COVID 19 Federal District

Data Analysis

1. Introduction & Business Problem

1.1. Problem Background:

The Federal District (in portuguese 'Distrito Federal') is one of the 27 federative units in Brazil. Located in the Midwest Region, it is the smallest Brazilian federative unit and the only one that has no municipalities, being divided into 31 administrative regions, totaling an area of 5,779,999 km². In its territory, it is located in the federal capital of Brazil, Brasília, and also at the headquarters of the federal government.

This region was created exclusively for the construction of the federal capital in the 1960s and today it is an important economic center, being the seventh federative unit with the largest gross domestic product (GDP) in Brazil (171.2 billion reais - 2012) and the highest GDP per capita in the country, 64,653 reais (2012). Despite this, Brasília has the highest income inequality among Brazilian capitals.

This inequality added to the advent of the global pandemic of COVID 19 resulted in an alarming scenario for the population of the Federal District. Because of this, it is necessary to have a greater understanding of which regions are having the greatest problems in dealing with the pandemic and which are most affected by COVID 19 so that it is possible to make strategic decisions to support the most affected regions.

1.2. Problem Description:

In this context, the Federal District is divided into 33 administrative regions, with the administrative region of Brasília being the main one. Administrative regions are territorial subdivisions of the Federal District, whose physical limits, established by the government, define the jurisdiction of government action for the purpose of administrative deconcentration and coordination of public services of a local nature. This action is exercised through each regional administration.

The administrative regions are:

Administrative Region	Estimated Population
Ceilândia	398285
Samambaia	254439
Taguatinga	222598
Plano Piloto	220393
Planaltina	189421
Águas Claras	148940

Recanto das Emas	145304
Gama	141911
Guará	132685
Santa Maria	125123
Sobradinho II	100775
São Sebastião	100161
Sol Nascente	91066
Vicente Pires	72879
Itapoã	68587
Sobradinho	68551
Sudoeste/Octogonal	53262
Brazlândia	52287
Riacho Fundo II	51709
Paranoá	48020
Arniqueiras	45851
Riacho Fundo	40098
SCIA	39015
Lago Norte	37455
Cruzeiro	33539
Lago Sul	29346
Jardim Botânico	27364
Núcleo Bandeirante	25072
Park Way	19824
Candangolândia	16848
Varjão	9215
Fercal	8746
SIA	1988

From these population data, it is necessary to gain an understanding of how these populations have been affected by COVID 19. To obtain a better awareness of the impact of COVID 19, it is necessary to look to investigate informations such as:

1. **Timeline of COVID 19**
2. **Age Group of the infected by COVID 19**
3. **Sex Group of the infected by COVID 19**

4. **Health condition of the infected by COVID 19**
5. **Distribution of the number of cases by administrative region**
6. **Distribution of the number of deaths by administrative region**
7. **Cases of COVID per region population**

This data analysis is fundamental for understanding the development of the COVID 19 pandemic in the Federal District. Based on this understanding, it is possible to monitor and develop support measures for regions that have been most vulnerable to the pandemic.

2. Data Description

2.1. Data Acquisition

The data acquire for this project is a combination form three major sources.

a. Official GDF Portal

The first source of data used was the **official COVID 19 database from the Government of the Federal District (GDF)**. This database is updated daily and the code used allows the analysis data to be constantly updated with a simple refresh on the console, on the currently date **(01/07/2020)** the total number was **49.290** total of cases. This database contains:

	Id	Data	Registration	Sex	Age	RA	UF	Condition	Pneumopatia	Nefropatia	Doença Hematológica	Distúrbios Metabólicos	Imunopressão	Obesidade	Outros	Cardiovasculopatia
0	1	24/06/2020	18/03/2020	Masculino	50 a 59 anos	Plano Piloto	DISTRITO FEDERAL	Recuperado	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	2	24/06/2020	18/03/2020	Masculino	50 a 59 anos	Lago Sul	DISTRITO FEDERAL	Recuperado	Sim	Não	Não	Sim	Não	Não	Não	Não
2	3	24/06/2020	18/03/2020	Masculino	40 a 49 anos	Lago Sul	DISTRITO FEDERAL	Recuperado	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	6	24/06/2020	18/03/2020	Masculino	>= 60 anos	Águas Claras	DISTRITO FEDERAL	Recuperado	Não	Não	Não	Sim	Não	Não	Não	Sim
4	8	24/06/2020	18/03/2020	Feminino	20 a 29 anos	Plano Piloto	DISTRITO FEDERAL	Recuperado	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Title	Info Description
Id	Patient Identification Number
Data	Update Date
Registration	Registration Date
Sex	Patient's Gender
Age	Age Range
RA	Administrative Region
UF	Federative Unit
Condition	Patient's Health Condition
Pneumopatia	Risk Factor COVID 19
Nefropatia	Risk Factor COVID 19
Doença Hematológica	Risk Factor COVID 19
Distúrbios Metabólicos	Risk Factor COVID 19
Imunossupressão	Risk Factor COVID 19
Obesidade	Risk Factor COVID 19

Outros	Risk Factor COVID 19
Cardio Vasculopatia	Risk Factor COVID 19

b. Latitude and Longitude Data

The second source of data is scraped from a csv file on github that contains the latitude and Longitude for the 33 Administrative region of Federal District. This dataset was build manually because of the lack of this type of informations in officials websites. The following are the columns:

	RA	Latitude	Longitude
0	Arniqueira	-15.8515	-48.0063
1	Brazlândia	-15.6701	-48.2005
2	Candangolândia	-15.8495	-47.9502
3	Ceilândia	-15.8219	-48.1021
4	Cruzeiro	-15.7841	-47.9417

Title	Info Description
RA	Administrative region
Latitude	Latitude from Administrative region
Longitude	Longitude from Administrative region

c. Federal District Population Data

The third source was was obtained from Wikipedia. However, the database was out of date and manual intervention was necessary to add two newly formed Administrative Regions (Arniqueiras and Sol Nascente).

	RA	Population
0	Ceilândia	398285.0
1	Samambaia	254439.0
2	Taguatinga	222598.0
3	Plano Piloto	220393.0
4	Planaltina	189421.0

Title	Info Description
RA	Administrative region
Population	Estimated Region Population

2.2. Data Cleaning and Preprocessing

The data preparation for each sources of data is done separately. I will try to describe all the changes that have been made to the data below.

a. Official GDF Portal

From the COVID 19 database, it was necessary to clean the columns that would not be used in the process and rename them according to their respective translation into English. In addition, it was also necessary to clear data from patients outside the Federal District - referred in dataset has 'Outros Estados' (translation 'Other states') and those who had not declared information and were listed as "Não Informado" (translation 'Not informed).

```
federal_district = df.drop(columns= ['Data', 'Id','Pneumopatia','Nefropatia','Doença Hematológica',
                                     'Distúrbios Metabólicos','Imunopressão','Obesidade','Outros',
                                     'Cardiovasculopatia'])
federal_district['Registration'] = pd.to_datetime(federal_district['Registration'])
federal_district = federal_district[federal_district.RA != 'Não Informado']
federal_district = federal_district[federal_district.RA != 'Outros Estados']
federal_district.head()
```

	Registration	Sex	Age	RA	UF	Condition
0	2020-03-18	Masculino	50 a 59 anos	Plano Piloto	DISTRITO FEDERAL	Recuperado
1	2020-03-18	Masculino	50 a 59 anos	Lago Sul	DISTRITO FEDERAL	Recuperado
2	2020-03-18	Masculino	40 a 49 anos	Lago Sul	DISTRITO FEDERAL	Recuperado

b. Latitude and Longitude Data

It was necessary to remove some lines that contained ‘nan’ data for regions that did not fit the analyzed samples of the 33 administrative regions.

```
data1 = "https://raw.githubusercontent.com/mathewspaes/COVID-19-Federal-District-Data-Analysis/master/Lat%20-%20Long%20DF.csv"
lat_long = pd.read_csv(data1, sep = ';')
lat_long.columns = ['RA','Latitude', 'Longitude']
Data = lat_long[lat_long.RA != 'Entorno DF']
Data = Data[Data.RA != 'Sistema Penitenciário']

Data.head()
```

	RA	Latitude	Longitude
0	Arniqueira	-15.8515	-48.0063
1	Brazlândia	-15.6701	-48.2005
2	Candangolândia	-15.8495	-47.9502
3	Ceilândia	-15.8219	-48.1021
4	Cruzeiro	-15.7841	-47.9417

c. Population Data

It was necessary to remove some columns that would not be used in the process and rename the columns according their respective data, in addition it was the items in the 'Population' column were transformed into floats for better use of the data.

```
pop_data = "https://raw.githubusercontent.com/mathewspaes/COVID-19-Federal-District-Data-Analysis/master/Population-DF.csv"
pop = pd.read_csv(pop_data)
pop = pop.drop(columns= ['Posição'])
pop.columns = ["RA", "Population"]
pop['Population'] = pop['Population'].astype(float)
pop.head()
```

	RA	Population
0	Ceilândia	398285.0
1	Samambaia	254439.0
2	Taguatinga	222598.0
3	Plano Piloto	220393.0
4	Planaltina	189421.0

3. Exploratory Data Analysis

In this segment, the data were analyzed according to each category of the COVID 19 dataset. Namely as:

1. **Cases Timeline**
2. **Age**
3. **Gender**
4. **Administrative Region (in Portuguese - RA)**
5. **Health Condition.**

To this end, several Python libraries - such as **Pandas, Plotly, MatPlot and SeaBorn** - will be used to build a comprehensive analysis of the COVID 19 pandemic data in the Federal District.

3.1. Evolution of the COVID 19 cases timeline

The first stage of the exploratory data analysis process is necessary to build the pandemic evolution curve in the federal district in order to understand the pandemic moment in the region. The graph was obtained from the GDF database using the Plotly library to generate an interactive graph of the accumulated number of cases.

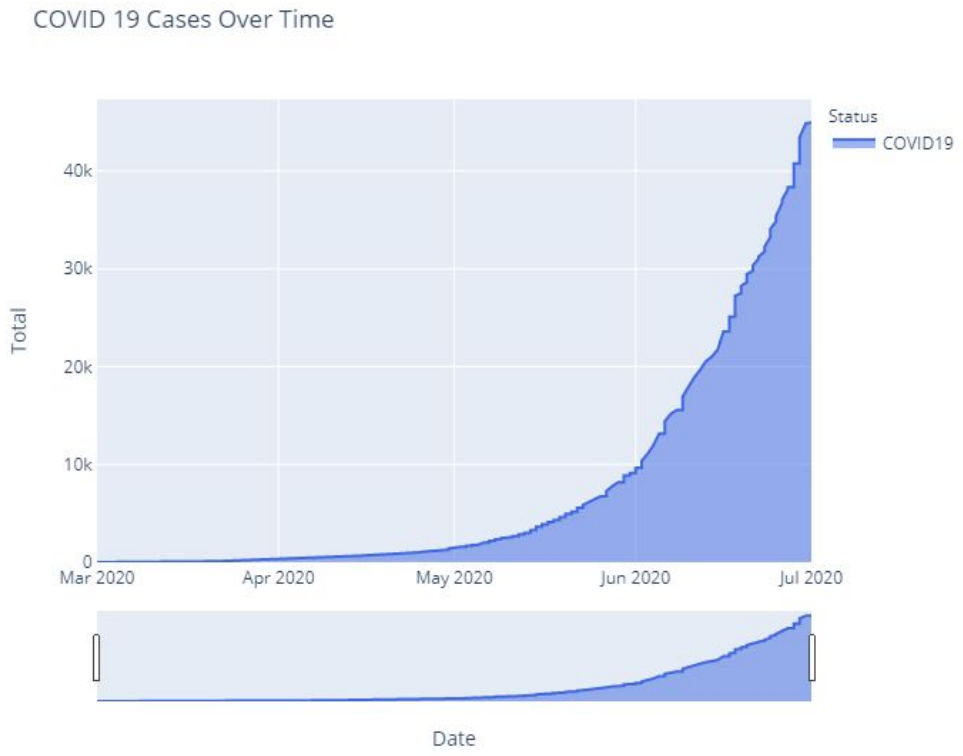

Figura 1: COVID 19 Cases Timeline

Thus, it is possible to observe that the pandemic is growing exponentially within the Federal District. It is interesting to note that this increase was catalyzed by the beginning of the relaxation of the quarantine and the easing of social isolation, as shown in the next graph, where the values of COVID 19 cases reported per day are represented.

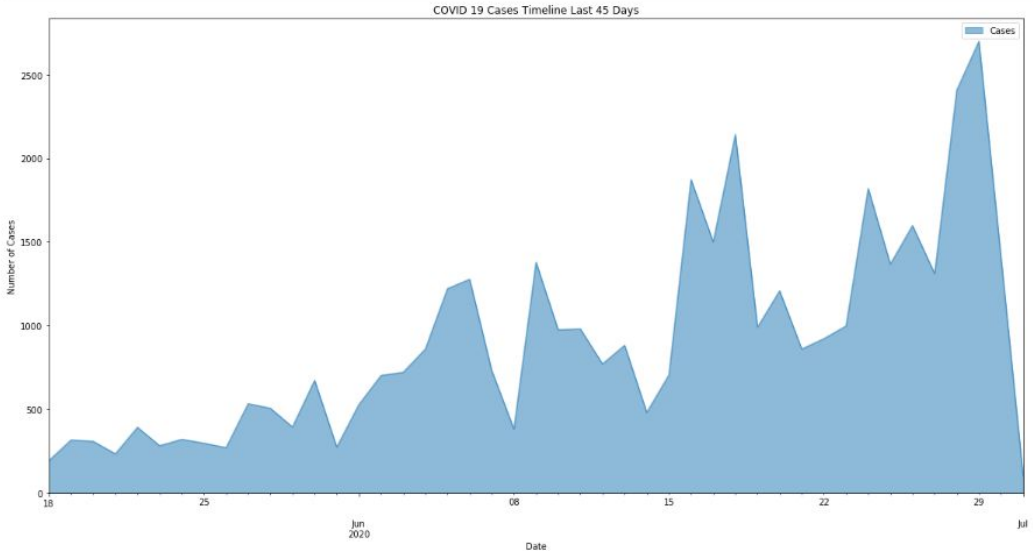

Figura 2: Covid 19 Cases Timeline Last 45 Days

Consequently, this generated a large increase in the number of deaths by COVID 19 that is represented by a curve still in the beginning of a potential exponential curve, less inclined than the curve of the number of cases.

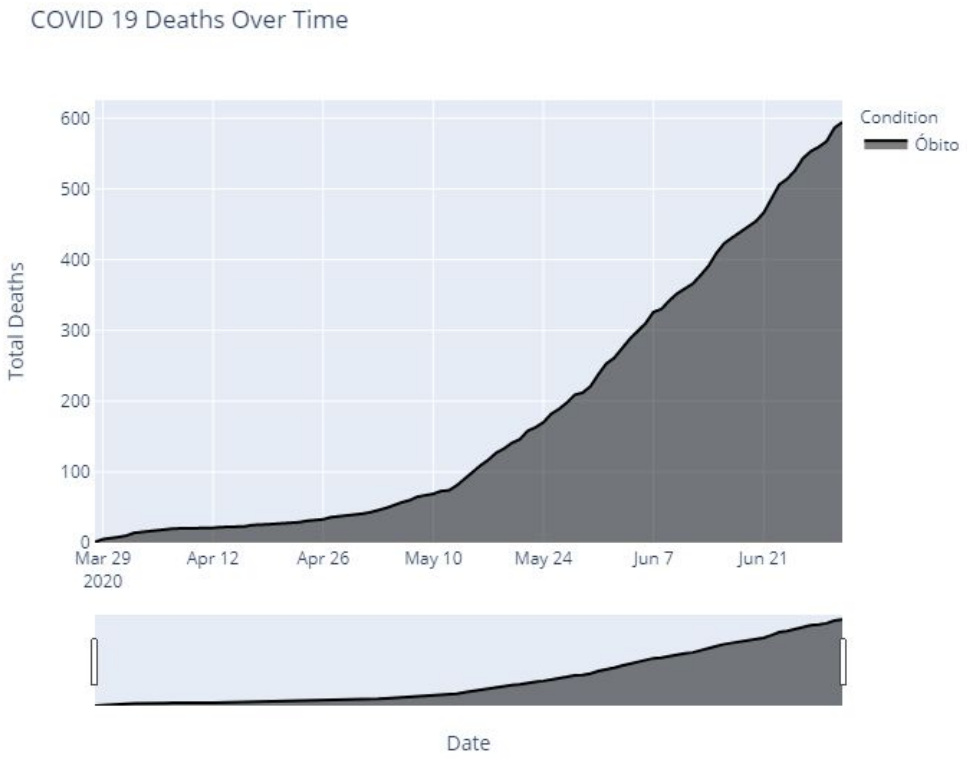

Figura 3: COVID 19 Deaths Timeline

3.2. Age Range

In this section, we will look at which age groups have been most affected by COVID 19 - both from the perspective of the number of cases and by the people who end up dying from the disease. The graph below shows the distribution of covid 19 cases by age group.

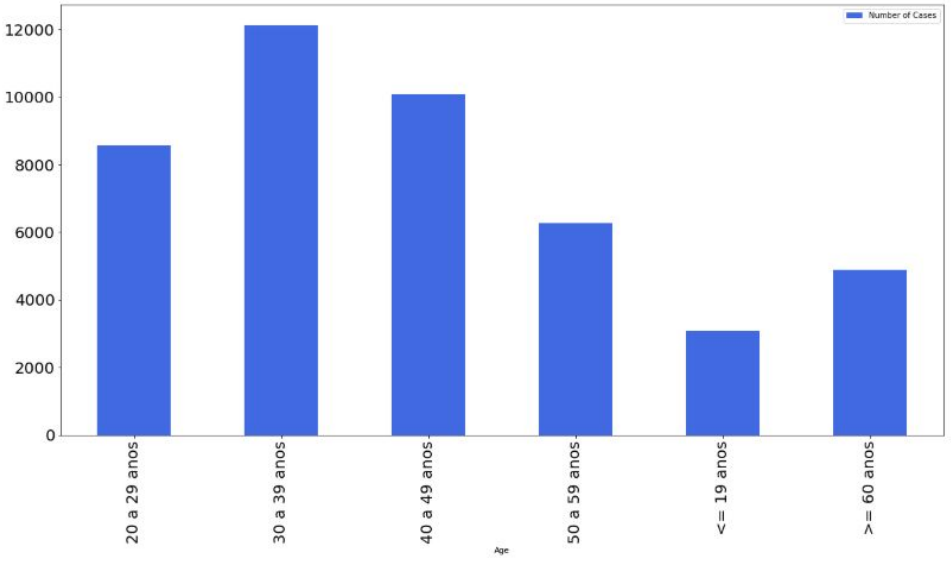

Figura 4: COVID 19 Cases Per Age Range

As we can see, there is a large concentration of cases among what is considered the active professional population (Between 20 and 50 years old). However, when we analyze the number of deaths we observed that **86%** (595 so far) of the numbers are in the age group older than 50 years.

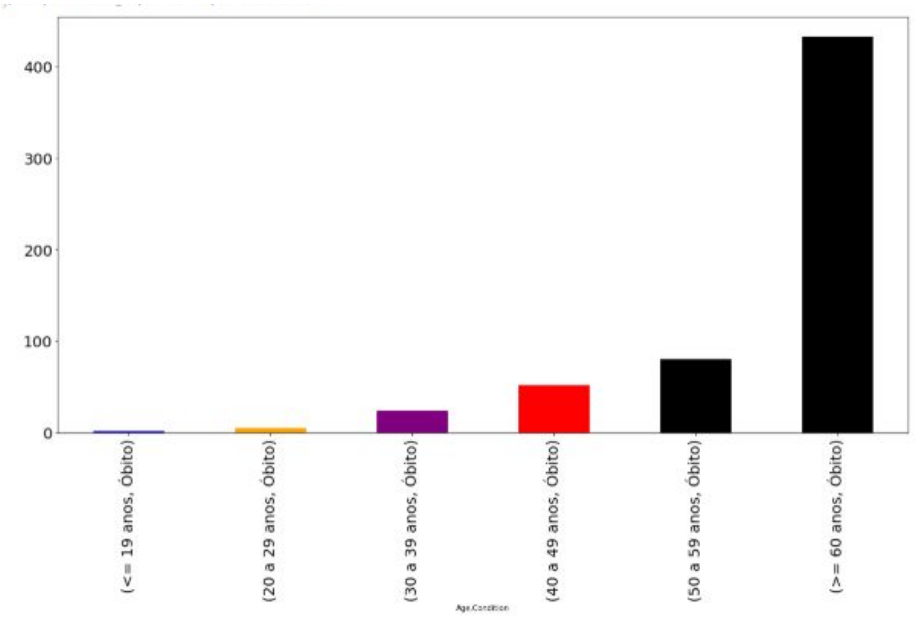

Figura 5: COVID 19 Age-Death Relation

3.3. Gender

Comparison between men and women with regard to both COVID 19 cases and deaths due to Pandemic.

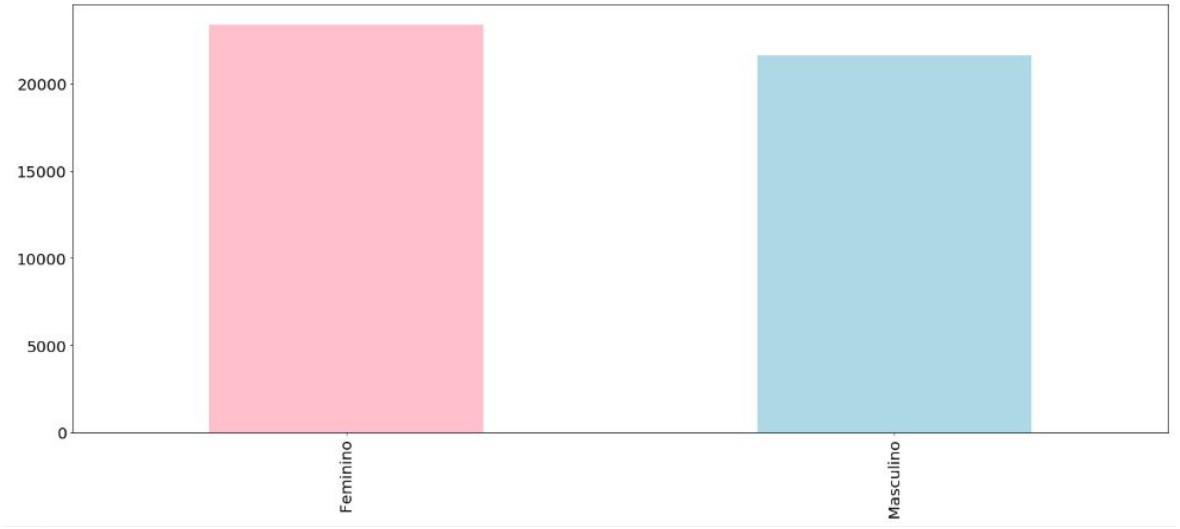

Figura 8: COVID 19 Cases Gender Distribution

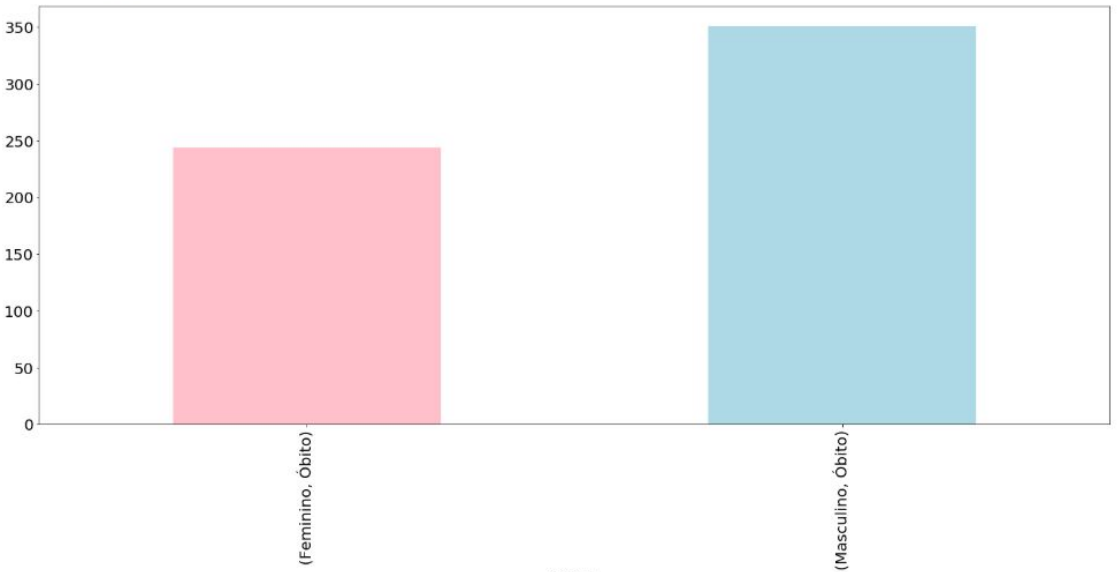

Figura 9: COVID 19 Deaths Gender Distribution

3.4. Administrative Regions (in portuguese, RA)

Now let's look at how the pandemic has affected the 33 Administrative Regions that make up the Federal District.

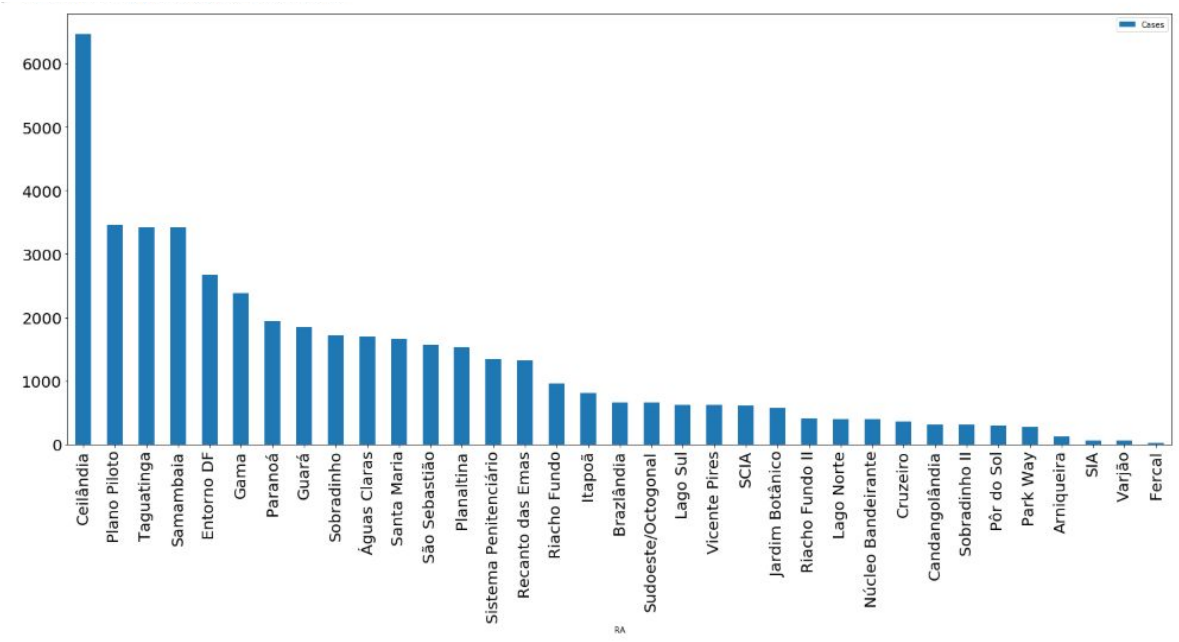

Figura 6: COVID 19 RA Cases Distribution

As we can see in this graph, the higher concentration of cases coincides with the Administrative Regions with the highest population densities. A similar pattern is observed when exposing the number of deaths by region.

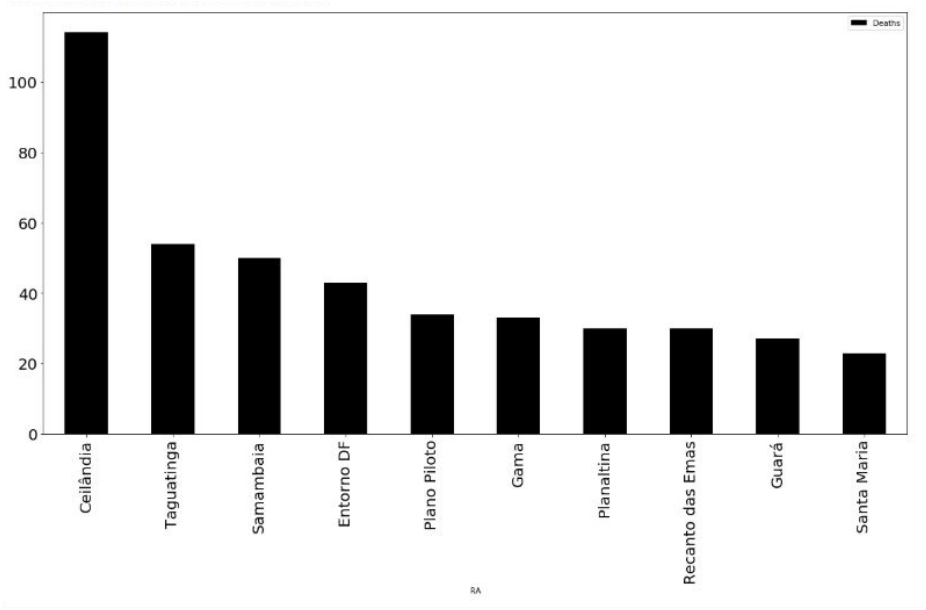

Figura 7: COVID 19 RA Deaths Distribution

In this graph we see a similar arrangement to the number of cases. However, it is worth noting that there is a greater tendency of death by COVID 19 for cases in peripheral regions - such as Ceilândia, Taguatinga and Samambaia - than for cases that occur in the economic center of Brasilia - represented by Plano Piloto.

3.5. Health Condition

In this section we look at the COVID 19 curve in the Federal District from another perspective, now focusing on the health condition of each patient. Through this view, we can have a greater understanding of how the number of current cases affect the health system of the Federal District and, also, we have a more precise view of the development of the pandemic.

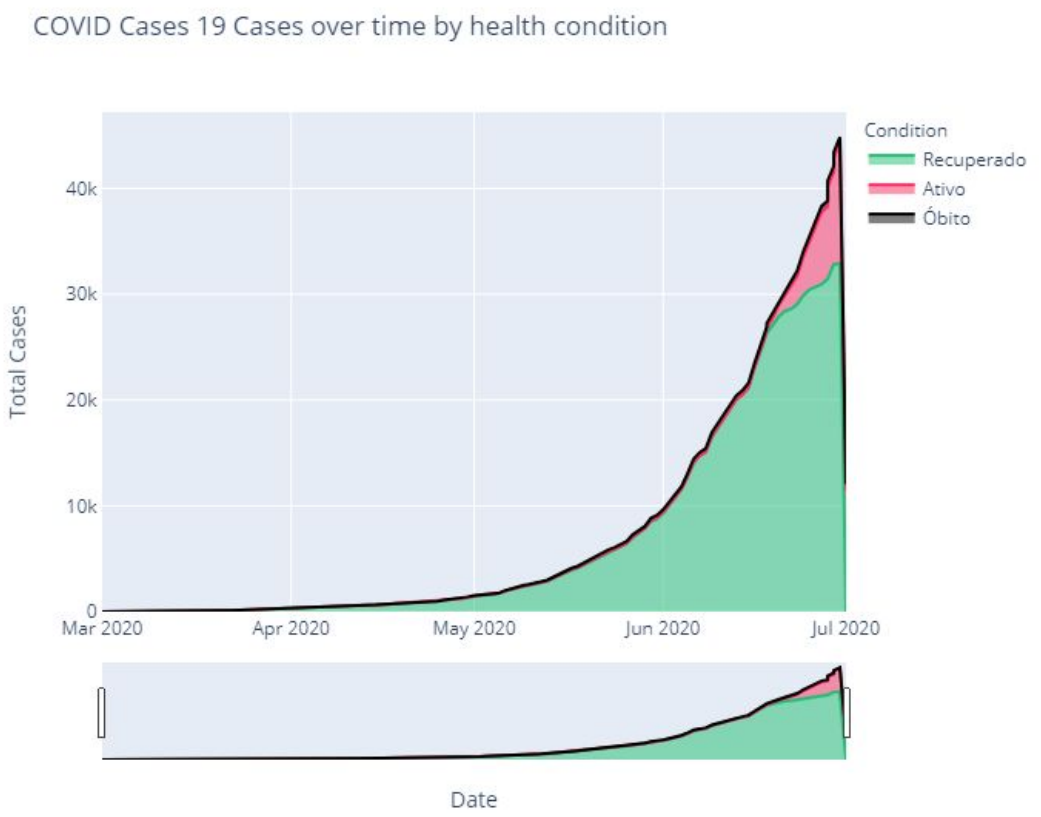

Figura 9: COVID 19 Health Condition Distribution

Within the value "Ativos", we have a compilation of 4 other values represented by their respective health status: "Light", "Moderate", "Severe" and "Not Informed".

Status	Cases
Light	13
Moderate	74
Severe	30
Not Informed	11.341

It is important to highlight the number of cases labeled "Not Informed", this can be considered an example of the difficulty of control and monitoring of cases by the GDF. Furthermore, caution is needed so that this does not lead to an even worse worsening of the pandemic in the Federal District.

4. Exploratory Map Analysis

In this step we will build an exhibition using the map of the Federal District, the intention is to expose the territorial distribution of COVID 19 cases with regard to both the number of cases per administrative region and the number of deaths per administrative region.

Based on the dataset that contains the longitudinal and latitudinal data for each Administrative Region as shown below:

	RA	Latitude	Longitude
0	Arniqueira	-15.8515	-48.0063
1	Brazlândia	-15.6701	-48.2005
2	Candangolândia	-15.8495	-47.9502
3	Ceilândia	-15.8219	-48.1021
4	Cruzeiro	-15.7841	-47.9417

To that end, I used the Python **Folium** library to build a detailed geographic view of the Federal District.

4.1. Map visualization of COVID 19 cases

The map below was constructed using the latitude and longitude values to have a better location of each administrative region.

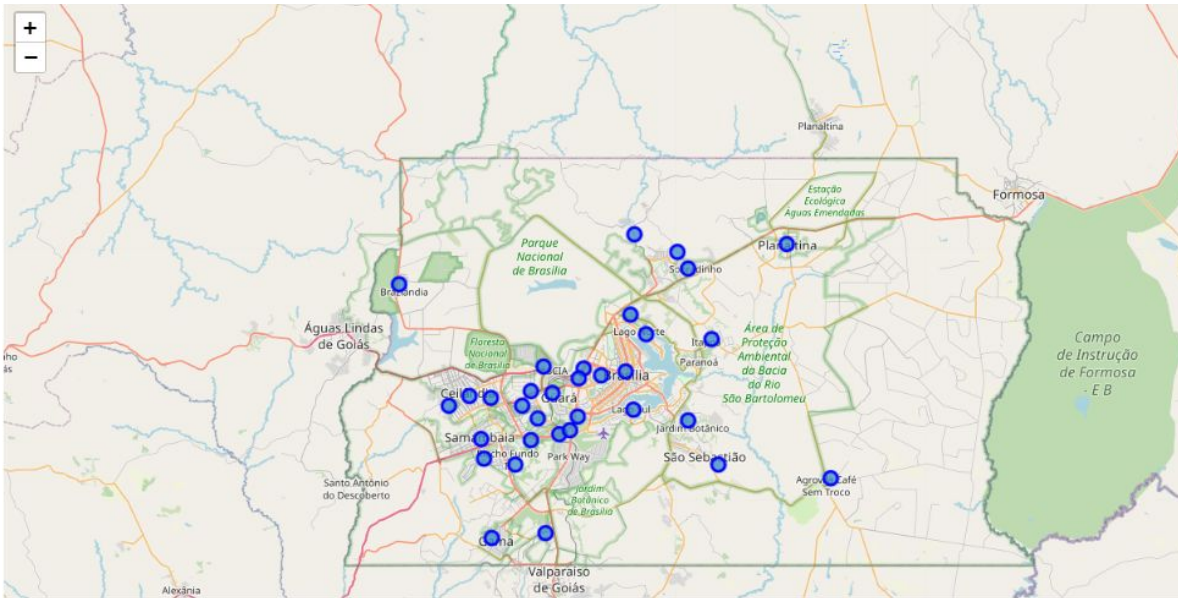

Figura 10: COVID 19 Cases RA Location Map

Based on the points marked on the previous map, a heat map was created with the distribution of COVID 19 cases by the Federal District to better visualize the geographical situation of each RA

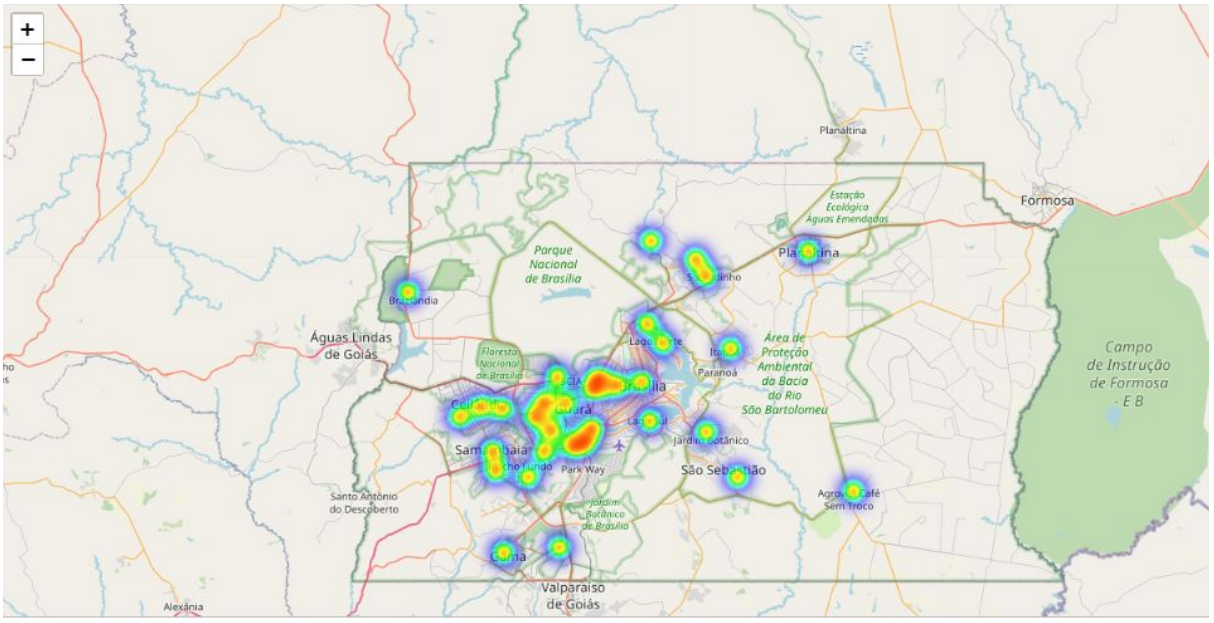

Figura 11: COVID 19 Cases RA Heat Map

In this way, the concentration of cases in the economic centers of the Federal District is more visual and reinforces the analysis on the distribution of the number of cases made in '**3.3.**'.

4.2. Map visualization of COVID 19 deaths

Likewise, a geographical view of the concentration of the number of deaths was developed following the same method as before.

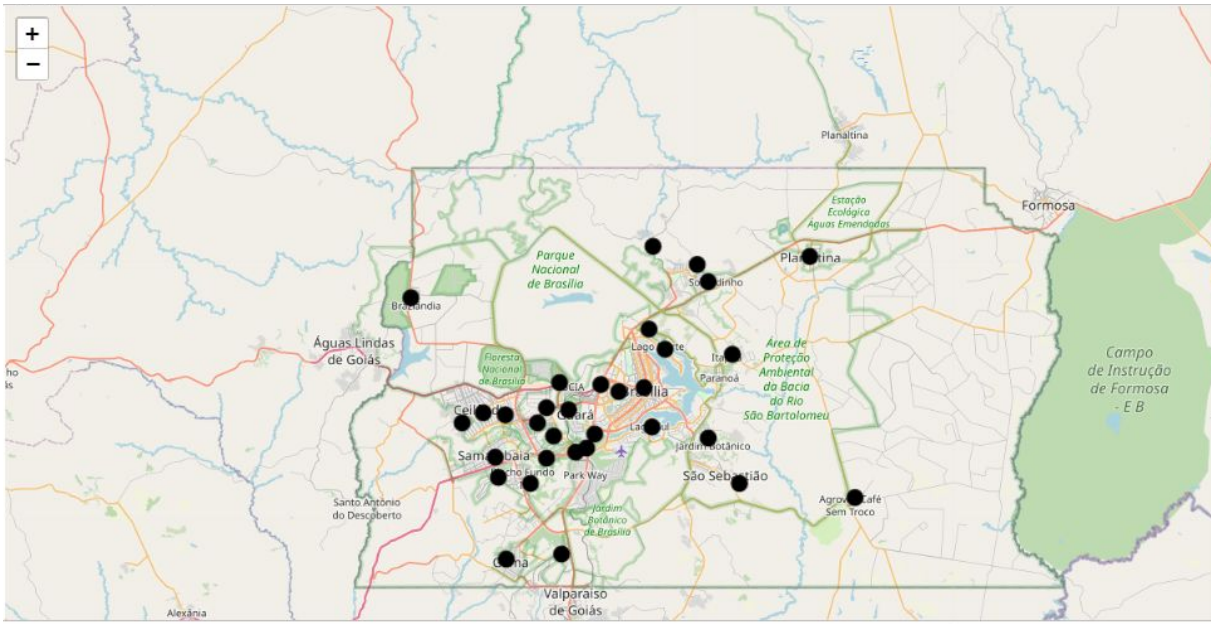

Figura 12: COVID 19 Deaths RA Location Map

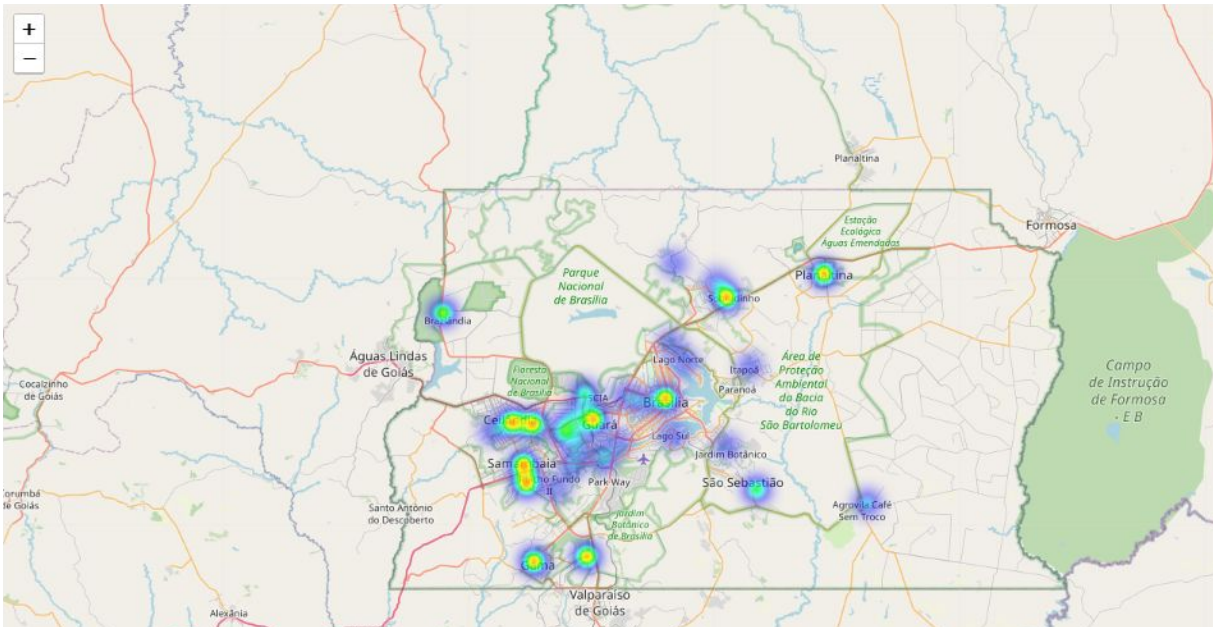

Figura 13: COVID 19 Cases RA Heat Map

The heat map of deaths caused by COVID 19 emphatically reinforces the situation pointed out in **'3.3.'** - where we get the perception that the surrounding regions concentrate the highest number of deaths, consequently this shows that people in these regions have a higher risk of dying due to COVID 19.

5. Conclusion

Based on the analysis of the available data, it was possible to raise important insights about the development of the COVID-19 Pandemic in the Federal District.

1. **Pandemic Evolution in the Federal District**

The Federal District is experiencing a exponential growth of the COVID-19 Pandemic in the last month. On 06/02/2020, the region accumulated about 11,256 cases of accumulated COVID-19. A month later, on 07/02/2020 we have about 51,123 cases of COVID - representing a 454% growth in the last 30 days.

This coincides with the relaxation of the quarantine and the distance measures in the region, which indicates that it would be necessary to make containment measures to try to flatten the curve again. From there, it is not recommended that the Federal District continue the process of opening shops and possible centers of agglomerations (such as gyms and shopping malls) as this can further accelerate the contamination of COVID-19

2. **Administrative Regions**

It is worth noting that the data show that contamination by COVID-19 has intensified in the satellite cities of Brasília (Ceilândia, Taguatinga and Samambaia standing out). Likewise, it is emphasized again that the surrounding regions concentrate the highest number of deaths, consequently this shows that people in these regions have a higher risk of dying due to COVID 19.

In addition, Ceilândia has stood out negatively as the center of the COVID-19 pandemic in the Federal District. This region needs to be observed carefully and, possibly, it would be interesting that it received more support from the government of DF to contain the pandemic.

3. Number of cases labeled "Not Informed"

It is necessary to monitor the development of cases that do not yet have more accurate information because this can cause a rapid change in the scenario for analysis. It is highlighted that, in addition to the slowness of the verification process, there is still a high probability of underreporting of cases that would further increase the results found by the analysis.