

Transformation et manipulation de données

Ce guide présente des transformations et des manipulations de données utiles lors du cours.

Les points traités sont les suivants :

- [Transformation de données](#);
- [Création d'une variable numérique](#);
- [Création d'une variable binaire](#);
- [Conversion de variables numériques en facteurs](#);
- [Création d'un sous-ensemble](#);
- [Fusion de deux jeux de données](#);
- [Juxtaposer deux séries de données \(données appariées\)](#).

Transformation de données

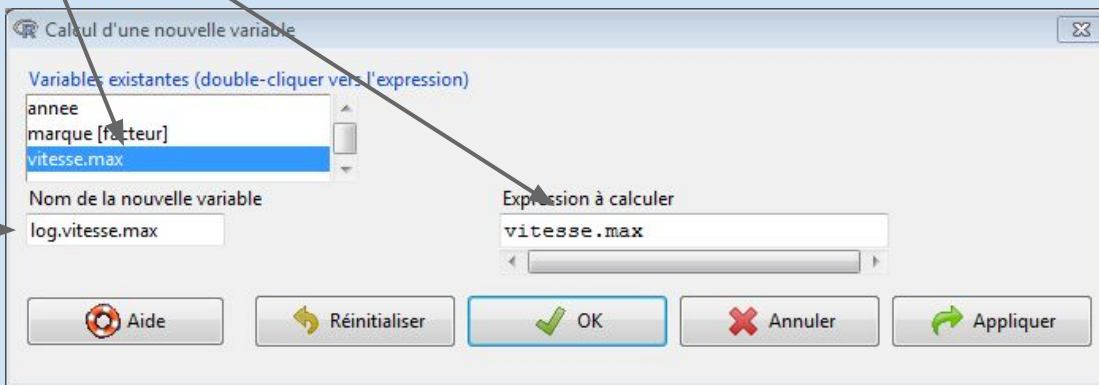
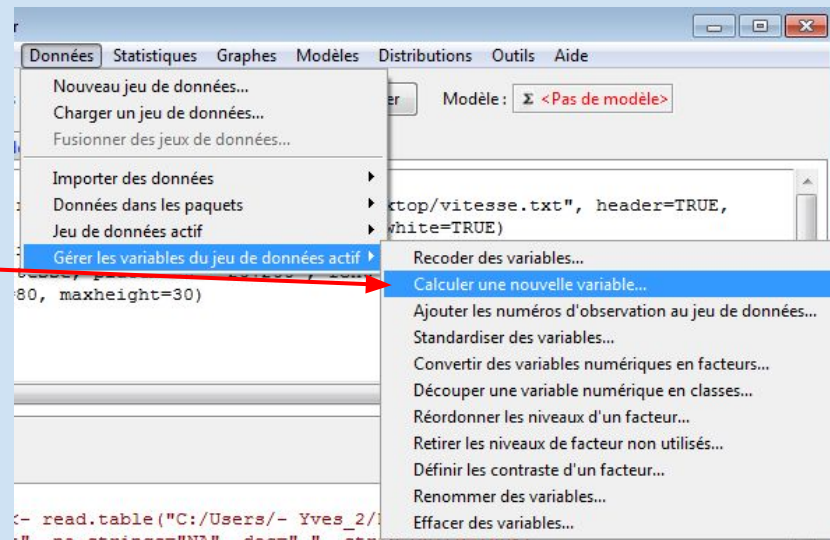
La transformation de données peut s'avérer nécessaire pour respecter les suppositions des tests statistiques. La procédure à suivre est décrite dans les prochaines animations*.

1) Cliquez sur le menu Données. Choisissez tout d'abord l'option Gérer les variables du jeu de données actif et ensuite l'option Calculer une nouvelle variable;

2) Double-cliquez sur la variable que vous souhaitez transformer. Ceci a pour avantage d'écrire le nom de la variable dans la cellule Expression à calculer;

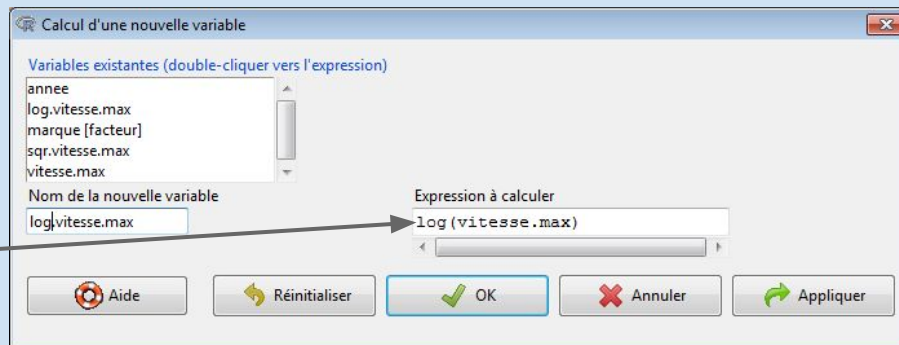
3) Saisissez le nom de la nouvelle variable. Il est préférable que le nouveau nom fasse référence à la transformation effectuée. Le suffixe "log" a été ajouté, car la nouvelle variable sera le logarithme de vitesse.max.

* Vous pouvez, à votre guise, faire les mêmes démarches. Veuillez cliquer sur ce [lien](#) et cliquer sur l'icône suivant en haut au centre de la page pour télécharger :



Transformation de données (suite)

4) Modifiez l'expression dans la cellule Expression à calculer afin d'indiquer à R Commander quelle transformation effectuer. Afin de réaliser la transformation logarithmique, veuillez ajouter "log(" avant le nom de la variable et ")" après le nom de la variable (sans les guillemets);



5) Une fois la modification effectuée, cliquez sur le bouton OK et la variable souhaitée sera créée.

Pour information, le bouton Réinitialiser permet d'effacer les informations saisies dans la fenêtre alors que le bouton Appliquer permet d'effectuer la transformation sans que la fenêtre Calcul d'une nouvelle variable ne se ferme.

	marque	annee	vitesse.max	vitesse.limite	log.vitesse.max
1	Benz	1894	19.00	100	2.914439
2	Jaguar	1949	200.50	120	5.300814
3	Mercedes-Benz	1955	225.00	140	5.416100
4	Jaguar	1961	245.00	100	5.501258

R Commander permet d'utiliser plusieurs autres transformations. Les tableaux 1 et 2 de la page 16 du document Programmation avec R – notions générales (dans la page Introduction du site Web du cours) présentent quelques exemples de transformations possibles.

abs(x)	valeur absolue des éléments de x
sqrt(x)	racine carrée des éléments de x
log2(x)	logarithme en base 2 des éléments de x
log10(x)	logarithme en base 10 des éléments de x
log(x, base)	logarithme en base base des éléments de x
log(x)	logarithme naturel des éléments de x équivaut à log(x, base = exp(1))
exp(x)	renvoie la valeur de e élevée à la puissance x
x^(y)	la valeur de x élevée à la puissance y

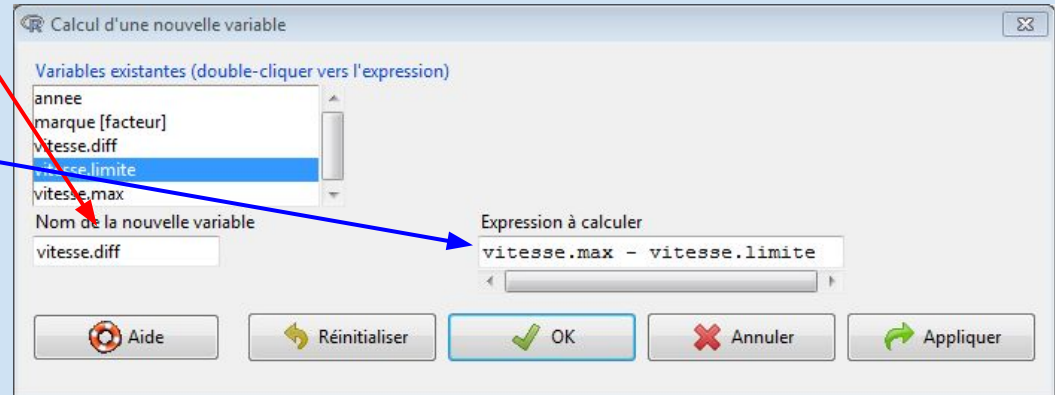
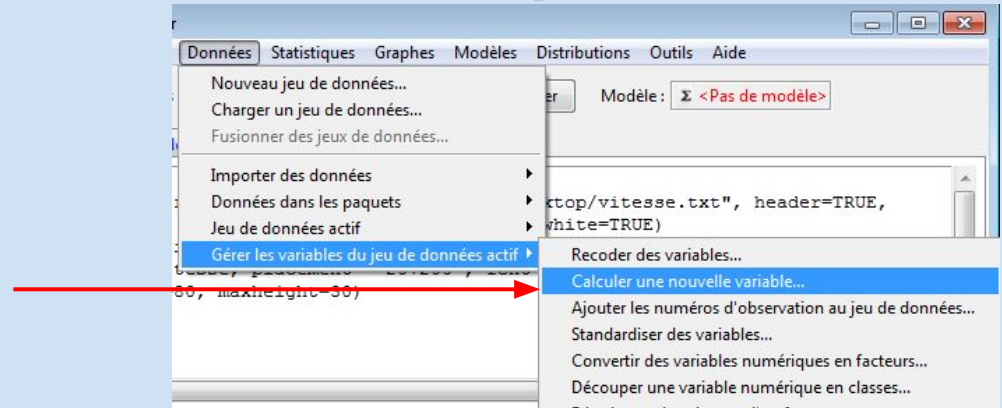
Création d'une variable numérique

La création d'une variable numérique suit à quelques détails près la procédure décrite pour la transformation de données. Voici un exemple de création d'une variable numérique :

1) Cliquez sur le menu Données. Choisissez tout d'abord l'option Gérer les variables du jeu de données actif et ensuite l'option Calculer une nouvelle variable;

2) Saisissez le nom de la nouvelle variable dans la cellule Nom de la nouvelle variable;

3) Écrivez l'opération que vous souhaitez réaliser dans la cellule Expression à calculer.



Création d'une variable numérique (suite)

Dans l'exemple présenté ici, R Commander a calculé la différence entre la vitesse maximale et la limite de vitesse sur les autoroutes.

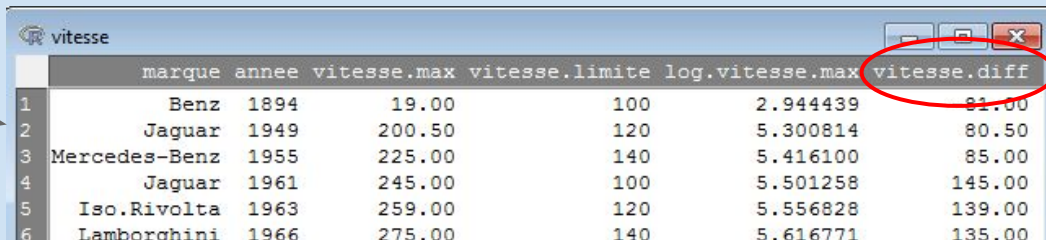
Il est possible d'utiliser plusieurs autres opérateurs arithmétiques. Voici une liste de ces opérateurs :

- + : addition;
- - : soustraction;
- * : multiplication;
- / : division;
- ^ ou ** : exponentiation.

Prenez note que :

- Ces opérations respectent les conventions de [priorité des opérations](#);
- Lors d'opérations mathématiques avec des variables ayant un nombre d'observations différent, les valeurs de la variable avec le moins d'observations sont recyclées ou réutilisées.

Pour de plus amples détails à ce sujet, veuillez consulter la section 6, page 14, du document [Programmation avec R – notions générales](#) (dans la page [Introduction](#) du site Web du cours).



	marque	annee	vitesse.max	vitesse.limite	log.vitesse.max	vitesse.diff
1	Benz	1894	19.00	100	2.944439	81.00
2	Jaguar	1949	200.50	120	5.300814	80.50
3	Mercedes-Benz	1955	225.00	140	5.416100	85.00
4	Jaguar	1961	245.00	100	5.501258	145.00
5	Iso.Rivolta	1963	259.00	120	5.556828	139.00
6	Lamborghini	1966	275.00	140	5.616771	135.00

Il est important de valider les résultats des calculs. Des erreurs peuvent toujours se glisser.

Il est aussi important de respecter l'utilisation des lettres majuscules et minuscules dans le nom des variables.

Création d'une variable binaire

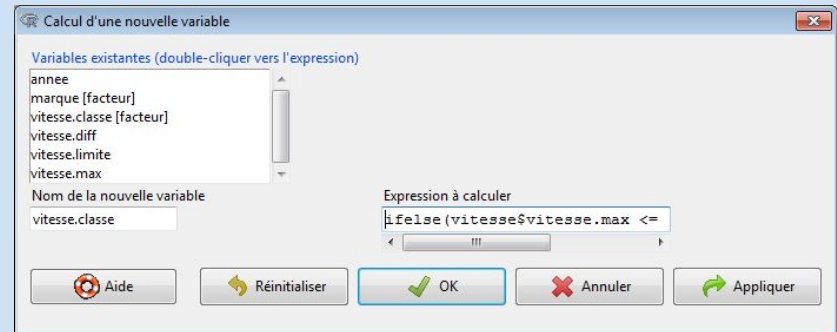
Il est parfois nécessaire de créer une variable qui n'est pas numérique. Par exemple, on peut souhaiter qualifier la vitesse maximale comme étant "lente" ou "rapide".

La marche à suivre est à quelques détails près similaire à celle présentée plus tôt. En fait, seule l'expression à calculer change. L'expression sera dans ce cas-ci :

```
ifelse(vitesse$vitesse.max <= 250, "lente", "rapide")
```

Où :

- `ifelse` : fonction permettant de réaliser des opérations selon certaines conditions;
- `vitesse$vitesse.max` : nom du jeu de données (`vitesse`) et de la variable utilisée (`vitesse.max`);
- `<= 250` : condition à respecter pour créer les classes. Ici, la vitesse maximale doit être plus petite ou égale à 250 km/h;
- `"lente"` : donnée inscrite dans le jeu de données lorsque la condition est respectée;
- `"rapide"` : donnée inscrite dans le jeu de données lorsque la condition n'est pas respectée.



	que	annee	vitesse.max	vitesse.limite	log.vitesse.max	vitesse.diff	vitesse.classe
1	enz	1894	19.00	100	2.944439	-81.00	lente
2	uar	1949	200.50	120	5.300814	80.50	lente
3	enz	1955	225.00	140	5.416100	85.00	lente
4	uar	1961	245.00	100	5.501258	145.00	lente
5	lta	1963	259.00	120	5.556828	139.00	rapide
6	ini	1966	275.00	140	5.616771	135.00	rapide
7	ari	1968	281.00	100	5.638355	181.00	rapide

Création d'une variable binaire (suite)

Il est possible d'utiliser différents opérateurs de condition. Ces opérateurs sont :

- < : plus petit que;
- <= : plus petit ou égal que;
- > : plus grand que;
- >= : plus grand ou égal que;
- == : égal à;
- != : différent de;
- ! : pas.

Il est également possible de combiner des conditions. Ces opérateurs sont :

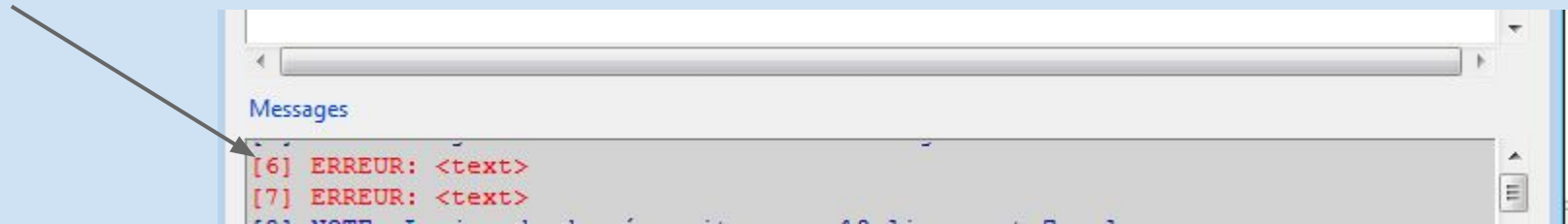
- & : et;
- | : ou.

On pourrait avoir par exemple l'expression suivante :

```
ifelse((vitesse$vitesse.max >= 250) & (vitesse$marque == "Ferrari"), "achat", "vente")
```


Création d'une variable binaire (suite)

Une erreur de programmation peut survenir lors de la création d'une variable. R Commander est discret et parfois très vague dans un tel cas. Les messages d'erreur apparaissent dans la zone Messages.



Pour information, on peut créer une variable binaire en utilisant l'option Recoder des variables... (option Gérer les variables du jeu de données actif du menu Données), mais cette option exige l'utilisation de valeurs prédéfinies lors de la définition des conditions.

Conversion de variables numériques en facteurs

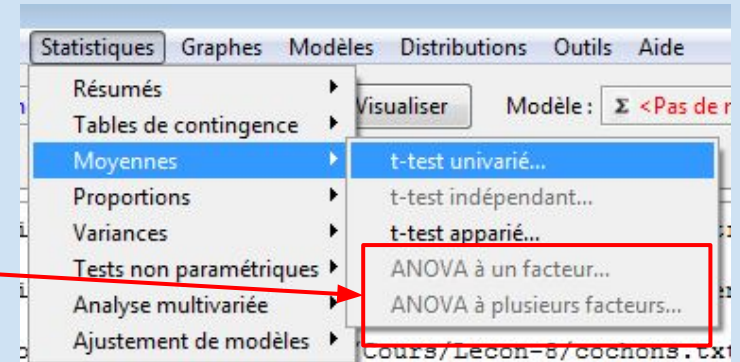
Les facteurs d'une ANOVA sont parfois indiqués à l'aide de chiffres. Ceci entraîne une mauvaise analyse des données, car le test traite cette information comme une variable numérique alors qu'il devrait la traiter comme un facteur.

La solution est alors de convertir ces chiffres en facteur. La marche à suivre sera illustrée à l'aide des [données](#) de l'Exemple 8.4 des notes de cours (Leçon 8, p. 11).

Il existe deux façons d'identifier cette situation :

1) R Commander ne permet pas d'ANOVA alors que le jeu de données permet de le faire (les options sont en caractères gris);

2) Soumettre la commande ci-dessous afin de connaître le type de variables présentes dans le fichier de données.



```
str(cochons)
```

Comme l'indiquent les résultats, les facteurs Bloc et Diète sont des entiers (int pour integer).

```
> str(cochons)
'data.frame': 20 obs. of 3 variables:
 $ Bloc : int  1 1 1 1 2 2 2 2 3 3 ...
 $ Diète: int  1 2 3 4 1 2 3 4 1 2 ...
 $ Masse: num  1.5 2.7 2.1 1.3 1.4 2.9 2.2 1 1.4 2.1 ...
```

Conversion de variables numériques en facteurs (suite)

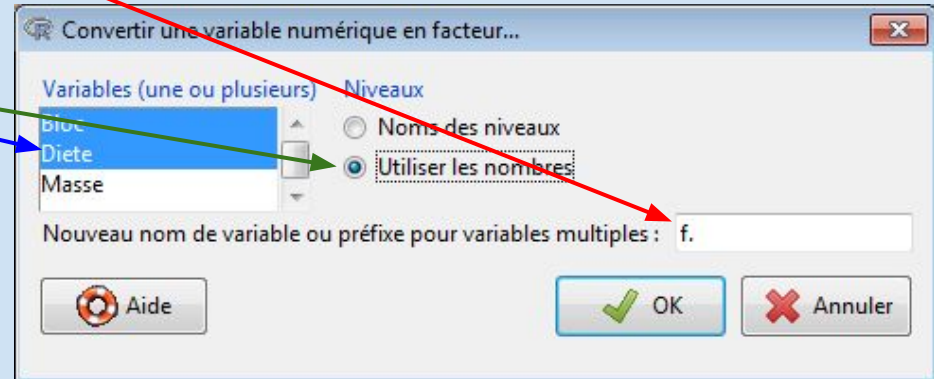
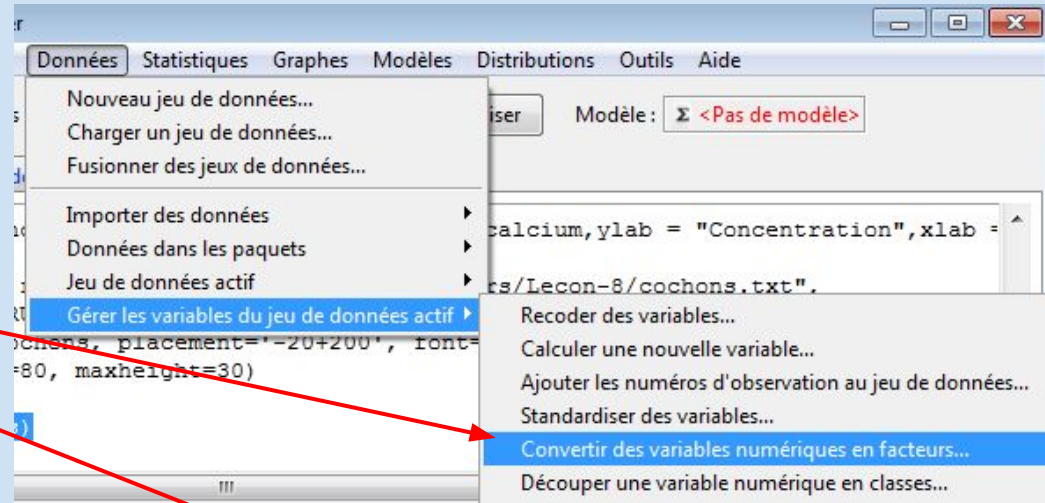
La conversion des variables numériques en facteur s'effectue de la façon suivante :

1) Cliquez sur le menu Données. Choisissez tout d'abord l'option Gérer les variables du jeu de données actif et ensuite l'option Convertir des variables numériques en facteurs...;

2) Dans la fenêtre qui apparaît, sélectionnez les variables à convertir (ici Bloc et Diete), choisissez l'option Utiliser les nombres et ajoutez "f." (sans les guillemets) dans la case Nouveau nom de variable ou préfixe pour variables multiples;

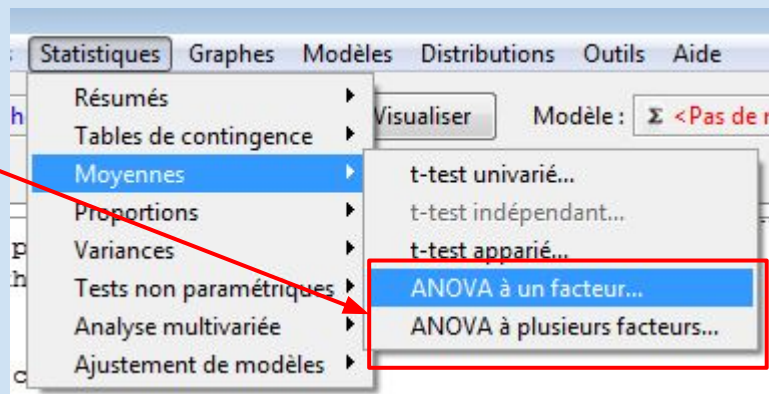
3) Cliquez sur le bouton OK afin de lancer la conversion.

Vous remarquerez à la Leçon 8 (page 11) que le rédacteur du cours utilise les noms originaux des variables lors de la transformation. L'approche retenue dans ce guide animé se veut plus prudente étant donné qu'elle conserve les données originales.



Conversion de variables numériques en facteurs (suite)

Les prises d'écran qui suivent démontrent que la conversion s'est effectuée avec succès.



```
> str(cochons)
'data.frame': 20 obs. of 5 variables:
 $ Bloc : int 1 1 1 1 2 2 2 2 3 3 ...
 $ Diete : int 1 2 3 4 1 2 3 4 1 2 ...
 $ Masse : num 1.5 2.7 2.1 1.3 1.4 2.9 2.2 1 1.4 2.1 ...
 $ f.Diete: Factor w/ 4 levels "1","2","3","4": 1 2 3 4 1 2 3 4 1 2 ...
 $ f.Bloc : Factor w/ 5 levels "1","2","3","4",...: 1 1 1 1 2 2 2 2 3 3 ...
```

Création d'un sous-ensemble

Il est parfois nécessaire d'analyser une partie d'un jeu de données. L'approche à prendre pour obtenir ce sous-ensemble sera illustrée à l'aide des [données](#) de vitesse présentées plus tôt :

1) Cliquez sur le menu Données. Choisissez tout d'abord l'option Jeu de données actif et ensuite l'option Sous-ensemble...;

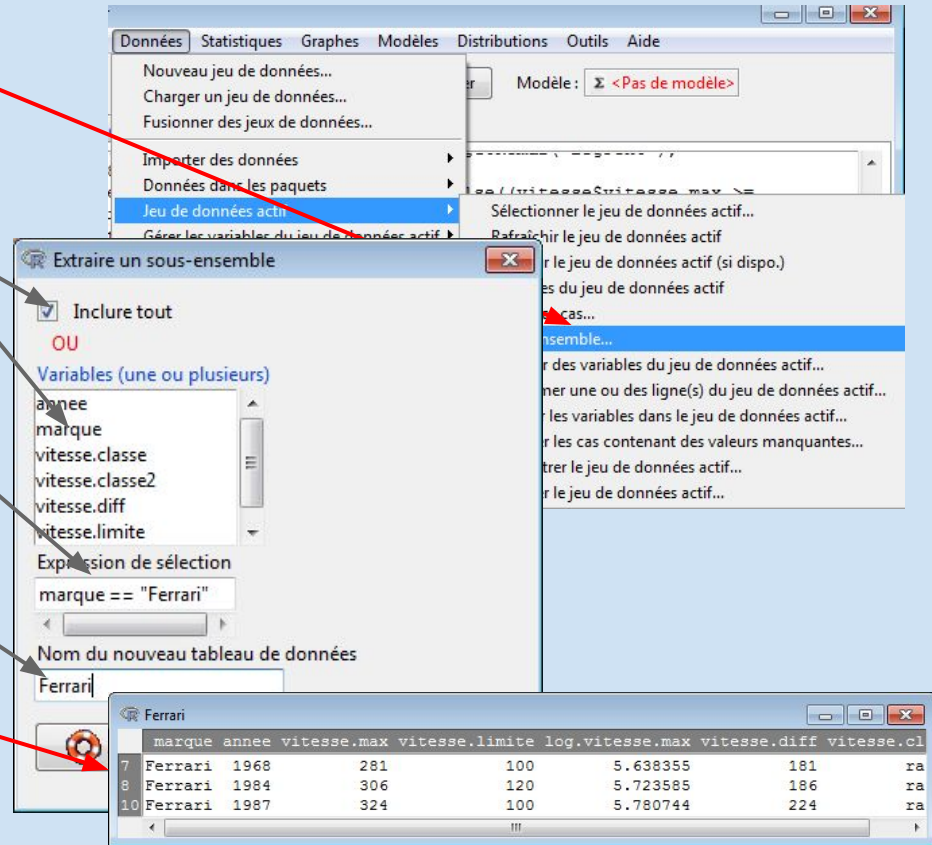
2) Spécifiez si vous souhaitez conserver toutes les variables du jeu de données ou une partie d'entre elles. Vous pouvez sélectionner plusieurs variables en cliquant sur le nom des variables tout en maintenant la touche Ctrl de votre clavier enfoncée;

3) Écrivez dans la cellule Expression de sélection les critères de sélection des données (p. ex. : `marque == "Ferrari"`). Vous pouvez utiliser les opérateurs de condition décrits précédemment.

4) Il est préférable de choisir un nouveau nom pour le jeu de données que vous êtes sur le point de créer;

5) Appuyez sur le bouton OK afin de créer le nouveau jeu de données..

Assurez-vous que le jeu de données de départ n'est pas un sous-ensemble. Sinon, vous pourriez obtenir le message d'erreur **ERREUR: data frame too wide**



Fusion de deux jeux de données

(Le cas des données appariées est traité à partir de la diapositive numéro 15.)

R Commander permet de fusionner deux jeux de données. À partir des [données](#) de vitesse, on souhaite retrouver dans un même jeu de données les voitures de marque Ferrari et Jaguar.

- 1) Importez les fichiers que vous souhaitez fusionner (ici *Ferrari* et *Jaguar*);
- 2) Cliquez sur le menu Données et choisissez l'option Fusionner des jeux de données...;
- 3) Écrivez le nom du nouveau jeu de données (ici choix);
- 4) Choisissez les deux jeux de données que vous souhaitez fusionner (ici *Ferrari* et *Jaguar*);
- 5) Sélectionnez le sens de la fusion :
 - Fusionner les lignes superpose les jeux de données;
 - Fusionner les colonnes place les jeux de données un à côté de l'autre.

5) Une fois les choix terminés, cliquez sur le bouton OK afin de fusionner les jeux de données.

The image shows three screenshots of the R Commander interface with arrows indicating the steps:

- Top screenshot:** The 'Données' menu is open, and 'Fusionner des jeux de données...' is highlighted.
- Middle screenshot:** The 'Fusionner des jeux de données' dialog box is shown. The 'Nom du tableau de données fusionné' field contains 'choix'. Under 'Premier jeu de données (en choisir un)', 'Ferrari' is selected. Under 'Second jeu de données (en choisir un)', 'Jaguar' is selected. The 'Sens de la fusion' section has 'Fusionner des lignes' selected.
- Bottom screenshot:** The resulting merged dataset 'choix' is displayed in a table. A red box highlights the first three rows (Ferrari 1968, 1984, 1987) and the next three rows (Jaguar 1949, 1961, 1992), demonstrating that the rows from both datasets are superposed.

	marque	annee	vitesse.max	vitesse.limite	log.vitesse.max	vitesse.diff	vitesse.cl
7	Ferrari	1968	281.00	100	5.638355	181.00	ra
8	Ferrari	1984	306.00	120	5.723585	186.00	ra
10	Ferrari	1987	324.00	100	5.780744	224.00	ra
2	Jaguar	1949	200.50	120	5.300814	80.50	l
4	Jaguar	1961	245.00	100	5.501258	145.00	l
13	Jaguar	1992	349.38	100	5.856160	249.38	ra

Juxtaposer deux séries de données (données appariées)

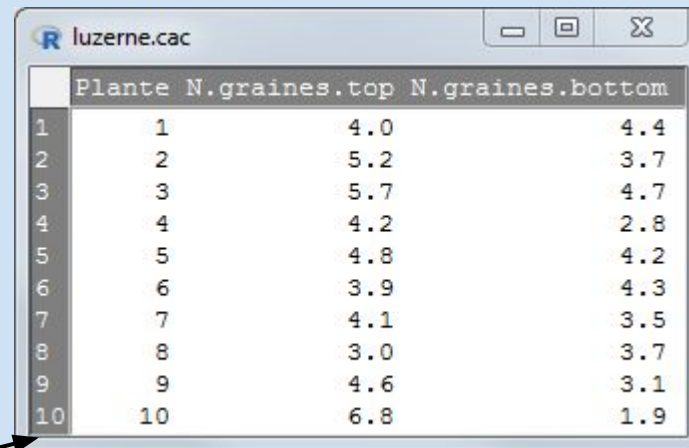
L'option *Fusionner les colonnes* ne permet pas juxtaposer deux séries de données provenant par exemple d'un même individu. Il est préférable de faire appel à la commande décrite ci-dessous. L'utilisation de cette commande est illustrée à l'aide d'un [jeu de données](#) tiré de [Steel et Torrie](#) (section 5.7, page 104).

Importez le fichier `luzerne.txt`. Assurez-vous de lui donner le nom `temp`. [Soumettez](#) ensuite les commandes suivantes :

```
luzerne <- temp
luzerne.cac <- reshape(luzerne, v.names="N.graines",timevar="Position",idvar="Plante",
direction="wide")
```

Où :

- `temp` : nom du jeu de données initial;
- `reshape` : fonction permettant de réorganiser un jeu de données;
- `luzerne.cac` : nom du jeu de données souhaité;
- `luzerne` : nom du jeu de données de travail;
- `v.names` : nom de la variable analysée;
- `timevar` : variable qui **différencie** plusieurs observations d'un même groupe ou d'un même individu (p. ex. : période de mesure, position sur la plante);
- `idvar` : variable(s) qui **identifient** plusieurs observations d'un même groupe ou d'un même individu (p. ex. : nom de l'individu);
- `direction` : format du jeu de données. Le jeu de données sera sur le sens de la largeur.



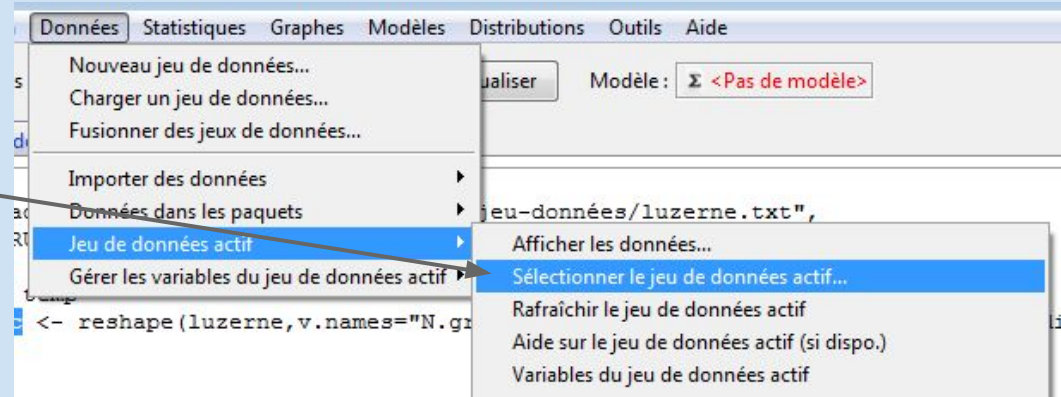
	Plante	N.graines.top	N.graines.bottom
1	1	4.0	4.4
2	2	5.2	3.7
3	3	5.7	4.7
4	4	4.2	2.8
5	5	4.8	4.2
6	6	3.9	4.3
7	7	4.1	3.5
8	8	3.0	3.7
9	9	4.6	3.1
10	10	6.8	1.9

Le résultat de cette commande est présenté à cette figure. Vous pourrez effectuer cette vérification en soumettant le nom du jeu de données souhaité (ici `luzerne.cac`).

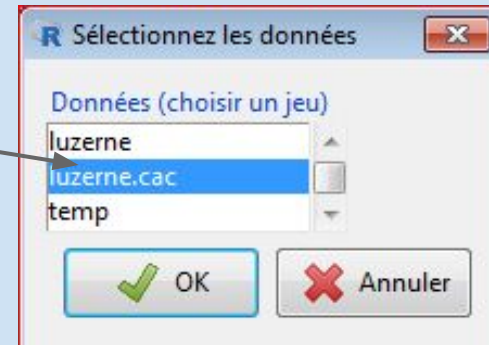
Juxtaposer deux séries de données (données appariées) (suite)

R Commander ne rend pas automatiquement disponible ce jeu de données. La marche à suivre pour le sélectionner est la suivante :

1) Cliquez sur le menu Données, descendez le curseur sur la ligne Jeux de données actif et sélectionnez l'option Sélectionner le jeu de données actif...



2) Dans la fenêtre qui apparaît, sélectionnez le jeu de données que vous avez créé et cliquez sur OK.



Il est maintenant possible d'analyser le jeu de données.