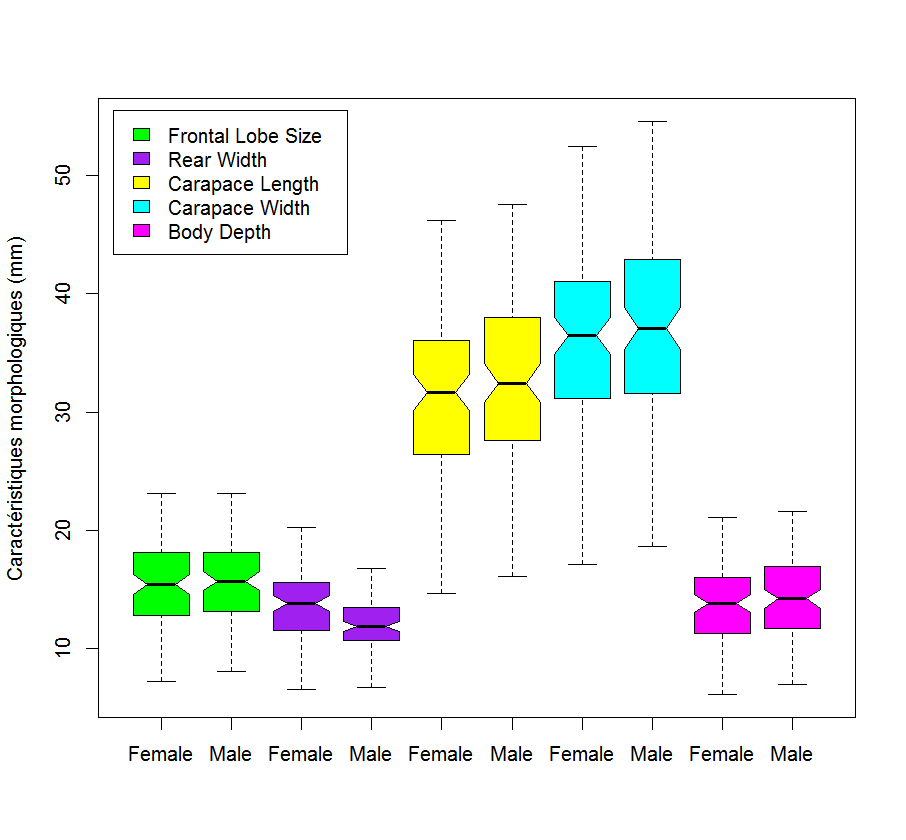
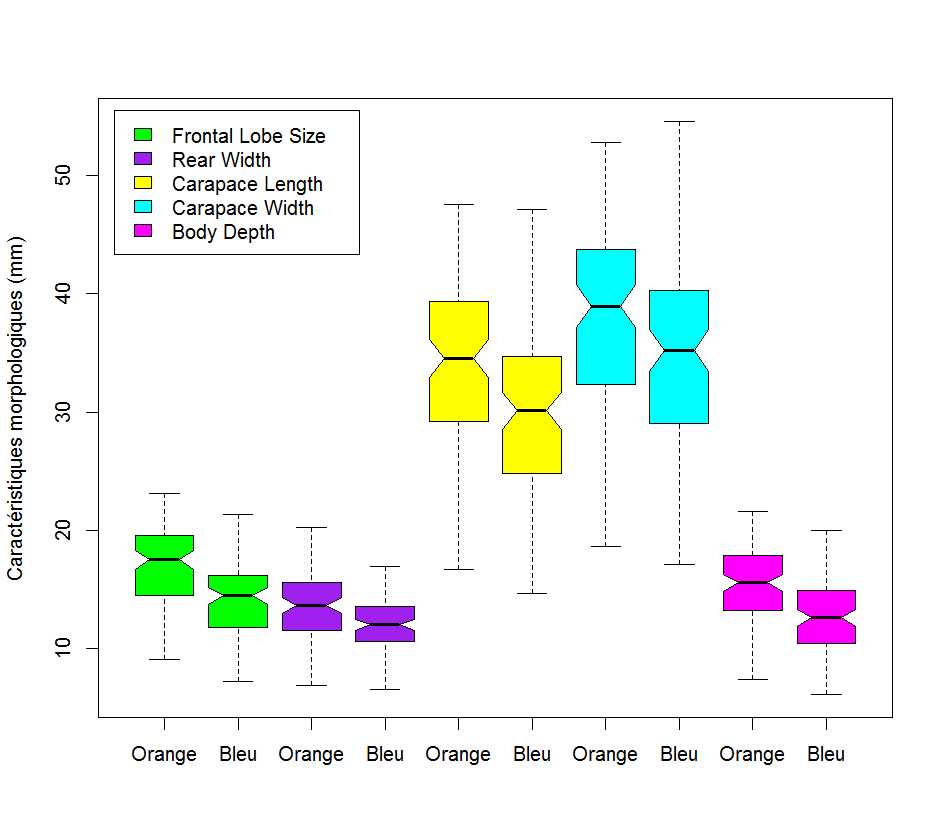
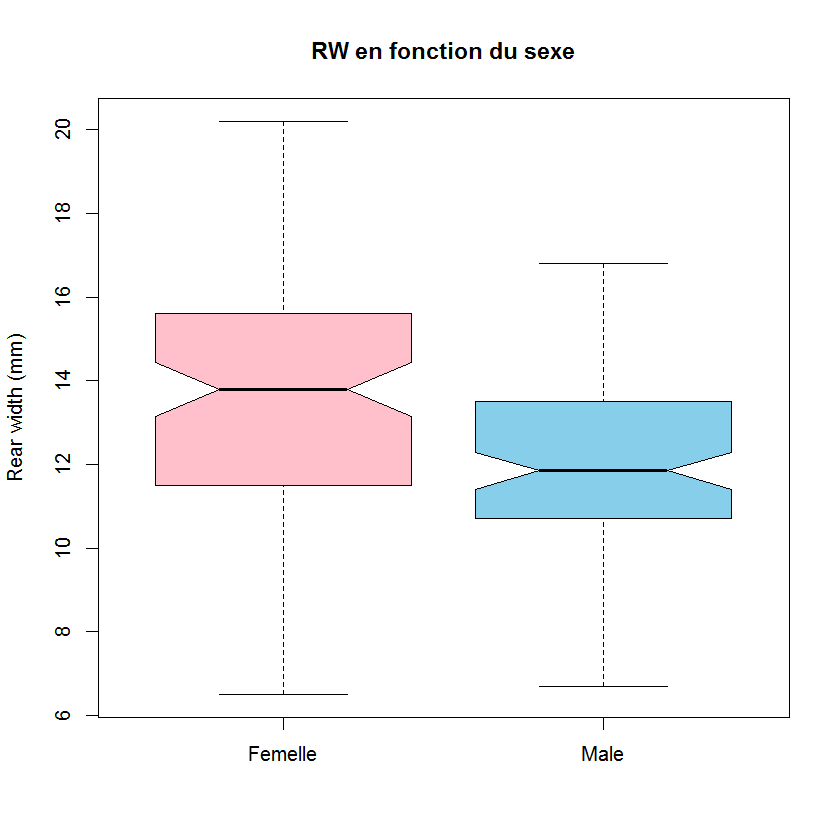
CRABS

Dans cet exercice nous allons analyser un jeu de données contenant 200 individus décrits par 8 variables : trois variables qualitatives et cinq variables quantitatives, dont en particulier l’espèce qui sépare les données en deux populations de 100 individus qui sont eux même identifiés en tant que femelle ou male (50 par espèce). Le but de l’étude est de déterminer s’il existe des différences morphologiques permettant d’identifier l’espèce ou le sexe d’un individu en fonction d’un ou plusieurs paramètres morphologique

Dans un premier temps, afin de visualiser s’il existant des différences morphologiques majeures entre les espèces ou les deux sexes, nous avons représenté les données sous forme de boxplot. Plus précisément, nous avons comparé tracé les boxplot de chaque paramètre en séparant soit selon le sexe soit l’espèce. Pour compléter l’étude, sur les boxplots, les intervalles de confiance à 95% ont aussi été représentés par un resserrement de la boite autour de la médiane, les points où la boîte se resserre représentent les bornes de cet intervalle.



(on ne mettra pas ces graphiques dans le compte rendu je pense)

Selon l’espèce, les intervalles de confiance des différents paramètres morphologiques ne se superposent pas ce qui indique bien qu’il y a des différences morphologiques significatives entre les deux espèces. En revanche, pour le sexe ce n’est le cas que pour une caractéristique : rear width, (largeur arrière)

Cependant, pour toutes les caractéristiques, que ce soit selon le sexe ou l’espèce, les extrémités des moustaches se superposent, ce qui nous empêche avec cette représentation d’identifier avec certitude le sexe ou l’espèce d’un individu selon une ou plusieurs caractéristiques morphologiques.

Le calcul de la corrélation entre les différentes variables est présenté dans le tableau.

FL RW CL CW BD

FL 1.00 0.91 0.98 0.96 0.99

RW 0.91 1.00 0.89 0.90 0.89

CL 0.98 0.89 1.00 1.00 0.98

CW 0.96 0.90 1.00 1.00 0.97

BD 0.99 0.89 0.98 0.97 1.00

Il est observé que toutes les variables sont très corrélées entre elles. En effet, les coefficients de corrélation sont tous supérieurs à 0.89. Ce résultat peut être expliqué par le fait que les variables ne sont pas indépendantes les unes des autres : la taille des membres des individus est proportionnelle à la leur taille

Pour s’affranchir de ce phénomène, il convient de diviser la taille de chaque membre de chaque individu par la somme de la taille de de ses membre. La matrice de corrélation obtenue après le traitement des données est représentée dans la matrice suivante.

FL RW CL CW BD

FL 1.00 -0.12 -0.24 -0.78 0.52

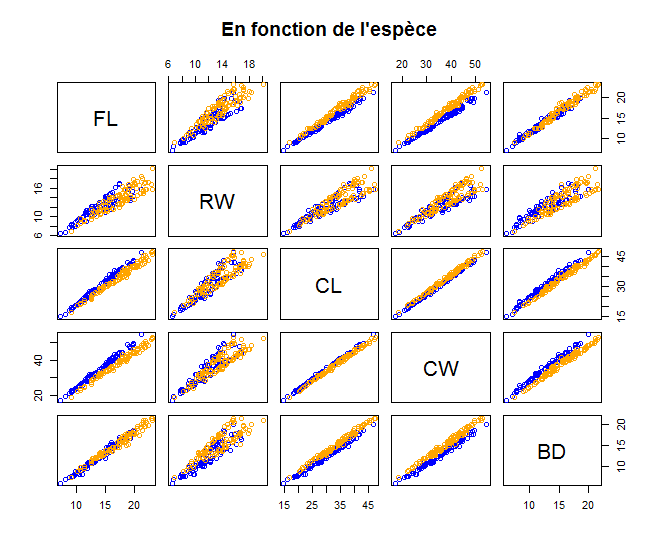
RW -0.12 1.00 -0.83 -0.20 -0.46

CL -0.24 -0.83 1.00 0.42 0.10

CW -0.78 -0.20 0.42 1.00 -0.65

BD 0.52 -0.46 0.10 -0.65 1.00

Les résultats obtenus indiquent que le traitement a permis de décorréler les variables entre elles. Les variables les plus corrélés sont désormais FL et CW (coefficient de corrélation de -0.78) et CL et RW (coefficient de corrélation de -0.83)

La figure représente les graphiques matriciels des données traitée, en les distinguant (a) selon le sexe et (b) selon l’espèce. Des nuages de points distincts sont observés ce qui nous permet maintenant de distinguer l’espèce ou le sexe d’un individu par ses caractéristiques morphologiques

