Penvins Cup 2017 : le cahier des charges

Comment votre futur script va être utilisé

Ne placez aucun menu interactif en tête de votre script. En effet nous considèrerons comme acquis que l'utilisateur a *déjà* chargé un tableau de données à vérifier dans R (via la fonction **`read.table()`**), et l'a placé dans un objet de classe **`data.frame`**. Il va ensuite charger le code de votre script soit "à la souris" (fichier / sourcer du code R) soit en plaçant votre script dans le répertoire de travail et en saisissant la commande : **`source("`**<nom de votre script>**`.R").`** Il saisira alors une seule ligne de commande :

`checkPenvins(`<nom de l'objet à vérifier>**`)`**

Toute la suite s'effectuera automatiquement sans autre intervention de sa part. Il lira simplement les messages d'information ou d'erreur produits par le script.

J'attire votre attention sur le fait que l'utilisateur fournit à la fonction **`checkPenvins`** le nom de l'objet de structure **`data.frame`** à vérifier : votre fonction peut donc utiliser cette information pour accéder aux colonnes de l'objet.

Bien sûr, le jury lira ensuite le code du script lui-même, pour juger de sa qualité et de la clarté des commentaires qu'il contient à l'attention des utilisateurs qui souhaiteraient comprendre son fonctionnement, par exemple dans l'optique de le modifier.

Liste des tâches que votre fonction **`checkPenvins`** doit accomplir

1) Vérifier que l'objet est bien de la classe **`data.frame`**. S'il ne l'est pas, afficher un message d'erreur signalant le problème et donnant le nom et la classe de l'objet fautif puis stopper la fonction. S'il n'y a pas d'erreur, afficher un message d'information donnant le nom et la classe de l'objet.

2) Compter les colonnes de l'objet à examiner. Supposons que le nombre de colonnes attendu pour un fichier de type quad soit N et pour un fichier de type ind soit N+8 (c'est à dire N + les colonnes sp, haut, larg, peri, pred, coul, text, masse). Tout autre nombre de colonnes que N ou N+8 doit provoquer un message d'erreur signalant qu'un nombre anormal de colonnes a été détecté, donner ce nombre, indiquer que le nombre de colonnes attendu était N ou N+8, puis stopper la fonction. Si le nombre de colonnes est correct, afficher un message d'information signalant le nombre de colonnes détecté et la nature supposée du fichier (données de quadrats ou données biométriques sur les individus).

3) Vérifier que les intitulés de colonnes correspondent à l'attendu (selon le type de fichier supposé d'après sa taille). Si non, afficher un message d'erreur signalant les noms des colonnes dont les intitulés posent problème en citant à chaque fois le nom correct attendu, puis stopper la fonction. Attention à ne pas stopper la fonction à la première erreur détectée mais bien à la fin de la vérification. S'il n'y a pas d'erreur, afficher un message d'information indiquant que les intitulés de colonnes sont corrects.

4) Vérifier que les colonnes sont bien de la classe attendue : factor (des lettres), integer (nombres entiers) ou numeric (nombres décimaux), selon le cas. Si non, afficher un message d'erreur signalant le nom et la classe des colonnes qui posent problème en citant à chaque fois la classe anormale détectée et la classe qui était attendue, puis stopper la fonction. S'il n'y a pas d'erreur, afficher un message d'information indiquant que les classes des colonnes sont correctes.

Si la procédure n'a pas été stoppée à une des étapes 1 à 4, elle ne doit plus s'arrêter maintenant jusqu'à la fin des autres vérifications, et se contenter d'afficher éventuellement des messages d'erreur :

5) pour toutes les colonnes de classe "factor", vérifier que la ou les modalités du facteur correspondent à l'attendu. En cas d'anomalie, afficher un message d'erreur donnant la colonne et les numéros et contenus des lignes qui posent problème ainsi que le contenu qui était attendu. Si aucune anomalie n'est détectée parmi les facteurs, afficher un message d'information pertinent. Voir ci-dessous le tableau d'attendu pour les colonnes de classe factor. Rappel : le n° de colonne est là pour vous aider à vous y retrouver mais il ne figurera nulle part dans votre tableau réel :

n°	**nom**	**attendu**
1	transect	même modalité pour toutes les lignes, ex.: "a10"
2	resp	même modalité pour toutes les lignes, ex.: "marie.theghost"
3	date	même modalité pour toutes les lignes, ex. "20septembre2017"
5	mode	même modalité pour toutes les lignes : "a" ou "b"
N+1	sp	abréviation parmi la liste "officielle" des espèces trouvées, ex.: "Litlit"
N+5	pred	NA ou "oui" (si présence de prédation) ou "non"
N+6	coul	NA ou "clair", "sombre", "rayures"
N+7	text	NA ou "lisse", "rugueux", "bosses"

6) pour toutes les colonnes de chiffres (de classe "integer" ou "numeric"), vérifier que le contenu ne sort pas de l'intervalle mini--maxi attendu. En cas d'anomalie, afficher un message d'erreur donnant les numéros et le contenu des lignes qui posent problème ainsi que l'intervalle mini--maxi qui était attendu. Si aucune anomalie n'est détectée parmi les variables, afficher un message d'information pertinent. Voir ci-dessous le tableau d'attendu pour les colonnes numériques

n°	**nom**	**attendu**
4	coef	même nombre entier pour toutes les lignes : 97 (sortie du mercredi) ou 99 (sortie du jeudi)
6	d.chenal	même nombre entier pour toutes les lignes. Valeurs admissibles : 0, 10, 20, 30, 40, 50, 70, 90, 110
7	d.mer	nombre entier entre 0 et une valeur max que vous choisirez
8	alt	nombre décimal entre 0 et une valeur max que vous choisirez
9	surf	nombre décimal entre deux valeurs min et max que vous choisirez
10-16	p.roc etc.	nombre entier entre 0 et 100
19-N	Litlit etc.	NA ou nombre entier entre 0 et une valeur max que vous choisirez. Ne confondez pas NA (donnée non relevée/non connue) et 0 (j'affirme, car je l'ai observé, qu'il n'y avait aucun individu de cette espèce).
17	s.flaq	NA (si ça n'était pas un quadrat choisi volontairement dans une flaque) ou nombre décimal entre deux valeurs min et max que vous choisirez
18	d.flaq	NA ou nombre décimal entre deux valeurs min et max que vous choisirez
N+2	haut	nombre décimal entre deux valeurs min et max que vous choisirez en fonction de chaque espèce
N+3	larg	nombre décimal entre deux valeurs min et max que vous choisirez en fonction de chaque espèce
N+4	peri	NA ou nombre décimal entre deux valeurs min et max que vous choisirez en fonction de chaque espèce
N+8	masse	NA ou nombre décimal entre deux valeurs min et max que vous choisirez en fonction de chaque espèce

On part du principe que vous avez au moins mesuré la hauteur et la largeur de chaque individu listé dans le fichier ind, donc pas de NA pour ces variables. Idem pour les données que vous *devez* avoir (coef, d.chenal...). Votre script devra donc signaler comme anormales les cases contenant un NA alors que ça n'est pas attendu.

7) vérifier que les ratios **`larg/haut`** et **`peri/larg`** du fichier biométrique qui ne sont pas NA sont situés respectivement dans des intervalles min-max réalistes que vous choisirez. Dans le cas contraire,

afficher autant de messages d'avertissement que nécessaire donnant le numéro de ligne, l'espèce, le ratio suspect et l'intervalle min-max attendu pour cette espèce. Si tous les ratios sont corrects, affichez un message d'information pertinent. Bien entendu les vérifications concernant les mesures biométriques ne doivent être effectuées que si le fichier a été identifié comme un fichier de type ind.

La fin de la procédure `checkPenvins` doit afficher un message de conclusion à votre goût (à supposer que tout n'ait pas été stoppé bien en amont...) signalant **soit** qu'aucune anomalie n'a été détectée **soit** au contraire que le fichier à vérifier ne satisfait pas le cahier des charges.

Il va sans dire (mais ça va mieux en le disant) qu'avant de nous le remettre, vous devez tester votre script face à un fichier contenant volontairement des fautes. Méfiez-vous particulièrement des modifications anodines de dernière minute, aussi minuscules soient elles : si vous en faites une, testez à nouveau votre script.

-- fin du cahier des charges ---