IBM Data Science Professional Certificate

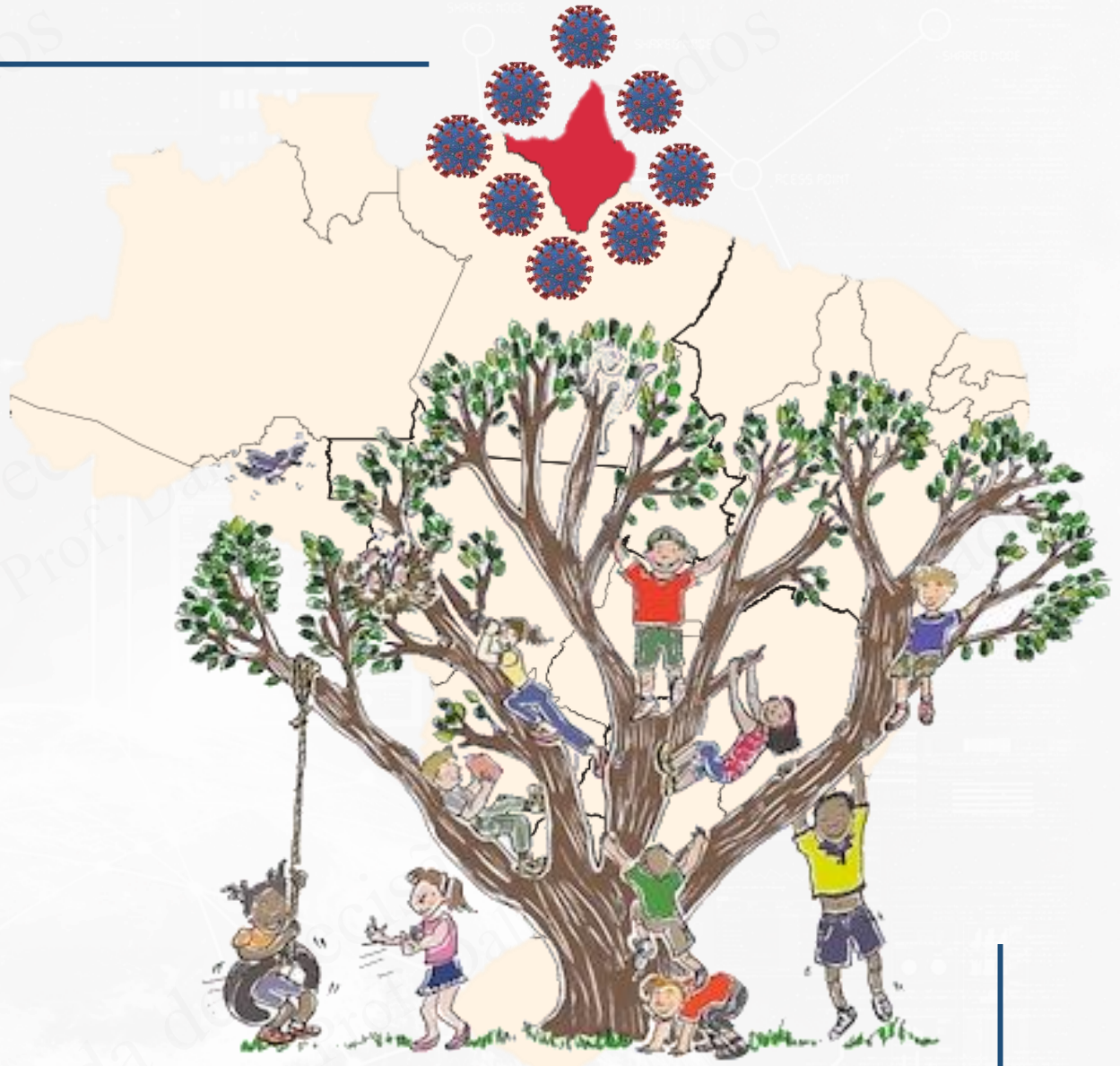
Capstone Project

Exploring Brazilian COVID-19 data and building a visualization maps according to the nacional reality.

Dalton Borges

Project context

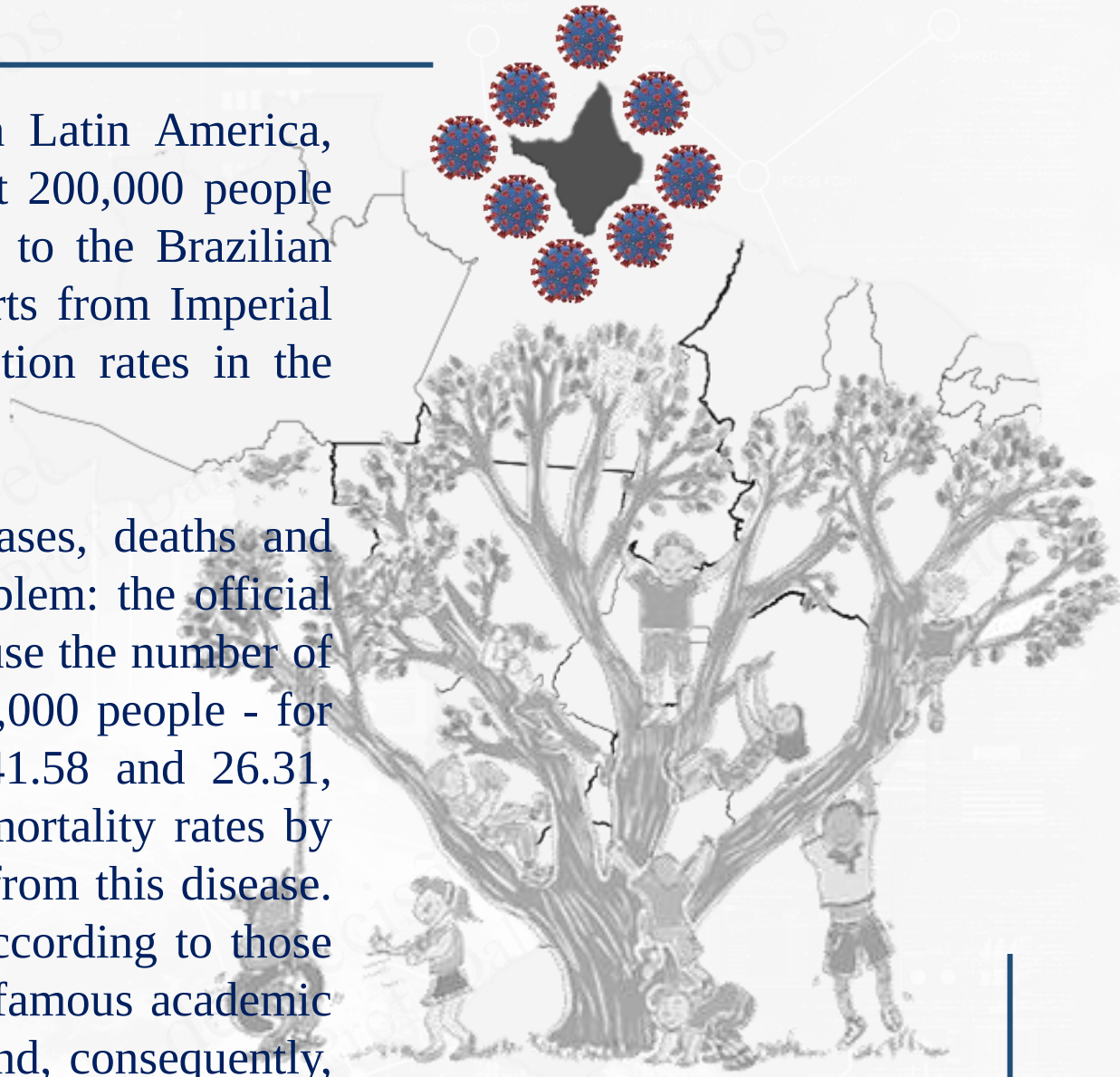
In March, the 2019 outbreak of the novel coronavirus disease has moved to a pandemic status and almost all countries in the globe are daily reporting new cases. All these data are generally collected by national health surveillance systems (HSS) and later compiled and reported by international organizations, research centers, and websites. According to those data, projections of future cases are made, in order to prepare and to adapt the health systems.



Project context

Today, Brazil is the epicenter of COVID-19 in Latin America, where SARS-COV-2 has already infected almost 200,000 people and caused more than 11,000 deaths, according to the Brazilian official reports. In fact according to recent reports from Imperial College, Brazil presents one of the worst infection rates in the planet.

However, despite the large number of daily cases, deaths and infection rates, Brazil is facing an other big problem: the official COVID-19 data is not trustable. All mainly because the number of COVID-19 tests available is to low: 0.063 per 1,000 people - for instance, the numbers in Italy and USA are 41.58 and 26.31, respectively. This shortage of tests inflates the mortality rates by COVID-19 and do not capture all people dying from this disease. Today, many COVID-19 projections are made according to those official data, including studies and reports from famous academic "brands", which may lead to inaccurate values and, consequently, to bad decisions.

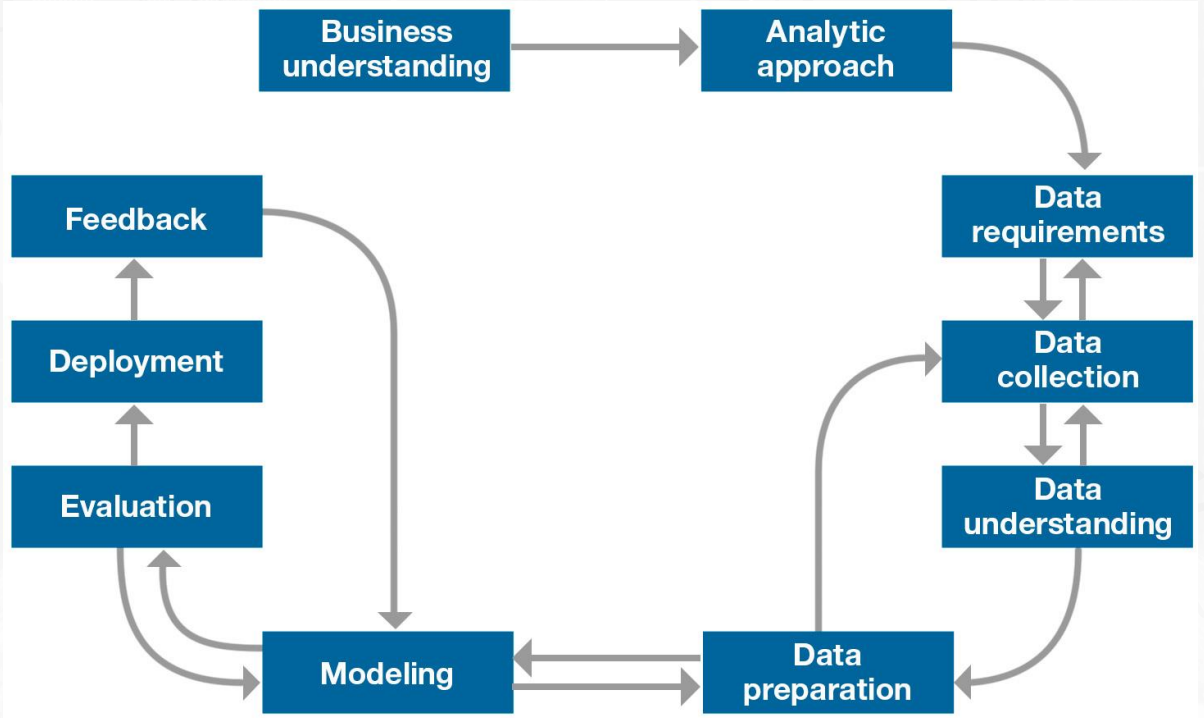


Project objective

Situational awareness is one of the main contributions of real-time analyses of epidemiological data. Thus, in this capstone short project I will answer the question: How bed is the situation of Brazil today on fighting COVID-19? In order to answer this question I will scrap and wrangle data that could be used to display the real-time Brazilian situation on facing COVID-19. Hence, a descriptive analytical approach will be used. Despite using many tools learned in the IBM courses, this project may serve as a first attempt to aid Brazilian decision-makers on fighting COVID-19.

Methodology

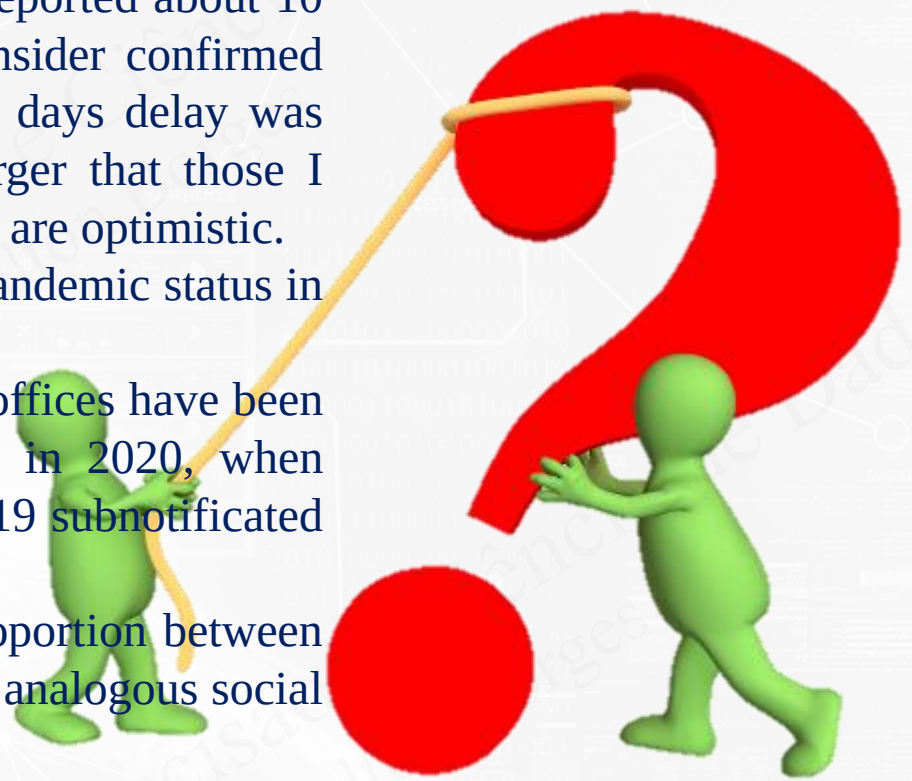
Just like a traditional scientific projects, data science projects also require methodology approaches, in order to guide the process and avoid common mistakes. The data science method used in this capstone project is the one proposed by John Rollins.



Assumptions

To complete this project I have assumed the following inferences as truth:

- While the number of deaths in day makes reference to a case reported about 10 days ago in Brazil, for the sake of simplicity, here I will consider confirmed cases and death cases as causalities of the same day. If a 10 days delay was condiered, the number of real COVID-19 would be even larger that those I calculate here. No keep in mind that the numbers presented here are optimistic.
- The Brazilian official data for COVID-19 does not reflect the pandemic status in the country;
- Since there are not enough COVID-19 tests available, notary's offices have been registering a surplus of deaths by other respiratory diseases in 2020, when compared to 2019. I assume this surplus is caused by COVID-19 subnotificated cases;
- When talking about viral pandemics, the fatality ratios (the proportion between death and confirmed cases) tends to be similar in locations with analogous social conditions.
- In Brazil, the percentage of infected people that will need a ventilator is 2% in average.
- In Brazil, the proportion of recovered cases and confirmed cases is 50% in average.



Data used in this project

In order to get around the Brazilian official data, other data must be used, in order to show the real Brazilian situation on facing COVID-19. Thus, we need the following data:

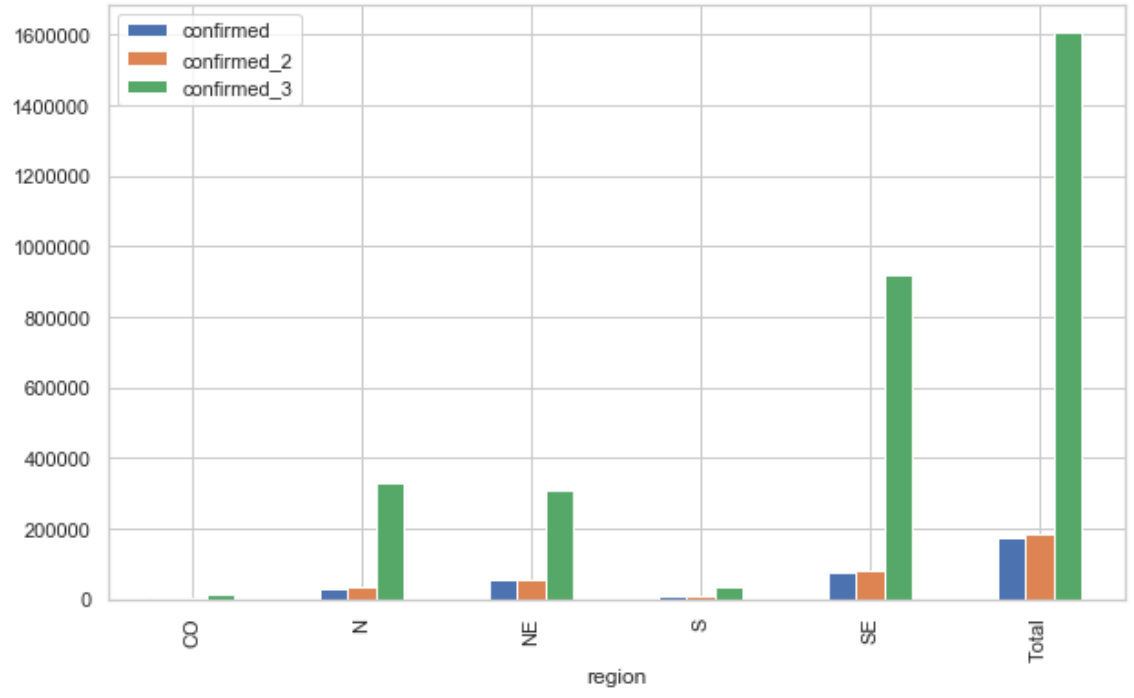
- A .csv or .xlsx file containing the number of ventilators available in each Brazilian city. Ventilators are the main equipment needed to fight severe COVID-19 occurrences. Contrasting this data to the real number of cases in a city will reveal how bed is the infrastructure of that city of fighting COVID-19.
- A .csv or .xlsx file containing the COVID-19 death and confirmed cases in each Brazilian state, day-by-day. Above this data we will perform out study.
- A .csv or .xlsx file containing the number of people that died from other respiratory diseases, such as pneumonia, in each Brazilian city day-by-day in 2020 and 2019. If the surplus exists, it could potentially be COVID-19, since there are not enought tests available.
- A .csv or .xlsx file containing social indicators of each Brazilian city. This data can be confronted to COVID-19 and infrastructure data, in order to get insights about the spreading of this disease.
- A .csv or .xlsx file containing geographic coordinates of Brazilian city centers. This data is needed to perform some data visualizations.
- A list of COVID-19 Infection Fatality Ratios for each Brazilian state. When confronted to official COVID-19 death and confirmed cases, this data may help to estimate the real number of COVID-19 cases in Brazil.
- A .json file containing the shape of each Brazilian municipality. This data is also needed to perform some data visualizations.

Findings

According to the government, Brazil has today around 170,000 (confirmed_1) confirmed cases. However, if we only add to the official number of confirmed cases the proportional number of confirmed cases, according to the Notary's Office, the number jumps to 182,000 (confirmed_2). Not that much.

Now, if we consider the number of cases by city that represents the correspondent IRF, the number is much bigger than the official ones. It goes to almost 1.6 million cases (confirmed_3). It makes Brazil so infected by COVID-19 as USA is right now.

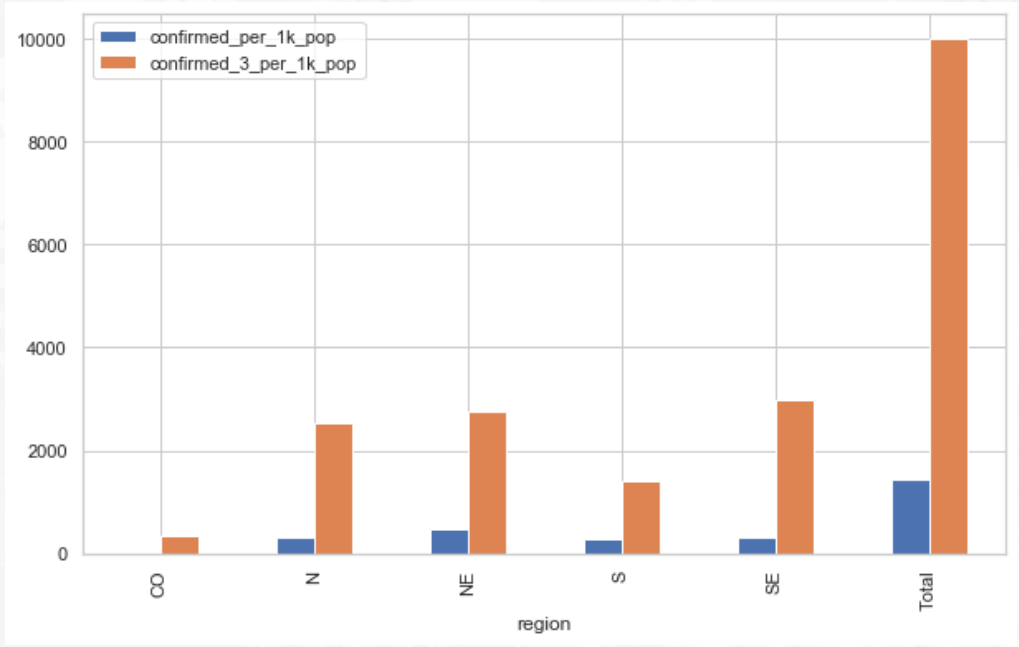
Let's look at those data with a bar plot. Pay attention the North (N) and Northeast (NE): while North has less cases than Northeast, according to the official data, when we consider the worst case, they have almost the same number of cases.

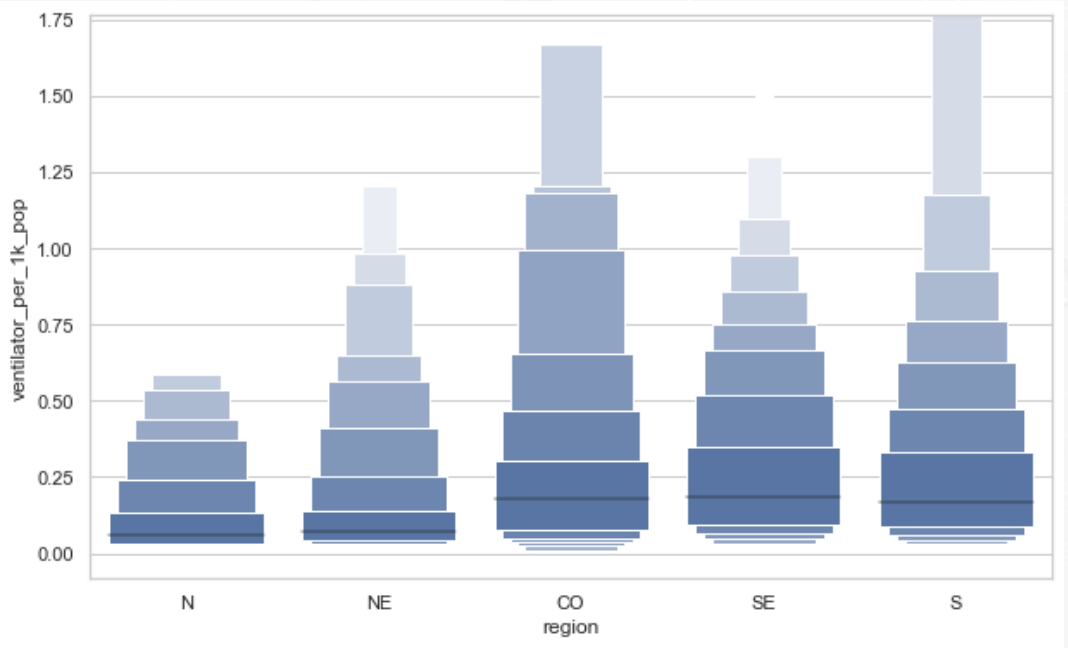


Findings

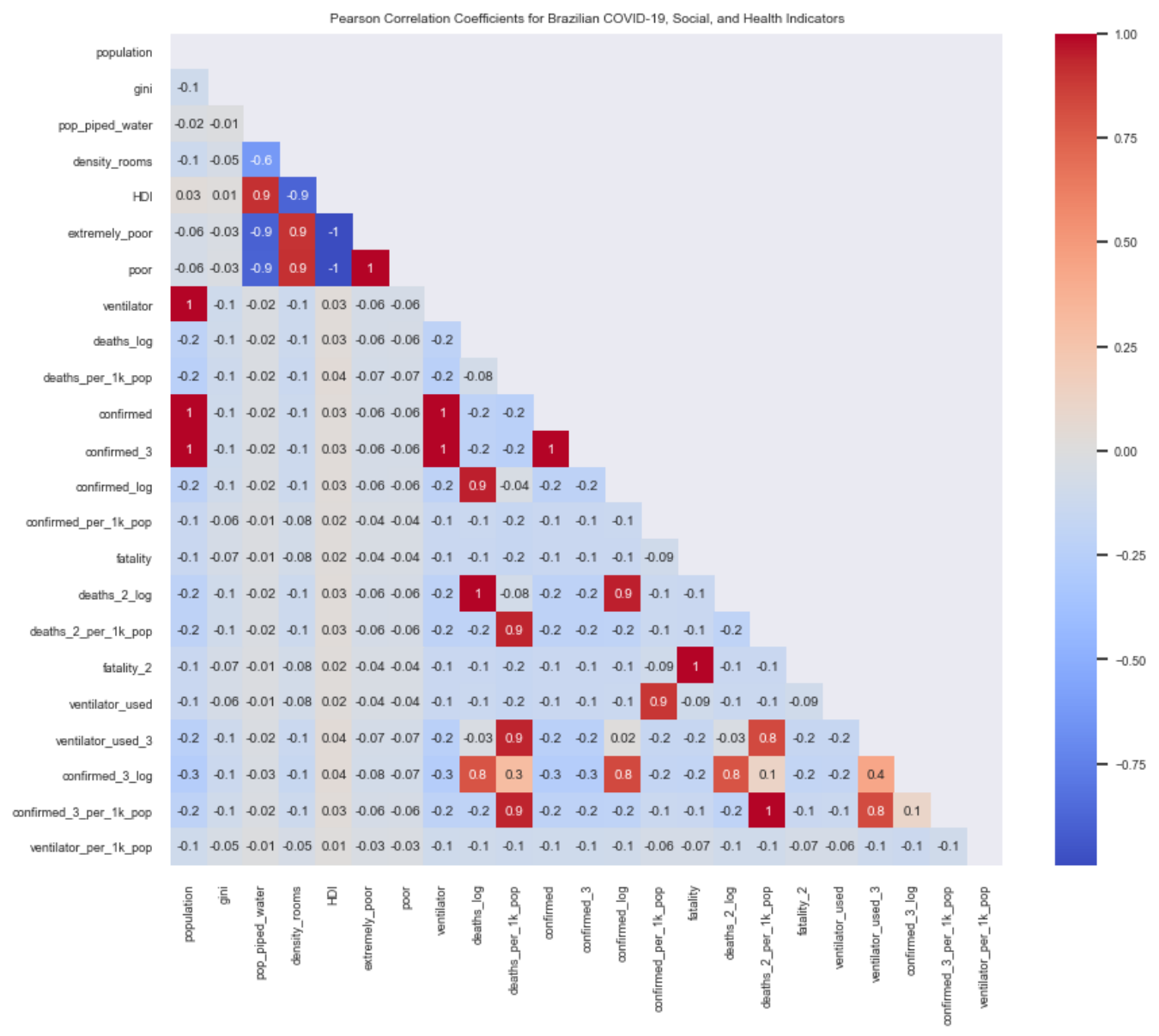
Now, if we take a look at the five Brazilian regions, we can see that the Southeast Region (SE) is the one with most ventilators per capita. North (N) and Northeast are the most susceptible ones, with less ventilators per capita.

Let's take a look at those data with a bar plot again. Pay attention that North (N) and Northeast (NE) has almost the same number of confirmed cases per 1,000 people. However, those regions receive much less attention.





Findings

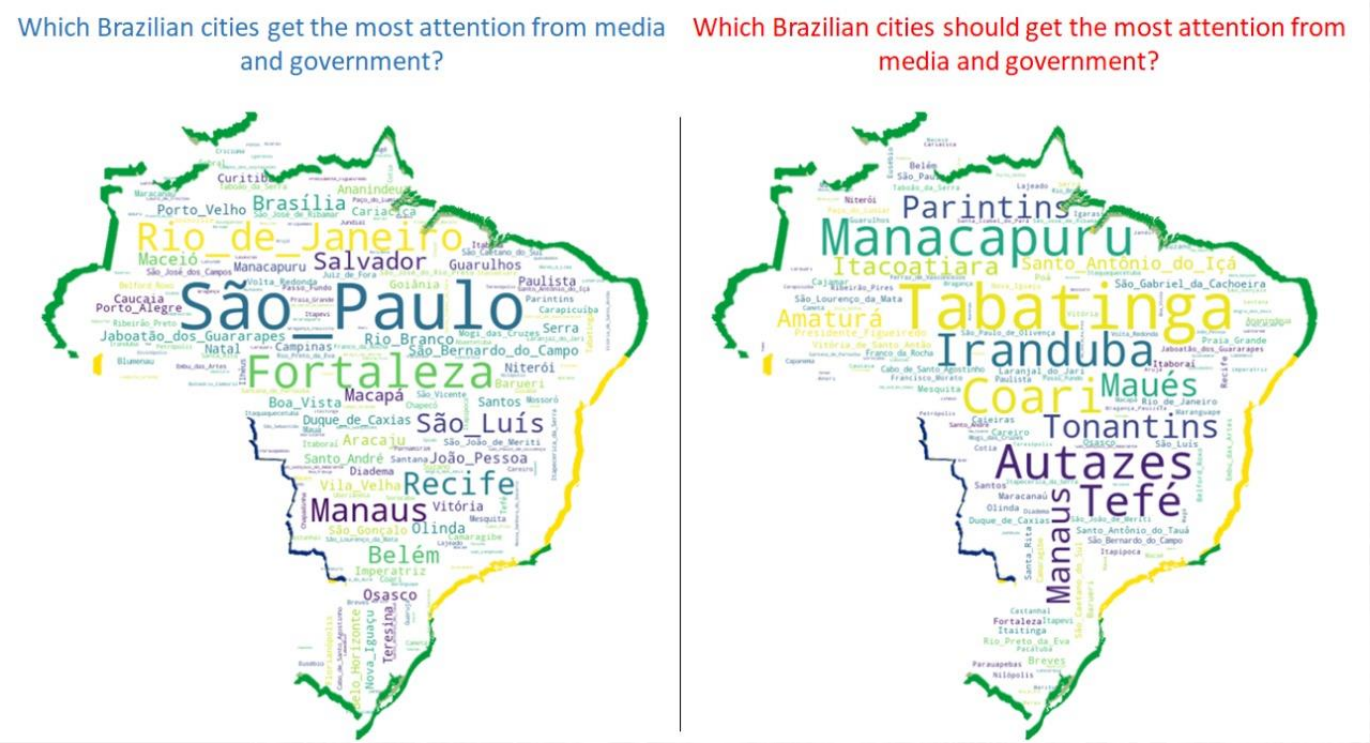


No relevant correlations were found. Relevant Correlation coefficients in the picture are just displaying correlations between dependent variables.

Findings

Generally, the Brazilian cities that get the most attention from media and government are those with more absolute number of cases, according to official data, provided by the Brazilian government. This is a good indicator. However, some smaller cities may be struggling with more cases per capita, according to the calculations we did.

Thus, lets create to word clouds. The first one will compute the name of the Brazilian cities that get most attention from media and government. The second one will compute the name of the Brazilian cities that should be getting most attention from media and government.

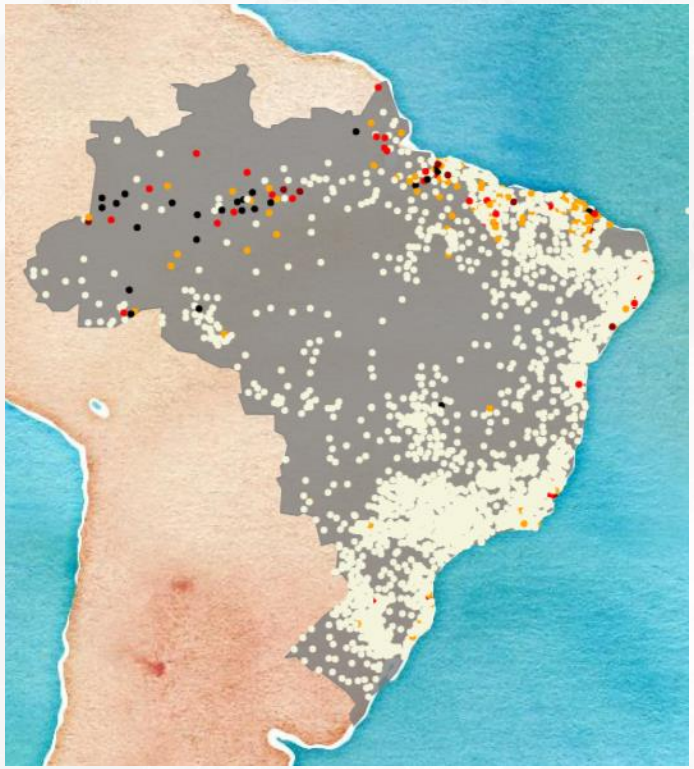


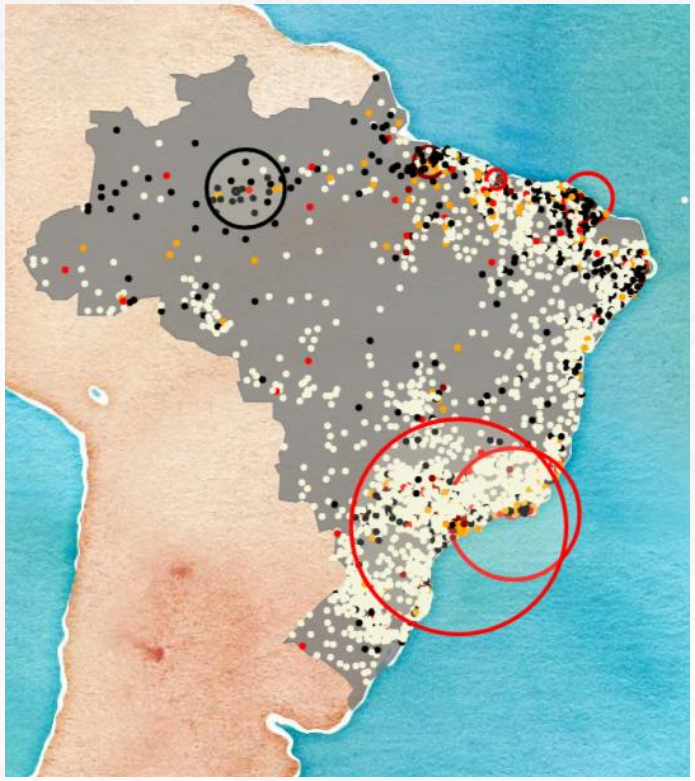
As you noticed, there is a huge difference between the clouds. While the first one may show you some common names, such as São Paulo, Rio de Janeiro and Brasília, the second one only shows us unknown names by most non-Brazilian people. That is because those cities are generaly outside the mainstream, a lot of them carved in the middle of amazon rainforest, with just a few resources to fight COVID-19.

Findings

As last representation, lest see two maps full of circles. In map 1, the radius represent the number of official confirmed cases, while the color represents the percentage of ventilators already used. Bigger the radius, more cases in each city. Complementally, darker is the color, greater is the percentage of ventilators already used.

The second map shows you the real Brazilian situation. The radius here are represented by the real number of cases, while the colors represent the real percentage of ventilators used.





Conclusions

COVID-19 case subnotification is a real problem in many countries, especially in Brazil. This leads to inaccurate predictions and do not help decision makers on selecting the most needed places to receive resources. Thus, this capstone projects intended to confront the scenario painted by the Brazilian official data with a scenario much closer to the real word.

According to official data, Brazil has today 171,000 confirmed COVID-19 cases, which represents a little more than 10% of the number of cases I've calculated, almost 1,6 million cases. The calculations were performed according to Notary's office data and Imperial College Report 25 findings.

Looking at the Brazilian regions we also see discrepancies with the data. According to the number of official confirmed cases, the Brazilian Southest (which is richest Brazilian region) is the more affected region. However, I've show that Northest and North are the most devastated regions, since they present greater infection rates per inhabitant and less ventilators available per capita.

I hope this project grabs the attention of those who read it. I also hope that other students, from other places, also feel the impetus of questioning data that sometimes may seem not questionable.


That's all Folks!