. Discussion

Regarding the Foursquare API, it did bring me the venues from the right category (Hospitals). However, the data was very messy and had to clean it a lot. Most of the hospitals also had their old names from before 2015. This makes me think that, at least in Lima, the app is not really being used as much as other countries. What's more interesting was that some hospitals retrieved had 2 to 3 different names in their database pointing to the same location. Those were hard to track, but they were also removed. The data used here is probably outdated and not so reliable.

Regarding COVID-19, the amount of cases keeps increasing by the day, so by the end of June the data will be outdated. This is a project that constantly needs to have new data to be relevant. The best way is to create a python file and call it every day to get updated data on this statistic. Even so, the data from the government of Peru might also be unreliable. This has been called out recently, because some statisticians in Peru have found out inconsistencies with our positive cases and deaths. Further investigation will be necessary to trust 100% this data.

6. Conclusion

The government of Peru now has the answer to their problem. The 4 districts of Lima, Peru have been set as a priority to build new hospitals and fight the pandemic. This was made possible by having the COVID-19 data and hospitals by district. In our case, Content-Based Filtering was the best choice.

This project has a lot of room for improvement. We can aggregate a lot more factors to do our decision like Pharmacies, population, transit of people, and much more. Considering my current level of expertise in Python programming, I feel satisfied with this project. I learned a lot about my city and how we are currently facing this pandemic.