TP 6 7 8 : Manipulation de tableaux de données statistiques (dataframe)

1. Valeurs manquantes

Importer les données du fichier echDecathlon.txt dans un dataframe.
Combien y a –t-il de variables ? d'individus ? Quel est le nom des variables ? Les lignes sont-elles nommées ?

- Repèrer et dénombrer les valeurs manquantes, au total, par variable, par individu. On pourra utiliser une écriture avec des boucles ou les fonctions rowSums, colSums

- Pour les variables Longueur et Hauteur, calculer moyennes, variances, covariance.

Calculer la matrice des corrélations pour 10 épreuves.

- Créer dans le dataframe une colonne indiquant le nombre de résultats manquants pour chaque athlète.

Sélectionner les athlètes qui ont au plus un résultat manquant.
Estimer la valeur manquante au 100mhaies par régression linéaire.

2. Renommer, fusionner les niveaux d'un facteur

Créer un dataframe à partir des données du fichier donnees.txt.
Obtenir le tableau des effectifs de la variable sexe.
Fusionner les modalités m et h en H, pour la variable sexe et recoder f en F.
Refaire le tableau des effectifs.

3. Calcul d'une variable. Découpage en classe

Reprendre le dataframe eleves créé au TP précédent.

- calculer la variable IMC : poids/taille² avec poids en kg et taille en m
- découper en classe et coder (créer une nouvelle variable : corpulence) , selon l'interprération suivante des classes d'IMC :
  - moins de 16,5 : dénutrition
  - 16,5 à 18,5 :maigreur
  - 18,5 à 25: corpulence normale
  - 25 à 30 : surpoids
  - 30 à 35 :obésité modérée
  - 35 à 40 : obésité sévère
  - plus de 40 : obésité morbide ou massive
- fusionner les 2 premiers niveaux et les 4 derniers. Il reste trois niveaux : maigre, normal, gros, pour la variable corpulence.
- faire le graphique en barre pour cette variable (fonction barplot) et vérifier l'ordre des modalités.

4. Exporter un data frame dans un fichier texte

Exporter le data frame eleves, (complété, donc avec 7 variables) dans un fichier texte nommé donneeseleves.txt du répertoire de travail. Vérifier le contenu du fichier texte avec un éditeur de texte.
Importer ce fichier texte dans un data frame nommé DF et vérifier l'ordre des niveaux de la variable corpulence. Les réordonner …

5. Sauvegarde des objets dans fichier Rdata

Sauvegarder les objets créés dans la session dans le fichier TP6.Rdata.

6. Supprimer les modalités absentes d'un facteur.

On définit x par x=factor(c("a","a","b","b","a"),levels=c("a","b","c"))
La commande table(x) donne c d'effectif nul.
Appliquer la fonction factor au facteur x.
Vérifier en appliquant la fonction table à ce facteur.

7. ** Regrouper les modalités rares d'un facteur.

Créer le dataframe actesMed:

```
actesMed=data.frame(cout=rnorm(100,10,2),specialite=factor(rpois(100,3),
        labels=c("angio","med","dent","ophtal","analyse","obst","radio","séance","reeduc")))
str(actesMed)  # 9 niveaux
```

Le facteur specialite devra avoir 9 modalités et il sera peut-être nécessaire de relancer la génération de nombres aléatoires.

Obtenir le tableau des effectifs de la variable spécialité.

On souhaite recoder les modalités rares en "autres":

Créer le vecteur des modalités de la variable.
Obtenir le vecteur des modalités rares (d'effectifs faible , par exemple<5)
Dans le vecteur des modalités modifier les modalités rares en "autres".
Recoder les modalités de la variable spécialité.