Master 2 Data Science pour la Santé

2019/2020

Rapport : Projet Covid-19

Ulrich TCHUENKAM

1) Recherche de différentes sources de données liées au Covid-19 (cas, décès, hospitalisés, admis en réanimation, mesures de santé publique [fermeture des écoles, confinement, déconfinement], ...)

The datasets used throughout the project originated from Santé Publique France, Google Mobility and Corona Virus Statistics. From Santé Publique France, I extracted hospital data in CSV and in Geographical formats. Google Mobility provided the datasets reporting movement to and some common public places like groceries, transport stations, work offices and pharmacies in France. From Corona Virus Statistics, I got an aggregated dataset containing the daily number of cases reported, the daily number of hospitalized people, the daily number of deaths and the daily number of recovered in France as a whole.

Data Sources:

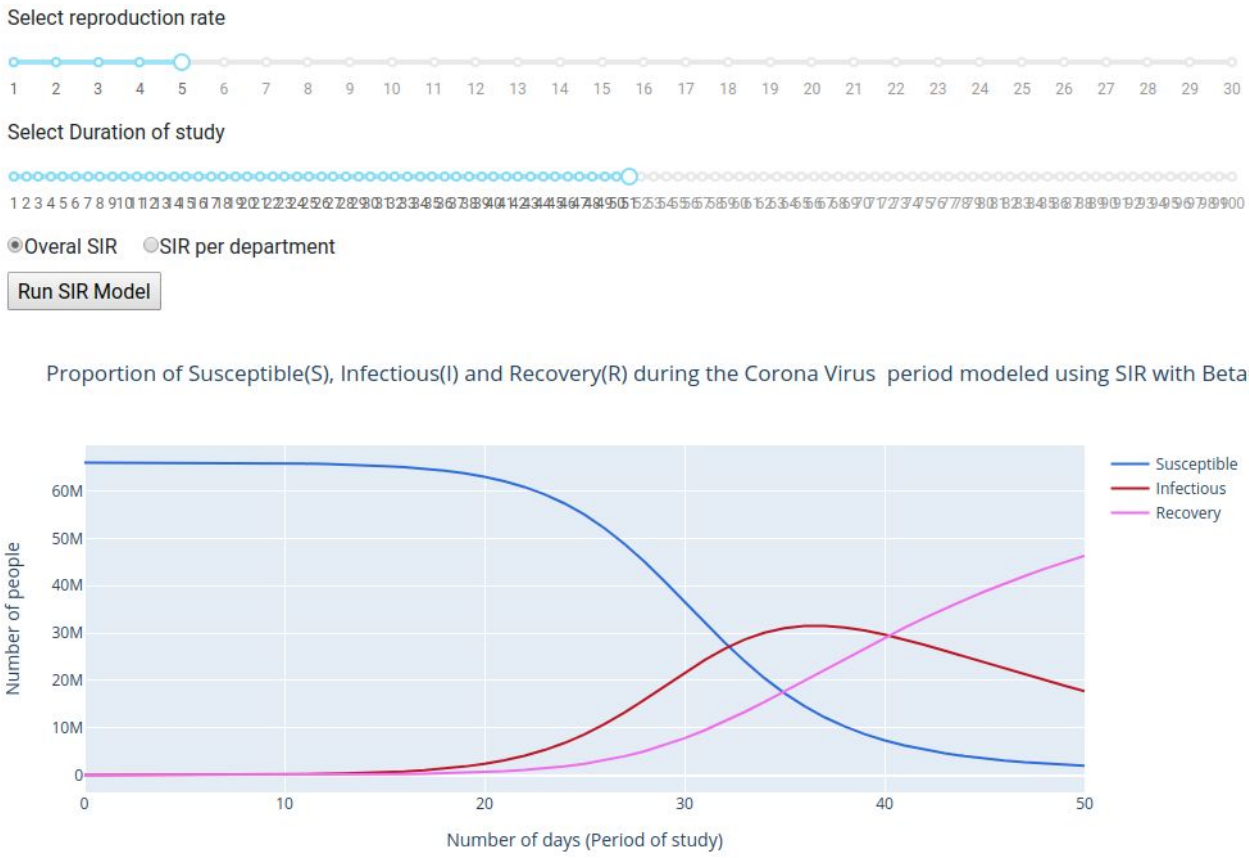
Santé Publique France CSV
Santé Publique France GÉODES
Google Mobility
Corona Virus Statistics

2) Mise en place d'un flux automatique de collecte des données pour alimenter une base de données (et transformation des données si nécessaire)

This section was addressed after manually collecting all the datasets used in the project. At the end of the project, I wrote a python script that scrapes all the pages where I downloaded my datasets. The script locates the required datasets and downloads into a predefined location. Preferably, the script should be run daily as it creates a folder named with the date of the day. To locate and download the data, the script primarily makes use of the python package called *BeautifulSoup.* The extracted data is saved in the CSV format. The script is accessible in the Github repository of the project. More details about the script is equally available in the GitHub repository of the project.

3) Développement et analyse d'une modélisation épidémiologique (à base d'un modèle compartimental de type SIR. C'est un modèle générique que vous avez toute la latitude de compléter, modifier et adapter compte-tenu du virus en circulation) sur la diffusion de l'épidémie, avec une comparaison régionale concernant les différents paramètres (i.e. évolution de la contagion, des taux de guérison, de décès, etc.)

I developed a SIR model using the python package *scipy.* Precisely, the function *scipy.integrate.odeint* was used to integrate the system of ordinary differential equations. Before developing the SIR model I did some data preprocessing. From the hospital data reporting the number of daily hospitalized patients, the number of daily patients under reanimation, the number of daily deaths and the number of daily recoveries, was able to compute the I and R components of the SIR model. From the initial value of I and the total population of France, I could compute the S components of the SIR model. The recovery rate for the model was obviously 14 days. The data was aggregated on a daily basis throughout the country. As you can see in the following screenshot taken from my dashboard, the user is free to define the reproduction rate and the duration of the model execution.



For example, considering the confinement period(70days, 7412 initial infected and 2056 initial recovered) and a reproduction rate of 5, we have a mean recovery rate(gamma) 0.07 of and a contamination rate(Beta) of 0.35. At these rates, we will have 5.1M infectious, 690K susceptibles and 60.2M recovered by the end of the period. You can visit the second tab of my dashboard and run some simulations for yourself. For the moment, the SIR model on the whole country is available meanwhile the SIR model at the level departments is not yet available.

Data preprocessing and SIR model codes

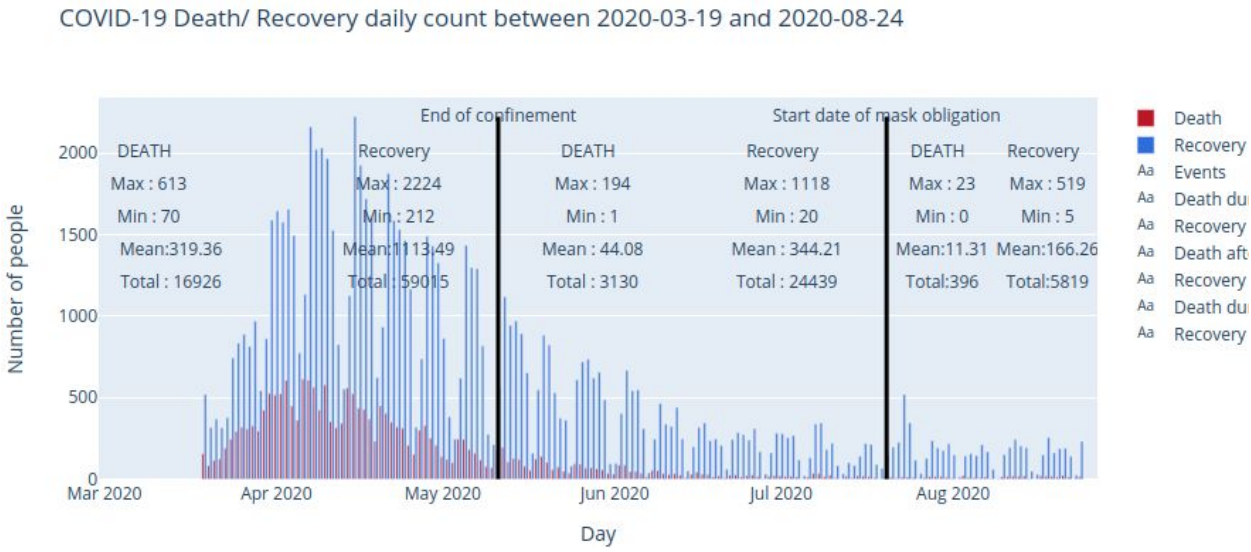
Dashboard

Source code of dashboard

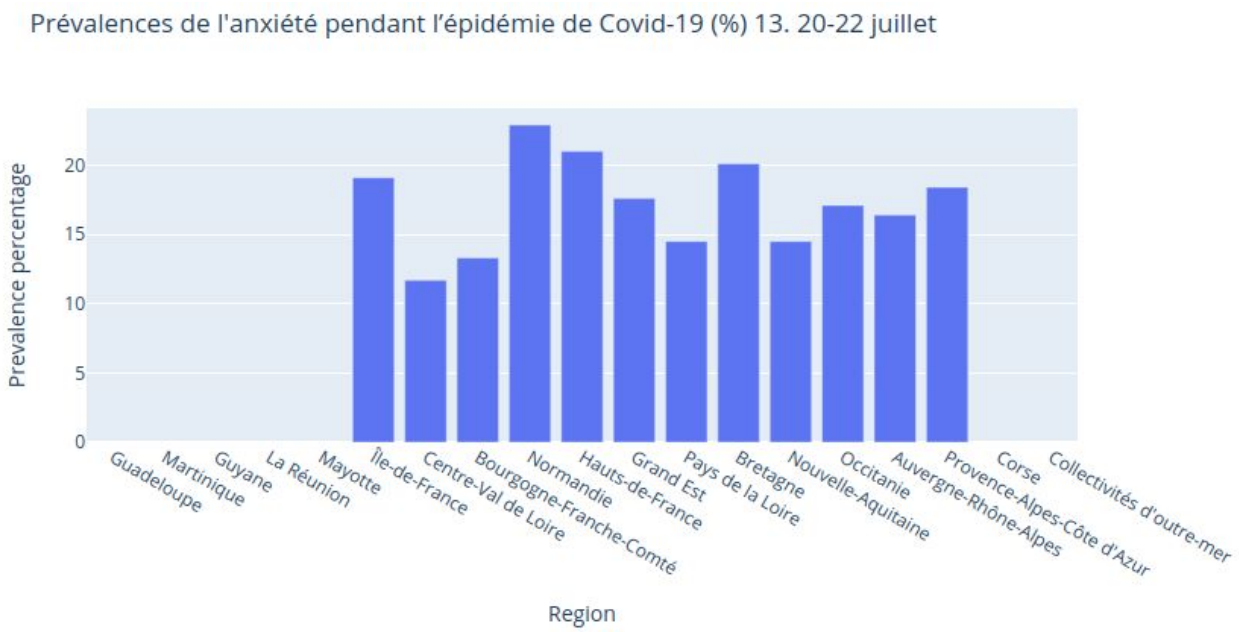
4) Évaluation de l'impact des mesures prises : distanciations sociales, fermeture des écoles, confinement, port de masque, déconfinement.

The first visualization in the 4th tab of my dashboard shows some figures separated by the starting dates of enforcement of the various measures. The visualization confirms the government's forecast of high death and contamination rates which resulted in a rigorous confinement to limit the spread of the virus. It can be seen that though the death rate is high during this period, we can also notice an important number of

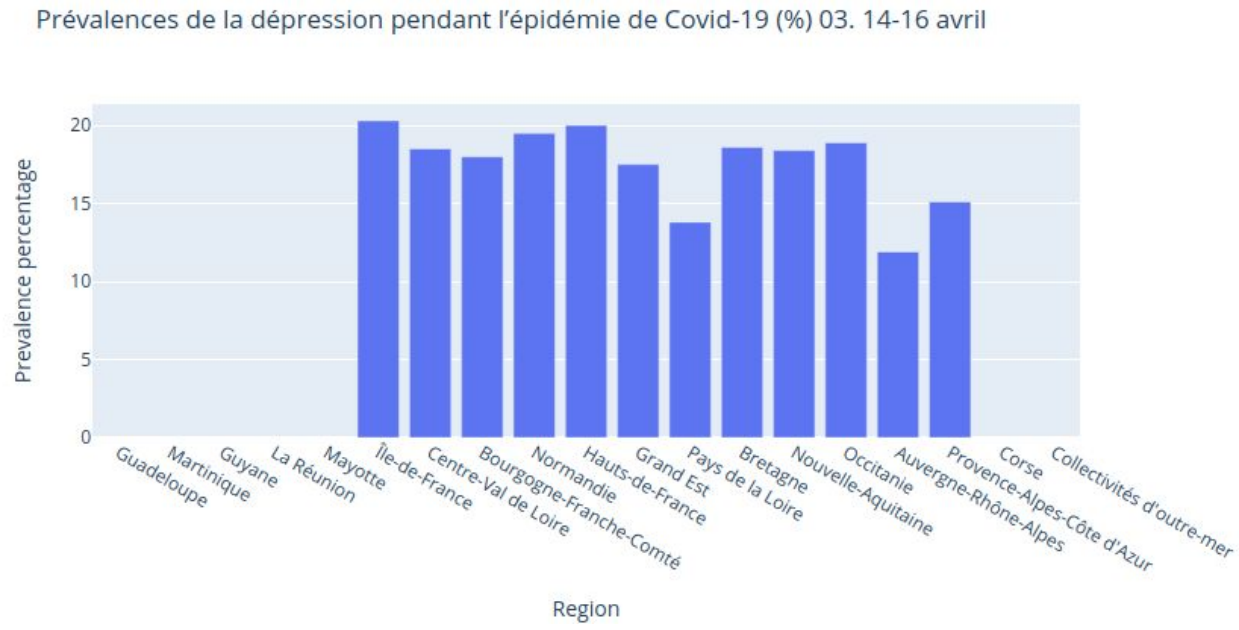
recoveries. At the end of the confinement period, new measures such as the obligation of social distancing and hygienic measures were enforced and the death rate greatly reduced with an increased number of recoveries. The last part of the visualization marks the period when the mask became an obligation in France, here we can notice a further decrease in the death rate.



The impact of the various measures could also be evaluated based on the indicators published by Sante Publique France. Six periodic indicators are published consisting of: anxiety, depression, sleep complication, adoption of mask, mean number of hygienic measures adopted out of the 4 prescribed and the proportion of adopted hygienic measures.



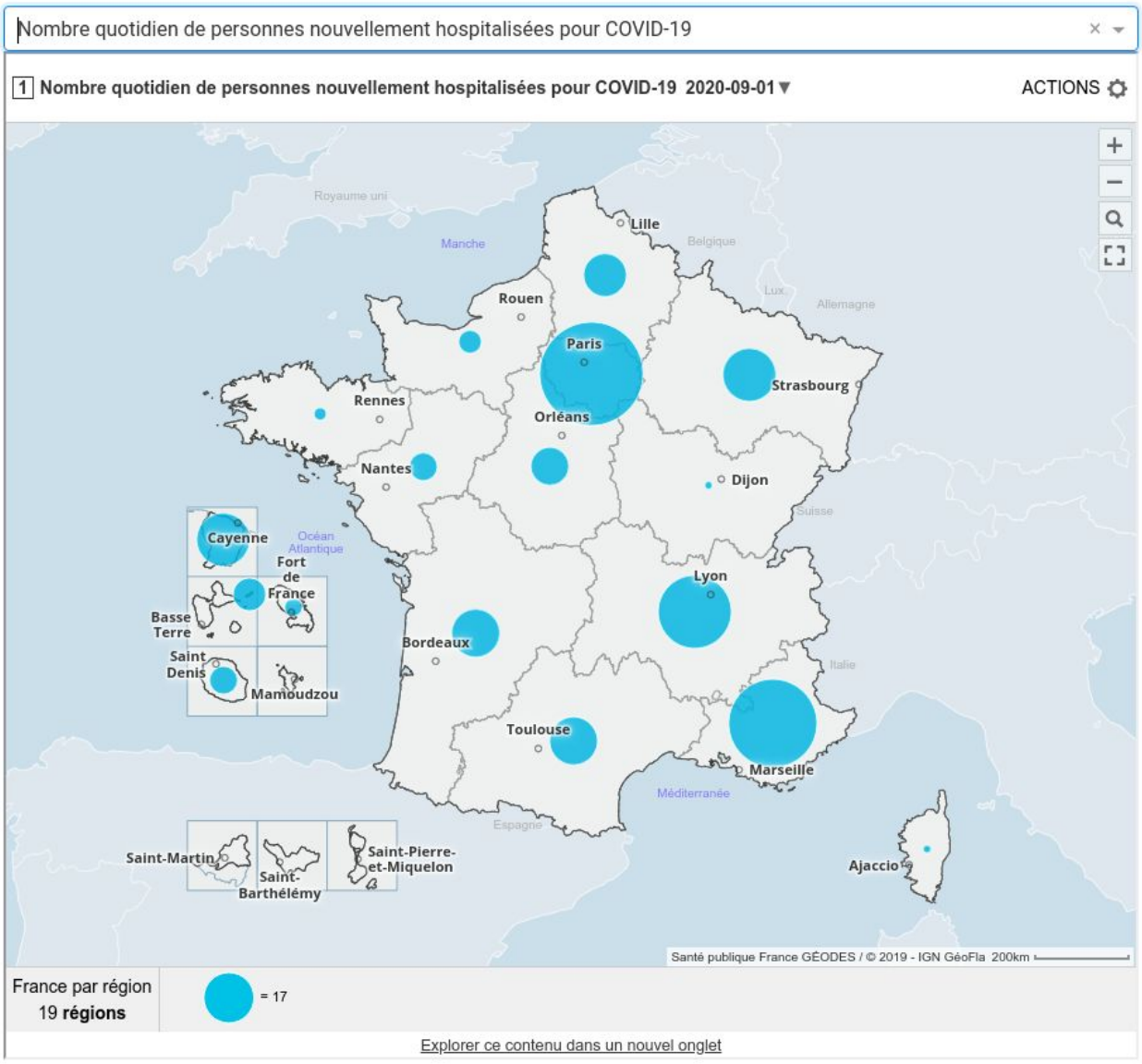
We can notice that the level of depression was generally high during the confinement period (e.g 14 - 16 Avril) compared to the post confinement period (e.g 20-22 July). The 3rd tab of my dashboard can allow you to analyse how the various indicators vary with time and know more about how the indicators were calculated.



[Code to preprocess dataset of indicators]

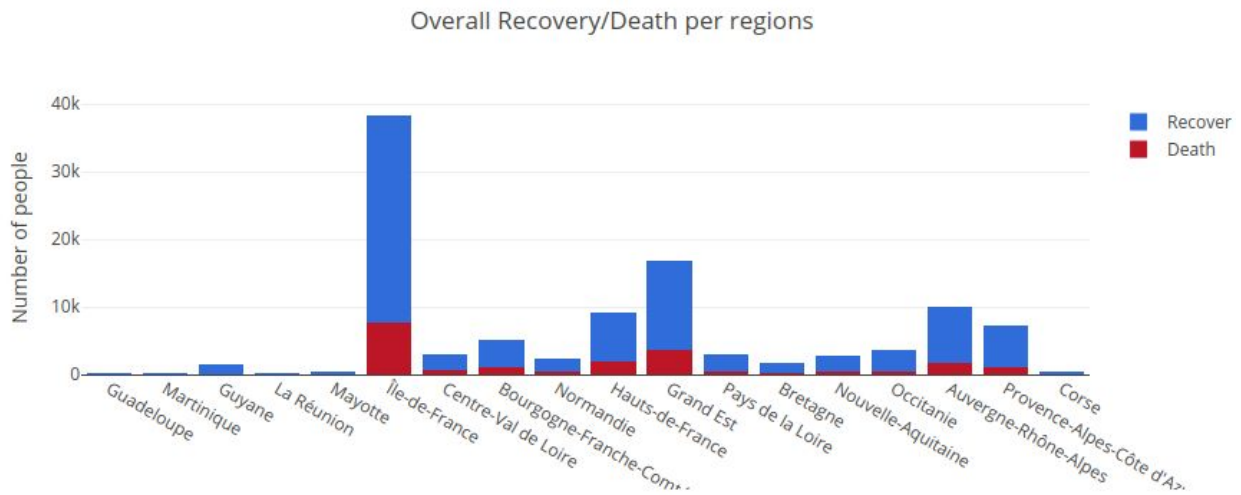
5) Proposition d'une visualisation graphique de type tableaux de bord, avec filtres et paramètres adaptés pour visualiser le nombre de cas, décès, guéris, éventuellement hospitalisés voire admis en réanimation par département/région.

From the 1st tab of my dashboard you can access a geographical representation to address this mission. You have a map of the regions with a bubble size equal to the variable under visualization. You have two filters, the first filter allows someone to select the question of interest and the second filter makes it possible to set the date of interest. For example, the next screenshot shows the number of hospitalized people per region on September 1st 2020. As in most of the cases, the region of Ile-de-France has the greatest number of hospitalized people.



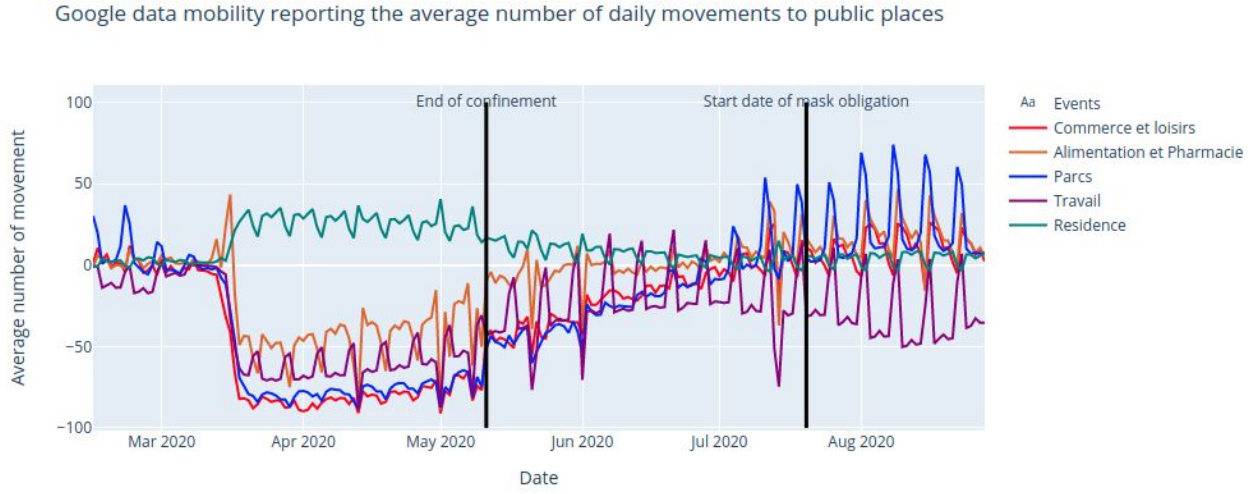
6) Superposer les courbes comme les cas, remis/guéris, décédés en comparant différentes régions (e.x. Hauts-de-France, Ile-de-France, Alsace).

Still on the 1st tab of my dashboard you will see a bar chart reporting the overall number of deaths and recovered per region.



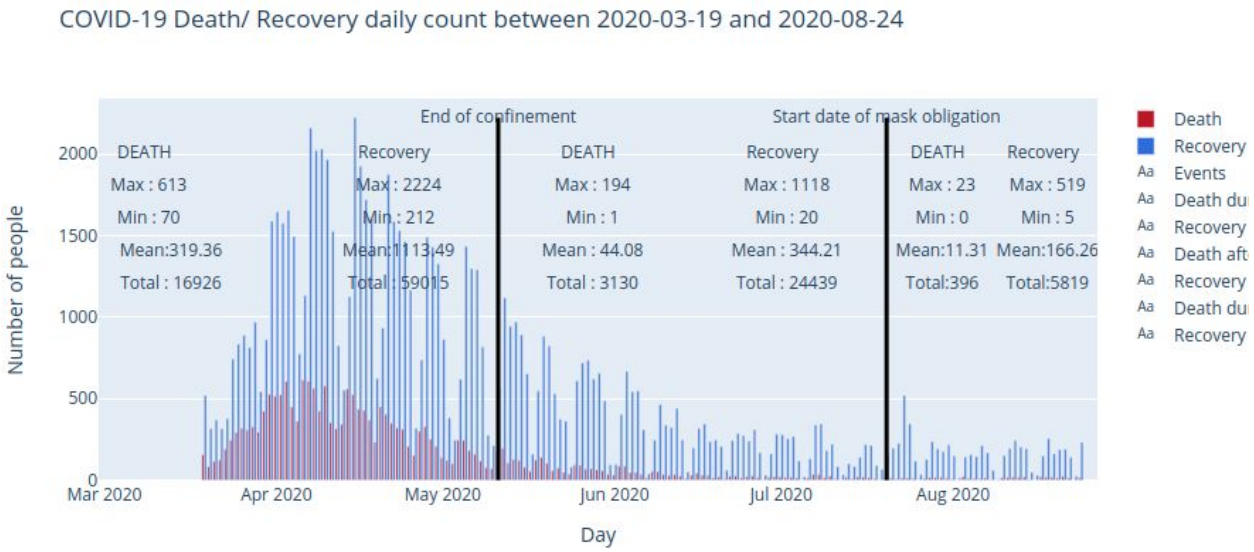
7) Construction d'une analyse montrant la nature des relations entre les données initialement utilisées avec d'autres sources de données (exemple: réseaux sociaux, mobilité via google et/ou autres propositions pertinentes)

To address this requirement, I focused my analysis on the data from Santé Publique France and Google Mobility. The data from Google Mobility correlates clearly with the events of pre-confinement, confinement and post confinement.



During the confinement period, we can notice that with the exception of high movements at home, movements to other places such as supermarkets, pharmacies and parks are greatly reduced as this the period that the government and health personals were actively looking for means to contain the virus. By the end of the confinement, we can notice a significant reduction in the number deaths and increase movement to places like parks, groceries and recreation areas. The movement to parks

and pharmacies rapidly increases following the obligation of masks. Notice that movements to work offices nearly remain low throughout the period.



The code for preprocessing the dataset from Google Mobility is available in my [GitHub repository] alongside any other code used to implement this project.