Université d'Orléans 2021–2022
UFR CoST Miage M1

Analyse de données – Initiation au logiciel R
TP n°1 : Manipulation de tables `data.frame`

1 Manipulations de base = révisions de L3

Il s'agit ici de suivre le tutorial R en testant les différentes commandes sur votre environnement RStudio au fur et à mesure de l'avancement. Bien connaître les différents objets structurés (`vector, matrix, list, data.frame`), et la manière de les manipuler.

2 Manipulations sur un `data.frame`

Voici quelques exercices de manipulation de tables individus-caractères. On utilise en exemple le jeu de données *StateFacts* utilisé dans le cours de Statistiques descriptives.

(1) Données importées avec `read.table` depuis un fichier texte :
Ce sera l'une des méthodes usuelles pour récupérer des données lors des séances. Il faut que `read.table` "voit" le fichier texte, pour cela plusieurs possibilités :

- vous copiez le fichier texte dans le répertoire de travail sur lequel pointe R, qui est visible dans l'interface RStudio.
- vous changez dans R le répertoire de travail pour le faire pointer vers le dossier où vous avez copié le fichier texte, cf menu
  `Fichier → Changer de répertoire de travail...`
- vous indiquez à `read.table` le chemin complet depuis la racine du système jusqu'au dossier contenant le fichier texte

1. Récupérer les données "texte" en local ou en ligne et sauver le fichier `StateFacts.txt`
2. lisez l'aide `?read.table` pour charger la table suivant l'une des méthodes ci-dessus

R récupère les noms des variables si ceux-ci sont en première ligne du fichier texte et que l'argument correspondant de la procédure `read.table` le précise.

(2) Manipulations élémentaires

1. Afficher les données (seulement pour n petit ! ! !), ou afficher le début
2. vérifier que les noms des variables ont bien été récupérées ; `StateFacts.txt` contient les variables utilisées en cours plus quelques autres (nom des Etats, Espérance de Vie, Nombre moyen de jours de gel, Aire).
3. vérifier que les codes états sont bien les labels individus
4. Comment a été chargée la variable `Etat` ? Même question pour la variable `Region`. *Note : la réponse dépend de la version de R utilisée.*
5. Expérimentez le méchanisme `attach, detach` (cf. tutoriel).

6. Comme dans une étude usuelle, sauver ensuite ces données au format interne de R pour usage ultérieur, fichier `StateFacts.Rdata`
7. calculer quelques statistiques élémentaires, d'abord avec `summary`, puis avec les fonctions statistiques `mean, var, sd, max,...`.
8. Utilisez `tapply` pour calculer des statistiques par niveaux d'un facteur qualitatif : Ici par exemple on se demande si la Region a un effet sur la criminalité moyenne, le revenu, etc. Mêmes questions avec la dispersion de ces mesures.
   Commentez.

NB : cette dernière question donne déjà une manière de croiser une variable quantitative avec un facteur qualitatif, en fournissant un résumé numérique.

(3) Révision : en une seule commande à chaque fois :

1. Afficher la moyenne empirique de la variable `Revenu`
2. Afficher l'écart-type de `Meurtre`
3. Combien y-a-t-il d'États de Revenu moyen > 5000 ?
4. Afficher les moyennes empiriques des variables quantitatives de la table `states`
5. le nom de la variable Analphabétisme est trop long (peu pratique) ; changez-le pour `Apb` (attention à `attach()`...)
6. Calculer $\sum_{i=1}^{n} X_i^2$ pour la variable Analphabétisme

(4) Sauvez à nouveau les données au format binaire `StateFacts.Rdata` pour prendre en compte les changements précédents.

(5) Quittez R puis relancez-le et chargez les données à partir de ce fichier binaire. Vérifiez en particulier que le nouveau nom de la variable Analphabétisme a été conservé.

(6) Obtenez en une seule commande l'affichage des données suivantes :

a) les 10 premières lignes du jeu de données ;
b) les 5 dernières lignes du jeu de données sans utiliser `50`
c) uniquement les 5 premières observations des variables `Etat, Crime, Region`, de plusieurs manières
d) uniquement les noms des états et population des états du sud de revenu > 4500\$.

(7) Création ou modification de variables

a) Convertir le Revenu en euros en créant une variable `RevenuE`.
b) Construire une variable `RG` identique au facteur Region, mais avec des noms de modalités plus concis : "NE", "S", "W" et "NC" (ceci sera utile pour les représentations graphiques). NB : utilisez `?factor`
c) Construire avec `?cut` un "facteur Diplôme" `FD` par recodage de Diplôme en 3 modalités :

faible	moyen	fort
moins de 47	entre 47 et 57	plus de 57

d) Créer un autre `data.frame` ne contenant que les labels individus et les variables : `RevenuE, Crime, RG, FD`.

(8) Sélection d'observations : Ecrire un programme qui, à partir de la table de départ construise les tables suivantes :

a) la table des états du sud avec un taux de diplome moyen.
b) la table des états de la région "NorthCentral" avec un revenu > 4000 euros.