



#LPQC

La Poule Qui Chante à l'International





#LPQC

Sommaire

01 Analyse en composantes principales

Composantes principales

Liasons entre les variables

02 Clustering

Classification Ascendante Hiérarchique

Clustering par K-Means

03 Analyse des clusters

Clustermap

Synthétisation des clusters



Partie 1 :

Analyse en composantes principales





Composantes principales

L'Analyse en Composantes Principales consiste à transformer les variables liées entre elles statistiquement en de nouvelles variables appelées "Composantes Principales".

Ces nouvelles variables permettent de résumer les informations en créant de nouveaux axes factoriels ou composantes plus facilement interprétables.

**Dans notre cas l'ACP sera normée, toutes les données seront centrées/réduites.*

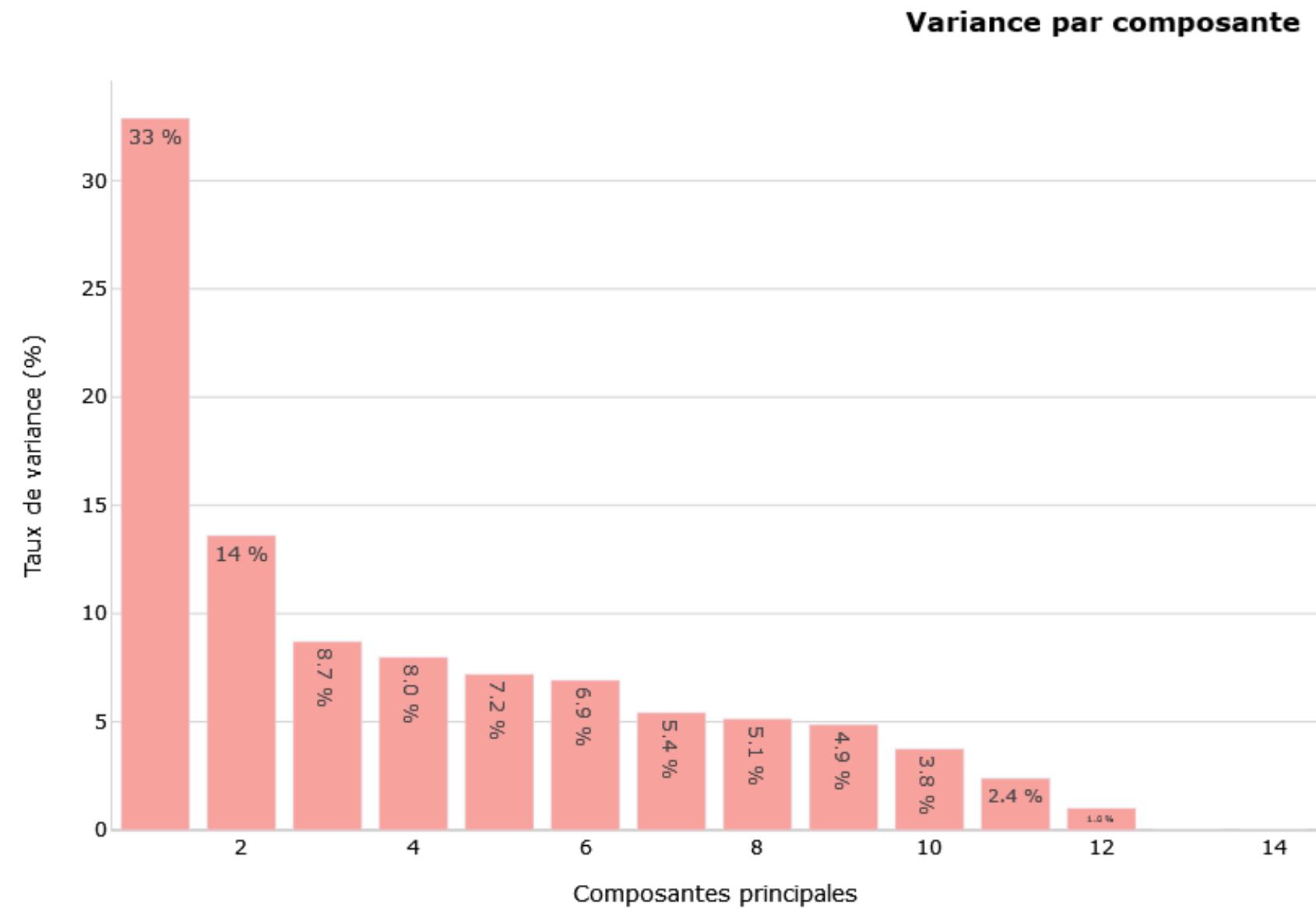


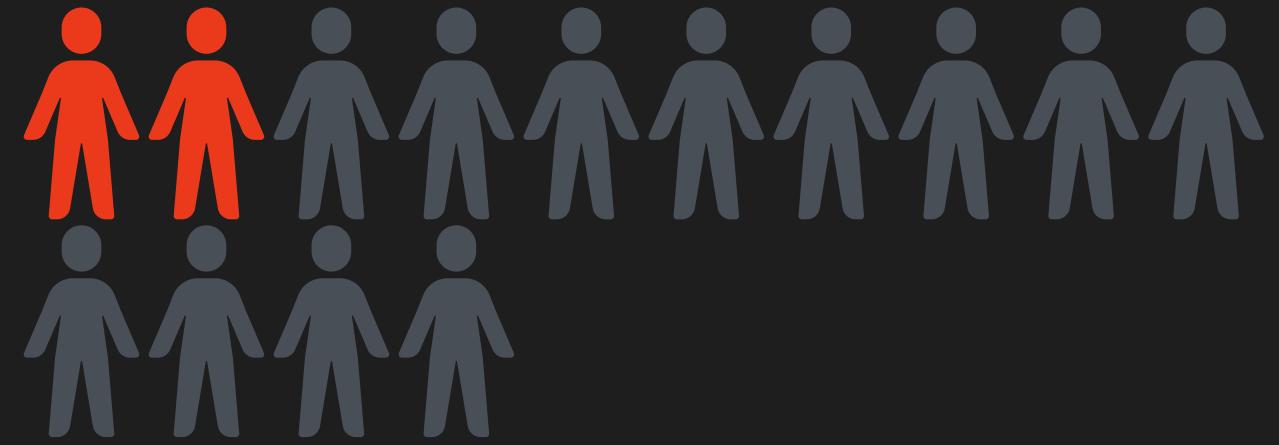
Variance par composante

Les composantes principales sont les vecteurs propres de la matrice de covariance des données.

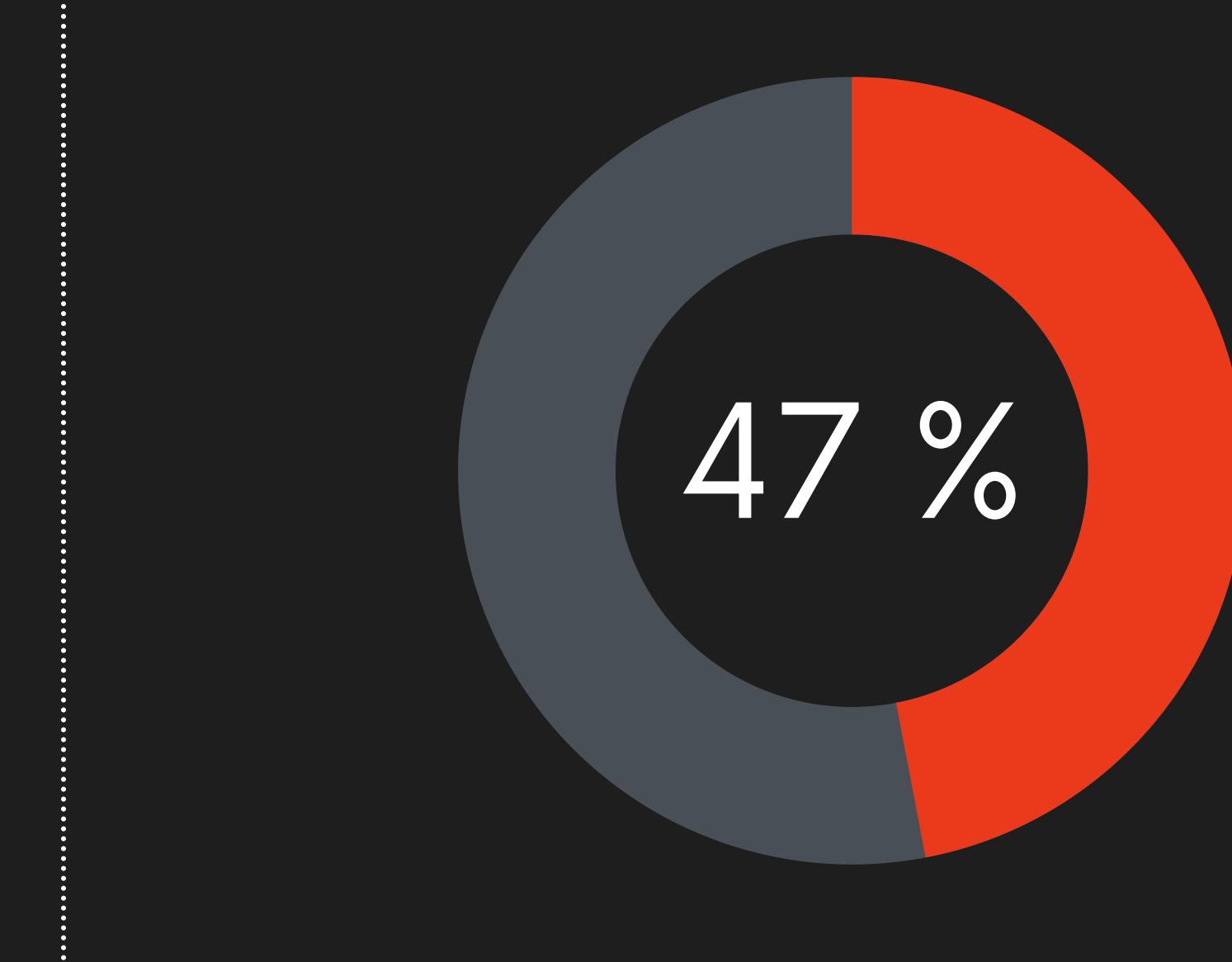
Elles forment une base orthonormée et sont construites pour que les données aient une variance maximale selon ces nouveaux axes.

Les 2 premiers axes forment le 1er plan factoriel qui nous permettra d'effectuer une projection 2D des informations avec le maximum de variance ou d'explications.





2 sur 14



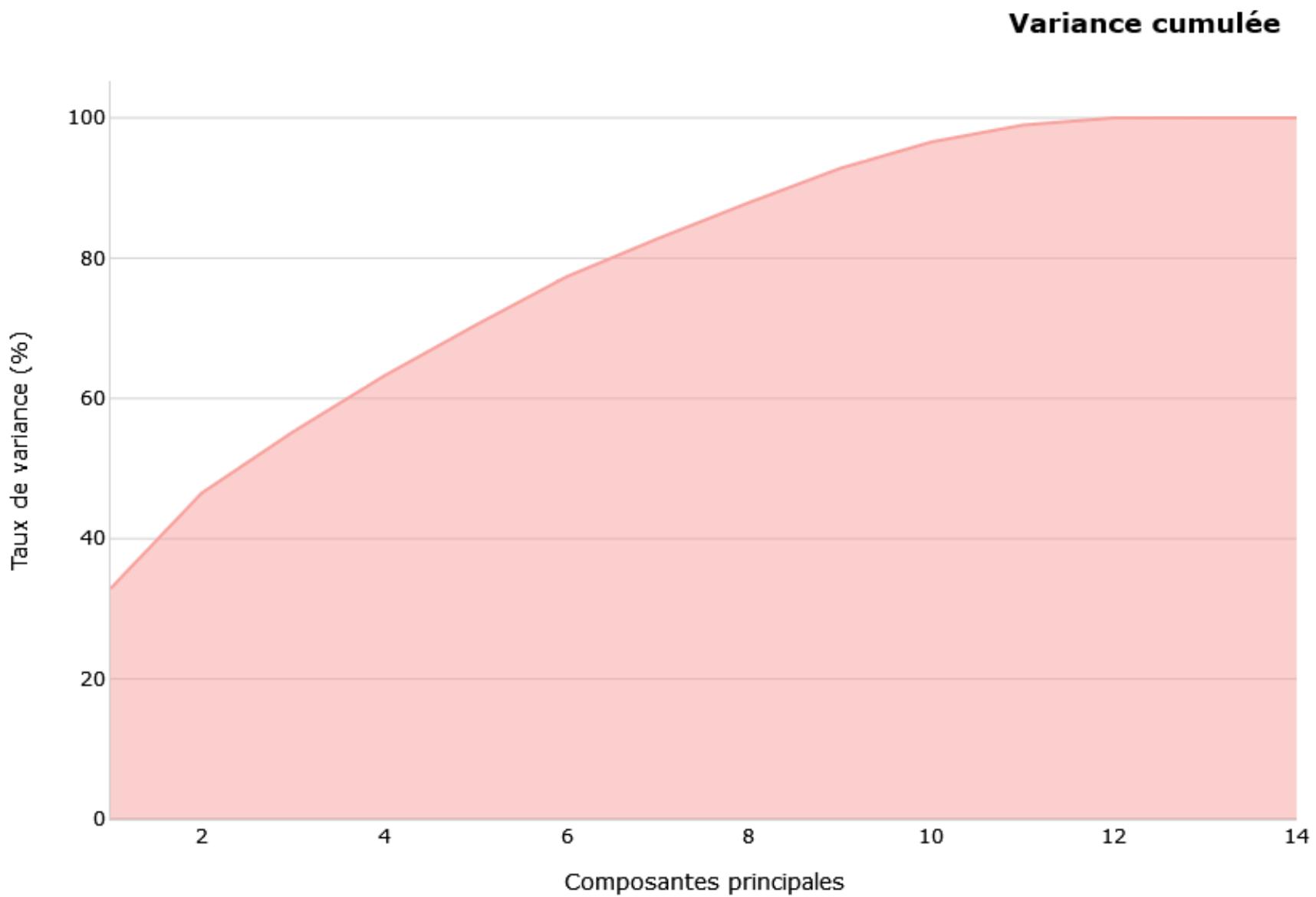
Proportion de contribution expliquée par les 2 premiers
axes qui définissent le 1er plan factoriel.

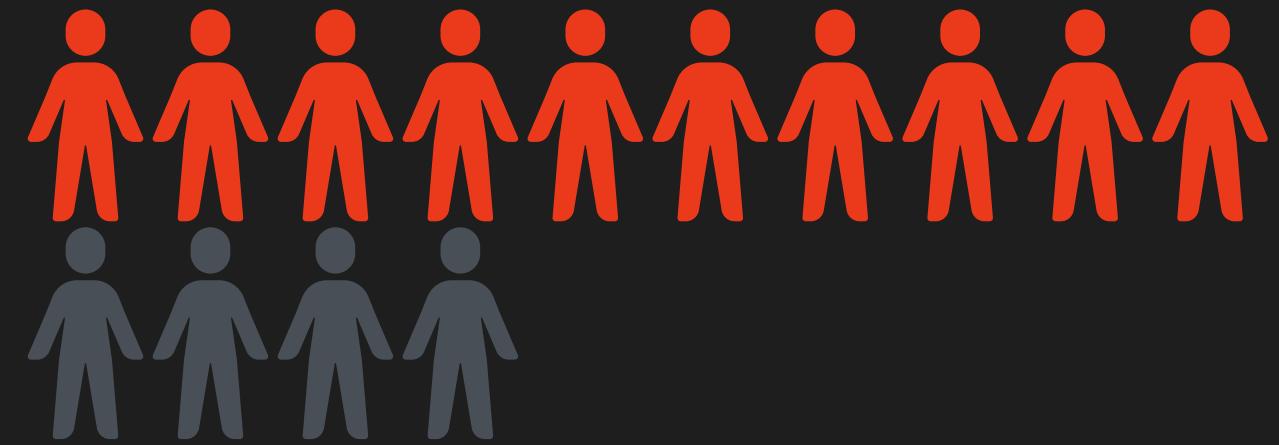
Variance cumulée

La variance cumulée est le cumul des contributions de chaque composante qui permettent d'expliquer les informations.

Pour choisir le nombre de composantes à utiliser, on regarde la proportion de la variance totale expliquée par le nombre de composantes.

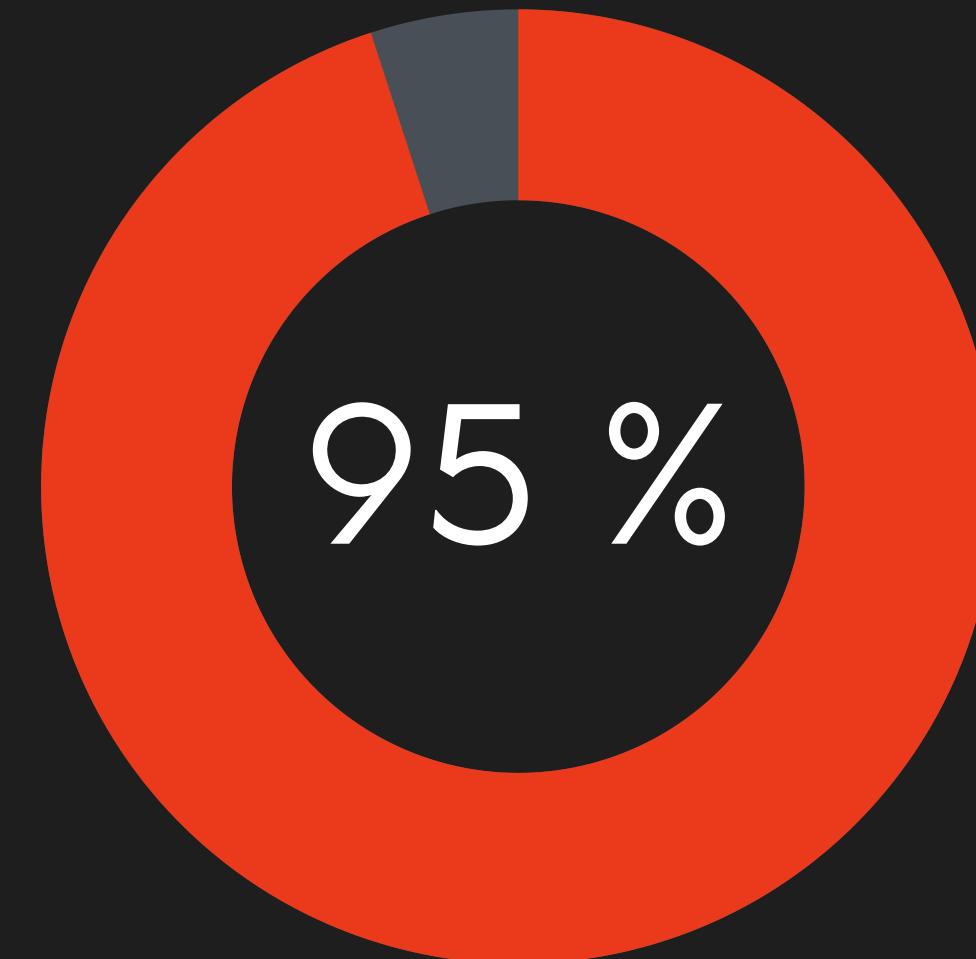
Cela permet de définir le nombre de composantes utiles pour une analyse.





10 sur 14

Nombre de composantes expliquant
95% de la variance cumulée



Proportion de variance cumulée



Liaisons entre les variables

Le 1er objectif consiste à étudier les liaisons entre les différentes variables en les projetant sur le 1er plan factoriel.

Les données étant normée, nous pourrons également nous appuyer sur un cercle de corrélations pour étudier les liens.

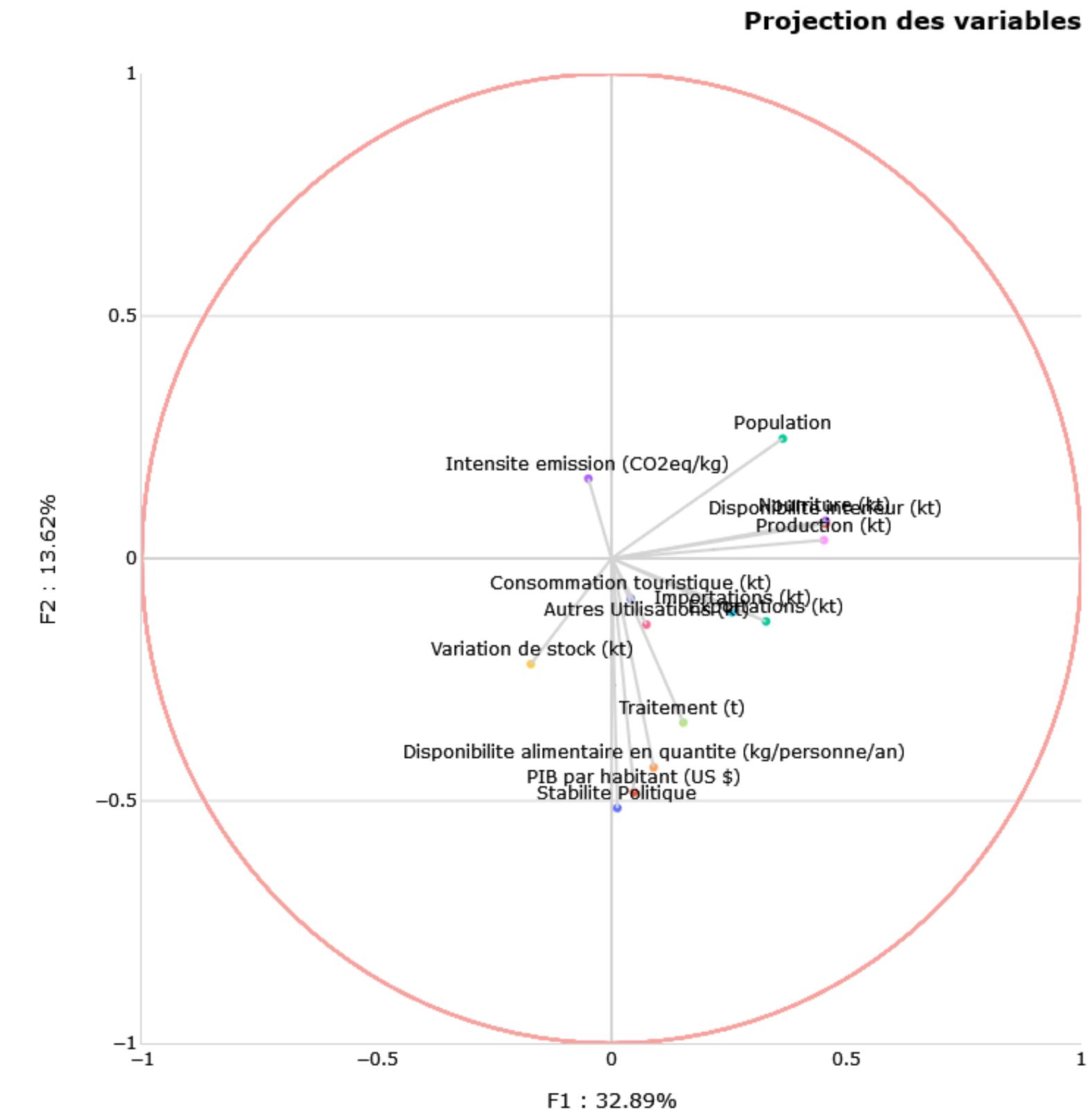
La recherche de ces liens entre les différents groupes de variables permettra aussi de les résumer par une variable synthétique.



Projection des variables

Projection des variables sur le 1er plan factoriel.

Quels sont les liens entre les variables et comment les résumer ?





Synthétisation axe F1

Les ressources

Corrélation positive

Disponibilité intérieur / Nourriture
Production / Population

Variable synthétique

L'axe F1 regroupe les variables ayant comme point commun les ressources. En effet, plus l'une de ces variables est importante, plus les autres augmentent. Cet axe permet aussi de faire le lien la production et la population. On dit qu'elles sont corrélées positivement.



Synthétisation axe F2

La situation géo-économique

Corrélation négative

Stabilité politique
PIB par habitant
Disponibilité alimentaire

Variable synthétique

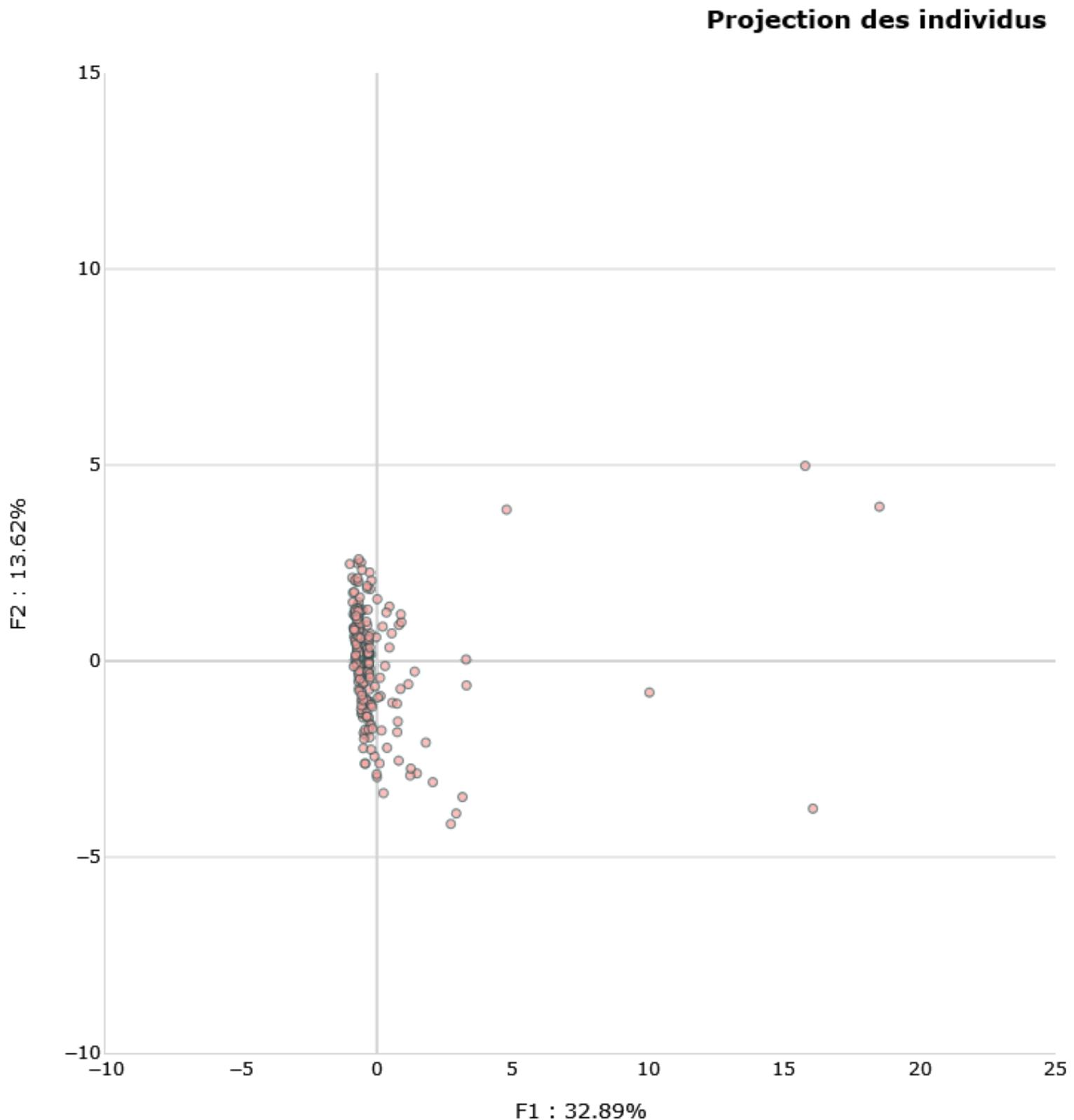
L'axe F2, regroupe les variables indiquant la situation géo-politique et économique des individus. Ici, on remarque que les individus ayant une bonne stabilité politique sont aussi ceux qui disposent le plus de ressources et bénéficient d'une bonne situation économique.

Projection des individus

Projection des individus sur le 1er plan factoriel.

Ici, on peut remarquer que nos individus sont relativement bien regroupés autour du centre de gravité avec 3 spécificités :

- 1 groupe corrélées positivement à F1
- 1 groupe corrélées négativement à F1
- 1 groupe excentrés du reste des individus





Partie 2 :

Clustering : Regroupement des individus



Classification Ascendante Hiérarchique



Le clustering hiérarchique (CAH) consiste à créer une arborescence de cluster pour représenter les données sous formes de noeuds liés entre eux.



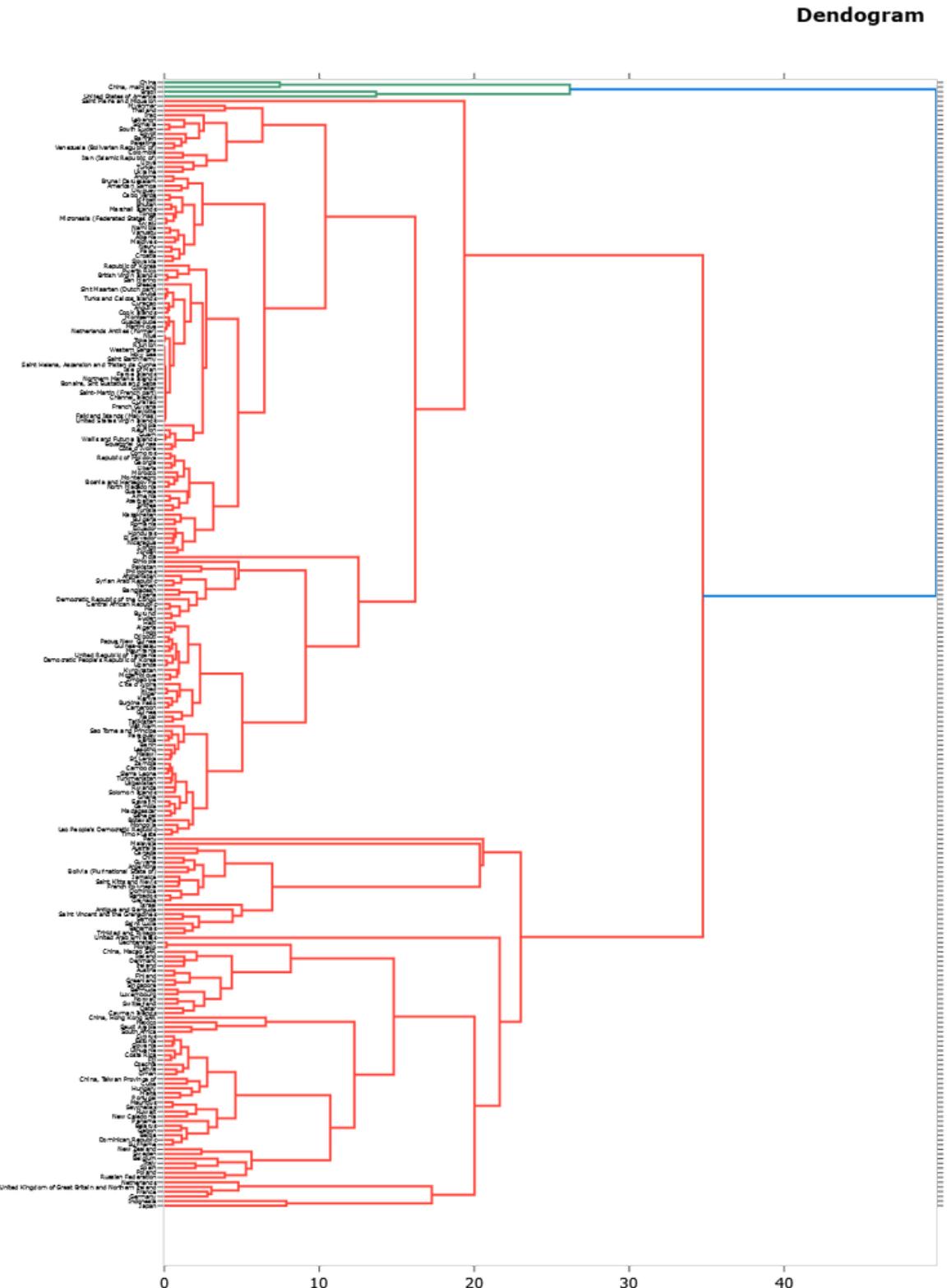


Dendogram

Le Dendrogramme permet d'organiser les données en arborescence en fonction de leurs similitudes.

Pour cela, nous allons regrouper les individus par paire et minimiser l'augmentation d'inertie intra-classe via la méthode de Ward, ce qui revient à resserrer les clusters sur eux mêmes à chaque itération.

Cela nous permet de mettre en évidence 3 groupes.

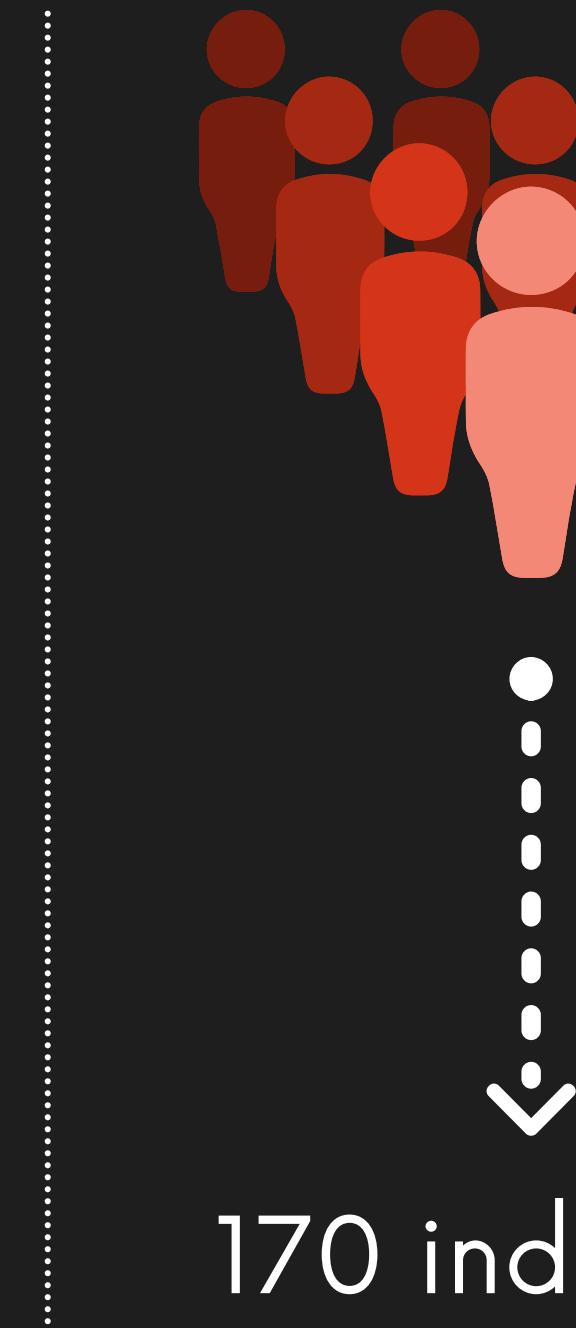




#LPQC



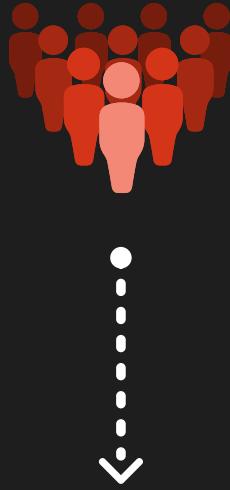
240 individus



170 individus



66 individus



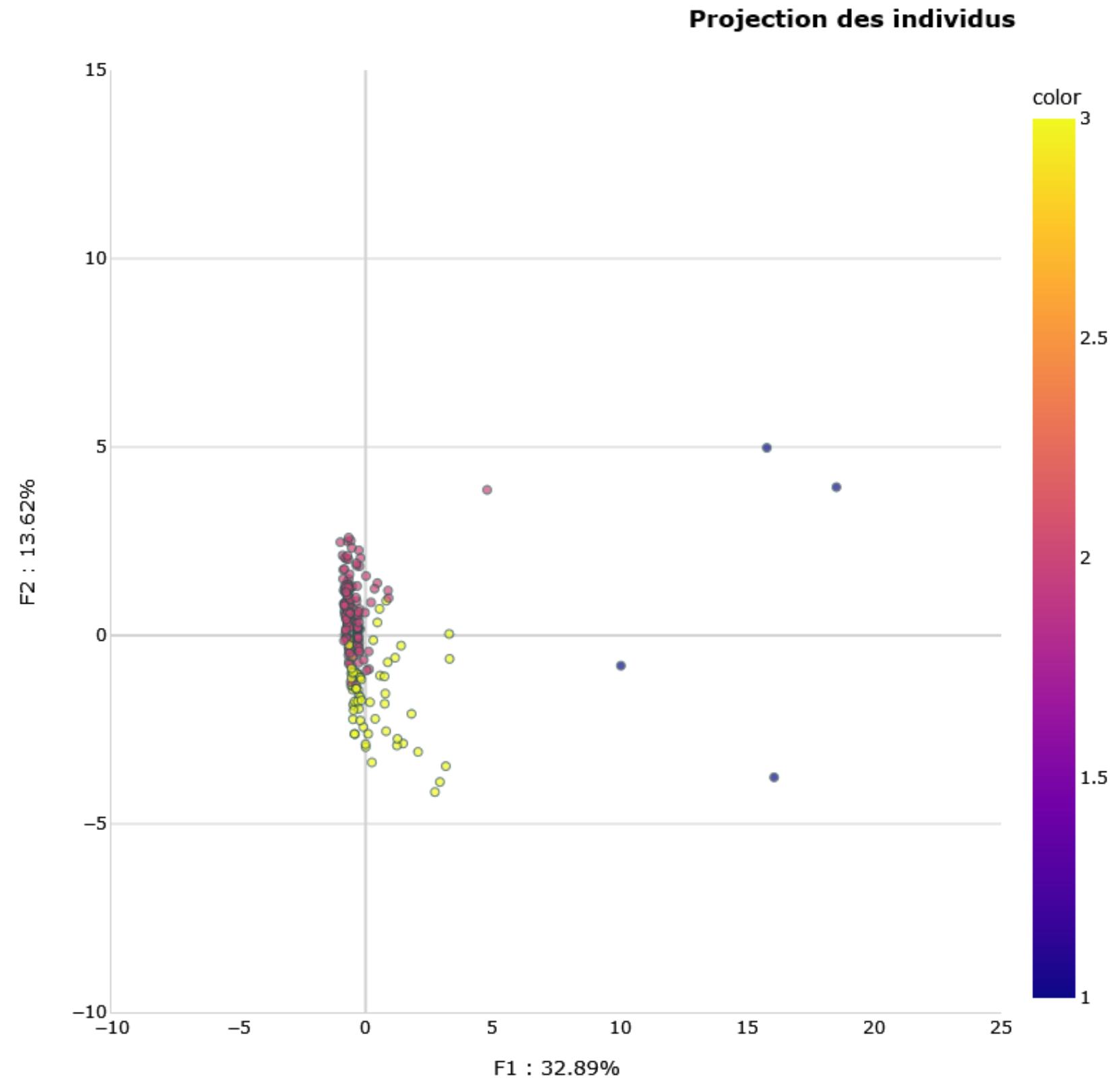
4 individus



Projection des individus

Après avoir effectuer la CAH, on remarque que la projection des clusters sur le 1er plan factoriel nous rappelle les 3 spécificités que nous avions mis en évidence lors de la définition des liens entre les variables soit :

- 1 groupe corrélées positivement à F1
- 1 groupe corrélées négativement à F1
- 1 groupe excentrés du reste des individus





Clustering par K-Means

La méthode des k-means permet de regrouper les individus en un nombre de clusters distincts.

Cette méthode repose sur la minimisation de la somme des distances euclidiennes au carré entre chaque individu et le centroïde (le point central) de son cluster.

Dans cette méthode, nous allons calculer à chaque itération le centroïde des clusters jusqu'à ce qu'il y ait convergence autrement dit jusqu'à ce que les centroïdes restent immobiles.

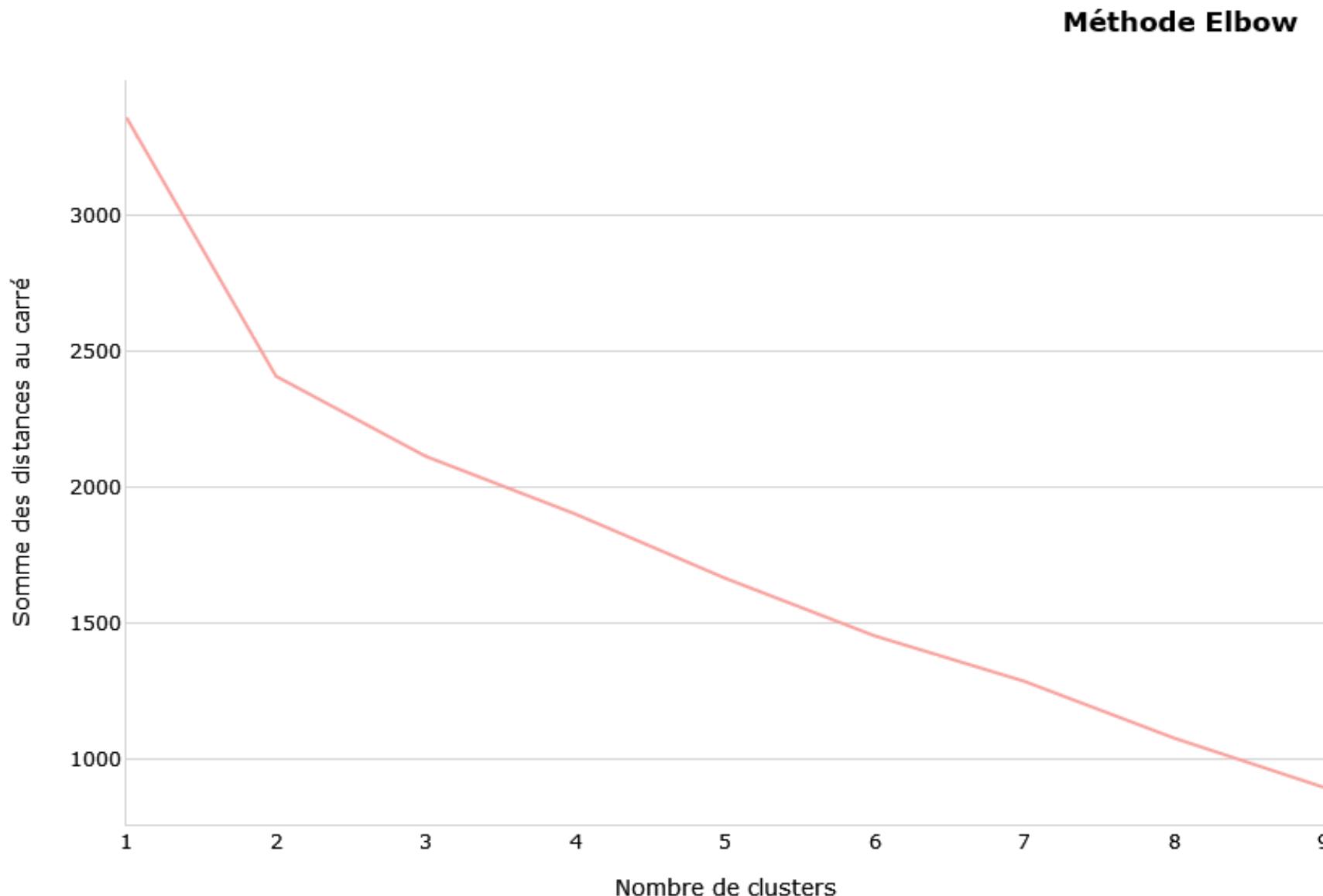


Méthode Elbow

La méthode Elbow consiste à tracer la variation expliquée en fonction d'un nombre de clusters donné.

Le coude de la courbe détermine le nombre de clusters optimal qui dans notre cas serait de 2.

Volontairement, je choisis 3 clusters ce qui permettra de lier les outliers dans le 3e cluster et de comparer les résultats ave la CAH.





#LPQC



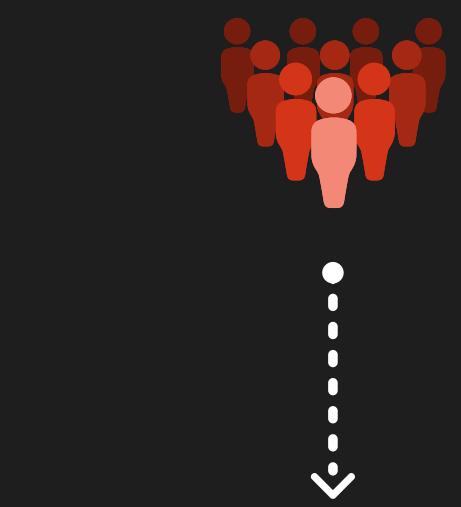
240 individus



168 individus



68 individus

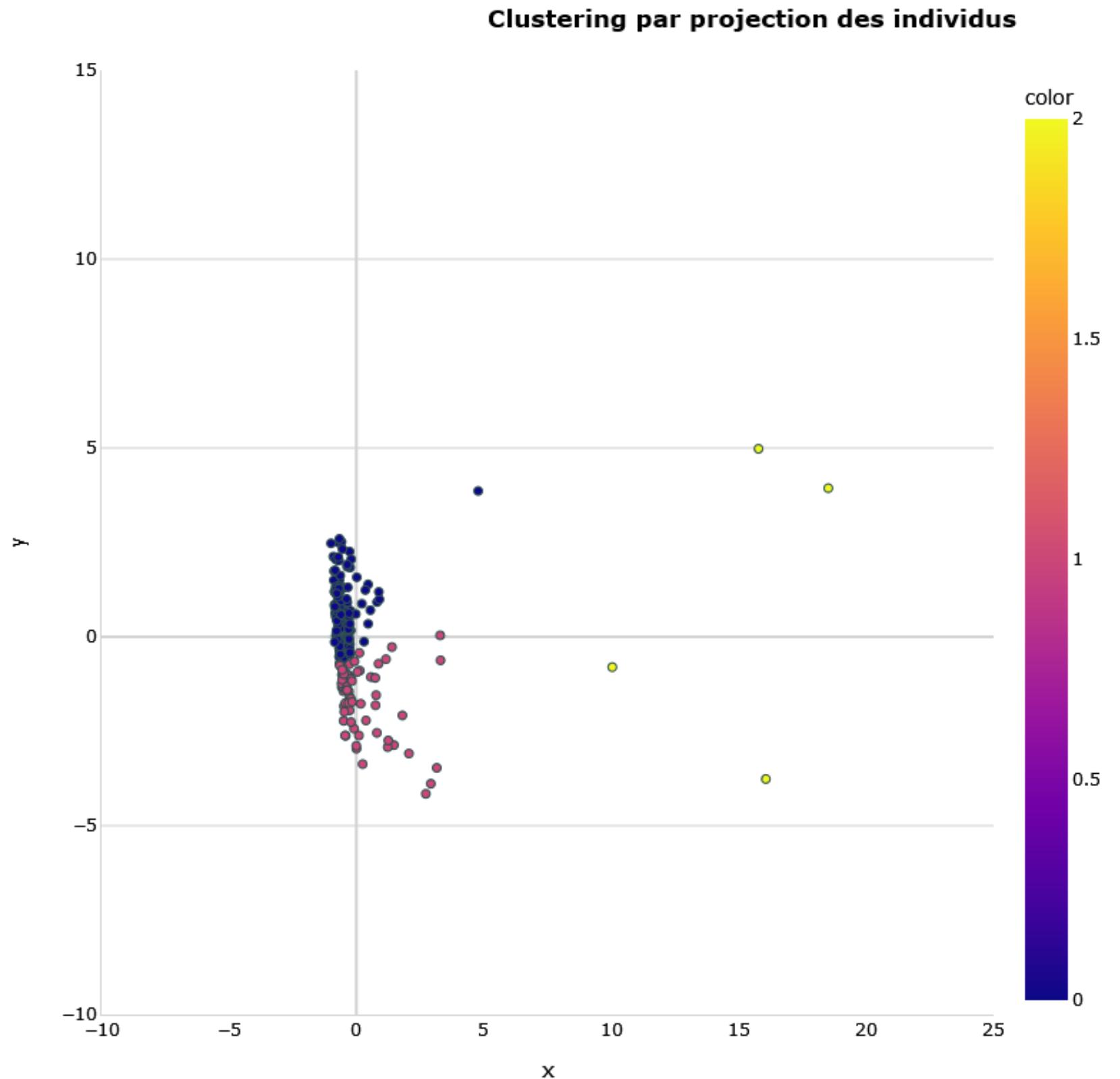


4 individus

Projection des individus

Avec la méthode des K-Means, comme pour la CAH, nous avons également nos 3 clusters avec nos 3 spécificités soit :

- 1 groupe corrélées positivement à F1
- 1 groupe corrélées négativement à F1
- 1 groupe excentrés du reste des individus

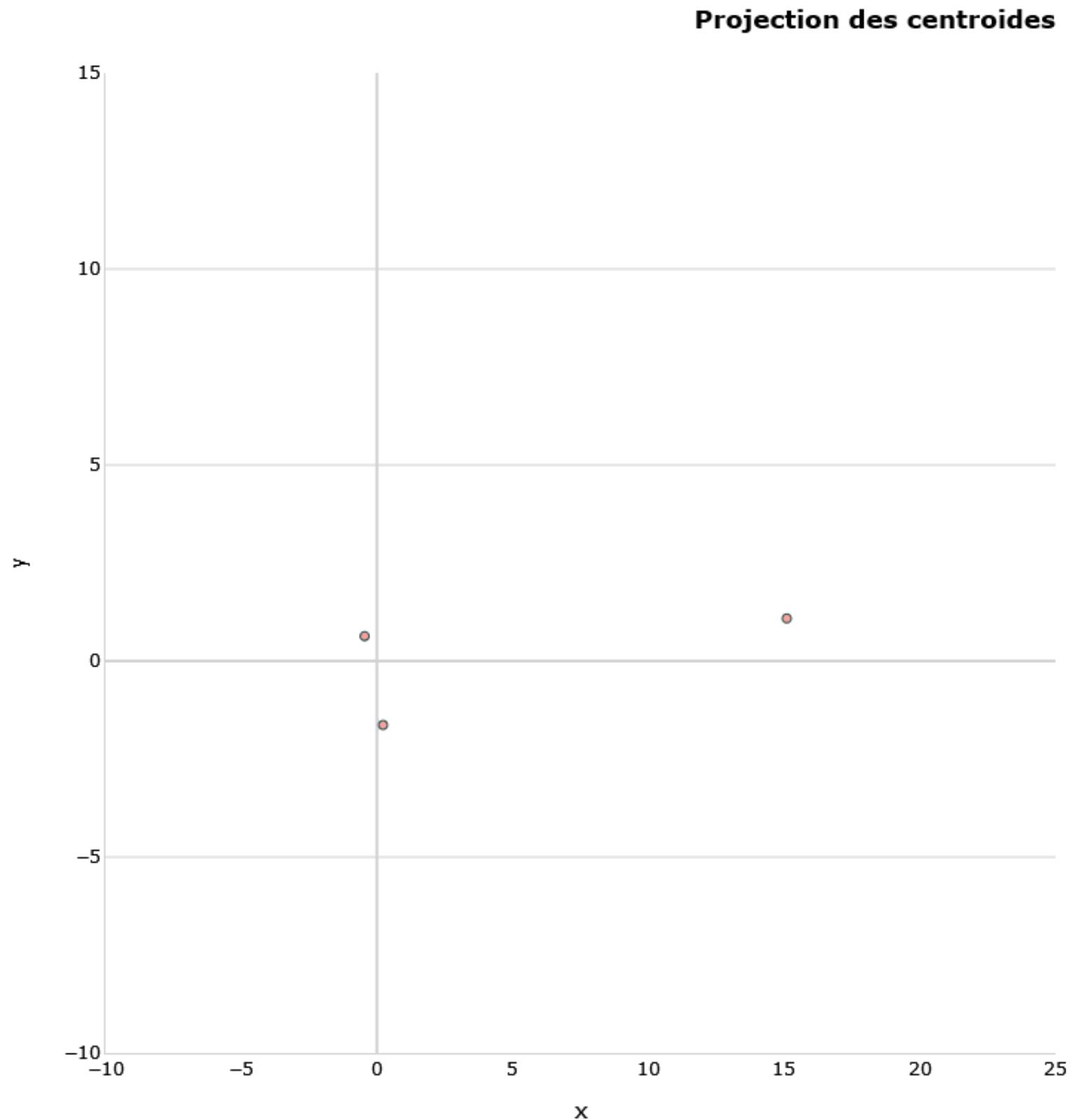




Projection des centroïdes

Les centroïdes sont représentatifs des clusters et gardent les spécificités des individus projetés sur le 1er plan factoriel soit :

- 1 centroïde corrélées positivement à F1
- 1 centroïde corrélées négativement à F1
- 1 centroïde excentrés du reste des centroïdes





#LPQC

Analyse des clusters

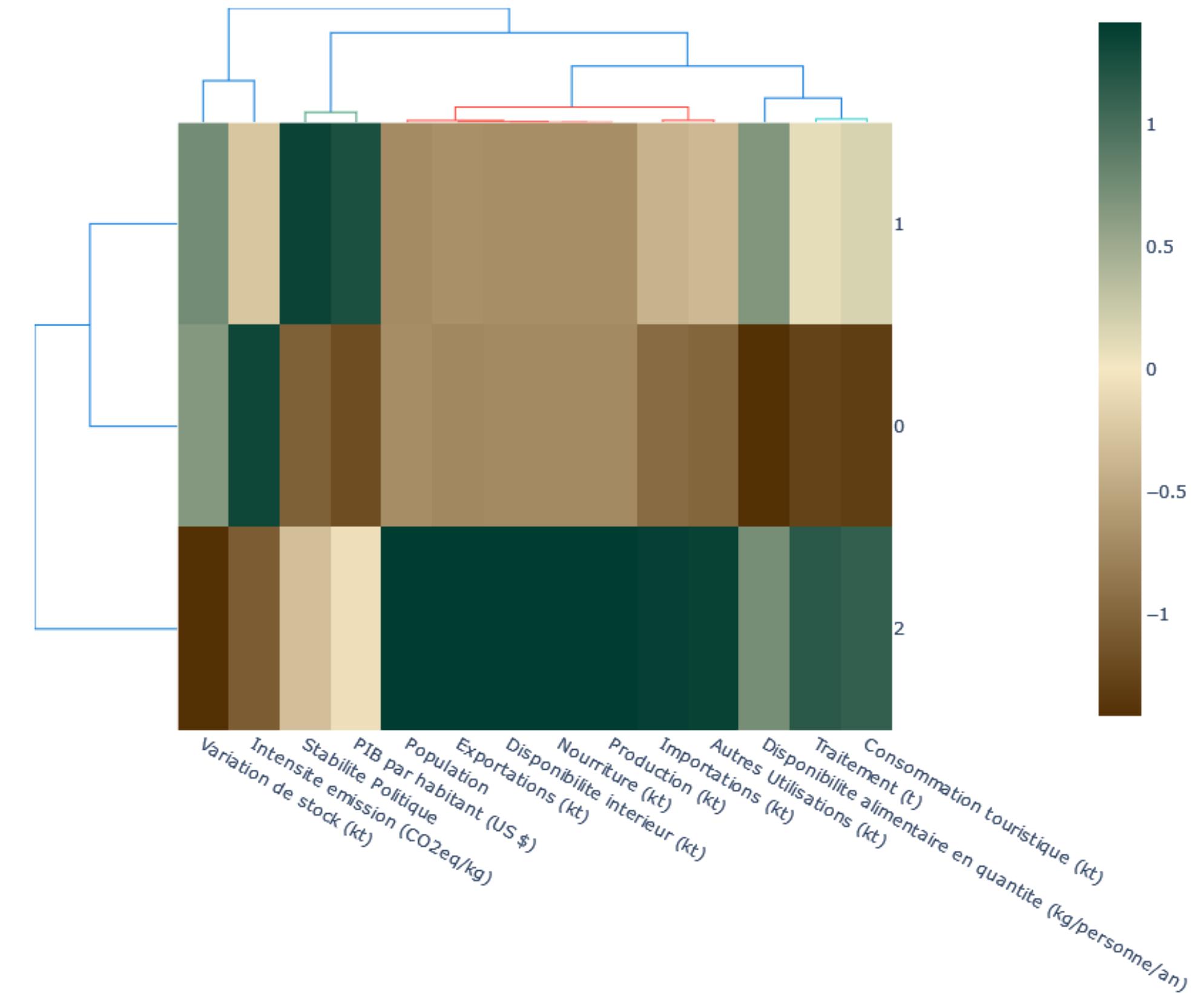
Partie 3:



Clustermap

Une clustermap est une carte qui rapproche et met en évidence les liens entre les différentes variables ainsi que les groupes auxquels elles appartiennent.

Comme pour le dendrogram, les données sont organisées en arborescence en fonction de leurs similitudes.





Synthétisation des clusters

Les développés

Cluster 0

Ce groupe correspond aux pays développés.

Il représente la plus grande partie des individus avec une situation géo-politique et économique stable.

Ce groupe a une gestion de sa disponibilité intérieur (production, nourriture ...) et du commerce associé (exportation, importations...) équilibrée.

Ces pays sont assez autonomes et le marché risque d'être concurrentiel tout en étant stable une fois implanté mais la demande y est présente.

La situation géopolitique y est aussi un facteur favorable au développement.



Synthétisation des clusters

Les en développement

Cluster 1

Ce groupe correspond aux pays en développement.

Il représente la plus faible partie des individus avec une situation géo-politique et économique instable.

Ce groupe a une gestion de sa disponibilité intérieur (production, nourriture ...) et du commerce associé (exportation, importations...) assez déséquilibrée.

Ces pays sont en demande pour la plupart avec une insuffisance en terme de disponibilité.

La situation géopolitique compliquée pourrait être un frein, seulement, parmi ces individus se trouvent des pays dont la stabilité ainsi que le PIB par habitant se rapproche des pays développés et pour lesquels il serait intéressant de pousser l'analyse.



Synthétisation des clusters

Les surdéveloppés

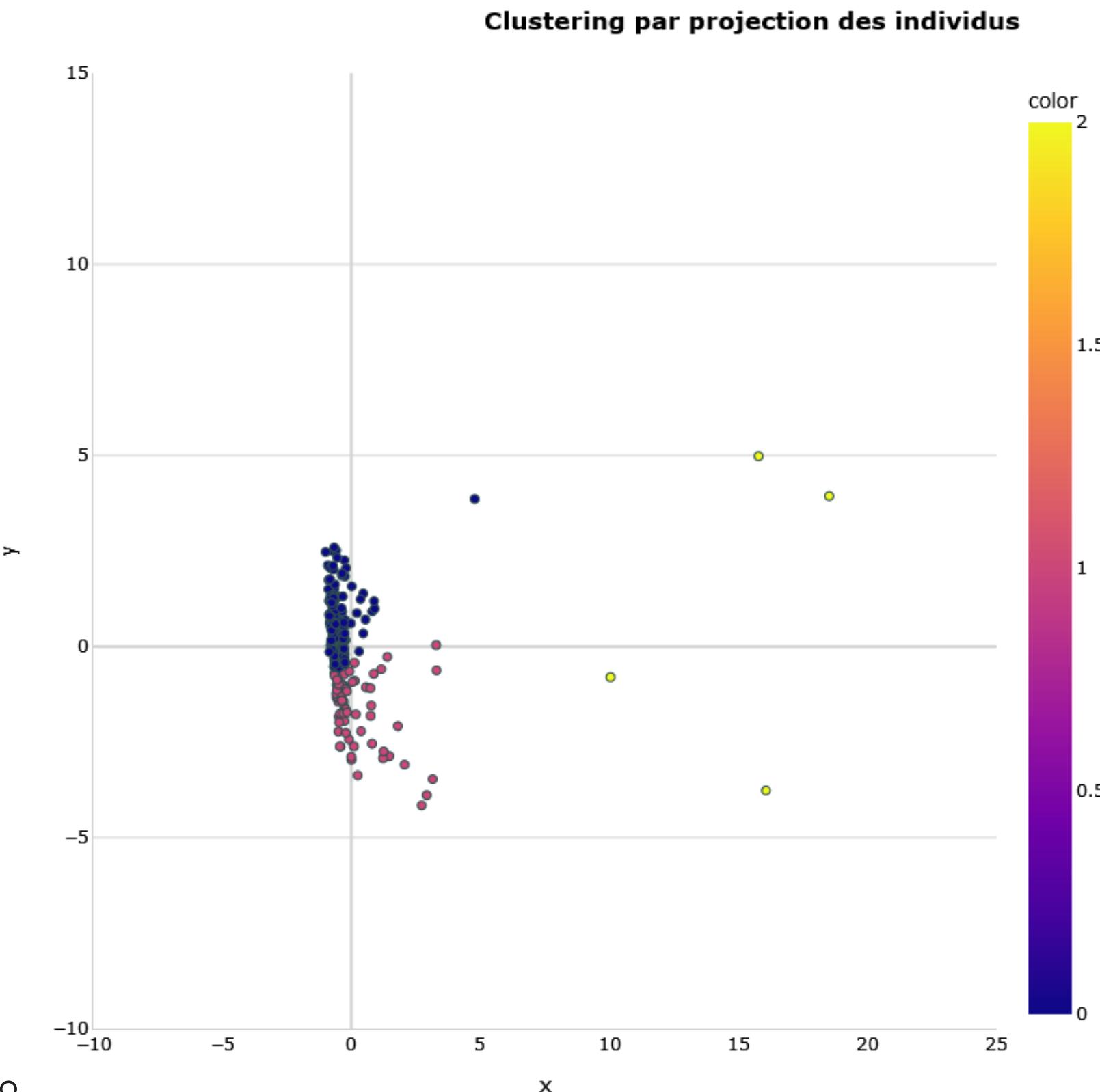
Cluster 2

Ce groupe correspond aux pays très développés.

Il représente une élite des individus avec une situation géo-politique et économique très stable.

Ce groupe a une très bonne gestion de sa disponibilité intérieur (production, nourriture ...) et du commerce associé (exportation, importations...).

Ces pays sont autonomes et le marché risque d'être très concurrentiel tout en étant stable une fois implanté.





#LPQC

Merci pour votre
attention

