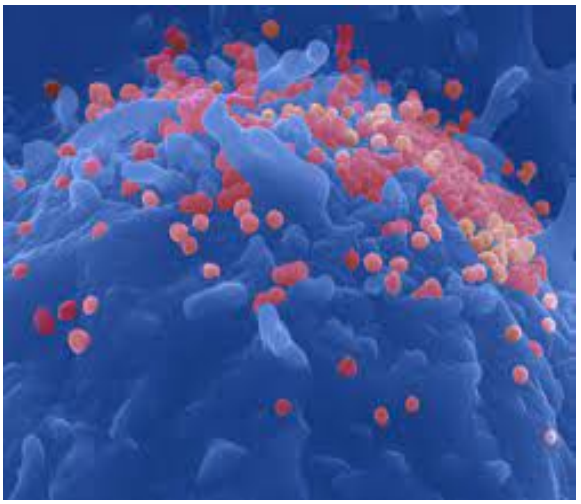
BUT Sciences des données

2023-2024

SAE croisée EMS/VCOD

Souches du virus



DIOP Coumba, LASSINA Fanni, SECK Fatoumata,
AUFRERE Thibault, MICHELS Cyprien

-

Loone Arthur

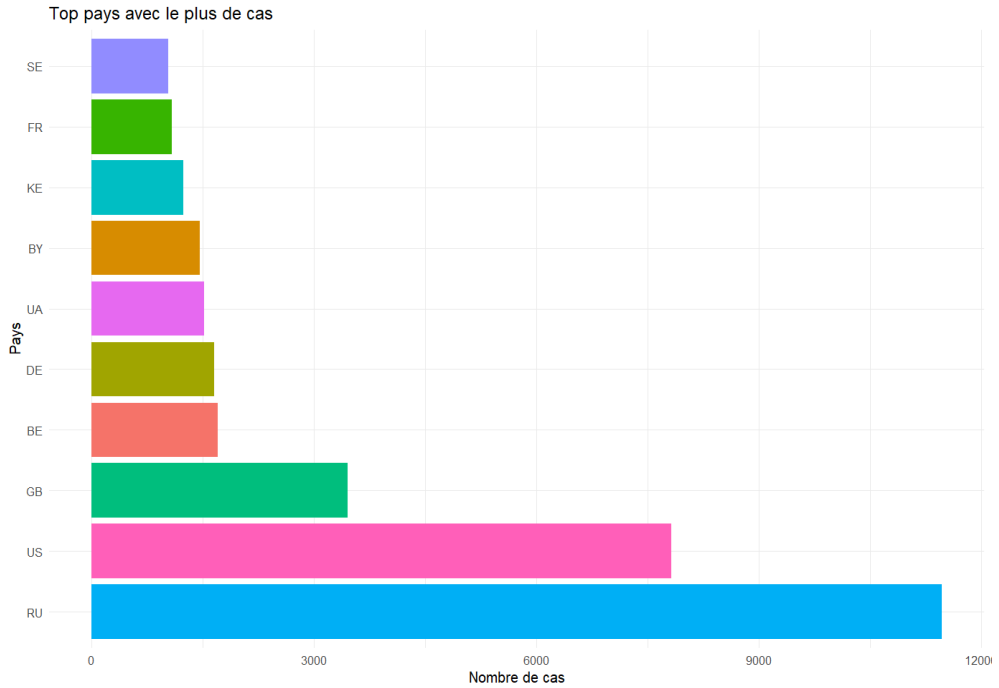
Introduction

Avant la réception de nouvelles données fournies par nos collègues de Vcod, notre équipe chargée de l'analyse a approfondi l'étude des anciennes données que nous avions déjà reçues auparavant. L'objectif était d'extraire le maximum d'informations pertinentes et d'automatiser nos scripts en R afin de préparer l'analyse des nouvelles données.

Après la réception des nouvelles données, nous avons peaufiné nos méthodes d'analyse pour mieux comprendre la distribution des souches du virus, étudier leur évolution dans le temps et évaluer leur corrélation éventuelle avec l'incidence du cancer du sein.

Distribution des Souches de VIH

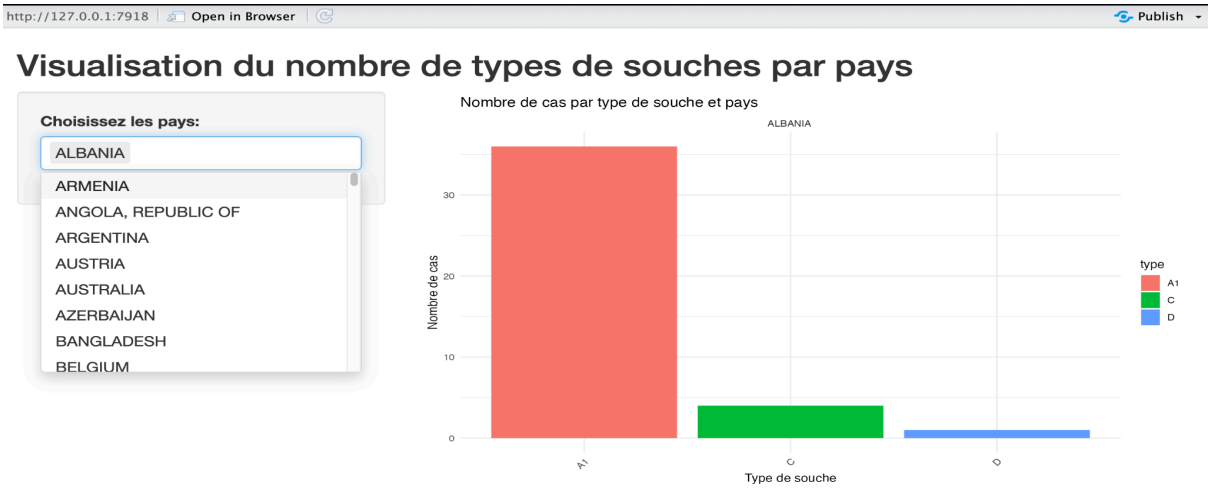
Notre analyse a débuté par une exploration des souches de VIH présentes dans notre base de données. Cette catégorisation virale est cruciale pour comprendre la répartition géographique du virus et pour étudier sa relation avec diverses maladies, dont le cancer du sein. Nous avons identifié plusieurs types de VIH, avec une prédominance notable de la souche A6, suivie des types B, C, A et A1. Nous avons d'abord identifié le nombre total de cas par pays, puis nous avons présenté les 10 pays comptant le plus grand nombre de cas.



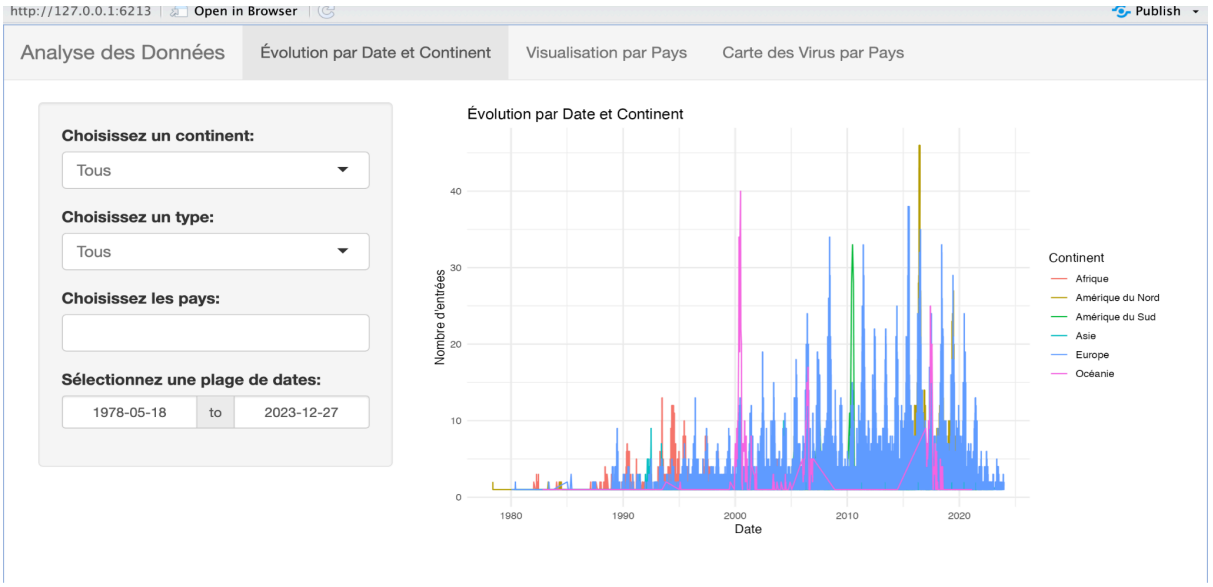
Ce graphique nous montre alors la répartition des personnes infectées par le SIDA dans notre échantillon. Comme nous pouvons le voir, la plus grande majorité de ces cas se trouvent dans des pays comme la Russie et les Etats-Unis qui sont les pays avec le plus de cas. Cette analyse peut être dû à la différence démographique entre les pays présents, comme la Grande-Bretagne ou encore la Belgique. Malgré ce déséquilibre, le fait que ce soit des pays aussi développés soient des points culminants, peut nous laisser envisager que de nombreux contaminants et/ou contaminés proviennent des ces environs. Ces pays serait

donc de bons points pour rechercher des similarités entres les différentes séquences de nos échantillons

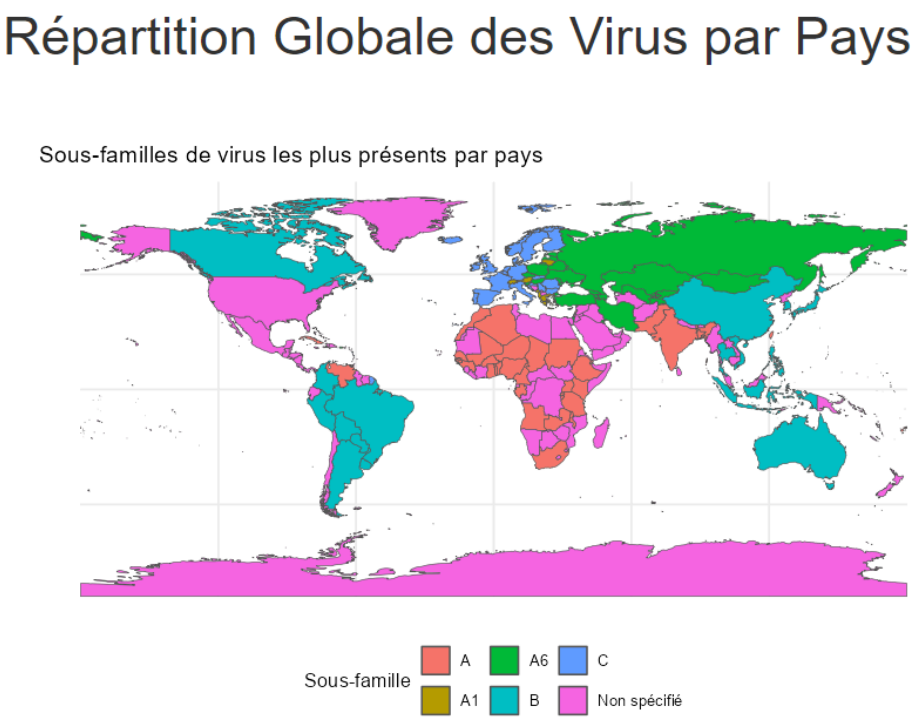
Les méthodes utilisées pour analyser ces données ont été variées. Nous avons utilisé R Shiny pour visualiser la distribution des cas de VIH par pays et fournir une interface interactive. Cette application permet aux utilisateurs de sélectionner des pays spécifiques et d'observer le nombre de cas par type de souche, facilitant ainsi la compréhension des schémas épidémiologiques:



Nos résultats préliminaires indiquent que certains pays présentent une concentration élevée de certains types de VIH, ce qui pourrait être associé à des facteurs socioculturels, économiques ou aux politiques de santé. Concernant la corrélation avec l'intensité du cancer du sein, bien que nos analyses soient encore en cours, nous commençons à discerner des modèles qui mériteront une investigation plus approfondie.



L'évolution temporelle du nombre de personnes par souche a été étudiée pour mettre en évidence des tendances au fil du temps.



La carte présentée illustre la répartition globale des différentes souches de VIH par pays. Chaque couleur représente une sous-famille du virus, indiquant la prédominance d'une souche spécifique dans chaque région. Les pays colorés en magenta montrent une prédominance de la souche A, tandis que ceux en vert foncé indiquent la prédominance de la souche A6. Le vert clair représente la souche C, le jaune la souche A1, et le bleu la souche B. Les pays en blanc ne disposent pas de données spécifiques ou ne présentent pas de prédominance claire d'une souche sur les autres.

Cette visualisation géographique offre une vue d'ensemble de la dispersion des souches principales du VIH, révélant des modèles régionaux qui pourraient être liés à des facteurs épidémiologiques, sociaux ou à la politique de santé publique de chaque pays. Par exemple, on peut observer que les souches sont non seulement spécifiques à certaines zones mais pourraient aussi refléter des voies de transmission historiques ou des politiques de prévention et de traitement locales.

La carte fournit également des indications précieuses sur des zones nécessitant plus d'attention et de ressources dans la lutte contre le VIH, en identifiant les régions où certaines souches prédominent. Cela peut permettre aux organisations de santé et aux chercheurs de mieux cibler leurs efforts et d'adapter leurs interventions à la souche prédominante dans chaque région.

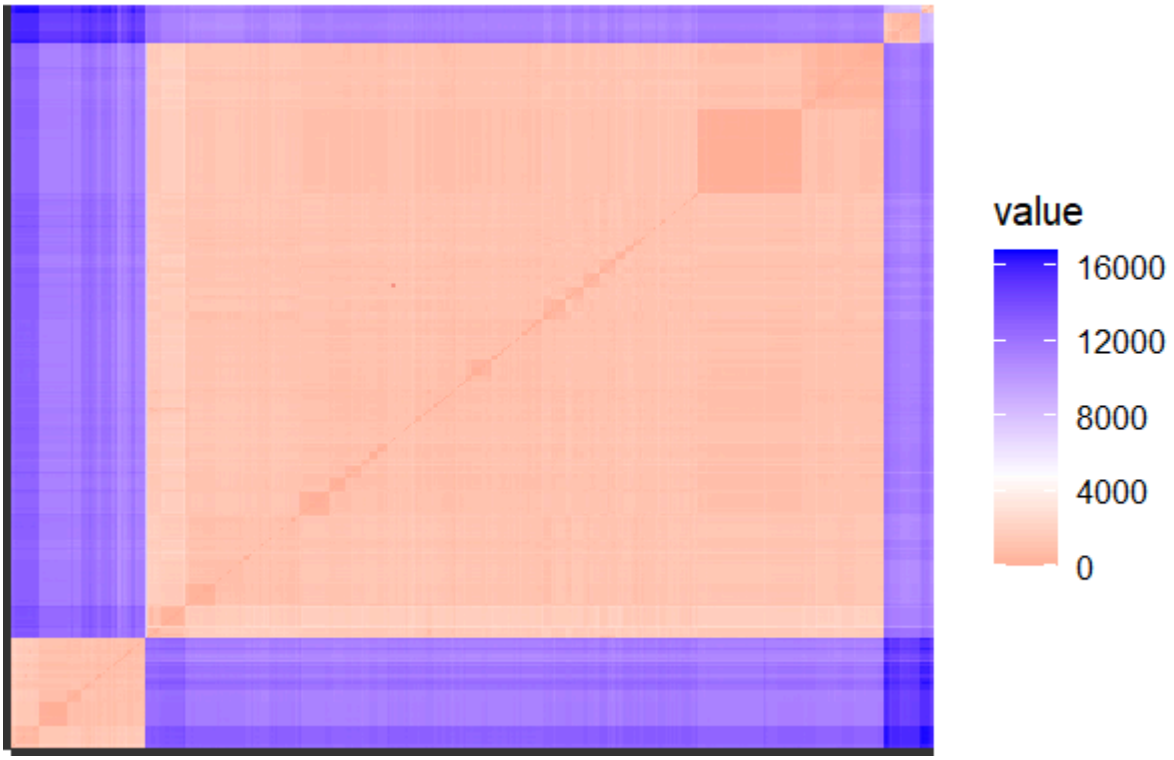
Dans ce tableau de bord que l’on avait, on a rajouté un onglet dédié à la représentation cartographique des différents types.

Homogénéité des Séquences ADN au sein des Souches

Dans le cadre de notre étude sur la variabilité génétique du VIH, nous avons mis au point une méthode pour calculer la distance entre les séquences ADN des différentes souches du virus. Notre choix s'est porté sur la distance de Hamming, une mesure qui quantifie le nombre de différences caractère par caractère entre deux chaînes de séquences ADN. Cette méthode est particulièrement pertinente dans le contexte de comparaisons génomiques où les séquences sont de même longueur, car elle permet de détecter les variations ponctuelles qui peuvent avoir un impact significatif sur la fonctionnalité et la pathogénicité du virus.

Pour les besoins de notre analyse, nous avons choisi de nous concentrer sur le type de souche B, conformément à la demande d'Arthur Loone. Cette souche, étant l'une des plus répandues dans certaines régions et associée à diverses dynamiques épidémiologiques, présentait un intérêt particulier. En appliquant la distance de Hamming aux séquences de la souche B, nous avons pu évaluer l'homogénéité génétique au sein de ce groupe spécifique.

Vu le nombre conséquent de données, nous avons choisi un échantillon de 1500 individus sur lequel nous avons commencé une analyse test.



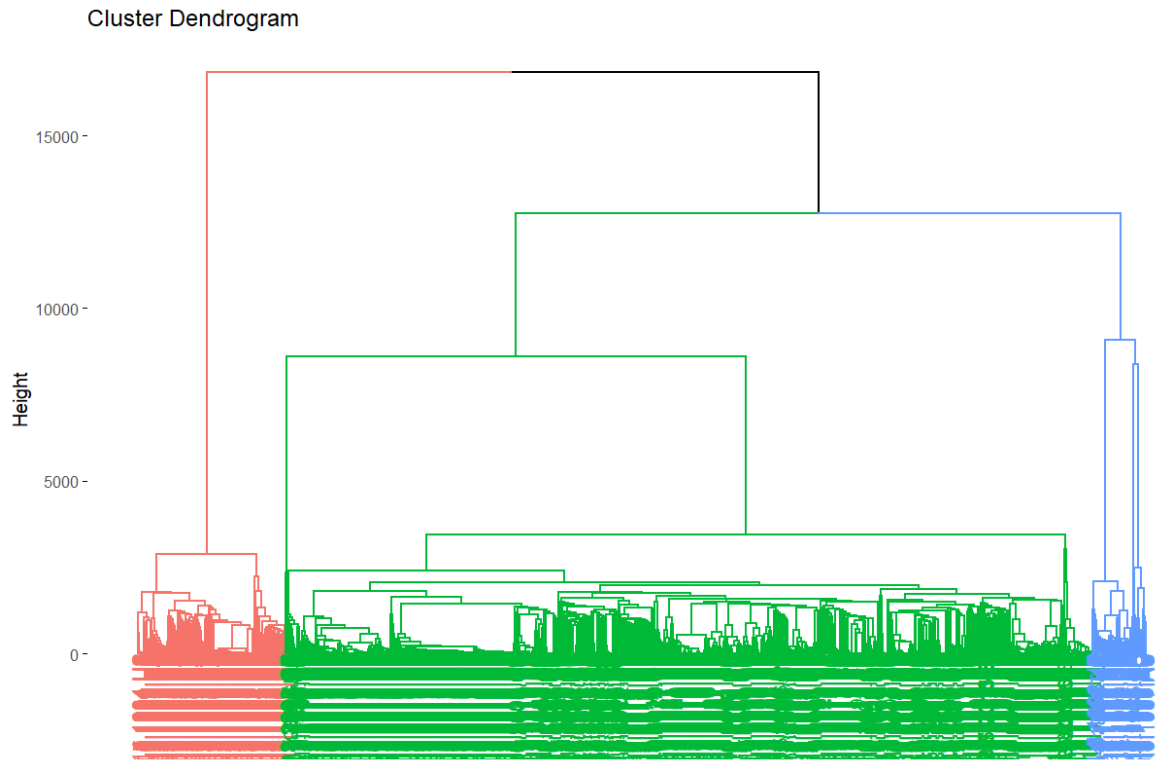
Groupe de distance

Nous disposons d'une visualisation sur la répartition des différents groupes. On constate que certains groupes ont une distance de zéro ou qui s'en approche. On peut en conclure que ces individus se sont contaminés entre eux. Cependant, cela est insuffisant pour tirer une telle conclusion. Nous nous sommes donc basés sur le calcul du temps entre les deux individus qui se sont infectés pour déterminer si ils sont réellement proches. Ensuite, nous avons un groupe d'individus en fonction des couleurs qui sont un peu proches, ce qui peut s'expliquer par l'évolution de la séquence. Donc, la séquence a peut-être muté, mais nous ne pouvons pas affirmer avec certitude qu'ils se sont contaminés. Nous avons également un groupe d'individus qui n'ont presque rien en commun, mais nous ne pouvons pas conclure qu'ils sont vraiment différents car ce groupe peut être constitué d'individus qui ne sont pas du même pays.

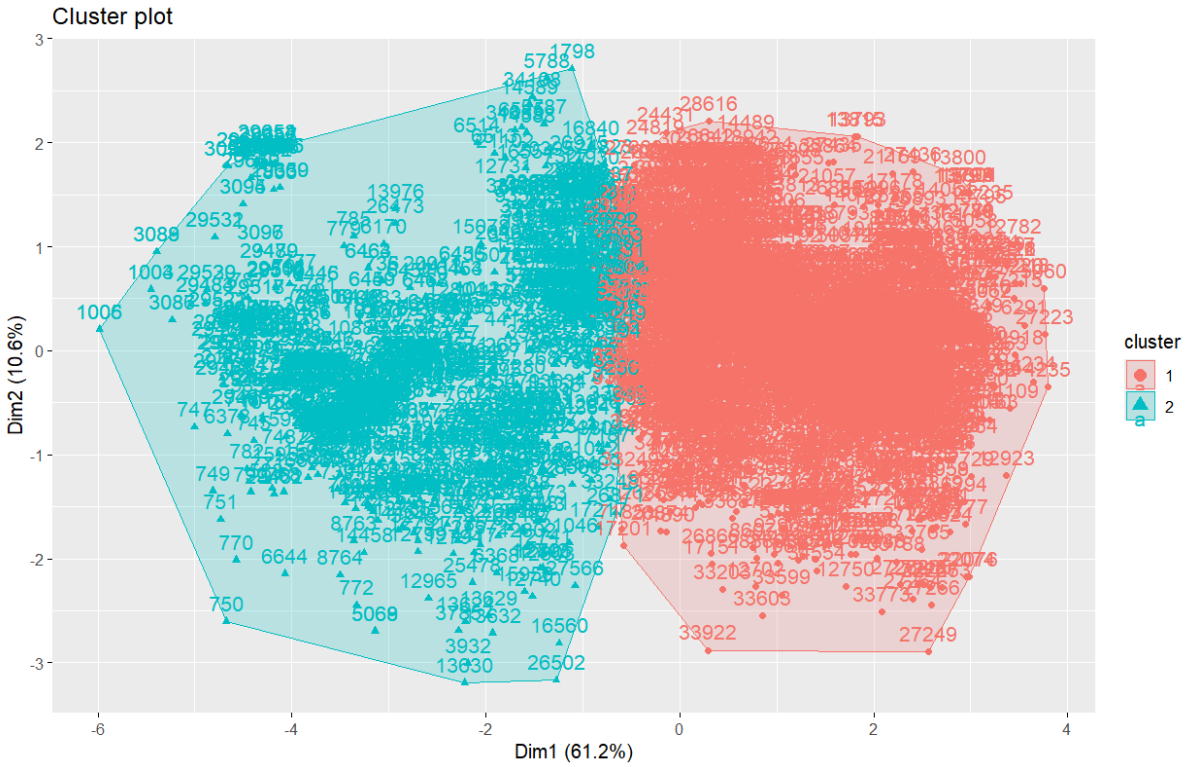
Cette approche nous permet d'affiner notre analyse sur les possibles contaminations au sein du type B, en prenant en compte à la fois la proximité spatiale, le temps et les caractéristiques génétiques des individus étudiés.

Détermination des sous groupes pour les contaminés du type B

Pour déterminer les sous groupes possible que les individu du type B pourraient être composés. Nous avons appliqué les méthodes de classification sur les individus de ce type. Plus précisément nous avons utilisé la méthode de CAH.



Après avoir appliqué cette méthode sur ces individus. Nous avons trois types de sous groupe.Cette méthode est appliquée sur les distances. On peut supposer que la méthode à classer en fonction de la distance entre les individus.



Grâce à la méthode des k-means nous avons été en mesure d'identifier jusqu'à 2 méthodes comparé aux 3 de la CAH. Nous pouvons cependant voir que cette répartition n'est pas optimale lorsque l'on observe que les interactions entre groupes sont trop proches. Cependant nous avons au moins la certitudes qu'au moins deux souches différentes de B existent, ils nous restent maintenant à déterminer la meilleure méthode afin de connaître le nombre de sous type optimal.

Conclusion:

Nous avons consacré cette journée à développer notre tableau de bord, mais avons été confrontés à de nombreux problèmes techniques. Ces défis, allant de simples erreurs de codage à des complications plus complexes dans le traitement des données, nous ont ralentis mais n'ont pas entamé notre détermination. Nous continuons à travailler activement à l'amélioration de notre tableau de bord, avec l'objectif de le rendre pleinement opérationnel pour notre soutenance.