

# Introduction à R Commander

## Importation et remaniement des données

Guyliann Engels & Philippe Grosjean

2021-10-14

---

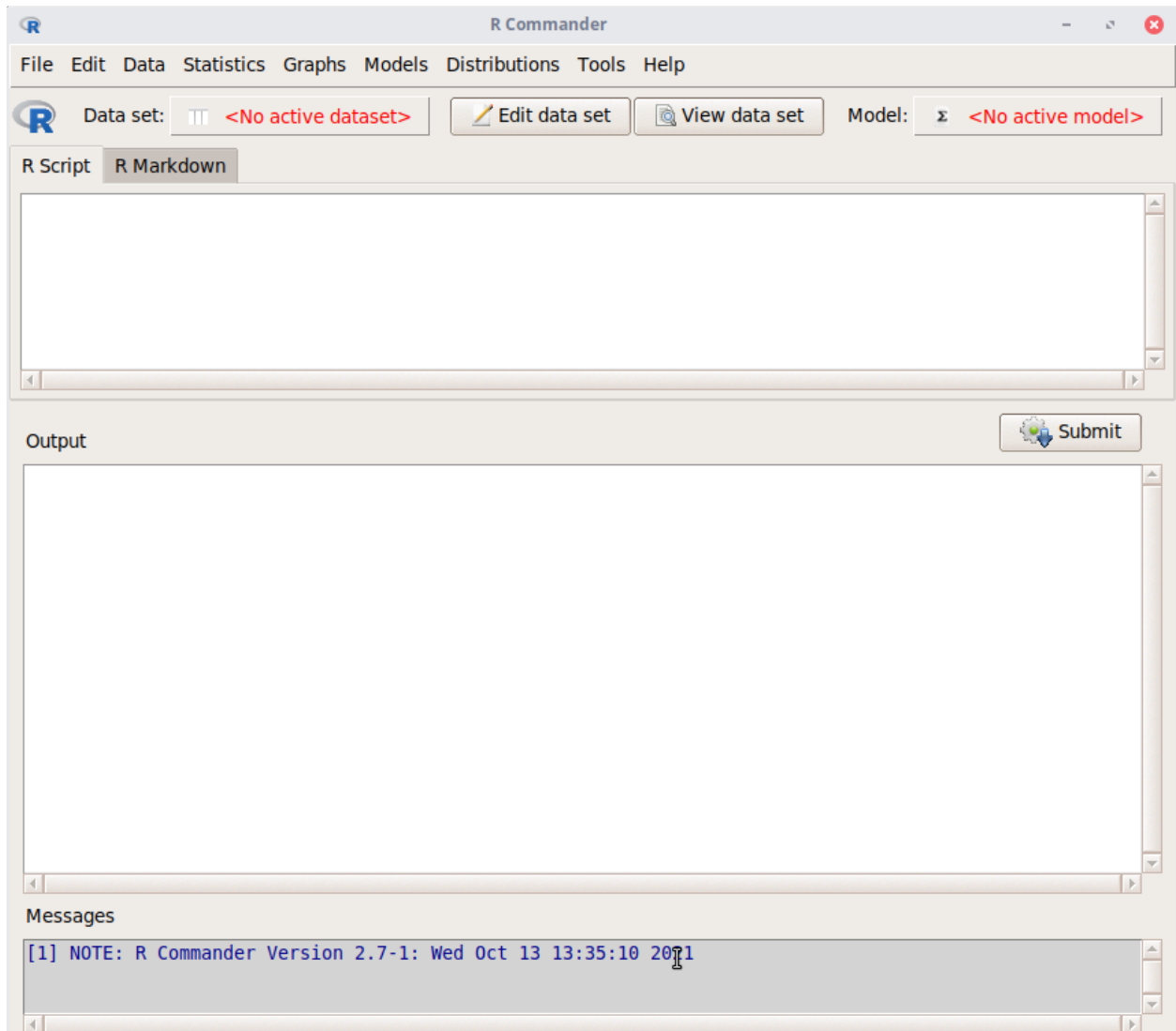
*Ce document est une contribution à STAT for U.*

**STAT** **for** **U**

# 1 Interface de R Commander

L'interface de R Commander se divise en quatre grandes parties.

1. les menus et sous-menus
2. **R Script** (ou **R Markdown**) : la section qui va contenir l'ensemble des instructions utilisées
3. **Output** : les résultats liés aux instructions utilisées
4. **Messages** : les messages associés aux instructions utilisées



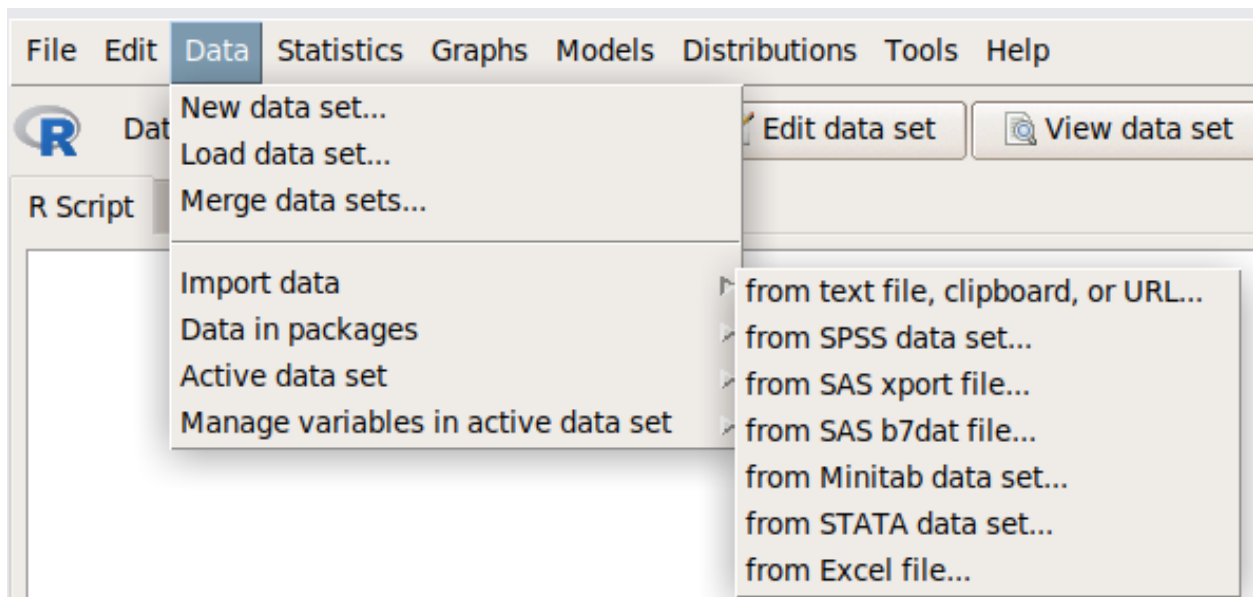
L'un des grands avantages de R Commander est de pouvoir garder l'ensemble des instructions réalisées sur un tableau de données afin d'avoir une analyse reproductible.

## 2 Importation de données

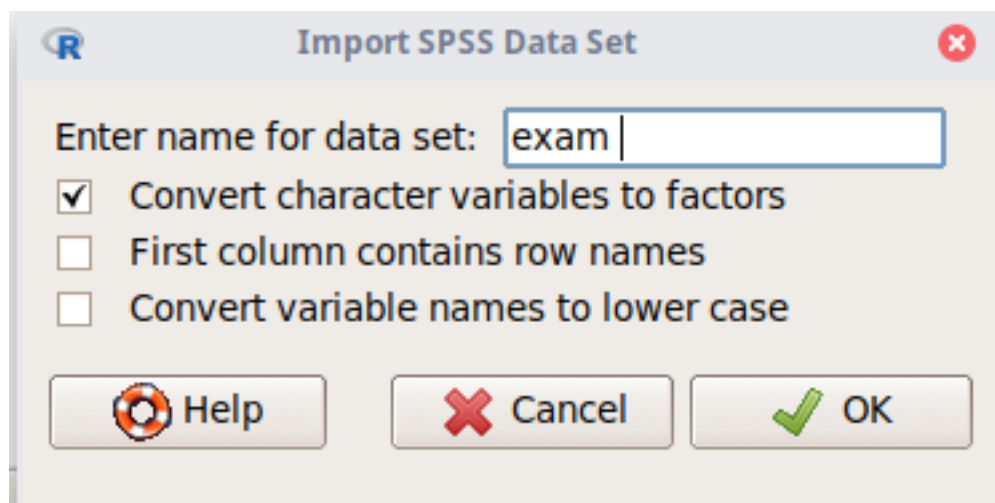
L'onglet **Data** comprend une série de sous-menus qui permettent d'importer et de remanier des données.

Les données sont disponibles dans de nombreux formats. Elles peuvent provenir d'un fichier au format varié (.txt, .xls, .sav, ...) ou même d'un package R (une sorte de boîte à outils qui comprend des fonctions qu'on utilise dans R et des jeux de données).

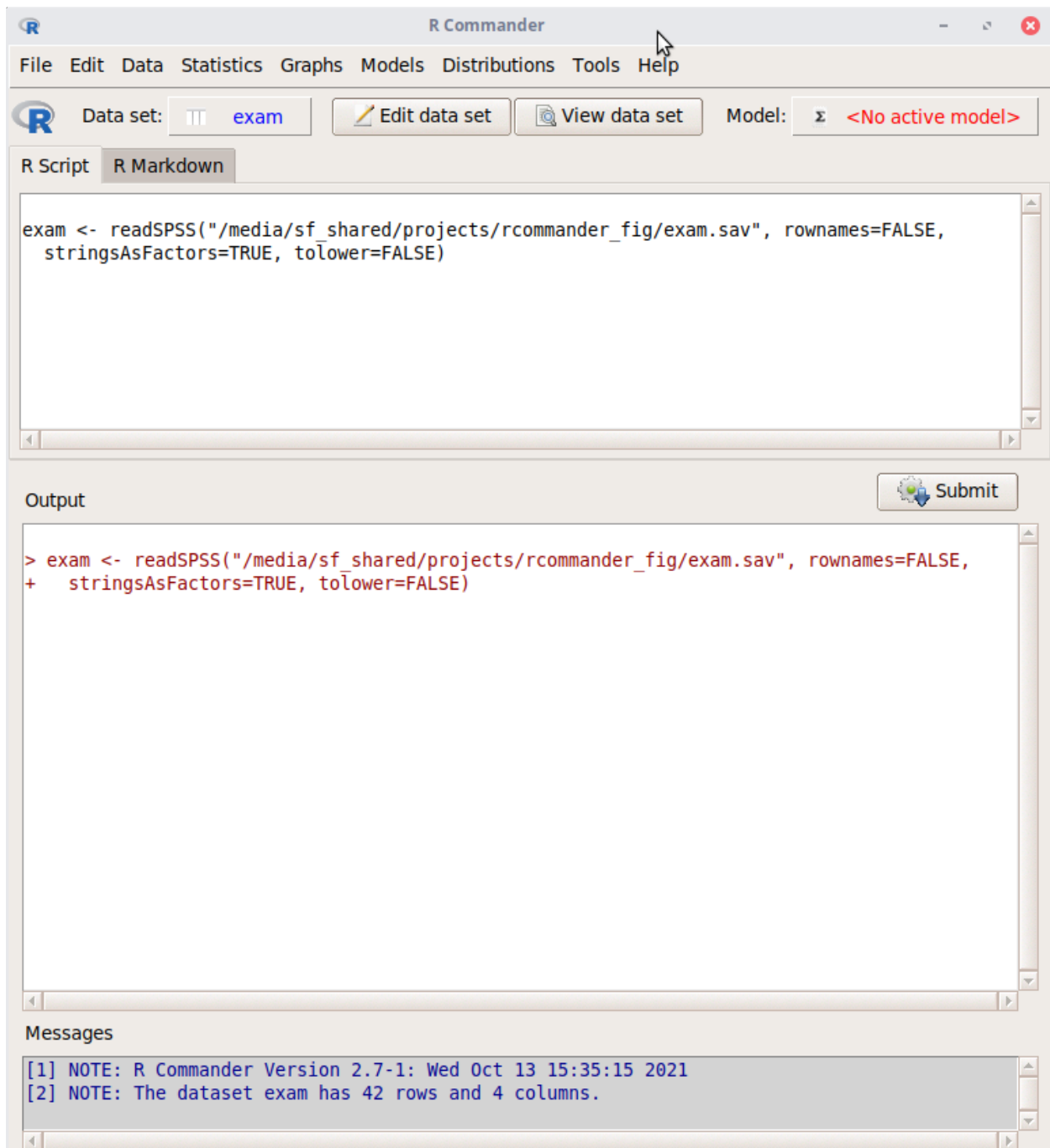
Nous utiliserons le tableau de données exam.sav (que vous pouvez télécharger ici). Afin d'importer ce tableau de données, nous allons naviguer dans les menus : **Data -> Import Data -> From SPSS data set...**



Nous allons nommer ce tableau par des données. Ce nom sera le nom de notre tableau de données dans R Commander. Différentes options sont disponibles afin de personnaliser l'importation des données.



Le tableau de données **exam** est importé avec succès. Ce tableau est le tableau de données actif. Il est possible d'éditer le tableau et de le visualiser (**Edit data set** ou **View dataset**). L'instruction R utilisée afin d'importer les données se trouve dans la section **R Script**. La section message nous informe sur le nombre de colonnes et de lignes du tableau.



Les étudiants ont passé 3 examens (variable **exam**) et ont obtenu une note sur 50 (variable **scores**).

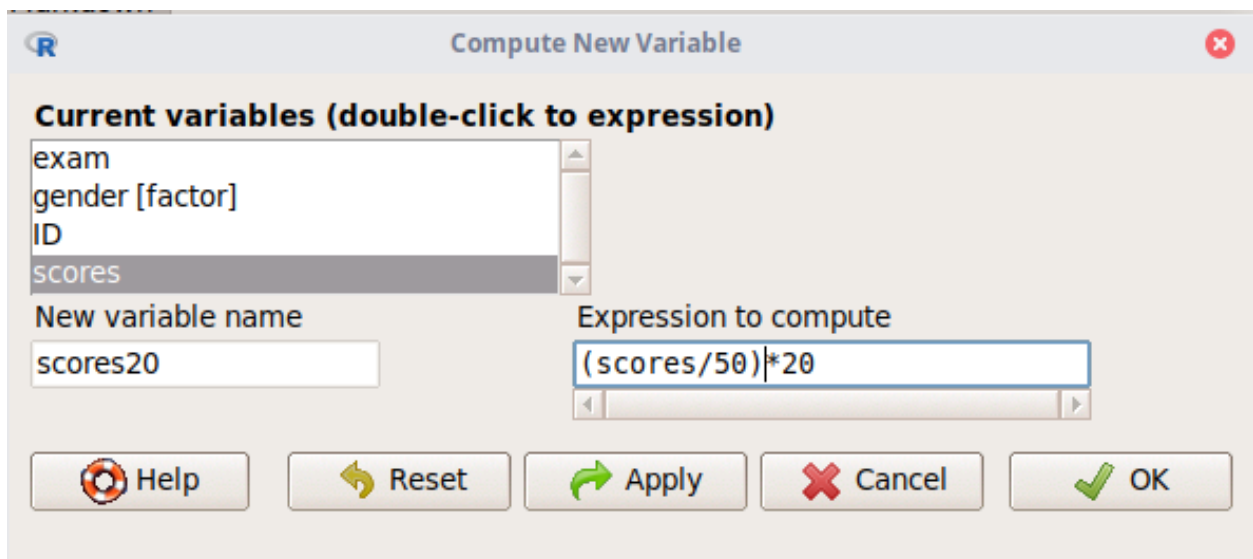
### 3 Remaniement des données

Les sous-menus **Active data set** et **Manage variables in active data set** sont dédiés à la manipulation sur le tableau de données. Le premier menu va s'intéresser à l'ensemble du tableau de données alors que le second va traiter une variable en particulier (une colonne du tableau est une variable dans R).

**Active data set** comprend les instructions afin de trier un tableau, de filtrer ou encore d'agréger un tableau. **Manage variables in active data set** permet de renommer une variable, de calculer une nouvelle variable, de changer le type de la variable (R est sensible au type de la variable)

La note des étudiants est sur 50. Nous voudrions l'avoir sur 20. Nous naviguons dans les menus afin de calculer cette nouvelle variable (Data -> Manage variables in active data set -> Compute new variable...).

La boîte de dialogue nous propose de nommer notre nouvelle variable (**New variable name**) et d'écrire le calcul associé à cette variable (**Expression to compute**).



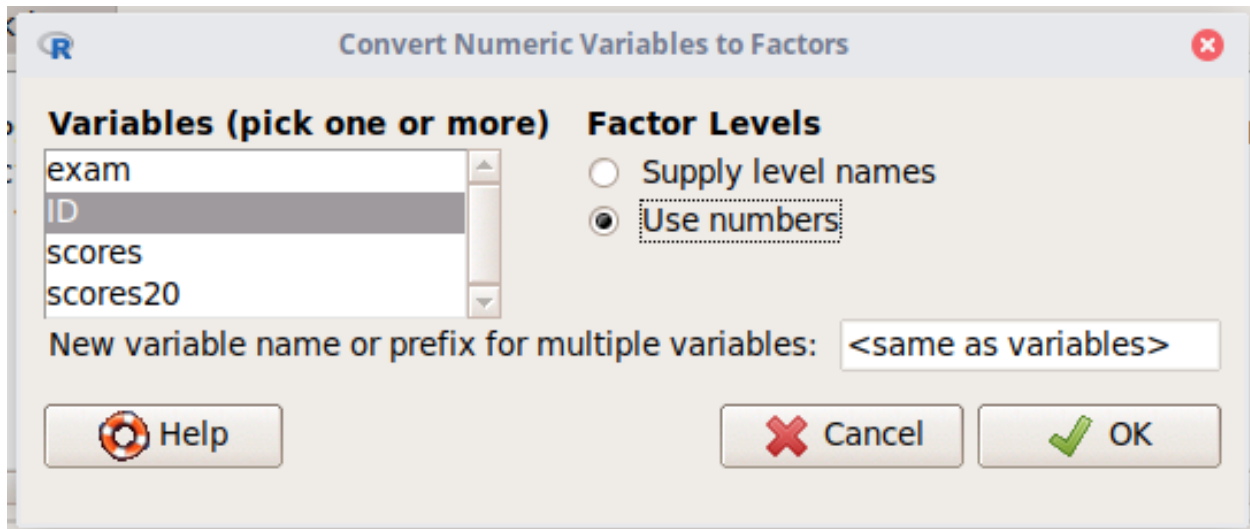
Le tableau de données s'adapte en temps réel afin de visualiser les modifications réalisées.

exam					
	ID	gender	exam	scores	scores20
1	1	f	1	45.0	18.0
2	1	f	2	33.0	13.2
3	1	f	3	29.5	11.8
4	2	f	1	44.0	17.6
5	2	f	2	33.0	13.2
6	2	f	3	27.0	10.8
7	3	f	1	44.0	17.6
8	3	f	2	40.0	16.0
9	3	f	3	44.5	17.8
10	4	m	1	37.0	14.8
11	4	m	2	39.0	15.6
12	4	m	3	34.0	13.6
13	5	f	1	47.0	18.8
14	5	f	2	46.0	18.4
15	5	f	3	45.5	18.2
16	6	m	1	45.0	18.0
17	6	m	2	42.0	16.8
18	6	m	3	39.5	15.8
19	7	m	1	46.0	18.4
20	7	m	2	38.5	15.4
21	7	m	3	42.0	16.8
22	8	m	1	40.5	16.2
23	8	m	2	36.0	14.4
24	8	m	3	30.5	12.2
25	9	f	1	33.5	13.4
26	9	f	2	42.5	17.0
27	9	f	3	41.0	16.4
28	10	f	1	39.0	15.6
29	10	f	2	19.5	7.8
30	10	f	3	32.0	12.8

Les étudiants ont réalisé 3 examens. Nous voudrions avoir une note moyenne par étudiant. La variable ID correspond à l'identifiant des étudiants (un nombre dans ce tableau). Nous convertissons ces valeurs numériques en valeurs facteurs (en facteur, nous précisons explicitement que les 3 premières lignes du tableau correspondent au même étudiant).

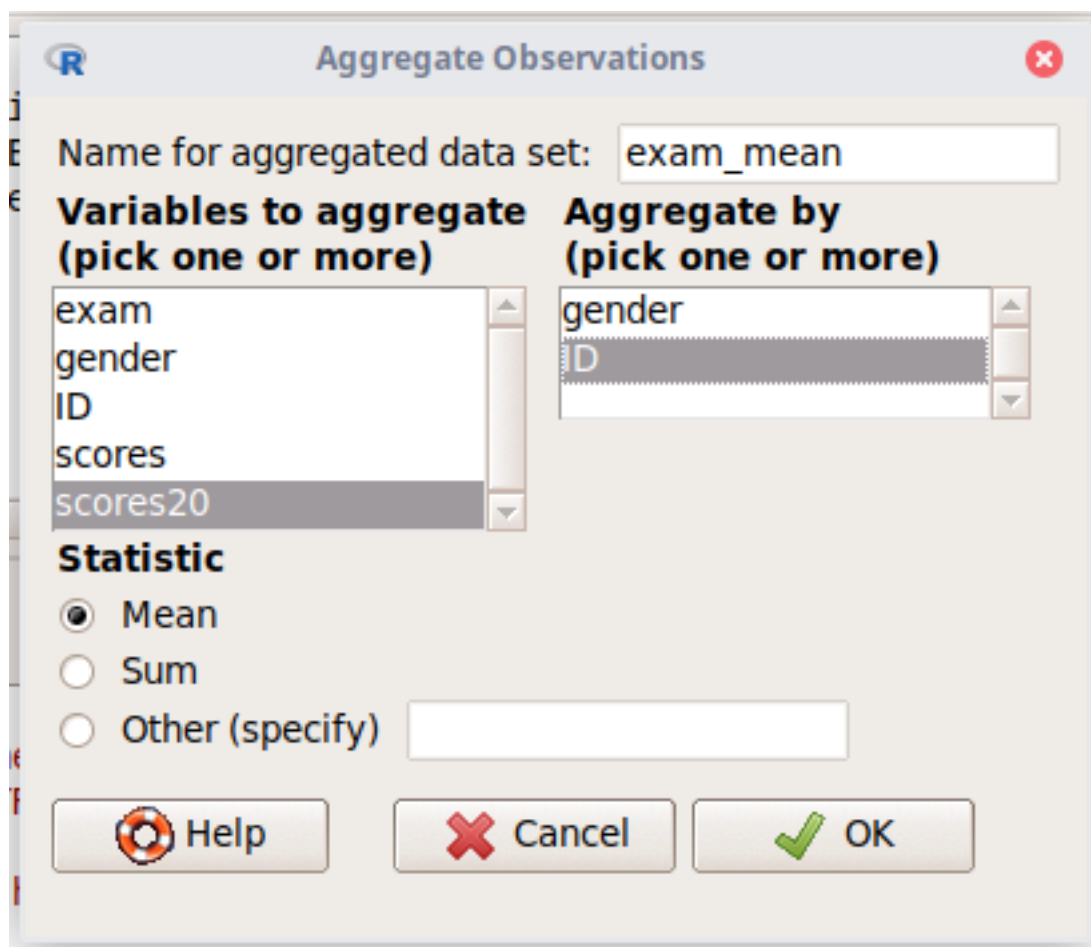
La suite d'instruction est la suivante : `Data -> Manage variables in active data set -> Convert`

numeric variables to factors.... À nouveau, nous pouvons choisir le nom de cette nouvelle variable et de définir les niveaux de cette variable.



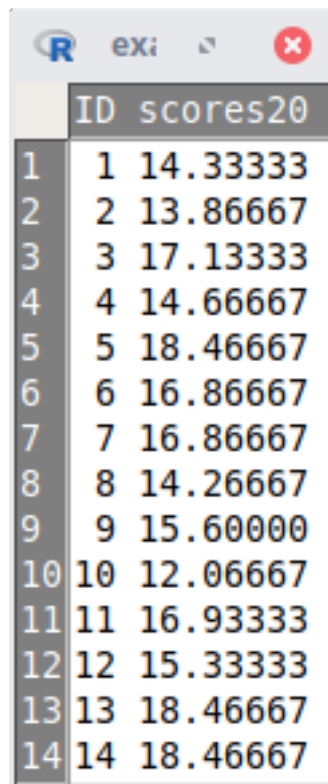
The dialog box is titled "Convert Numeric Variables to Factors". It has a list of variables on the left: "exam", "ID", "scores", and "scores20". The "ID" variable is selected. On the right, under "Factor Levels", there are two radio buttons: "Supply level names" (unselected) and "Use numbers" (selected). Below this, there is a text field labeled "New variable name or prefix for multiple variables:" with the value "<same as variables>". At the bottom, there are three buttons: "Help" (with a lifebuoy icon), "Cancel" (with a red X icon), and "OK" (with a green checkmark icon).

L'agrégation va modifier l'ensemble du tableau, il faut donc changer de menu (Active data set -> Aggregate variables in active data set...). Nous avons la possibilité de donner un nom à ce tableau résumé, de choisir les variables d'intérêt et la formule mathématique utilisée pour agréger les observations.



The dialog box is titled "Aggregate Observations". It has a text field at the top labeled "Name for aggregated data set:" with the value "exam\_mean". Below this, there are two lists: "Variables to aggregate (pick one or more)" and "Aggregate by (pick one or more)". The first list contains "exam", "gender", "ID", "scores", and "scores20", with "scores20" selected. The second list contains "gender" and "ID", with "ID" selected. Below these lists, there is a section labeled "Statistic" with three radio buttons: "Mean" (selected), "Sum" (unselected), and "Other (specify)" (unselected). To the right of "Other (specify)" is an empty text field. At the bottom, there are three buttons: "Help" (with a lifebuoy icon), "Cancel" (with a red X icon), and "OK" (with a green checkmark icon).

Un nouveau tableau est disponible avec la note moyenne par étudiant.



The image shows a screenshot of an R console window. The window title is 'exi'. The output is a table with two columns: 'ID' and 'scores20'. The table contains 14 rows of data. The first column 'ID' has values from 1 to 14. The second column 'scores20' has values ranging from 12.06667 to 18.46667.

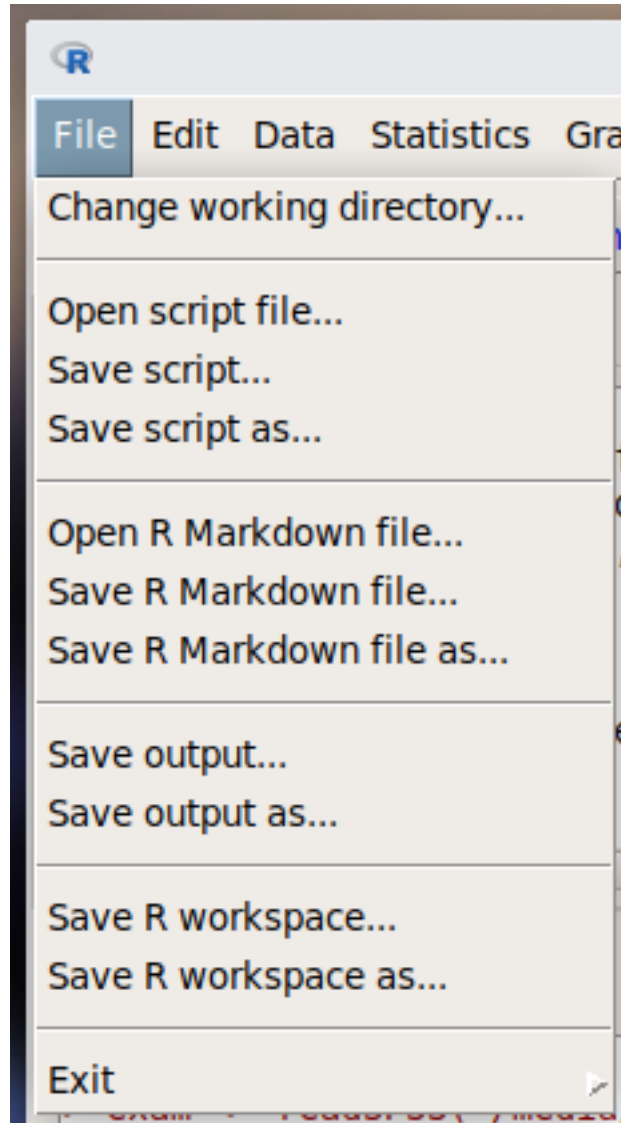
	ID	scores20
1	1	14.33333
2	2	13.86667
3	3	17.13333
4	4	14.66667
5	5	18.46667
6	6	16.86667
7	7	16.86667
8	8	14.26667
9	9	15.60000
10	10	12.06667
11	11	16.93333
12	12	15.33333
13	13	18.46667
14	14	18.46667



## 4 Sauvegarde du script R

Afin d'être reproductible, il est important de sauver l'ensemble des instructions que nous avons réalisés sur un tableau de données.

Dans l'onglet file, il est possible de sauvegarder notre script R (File -> Save script...)



## 5 Exercices supplémentaires

Entraînez-vous sur 2 jeux de données supplémentaires.

### 5.1 Les pétales d'iris

- Importez le tableau de données `iris` provenant du package `datasets` (consultez la page d'aide afin d'en apprendre davantage sur ces données).
- Visualisez le tableau de données
- Triez le tableau de données en fonction de la longueur des pétales par ordre décroissant.
- Calculez le ratio de la longueur des Pétales (`Petal.Length`) par la longueur de Sépales (`Sepal.Length`). Nommez cette variable `ratio`.
- Calculez la moyenne par espèces de la variable `ratio`.

Vous devriez obtenir un tableau similaire à ce dernier

Species	ratio
setosa	0.29
versicolor	0.72
virginica	0.84

### 5.2 La note des étudiants aux examens

- Importez le tableau de données `exam.txt` (disponible ici). Il s'agit du même tableau que nous avons traité précédemment.
- Intéressez-vous uniquement au garçon et déterminez quel est l'étudiant qui a eu la note moyenne sur ces 3 examens la plus faible.