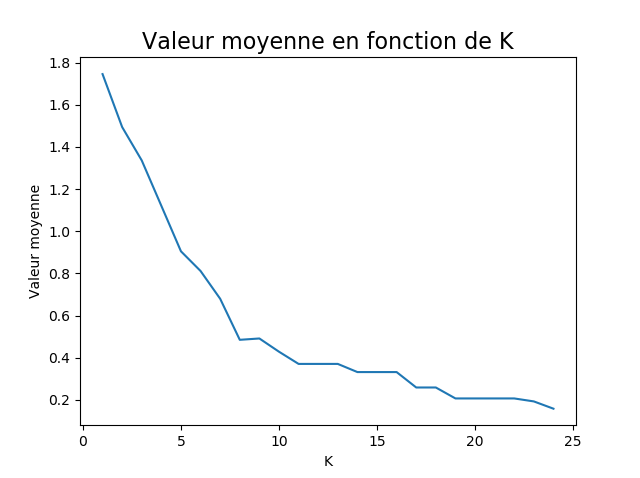
Partie 1 :

Valeur optimale pour K

Nous avons utilisé la méthode des K-Means avec un nombre de clusters K qui varie entre 1 et 25 et comme paramètres : *init=’random’* pour que l’algorithme choisisse comme barycentres initiaux des données aléatoires et *random\_state=0* comme seed pour le générateur de nombres aléatoires.

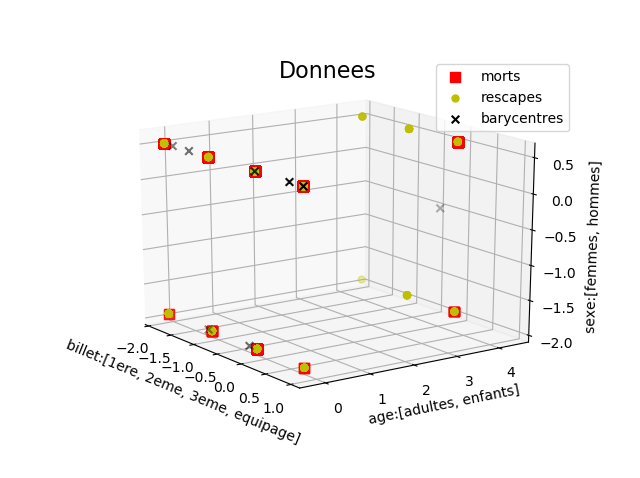
Pour chaque valeur de K, nous avons déterminé la moyenne des distances moyennes entre les barycentres et leurs points. Nous obtenons comme résultats:



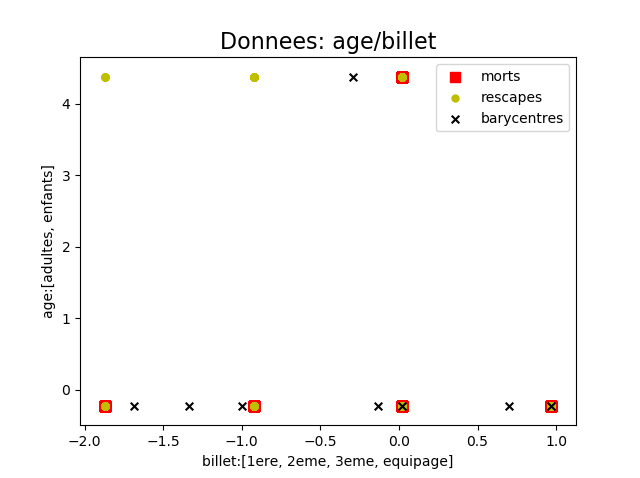
Nous pouvons voir que le nombre optimal de clusters à choisir est 8.

En effet, plus K augmente plus la valeur moyenne diminue mais on peut voir qu’à partir de K = 8 la distance entre les barycentres et leurs points diminue beaucoup moins vite. Choisir un nombre de clusters plus élevé n’apporterait donc pas beaucoup plus d’information sur les données.

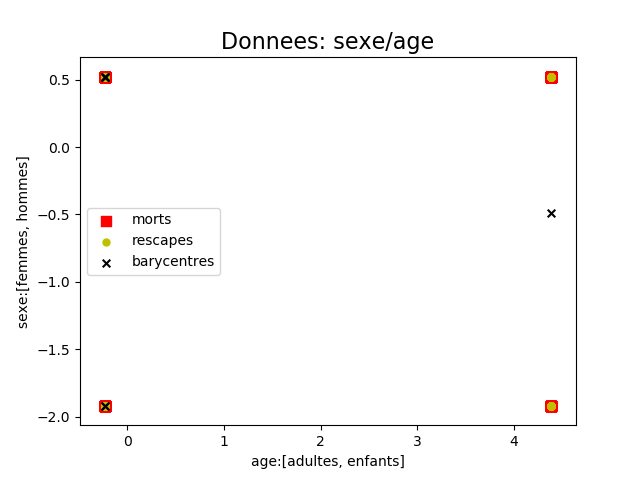
Représentation des données :



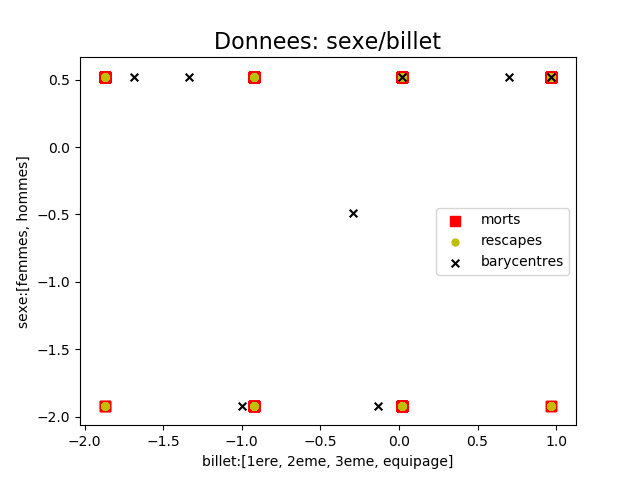
A partir de la représentation en 3 dimensions, nous pouvons déjà constater qu’aucun enfant en 1ère et 2ème classes n’est mort et que le sexe n’a que peu d’influence sur le taux de survie des enfants.



Nous pouvons voir par la position des barycentres que la grande majorité des passagers étaient des adultes et que parmi eux la plupart étaient en 3ème classe ou faisaient partie de l’équipage. C’est dans ces groupes que se trouvaient la majeure partie des personnes qui n’ont pas survécu.



Ici nous constatons (mieux avec la représentation 3d) qu’il y avait bien plus d’hommes que de femmes sur le titanic et que leur taux de survie était plus faible.



Finalement ici, nous pouvons voir que les femmes étaient presque pas présentes parmi les membres de l’équipage. C’est surtout le fait d’être une femme qui offrait une meilleure chance de survie mais aussi le fait d’être en 1ère ou 2ème classe.