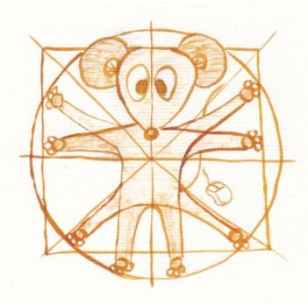

M1 EDBS

Data Mining

TD 1 : compléments de R
Exploration de tableaux de données

Robert Sabatier
Christelle Reynès

Pour ce TP on se propose de charger dans votre session de R, deux tableaux de données (que vous trouverez sur l'ENT) :

crabes.txt qui contient un mélange de variables quantitatives et qualitatives,
Fishertot.txt qui est un tableau de 2 variables qualitatives.

Enregistrer les données dans un dossier TP1.

Il y a également un fichier : *qq_fonctions_TP1.txt*, contenant des fonctions R que l'on utilisera plus loin, mais qui sont à charger dans votre session R. On utilisera également la library *plotrix*.

Travail sur un tableau de variables quantitatives avec des groupes:

1. A l'aide de la fonction *getwd*, cherchez quelle est la direction de travail actuelle.
2. Si nécessaire, à l'aide du menu ou de la fonction *setwd*, changez la direction de travail vers ce répertoire.
3. Appliquez la fonction *source* au fichier *qq_fonctions_TP1.txt*.
4. Importez le jeu de données *crabes*.
5. Quel type d'objet obtenez-vous ?
6. Explorez ce jeu de données à l'aide de la fonction *summary*. Que remarquez-vous ?
7. Réalisez un histogramme de la variable RW en effectifs et en probabilités, ajoutez l'étiquette « RW » sur l'axe des abscisses.
8. Utilisez la fonction *density* pour réaliser une estimation de la densité sous-jacente et représentez-la sur l'histogramme.
9. Réalisez un boxplot de la variable RW.
10. Représentez sur un même graphique les boxplots de toutes les variables quantitatives. Commentez.
11. Calculez la moyenne et l'écart-type de toutes les variables quantitatives (utilisation de la fonction *apply*).
12. Réalisez un centrage/réduction du jeu de données à l'aide de la fonction *scale*. Observez graphiquement l'effet de cette transformation. Commentez.

13. Utilisez la fonction *cor* pour calculer la matrice de corrélation entre les variables quantitatives.
14. Utilisez la fonction *image* pour représenter cette matrice.
15. Utilisez la fonction *pairs* pour représenter les informations de cette matrice.
16. Appliquez le code suivant et commentez :
    pairs(X,lower.panel=panel.smooth,upper.panel=panel.cor,diag.panel=panel.hist)
17. Enregistrer le jeu de données centrées-réduites à l'aide de la fonction *save*, supprimez l'objet R de votre environnement (fonction *rm* et *ls* pour vérifier) et rechargez-le dans R à l'aide de la fonction *load*.

Les données de Fisher :

1. Importez le fichier *Fishertot.txt*.
2. Utilisez la fonction *table* pour décrire la distribution de chaque variable.
3. Avec la même fonction, construire la table de contingence croisant ces deux variables.
4. A l'aide de la fonction *which.max*, identifier le mode de chaque variable.
5. Utilisez la fonction *barplot* pour représenter graphiquement la distribution de la variable « couleur des yeux ».
6. Appliquez la fonction *barplot* à la table de contingence obtenue précédemment. Que représentez-vous ? Ajoutez une légende.
7. Représentez la distribution des couleurs de cheveux en fonction de la couleur des yeux.
8. Explorez l'effet de l'argument *beside* de la fonction *barplot*.
9. Utilisez la fonction *pie* pour représenter la distribution de la variable « couleur des cheveux ».
10. Utilisez la fonction *text_pie* pour ajouter les pourcentages de chaque catégorie.

Sauvegardez l'environnement de travail.