(https://cognitiveclass.ai/)

Capstone Project

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

Dalton Borges

---

1. From Problem to Approach

1.1. Project context

Coronavirus, an enveloped non-segmented positive-sense RNA family of viruses (Coronaviridae), are widely spread among mammals, including humans. Yet those viruses generally cause mild infections, some of them have the potential to induce several human causalities, such as the severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-COV), with mortality rates of 10\% and 37\%, respectively **[1, 2]**. At the very end of 2019, Chinese researchers discovered a novel coronavirus, a seventh known type and provisionally called by that time 2019-nCoV, today called SARS-COV-2, that was causing atypical viral pneumonia. This novel virus was firstly diagnosed in Wuhan City, Hubei province, China. Since them, it called the attention of researchers and governments, not just because it causes considerable mortality rates among certain groups, but also because many countries across the world may no be well prepared to effectively respond to epidemic outbreaks **[3]**.

The earliest cases were identified as "pneumonia of unknown etiology", which is an illness without causative pathogen with the following symptomatic characteristics: fever, radio-graphic evidence of pneumonia, low or normal white-cell count and no symptomatic improvement after antimicrobial treatment up to 5 days **[4]**. However, it quickly turned into a pandemic, with many cases reported throughout China and allover the world **[2, 5]**. This rapid spread contradicted the first reports that indicated that human-to-human transmissions was limited or nonexistent **[1]**. By April 13th, almostfive months after 2019-nCoV first evidences, World Health Organization (WHO) reported a total 4.31 million cases worldwide, and a total of 294 thousand people already died from this coronavirus disease **[6]**. Yet according to WHO, almost all most reported regions/countries are in the stage of local transmission.

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.

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.

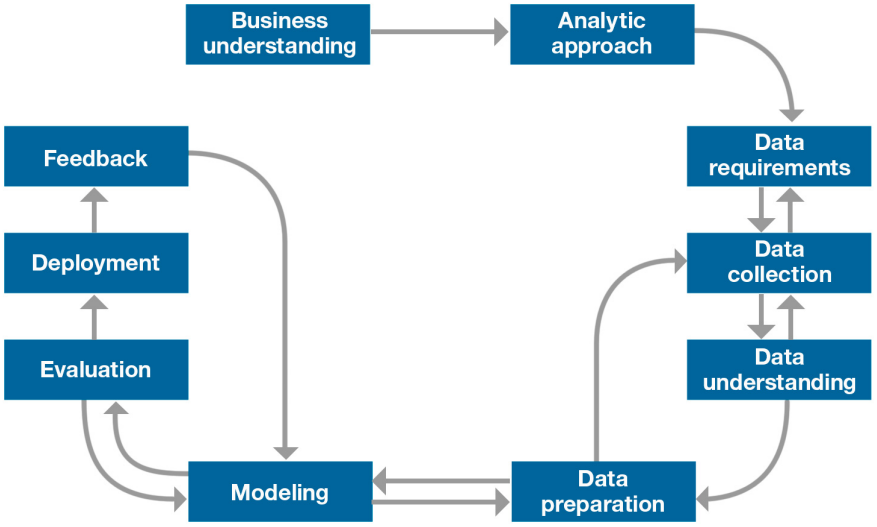
However, despite the large number of daily cases, deaths and infection rates, Brazil is facing an other big problem: the official COVID-19 data is not trustable. All mainly because the number of COVID-19 tests available is to low: 0.063 per 1,000 people - for instance, the numbers in Italy and USA are 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.

1.2. Project objective

According to Sun et al. **[5]**, 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 realtime 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.

1.3. 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.



(https://www.ibmbigdatahub.com/blog/why-we-need-methodology-data-science)

This methodology seeks to answer 10 well difined questions, each one related to a speficific step in the framework above.

From problem to approach:

1. What is the problem that you are trying to solve?
2. How can you use data to answer the question?

Working with the data:

1. What data do you need to answer the question?
2. Where is the data coming from (identify all sources) and how will you get it?
3. Is the data that you collected representative of the problem to be solved?
4. What additional work is required to manipulate and work with the data?

Deriving the answer:

1. In what way can the data be visualized to get to the answer that is required?
2. Does the model used really answer the initial question or doest it need to be adjusted?
3. Can you put the model in practice?
4. Can you get constructive feedback into answering the question?

In order to answer those questions, this project is structure as follows:

- Section 1, Introduction, presents the problem, the context, and how they will be explored. Thus, questions 1 and 2 from the methodology were already answered on subsections **1.1.** and **1.2.**
- Section 2, Assumptions and requirements needed to perform this project, presents the assumptions we assume to conduct this project and the libraries we need to import in order to perform all analysis needed.
- Section 3, The Data, presents the data required to perform the project, their sources and their prepation. Thus, questions 4, 5, 6, and 7 are answered on subsections **3.1.**, **3.2.**, **3.3.**, and **3.4.**.
- Section 4 answers our research question an the methodology questions 8 and 9.
- Section 5 presents the conclusions and recommendations for future projects.

---

2. Assumtions and requirements needed to perform this projet

2.1. 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 casualities 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 optimitic.
- The Brazilian official data for COVID-19 does not reflect the pandemic status in the country;
- Since there are not enought COVID-19 tests available, notary's offices have been registring 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.

2.2. Libraries needed to perform this project

The following libraries were used to perform this project:

- Pandas
- Numpy
- Folium
- Json
- Matplotlib
- Seaborn
- Wordcloud
- PIL

```
Libraries imported.
```

---

3. Working with the data

3.1. Data requirements

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

1. 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**.
2. 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.
3. 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 existis, it could potencialy be COVID-19, since there are not enought tests available.
4. A .csv or .xlsx file containing **social indicators** of each Brazilian city. This data can be confronted to COVID-19 and infrustrucuture data, in order to get insights about the spreading of this disease.
5. A .csv or .xlsx file containing **geographic coordinates** of Brazilian city centers. This data is needed to perform some data visualizations.
6. 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.
7. A .json file containing the **shape of each brazilian municipality**. This data is also needed to perform some data visualizations.

3.2. Data collection

In this sub-section I will get all data needed to perform this project. Notice that all the data must be wrangled and/or adjusted before used.

3.2.1. Number of ventilators available in each Brazilian city

First, let us import a dataframe (**df_resp**) that lists the number of ventilators for each Brazilian city. Ventilators are the main equipaments used to fight SARS-COV-2. The data is taken from the Brazilian Health Ministry website. This number displayed makes reference to the summation of ventilators present in private and public hospital. By far, public hospital are the majority of hospitals in Brazil. For curiosity, the Brazilian Public Healthcare System (SUS) is the biggest public healthcare system in the world.

```
b'Skipping line 2354: expected 1 fields, saw 3\nSkipping line 2357: expected 1 fields, saw 3\nSkipping line 2358:
expected 1 fields, saw 2\n'
```

Out[3]:

	IBGECode6	ventilator
5	110006	2.0
6	110009	3.0
7	110100	1.0
8	110010	3.0
9	110011	2.0

3.2.2. Number of COVID-19 confirmed cases and deaths in each Brazilian city

Now, let us import a dataframe (**df_gov**) from the website Brasil.io some data about the number of COVID-19 deaths and confirmed cases in Brazil. This website put together information about COVID-19 for all affected Brazilian city. They do it daily, so we have data for each city since the beggining of the outbreak in Brazil.

Out[4]:

	date	city	state	confirmed	deaths	population	IBGECode7	IBGECode6
0	2020-05-12	Acrelândia	AC	31.0	1.0	15256.0	1200013.0	120001
1	2020-05-12	Assis Brasil	AC	1.0	0.0	7417.0	1200054.0	120005
2	2020-05-12	Brasiléia	AC	3.0	0.0	26278.0	1200104.0	120010
3	2020-05-12	Bujari	AC	8.0	0.0	10266.0	1200138.0	120013
4	2020-05-12	Capixaba	AC	1.0	0.0	11733.0	1200179.0	120017

3.2.3. Number of deaths by respiratory diseases (COVID-19 Excluded) in 2019 and 2020

This is another important dataset (df_car) made available by Brasil.io. This dataset is set by Brasil.io by puting together data from many notary's offices. Notice that this dataframe shows us state-level data.

Out[5]:

	date	state	deaths_resp_plus
0	2020-05-13	AC	0
1	2020-05-13	AL	0
2	2020-05-13	AM	531
3	2020-05-13	AP	0
4	2020-05-13	BA	0

3.2.4. Brazilian social indicators

Now we will import a datafame (**df_soc**) containing Brazilian social indicators. These indicators will be used to investigate possible correlations in the data. The indicators are:

- Gini index;
- Percentage of the population with piped water;
- Percentage of rooms that house two ore more people;
- Human Development Index;
- Percentage of extremly poor people;
- Percentage of poor people.

Out[6]:

	IBGECode6	gini	pop_piped_water	density_rooms	HDI	extremely_poor	poor
1	520005	0.42	93.06	19.52	0.708	1.97	6.18
2	310010	0.47	88.50	8.62	0.689	1.85	7.94
3	520010	0.43	94.50	14.99	0.689	2.00	8.45
4	310020	0.54	98.40	10.34	0.698	1.61	6.69
5	150010	0.53	68.86	62.09	0.628	18.98	38.95

3.2.5. Database containing latitude and longitude for each Brazilian city

This is an official dataset from the Brazilian Institute of Geography and Statistics (IBGE).

Out[7]:

	IBGECode6	Latitude	Longitude
0	110001	-11.928729	-61.995865
1	110002	-9.906109	-63.033026
2	110003	-13.462377	-60.628509
3	110004	-11.434693	-61.456688
4	110005	-13.210987	-61.280200

3.2.6. Infection Fatality Ratio for each Brazilian state

The numbers above were taken from Imperial College COVID-19 Response Team (Report 21). The numbers make reference to estimated Infection Fatality Ratios (IRF) for each Brazilian state, based on their age pyramid and social indicatiors. I've made extrapolations for states that do not have IRF calculated by the report. In these cases, the average IRF for the region were considered. Since no states from the Central-West (CO) part of Brazil were not studied, for this states the average IRF between NE and SE was calculated.

Out[8]:

	region	state	IRF
0	N	AC	0.009
1	NE	AL	0.011
2	N	AM	0.008
3	N	AP	0.009
4	NE	BA	0.011

3.2.7. Boundaries for each Brazilian municipality.

This is an official file from the Brazilian Institute of Geography and Statistics (IBGE).

3.3. Data understanding

In this subsection we explore tha data looking for inconsistencies and insights. Since I don't want to create many cells in this notebook, I've already performed some **data understanding** and **data preparation** steps on previous cells. Such as:

- The shape and data types in all datasets were verified.
- I've also looked for inconsistencies on the IBGECodes.
- If we take a look at the dataset that contains the number of ventilators by municipality (**df_vent**), we see that its IBGECode contains only six numbers. On ther other hand, all other dataset have a IBGECode with 7 numbers. Since the IBGECode is a kind of ID and unique for each city, it is the best parameter to merge columns. Thus, many IBGECode7 were converted.
- Also in order to merge the dataframes, all IBGECode's must have the same type. Thus, we should convert all of them to **int**. I've already done this on previous cells.
- In **df_car**, the dataset that contains the number of deaths by respiratory diseases, I've already performed some calculations and reductions in the dataset. Notice that some logic operations had to be performed, in order to avoid negative and inf values.

3.4. Data Preparation

In this subsection I will construct the dataset that will be used in the modeling stage. Data preparations include data cleaning, combining data from multiple sources and transforming data into more useful variables. Notice that many data preparations steps were already performed in previous cells.

You may not notice but this step, along with data collecting, took most of the time I've spent on this capstone project.

First of all, I will create an index that compares the number of official and extra-official COVID-19 deaths. This index will be calculated for each Brazilian state. The index should assume values grater or equal to one. The values represents the percentage of death subnotification in the state: **1** means there is no subnotification, the subnotification percentage is given by the **index - 1**.

Out[10]:

	date	state	deaths_covid	deaths_resp_plus	possible_deaths	adj_index
1156	2020-03-30	MG	1	466	467	467.0
1444	2020-03-19	RJ	2	494	496	248.0
1129	2020-03-31	MG	2	463	465	232.5
1296	2020-03-25	RS	1	218	219	219.0
1289	2020-03-25	PE	1	217	218	218.0

Now we will merge all datasets together and create the dataset that will be used to model our problem.

Out[11]:

	date	city	state	confirmed	deaths	population	IBGECode7	IBGECode6	gini	pop_piped_water	...	poor	ventilator	Latitude
0	2020-05-12	Acrelândia	AC	31.0	1.0	15256.0	1200013.0	120001	0.54	93.87	...	36.56	0.0	-10.076392
1	2020-05-12	Assis Brasil	AC	1.0	0.0	7417.0	1200054.0	120005	0.61	80.31	...	44.09	0.0	-10.757958
2	2020-05-12	Brasiléia	AC	3.0	0.0	26278.0	1200104.0	120010	0.58	91.48	...	35.22	0.0	-10.705993
3	2020-05-12	Bujari	AC	8.0	0.0	10266.0	1200138.0	120013	0.58	81.12	...	37.61	0.0	-9.815664
4	2020-05-12	Capixaba	AC	1.0	0.0	11733.0	1200179.0	120017	0.56	86.19	...	35.00	0.0	-10.572112

5 rows × 23 columns

For the sake of simplicity, let us assume that all Brazilian municipalities have at least 1 ventilator each.

Now I will create more usefull variables.

Out[13]:

	date	city	state	confirmed	deaths	population	IBGECode7	IBGECode6	gini	pop_piped_water	...	confirmed_2_per_1k_pop	cor
0	2020-05-12	Acrelândia	AC	31	1	15256	1200013.0	120001	0.54	93.87	...	2.031987	
1	2020-05-12	Assis Brasil	AC	1	0	7417	1200054.0	120005	0.61	80.31	...	0.134825	
2	2020-05-12	Brasiléia	AC	3	0	26278	1200104.0	120010	0.58	91.48	...	0.114164	
3	2020-05-12	Bujari	AC	8	0	10266	1200138.0	120013	0.58	81.12	...	0.779271	
4	2020-05-12	Capixaba	AC	1	0	11733	1200179.0	120017	0.56	86.19	...	0.085230	

5 rows × 42 columns

The dataset is all adjusted. Let us proceed to the next sections.

---

4. Deriving the answer

Since this is a capstone project, many subsections of this project phase are out of our scope, such as deployment and feedback. Thus, this section will not be subdivided accordig to the methodology proposed by John Rollins. However, formal rigor stablished anyway.

First, I will explored the data we have, than perform some correlation analysis with the dataset. Then, maps will be displayed, in order to address and answer the research question.

4.1. Exploring the data to answer some questions

Let me slide our dataframe and take a look into some data.

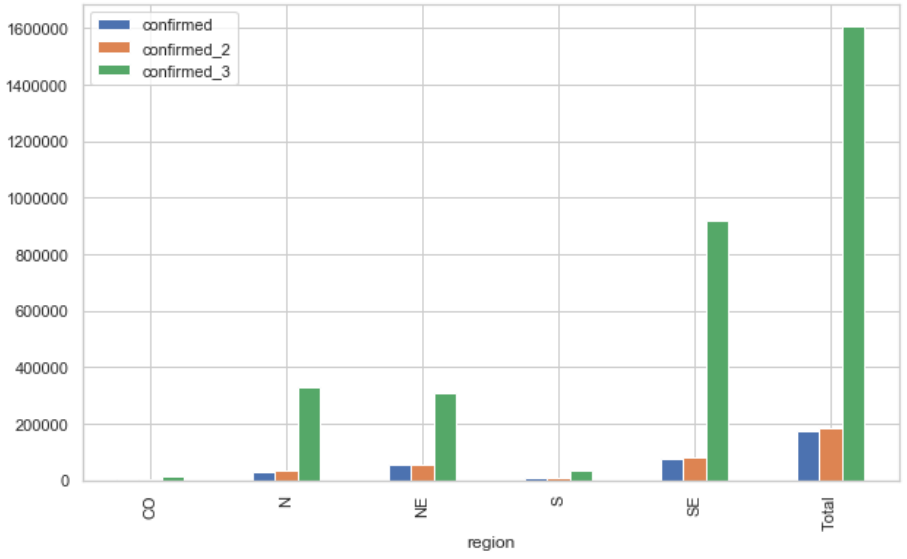
Out[14]:

region	confirmed	confirmed_2	confirmed_3
CO	4841.0	4841.0	13311.0
N	29187.0	36100.0	328109.0
NE	54962.0	54962.0	308700.0
S	8527.0	8527.0	36616.0
SE	74426.0	80238.0	920174.0
Total	171943.0	184668.0	1606910.0

In the dataframe above we can wee that according to the government, Brazil has today around 170,000 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. Not that much.

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

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



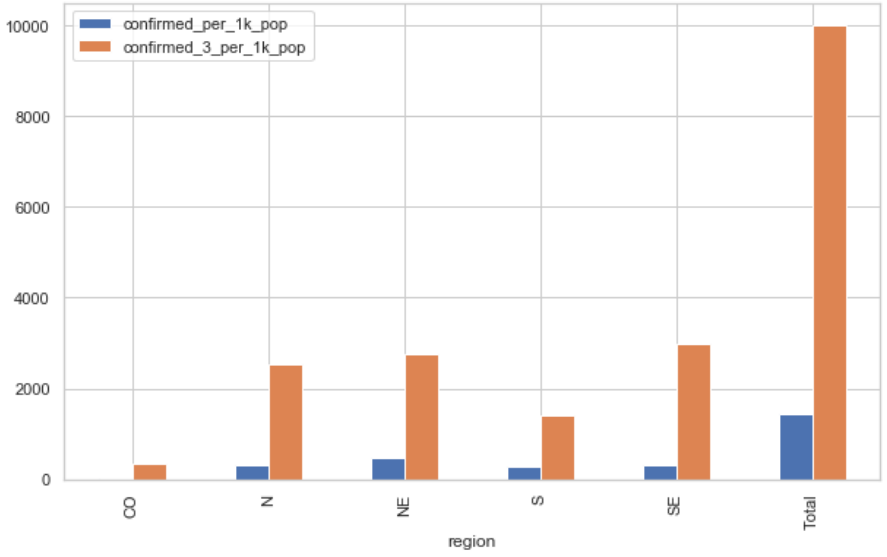
Lets take a look at the confirmed cases per 1.000 people now.

Out[16]:

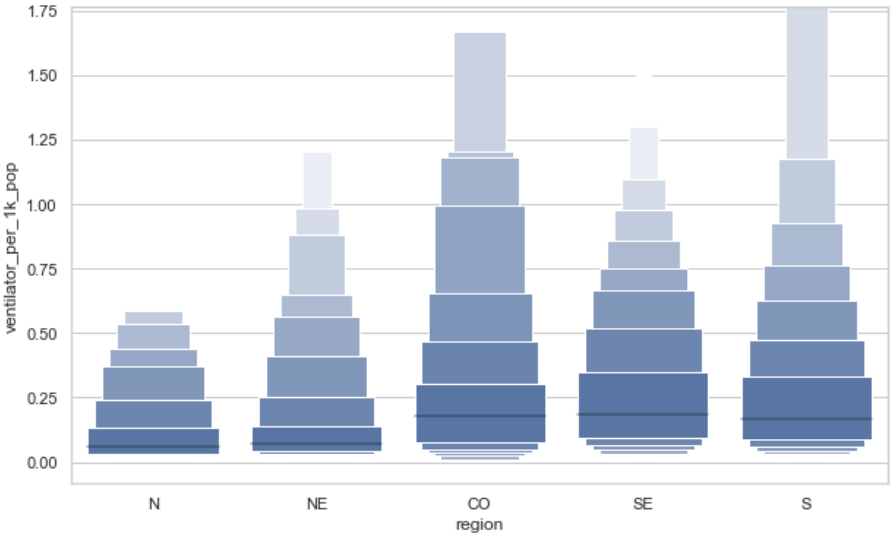
region	confirmed_per_1k_pop	confirmed_3_per_1k_pop
CO	38.30	348.39
N	319.97	2517.30
NE	482.32	2758.88
S	287.24	1395.82
SE	315.30	2977.27
Total	1443.13	9997.66

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 Northest are the most susceptible ones, with less ventilators per capta.

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



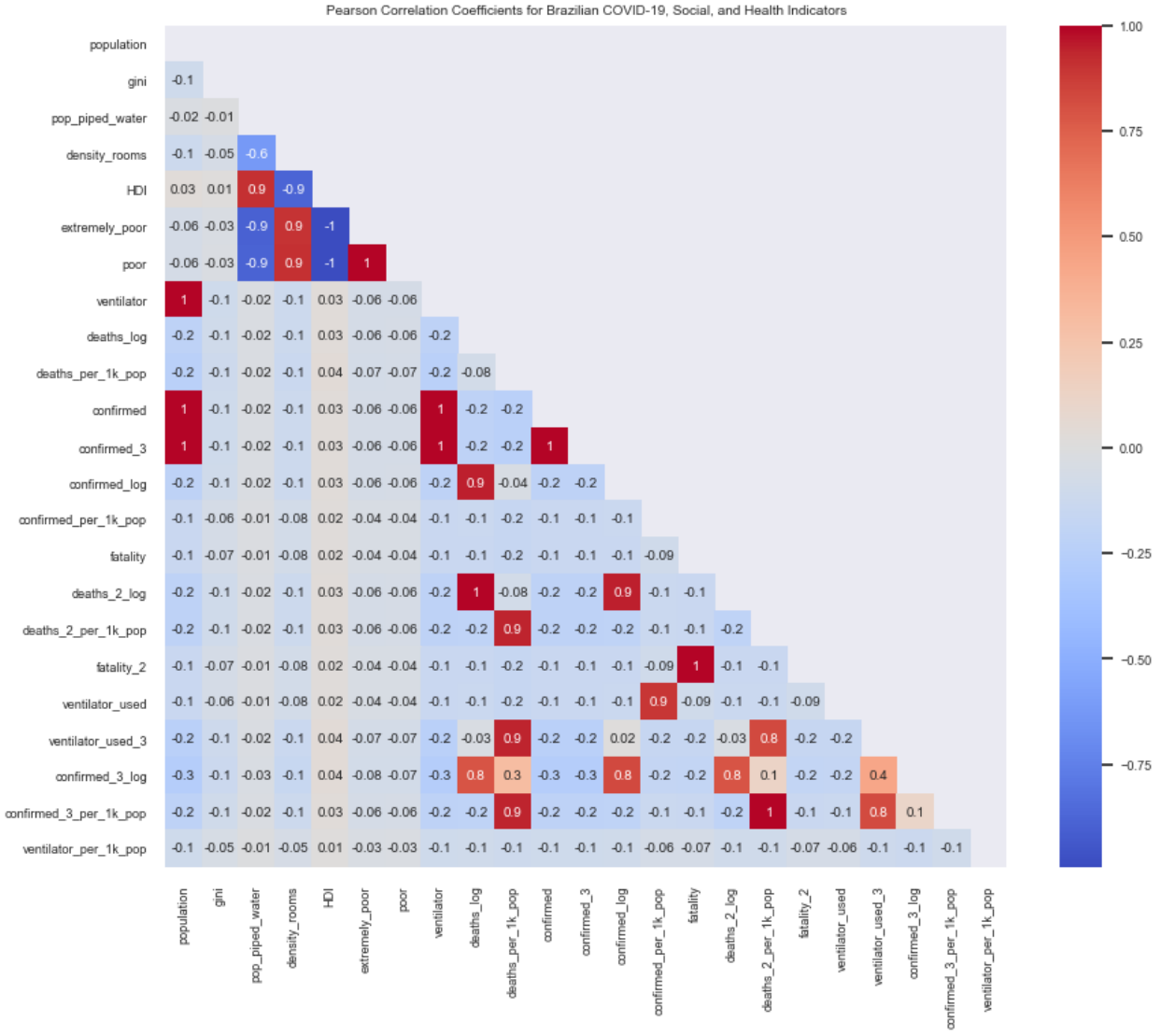
For instance, North and Northest are the regions with less ventilator available per 1.000 people.



4.2. Investigating possible correlations

First, in order to perform some correlations, let me slice the dataset and take only the columns that will be used.

Now, lets visualize the Pearson correlation coefficients for the vartiables above. Only coeficients above 0.6 and below -0.6 will be highlighted.



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

4.2. Maps

4.2.1. Word cloud maps

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 strugling 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 sould be getting most attention from media and goverment.

Which Brazilian cities get the most attention from media and government?

Which Brazilian cities should get the most attention from media and government?

Before, let us create a new dataframe, containing only the data needed to perform this task. Some arrangements will also be done.

Out[21]:

city	confirmed	confirmed_3_per_1k_pop
Acrelândia	30	7.28
Assis_Brasil	1	0.13
Brasiléia	1	0.04
Bujari	4	0.39
Capixaba	1	0.09

Now, let us plot our first word cloud.


```
<Figure size 300x300 with 0 Axes>
```

Now let us see the second one.


```
<Figure size 300x300 with 0 Axes>
```

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.

Yet those are cool representations, let us move to more informational representations.

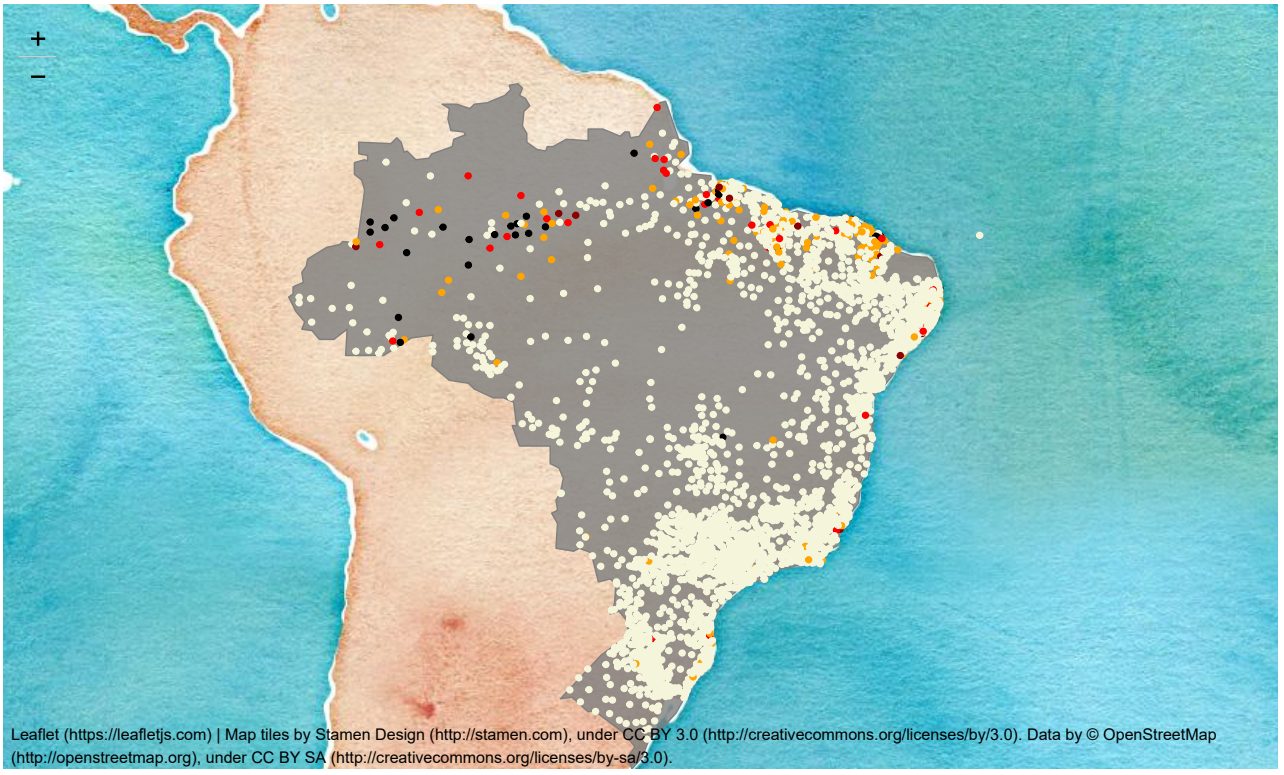
4.2.2. Creating a bubble map with folium

In this section I'm gonna show you how the official and real data differ from each other. But first, let me slice the dataframe.

Now let me show you a first map. This map is full of circles. 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 a given city. Complementatly, more darker is the color, greater is the percentage of ventilators already used.

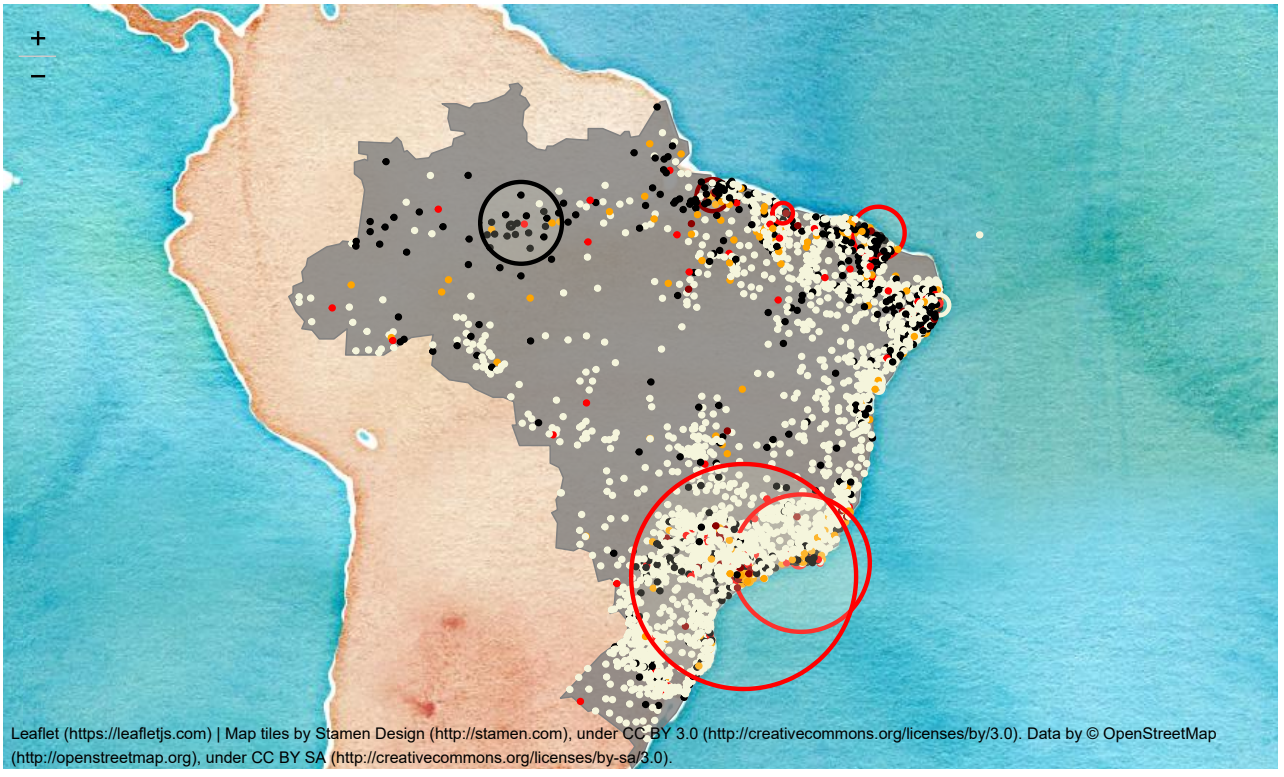
You can click on the circles on the map, in order to get informations about the city.

Out[27]:



Now let me show you the real Brazilian sitituation. The radius here are represented by the real number of cases, while the colors represent the real percentage of ventilators used.

Out[26]:



As you may have noticed, the real situation is much worse than the one disclosed by the government. The cities in the North (Amazon region) and Northest are those where the population is suffering more, due to the shortage of resources.

5. 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.

References

- [1] S. Perlman, “Another decade, another coronavirus,” 2020.
- [2] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan,J. Xu, X. Gu et al., “Clinical features of patients infected with 2019 novelcoronavirus in wuhan, china,” The Lancet, vol. 395, no. 10223, pp. 497–506, 2020. [3] J. R. Ortiz, V. Sotomayor, O. C. Uez, O. Oliva, D. Bettels, M. McCarron,J. S. Bresee, and A. W. Mounts, “Strategy to enhance influenza surveil-lance worldwide,” Emerging infectious diseases, vol. 15, no. 8, p. 1271,2009.
- [4] Q. Li, X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K. S. Leung,E. H. Lau, J. Y. Wong et al., “Early transmission dynamics in wuhan,china, of novel coronavirus–infected pneumonia,” New England Journalof Medicine, 2020.
- [5] K. Sun, J. Chen, and C. Viboud, “Early epidemiological analysis ofthe coronavirus disease 2019 outbreak based on crowdsourced data: apopulation-level observational study,” The Lancet Digital Health, 2020.
- [6] W. H. Organization et al., “Coronavirus disease 2019 (covid-19): Situationreport – 64,” 2020


That's all Folks!