

Projet BIUM 2020

Comment gérer au mieux une épidémie ? Exemple du Covid-19

Vincent Jouve

Antoine Cadiou





Sommaire

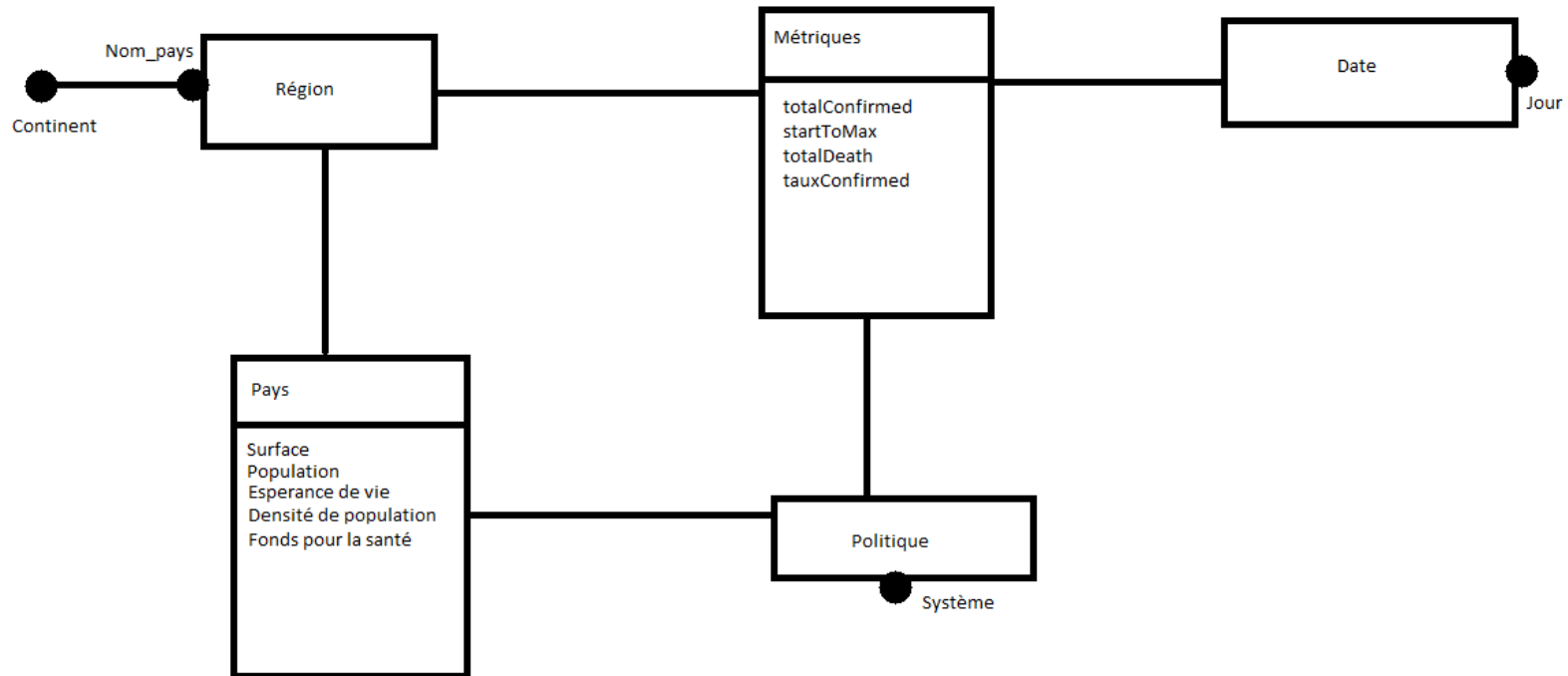
- Problématique
- Schéma en constellation
- Modélisation du Data Warehouse
- Analyses réalisées
- Conclusions

Problématique

Comment gérer au mieux une épidémie (avec le cas du coronavirus) ?

- Sous-question 1: Quels pays ont le mieux pris en charge l'épidémie ? (selon différentes analyses de métriques)
- Sous-question 2: Comment expliquer leur gestion de l'épidémie ? (système politique, démographie, budget pour la santé, ...)

Schéma en constellation



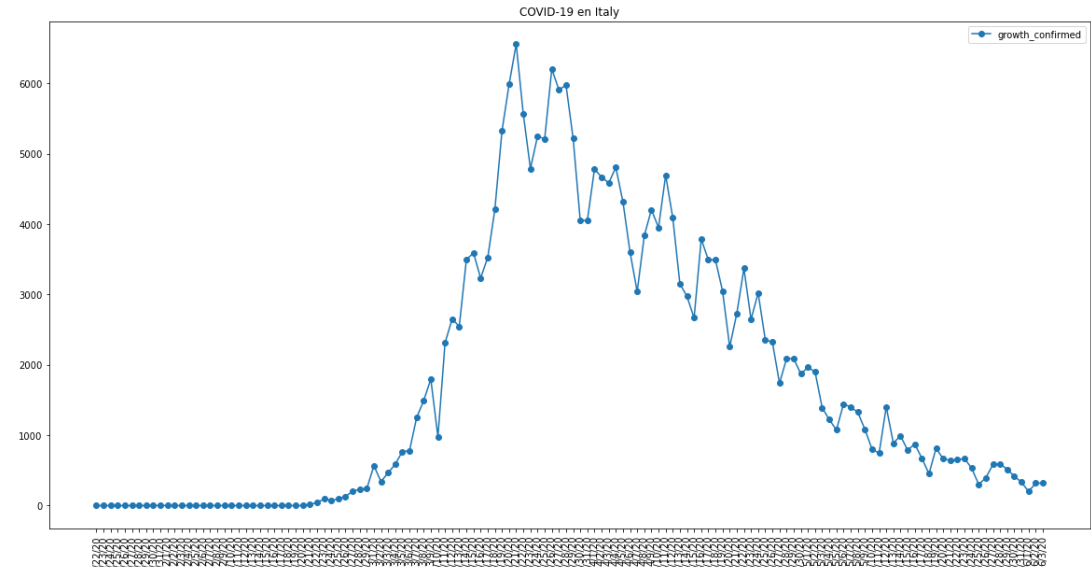
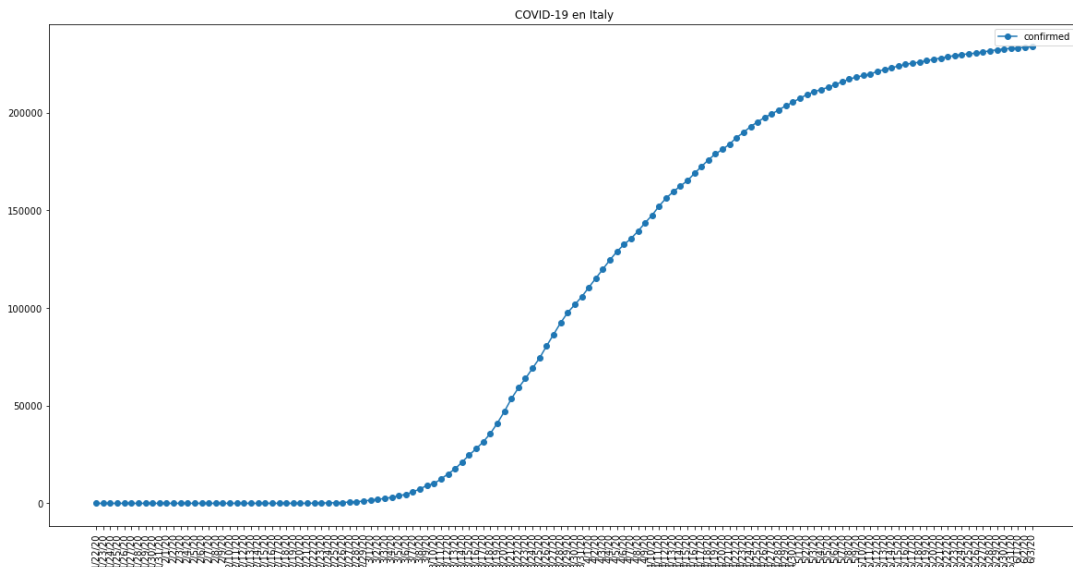
Modélisation du Data Warehouse

Sources des datasets utilisés :

1. world.sql (obligatoire pour le projet)
2. Séries temporelles d'évolutions des cas confirmés et des décès par pays <https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases>
(Séries du 22/01/20 au 30/05/20)
3. Dépenses des états dans le domaine de la santé(données de l'OMS)
<https://apps.who.int/nha/database/Select/Indicators/en>

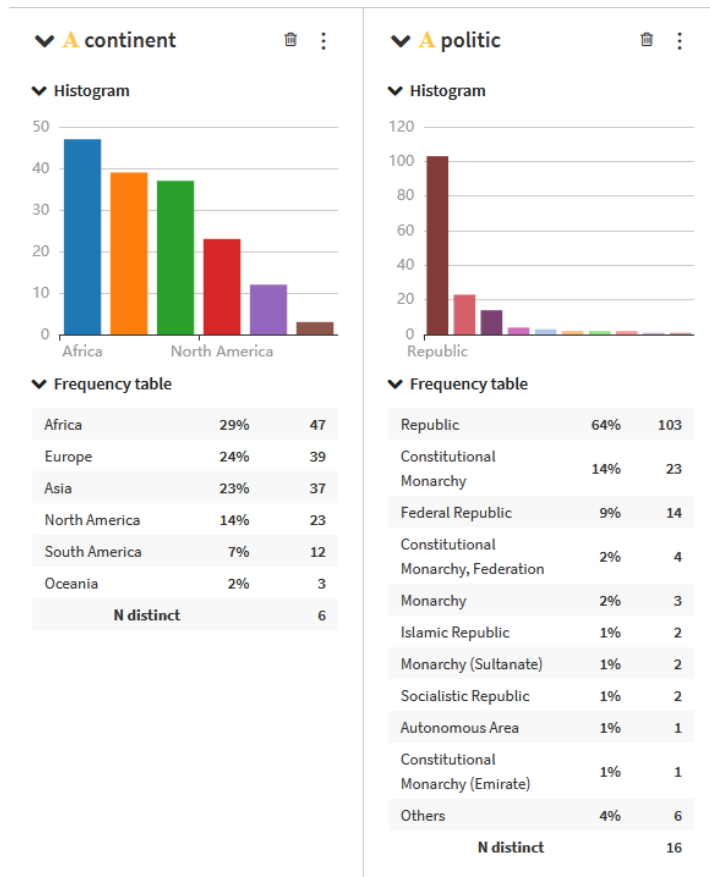
Modélisation du Data Warehouse

- Les séries temporelles dont nous disposons représentaient le nombre de cas confirmés totaux jour après jour, Nous avons décidé de transformer la série pour avoir la croissance du nombre de cas. Cela permet d'obtenir la dérivée de la série initiale (qui représente la vitesse).



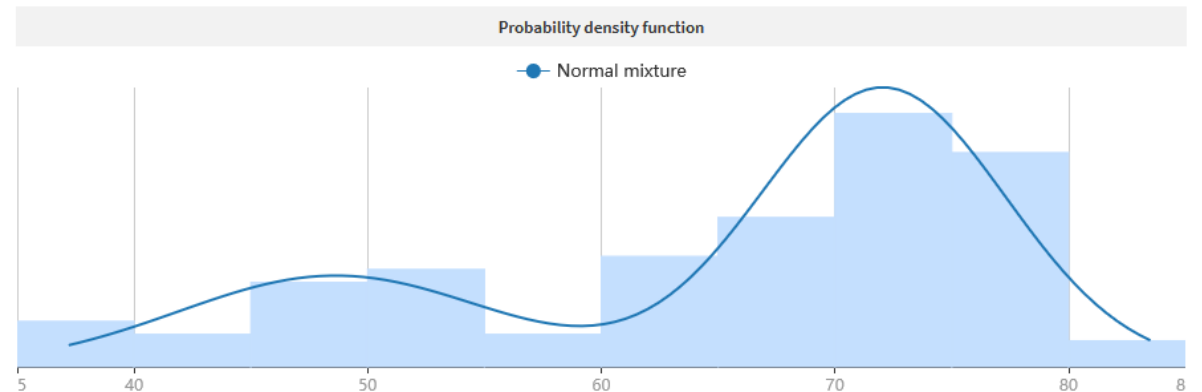
Modélisation du Data Warehouse

Une fois le Data Warehouse terminé, nous pouvons voir quelques données:

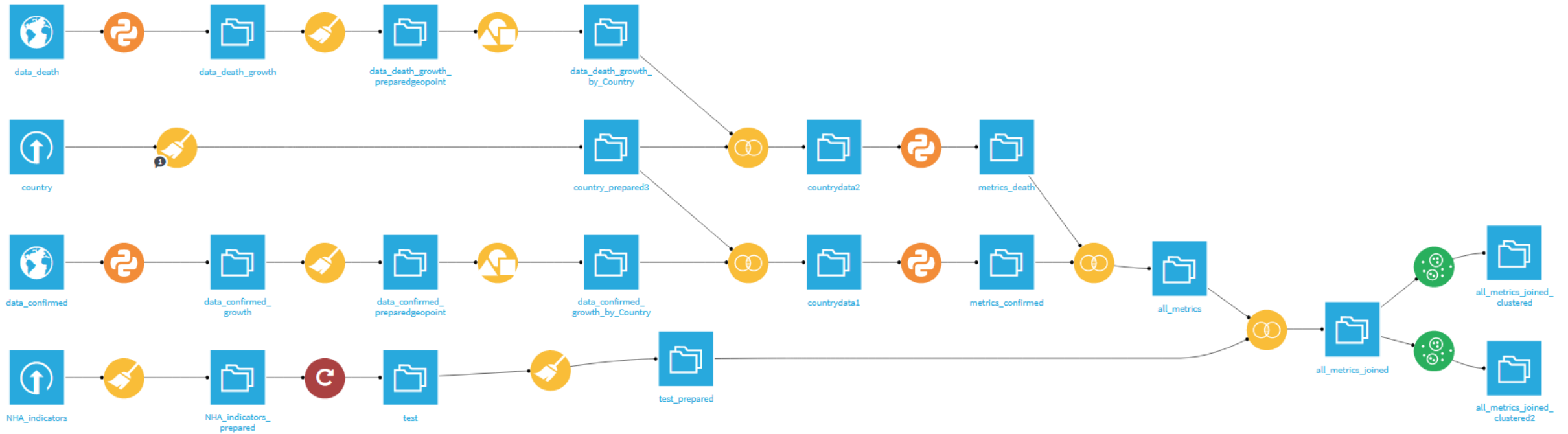


Lignes : 161
Colonnes : 16

▼ Fit distribution on lifeExpectancy No split ▼



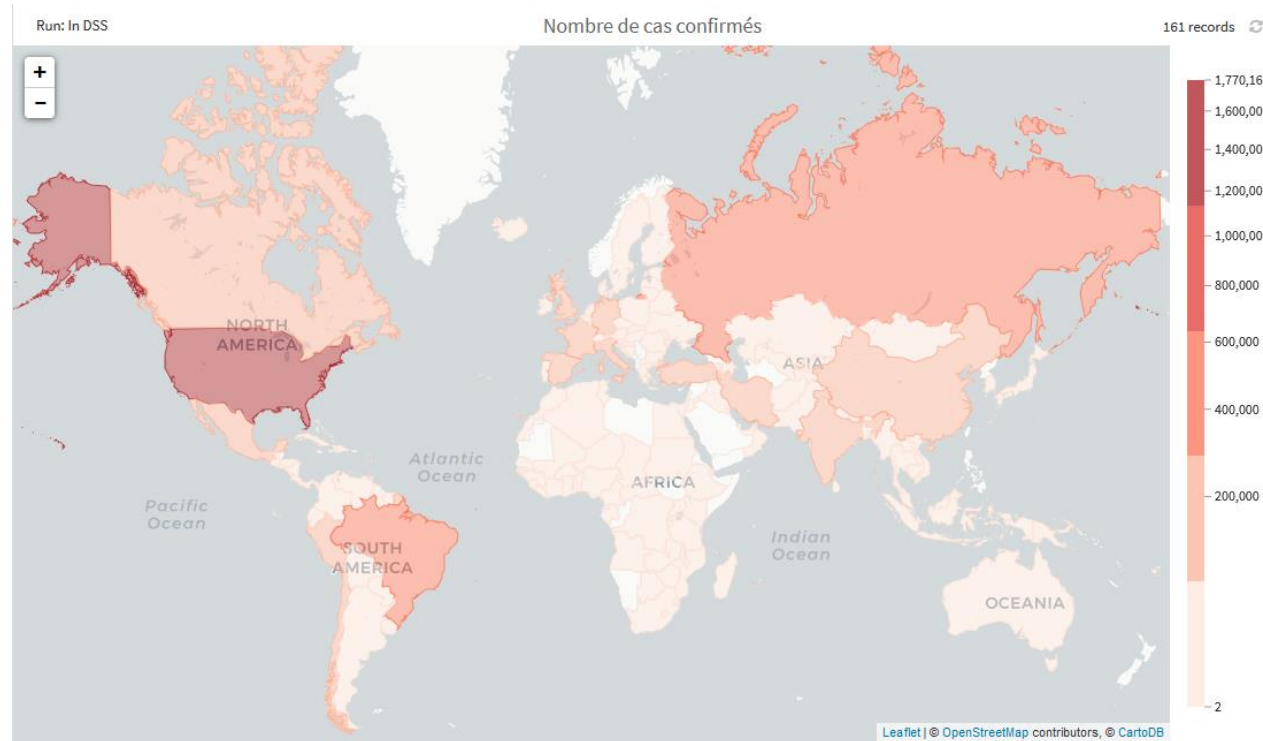
Modélisation du Data Warehouse



Analyses réalisées

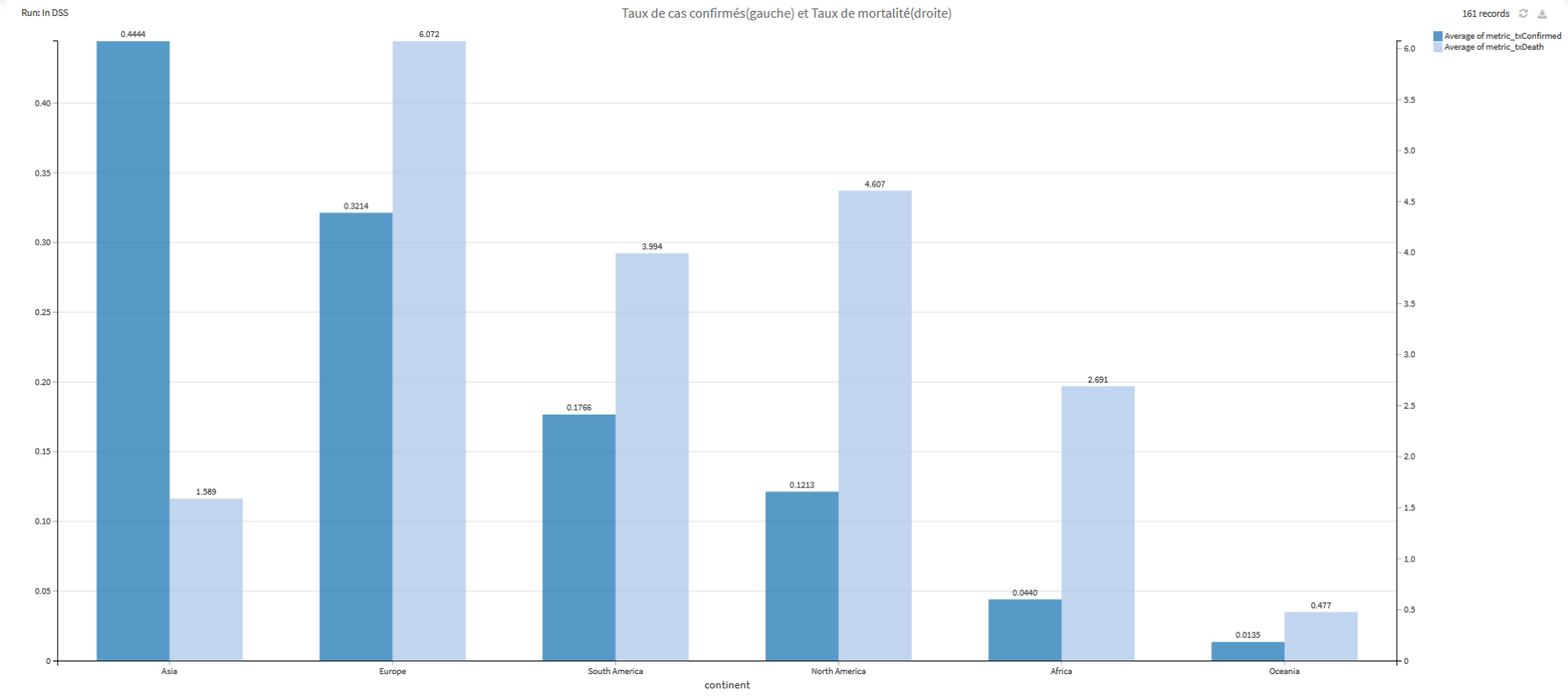
- Dans un premier temps, pour répondre à la sous problématique 1, nous avons défini plusieurs métriques qui pourraient permettre de dire qu'un pays a mieux géré l'épidémie par rapport à un autre.
 - startToMax : qui représente le nombre de jours entre le premier cas détecté et le jour du maximum de cas supplémentaires atteints en 1 jours (avec la plus forte croissance).
 - totalConfirmed : le nombre total de cas confirmés
 - totalDeath : le nombre total de décès
 - tauxConfirmed : le taux des cas confirmés divisé par la population
 - tauxDeath : le taux de mortalité parmi les cas confirmés
 - confirmedBySurface : le nombre de cas par km^2 (le nombre de cas total, pas en simultanés)

Analyses réalisées



On observe qu'il y a des pays qui n'ont eu que très peu de cas, on décide par la suite de ne s'intéresser qu'aux pays ayant eu plus de 10 000 cas car nous voulons étudier les pays qui ont 'subit' l'épidémie.

Analyses réalisées

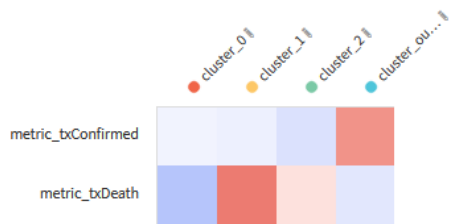


On observe que le taux de mortalité est le plus élevé en Europe.

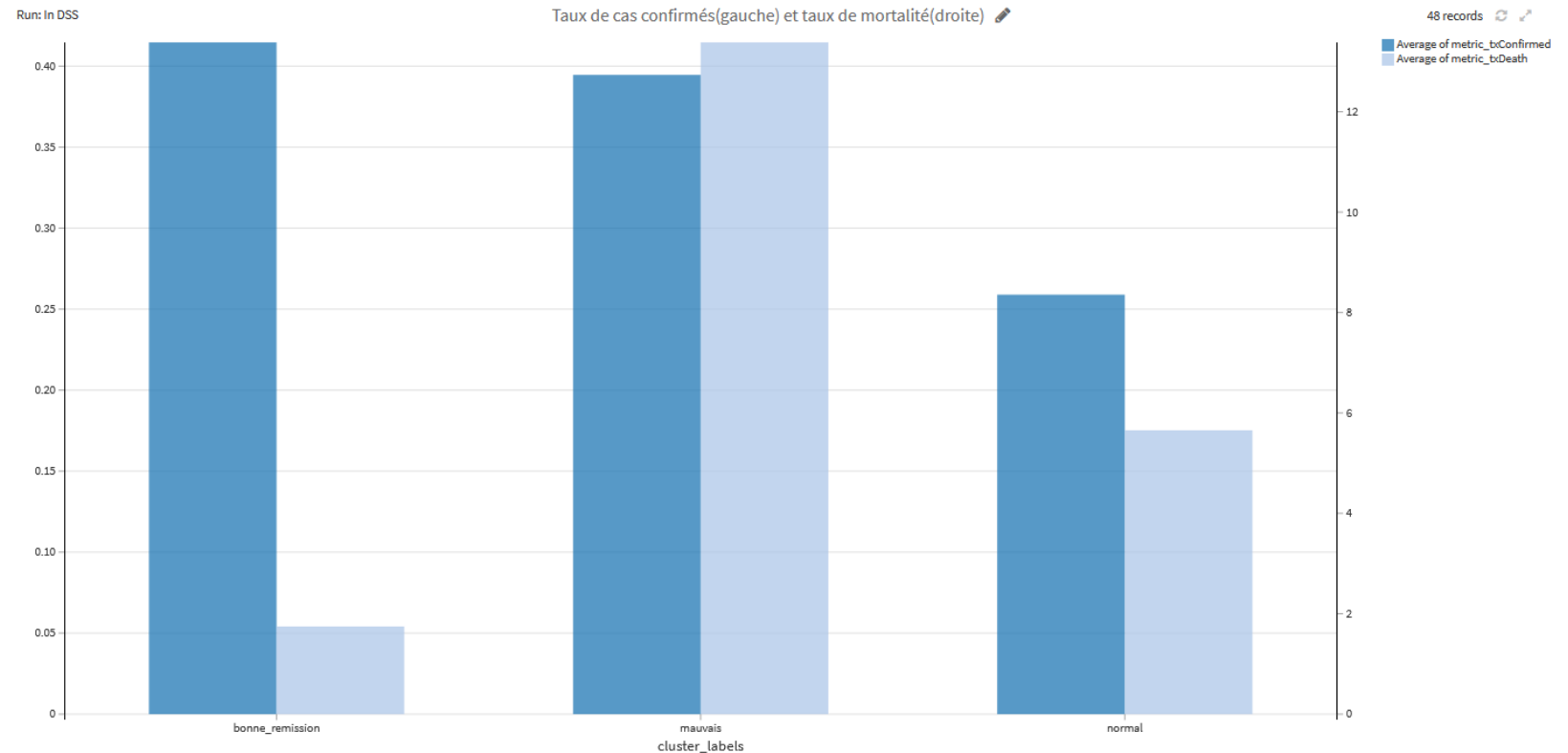
Cependant, on voit qu'en Asie on a le taux de cas confirmés le plus élevés mais un taux de mortalité faible, ce qui est étrange.

Analyses réalisées

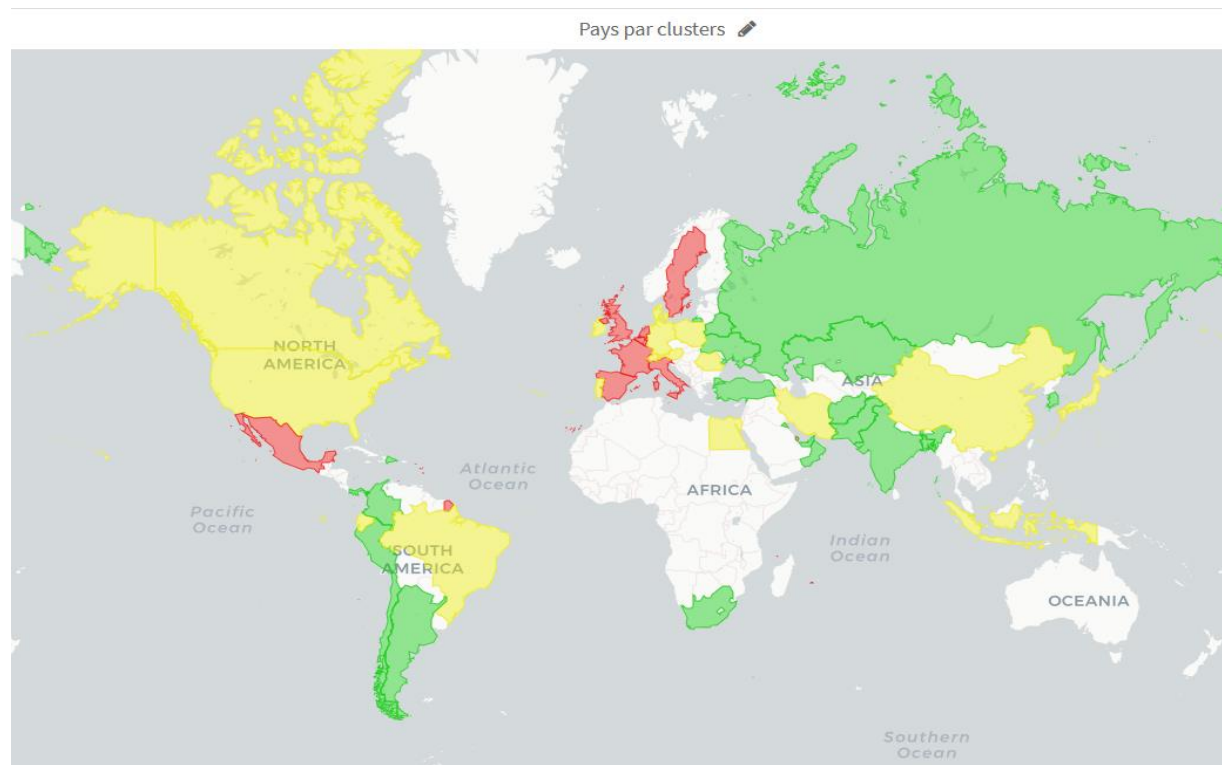
- On a fait du clustering (K-Moyennes) et nous avons gardé 3 clusters (et un outlier) que nous avons essayé de nommer. Le clustering a été effectué sur les deux métriques : taux de cas confirmés et taux de mortalité du virus.



- Le cluster_0 a un faible taux de mortalité mais un grand taux de cas confirmés, donc il y a une bonne rémission des patients dans les pays du cluster.
- Le cluster_1 possède un taux élevé à la fois pour les cas confirmés et pour la mortalité, donc ces pays ont mal géré l'épidémie
- Le cluster_2 suit une tendance normale.



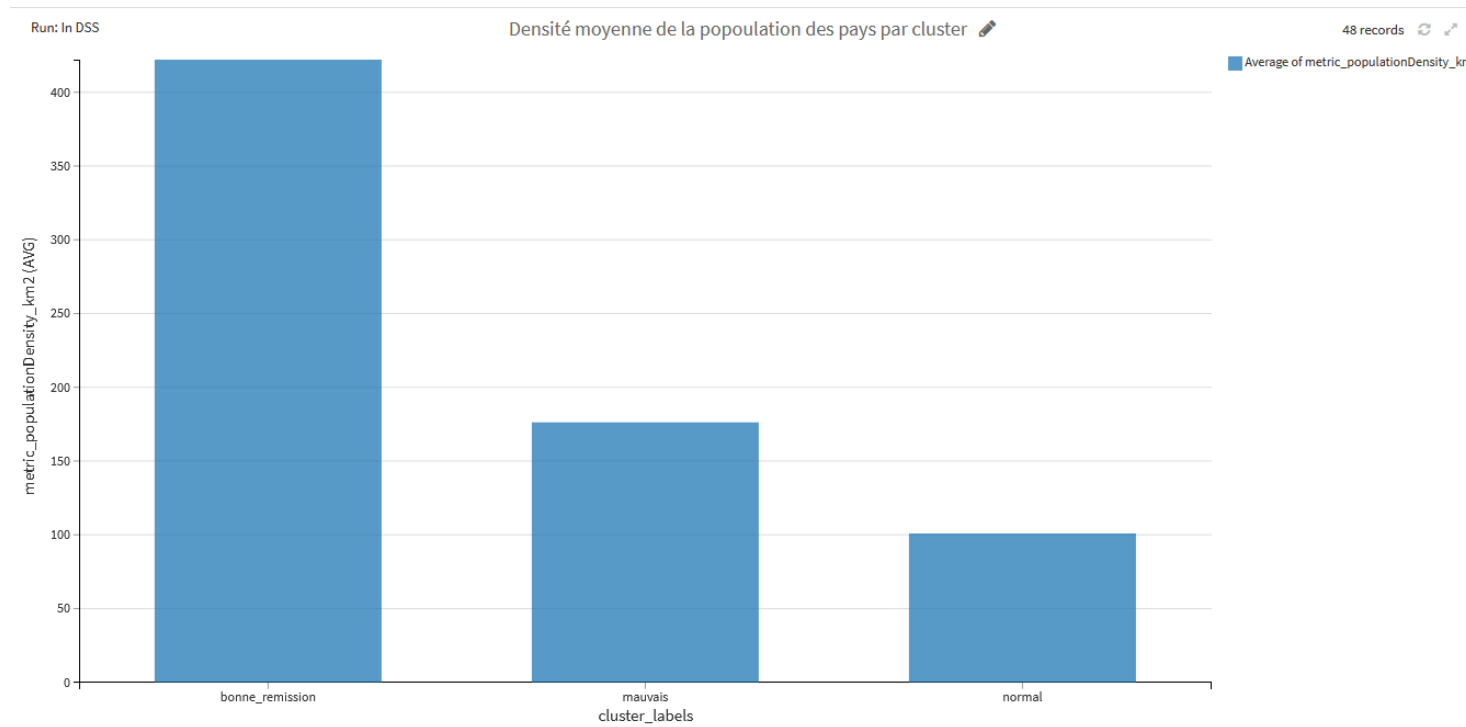
Analyses réalisées



Vert: cluster bonne_remission
Jaune: cluster normal
Rouge: cluster mauvais

On peut visualiser les pays en fonction de leur clusters d'appartenance. Cela permet de répondre à notre première question. Maintenant, pour répondre à la deuxième question, nous pouvons essayer de comprendre pourquoi ces pays appartiennent à leur cluster respectif.

Analyses réalisées



Nous remarquons que le taux de cas confirmés est corrélé avec la densité.
Les résultats sont cohérents, cependant pour répondre à la problématique on ne peut pas jouer sur cette variable.

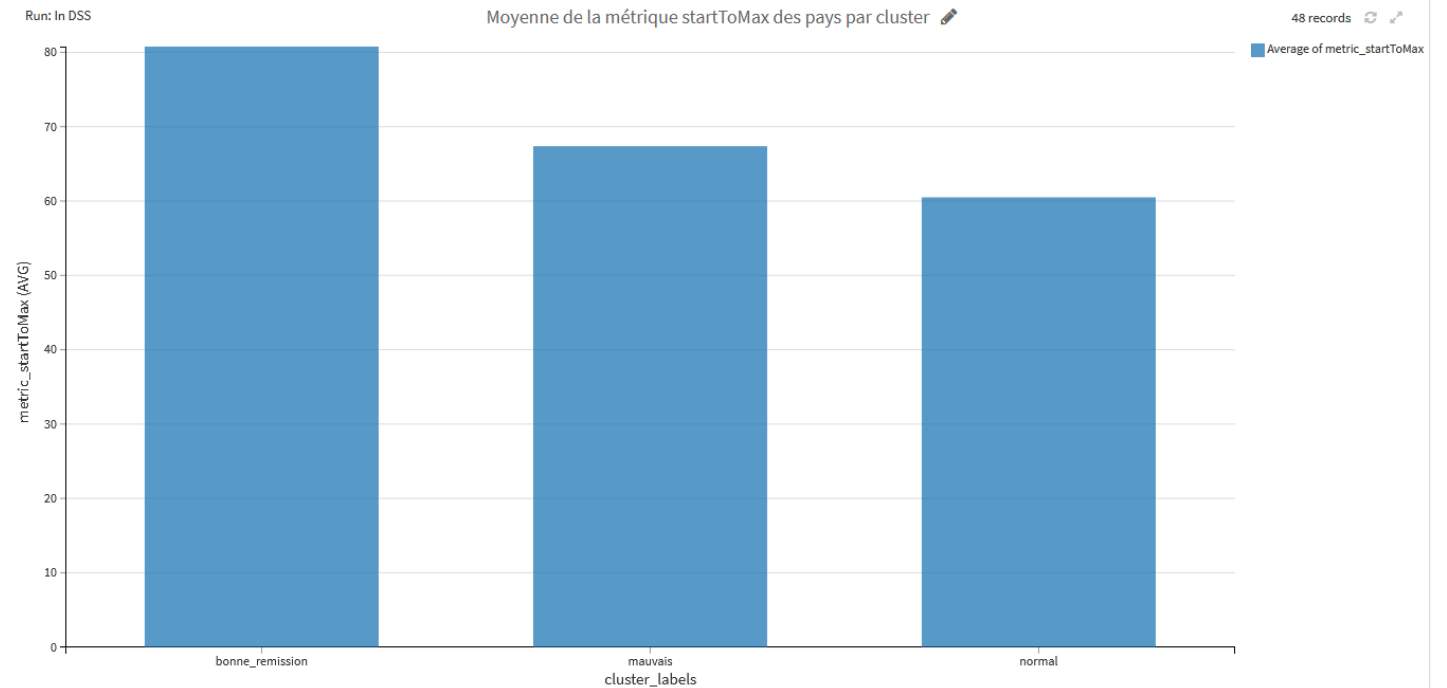
Analyses réalisées

Nous observons que la métrique startToMax témoigne de l'aplatissement de la courbe de croissance.

Pour le cluster ayant une bonne rémission, la courbe était bien aplatie.

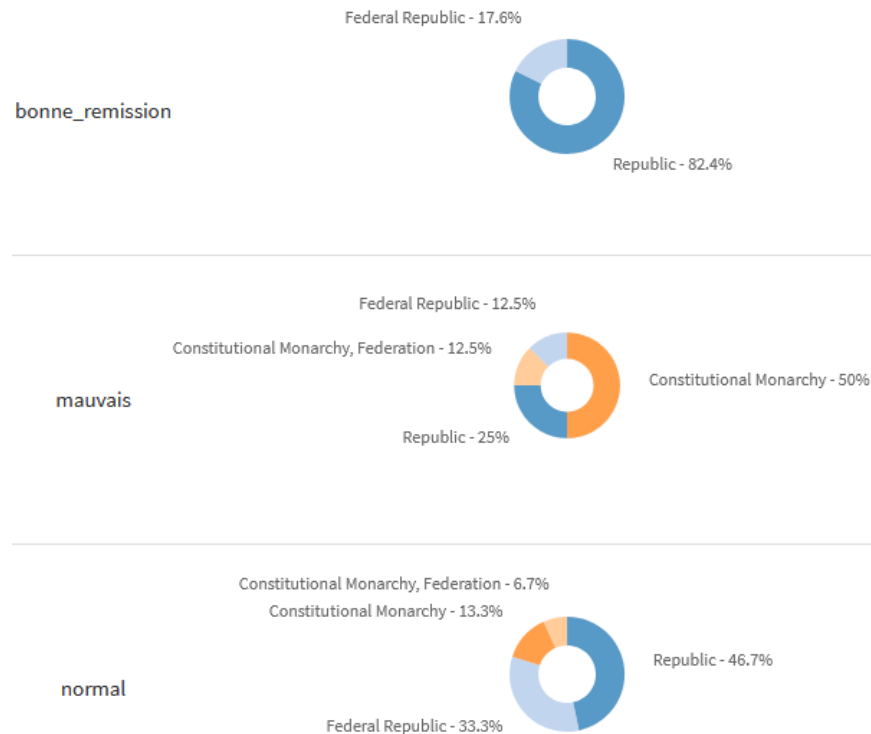
Une métrique encore plus intéressante aurait été d'avoir l'accélération (la dérivée de la croissance) car on aurait vraiment pu savoir si la courbe était aplatie ou non.

Est-ce que le système politique joue un rôle dans l'aplatissement de la courbe, par exemple en mettant en place un confinement ? (voir slide suivante)



Analyses réalisées

Répartition des systèmes politiques dans les clusters

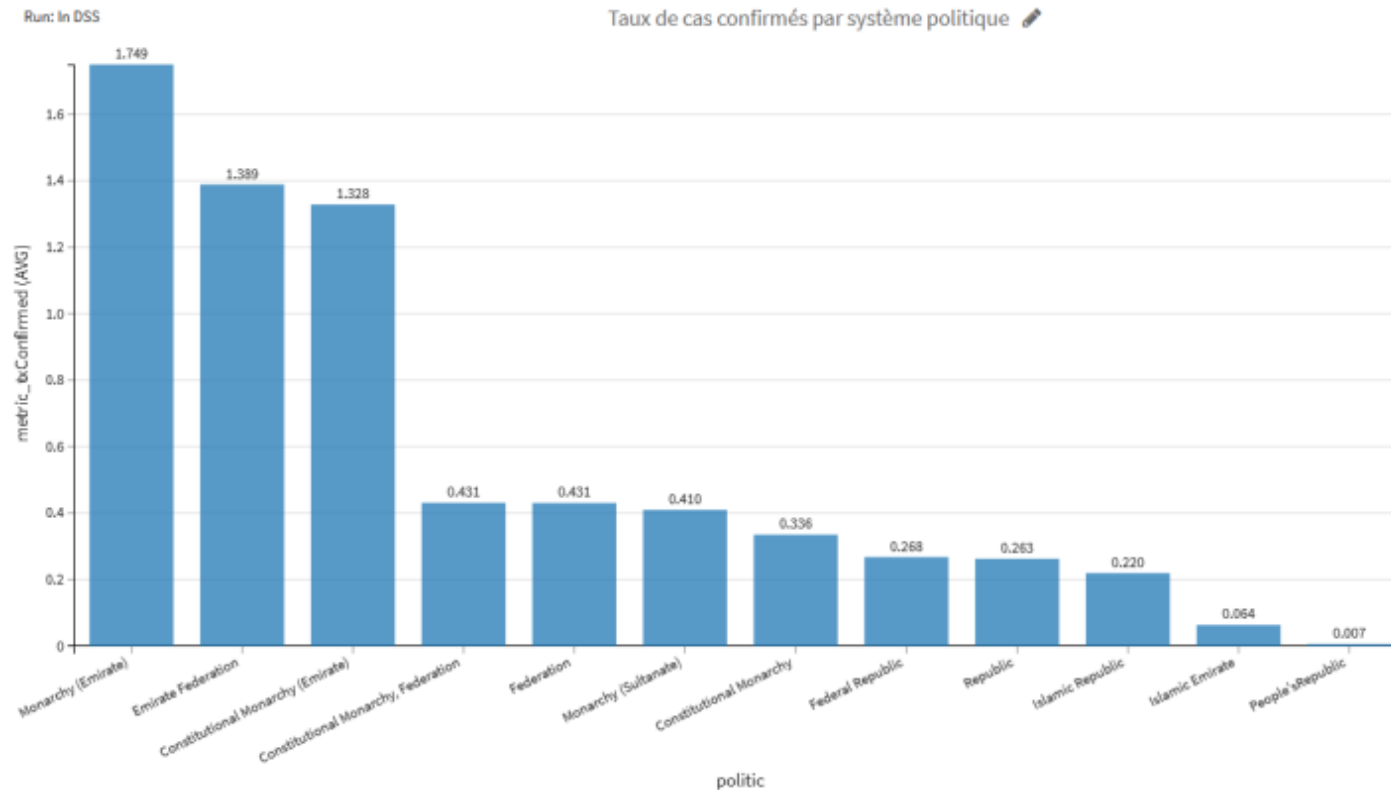


Existe-t-il un système politique plus adapté qu'un autre pour gérer une épidémie ?

Ici nous constatons que les Monarchies sont significativement plus présentes dans le mauvais cluster que dans les autres

(nous avons enlevé les systèmes politiques qui n'étaient représentés que par 1 pays, car nous n'avions pas assez de données pour conclure sur eux).

Analyses réalisées



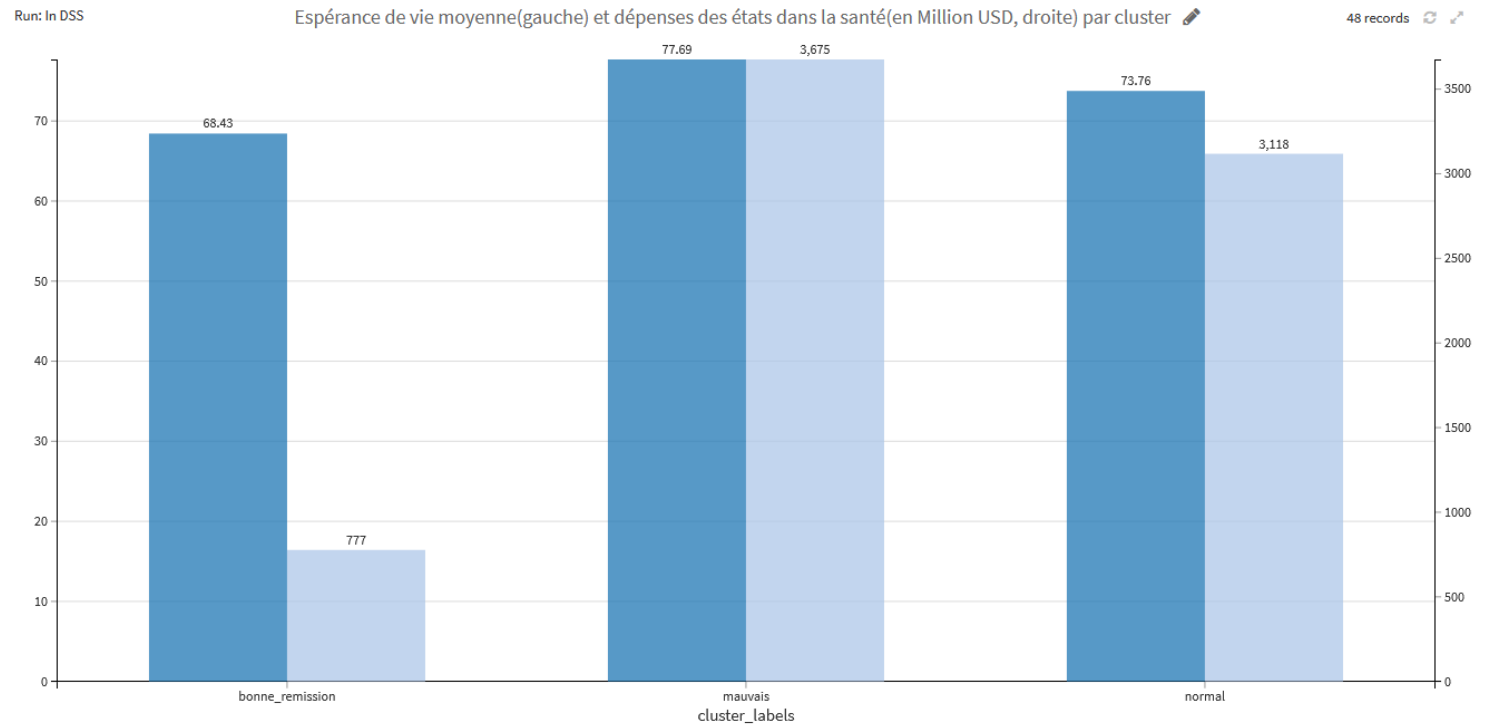
Ce graphe confirme ce que nous avons dit précédemment, les Monarchies ont un taux de cas confirmés plus haut que les autres systèmes politiques.

Analyses réalisées

Les dépenses dans la santé sont bien corrélées avec la durée de vie moyenne (mais pas linéairement car on vit forcément un minimum même avec un système de santé inexistant).

Cependant, ce qui est étrange est que là où on a une bonne rémission, on dépense moins pour le système de santé. Et à l'inverse, on en dépense le plus dans le cluster mauvais.

On imagine qu'il existe d'autres facteurs qui entrent en compte, comme par exemple la mondialisation. En effet, un pays fortement mondialisé aura à la fois plus d'argent à mettre dans le système de santé mais aussi, il sera plus susceptible d'être touché par l'épidémie.



Conclusions

Dans un premier temps, nous avons pu observer 3 groupes différents de pays ayant des caractéristiques propres sur l'impact d'une épidémie.

Ensuite nous avons cherché à comprendre ce qui explique cet 'impact'.

D'abord, nous avons remarqué que la densité de la population était corrélées avec le taux de cas confirmés. Cependant ce facteur n'est pas manipulable, il faut trouver une solution à la problématique qui ne dépende pas de la densité.

En revanche, une variable manipulable serait le système politique. Ce qui aurait été intéressant d'étudier est plus ce qu'un pays a mis en œuvre pour aplatir la courbe de croissance du virus plutôt que le système politique qu'il possède (même si les deux sont liés).

Enfin, nous avons étudiés des données sur les dépenses dans le domaine de la santé.

Les résultats paraissent surprenants mais on peut avoir un doute sur la véracité des données d'origine (par exemple certains pays qui ne communiqueraient pas de vraies données).

Aussi, on est amené à se demander si la mondialisation (ou d'autres données) expliqueraient davantage ce résultat.

Quoi qu'il en soit, nous aurions aimé explorer cet axe.